Sports Analytics

by

Rajitha Minusha Silva

M.Sc., Sam Houston State University, 2013
B.Sc.(Hons.), Rajarata University of Sri Lanka, 2008

Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Fall 2016

Approval

Name:	**Rajitha Minusha Silva**
Degree:	**Doctor of Philosophy (Statistics)**
Title:	***Sports Analytics***
Examining Committee:	**Chair:** Yi Lu Associate Professor

Tim Swartz
Senior Supervisor
Professor

Boxin Tang
Supervisor
Professor

Oliver Schulte
Internal Examiner
Professor
School of Computing Science

Michael Schuckers
External Examiner
Professor
Department of Mathematics, Computer Science and Statistics
St. Lawrence University, USA

Date Defended: 08 December 2016

Abstract

This thesis consists of a compilation of four research papers.

Chapter 2 investigates the powerplay in one-day cricket. The form of the analysis takes a “what if” approach where powerplay outcomes are substituted with what might have happened had there been no powerplay. This leads to a paired comparisons setting consisting of actual matches and hypothetical parallel matches where outcomes are imputed during the powerplay period. We also investigate individual batsmen and bowlers and their performances during the powerplay.

Chapter 3 considers the problem of determining optimal substitution times in soccer. An analysis is presented based on Bayesian logistic regression. We find that with evenly matched teams, there is a goal scoring advantage to the trailing team during the second half of a match. We observe that there is no discernible time during the second half when there is a benefit due to substitution.

Chapter 4 explores two avenues for the modification of tactics in Twenty20 cricket. The first idea is based on the realization that wickets are of less importance in Twenty20 cricket than in other formats of cricket (e.g. one-day cricket and Test cricket). The second idea may be applicable when there exists a sizeable mismatch between two competing teams. In this case, the weaker team may be able to improve its win probability by increasing the variance of run differential. A specific variance inflation technique which we consider is increased aggressiveness in batting.

Chapter 5 explores new definitions for pace of play in ice hockey. Using detailed event data from the 2015-2016 regular season of the National Hockey League (NHL), the distance of puck movement with possession is the proposed criterion in determining the pace of a game. Although intuitive, this notion of pace does not correlate with expected and familiar quantities such as goals scored and shots taken.

Keywords: Bayesian logistic regression; WinBUGS software; Cricket; Soccer substitutions; Variance inflation

Dedication

To my loving wife, my son, and my parents in Sri Lanka...

Acknowledgements

I am grateful to have a chance to be a PhD student under the supervision of Dr. Tim Swartz. I would like to thank Dr. Swartz for his endless support on both a professional and personal level. I constantly benefited from his support and guidance throughout my life at SFU. Thank you Dr. Ananda Manage for introducing me to Dr. Swartz.

I wish to express my sincere thanks to my examining committee: Dr. Michael Schuckers, Dr. Tim Swartz, Dr. Boxin Tang, Dr. Oliver Schulte. I appreciate their time to evaluate my PhD dissertation. I also appreciate their constructive input and challenging questions during my thesis defence.

I take this opportunity to record my gratitude to all the faculty members of the Department of Statistics and Actuarial Science. I extend my gratitude to Sadika, Kelly, and Charlene for their assistance.

Special thanks are due to my research colleagues Dr. Jack Davis and Dr. Harsha Perera for their support and their genuine interest in my research.

My whole stay in Vancouver has been a rewarding experience, mostly, because of my Sri Lankan graduate and undergraduate friends at SFU. Vancouver is one of the most enjoyable cities I have lived in.

Finally, my warmest thanks go to my family for their understanding and patience during my stressful times.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Statistics in Sports

Statistics in Sports is a growing field in Statistics that provides specialized methodology for collecting and analyzing sports data in order to make decisions for successful planning and implementation of new strategies. Prior to the twenty-first century, decision making in sports was primarily based on the information acquired by observation. This has changed with technological advances, mainly related to data acquisition and the availability of personal computing.

The term "sports analytics" has been more popular than the term, "Statistics in sports", perhaps due to the fact that the expertise is borrowed from different fields such as Statistics, Computer Science, Management, and the Health Sciences. Therefore, sports analytics is broadly described as the process of data management, predictive model implementation, and the use of information systems for decision making to gain a competitive advantage on the field of play (Alamar and Mehrotra 2011). There are many areas where sports analytics have been implemented. For example, sports teams use statistical analysis to evaluate players in order to determine the best game strategy. Sports associations develop rankings of players and teams, evaluate existing rules and study the feasibility of introducing new rules. Sports health professionals use statistical methods to understand players' physical and mental conditions.

Beginning in 1977, Bill James self-published an annual book titled "Baseball Abstract" that is viewed by many as the beginning of sports analytics. Though he did not utilize even basic tools of Statistics such as model fitting or graphical displays for most of his writing, the work was influential in contesting long-held beliefs in baseball. The main characteristic of his work was that his opinions were based on evidence contained in data. Appreciation for his work came to the peak with a 2011 movie called "Moneyball" which is based on the book titled "Moneyball: The Art of Winning an Unfair Game" by Michael Lewis (Lewis 2004). In the book, Lewis explores some of James' innovative concepts and illustrates how Billy Beane, the general manager of a Major League Baseball (MLB) team called the Oakland Athletics, adapted these concepts into practical application. One of the most recent academic books written about sports analytics is “Handbook of Statistical Methods and Analyses in Sports” co-edited by Jim Albert, Mark E. Glickman, Tim B. Swartz and Ruud H. Koning. It provides overviews of statistical methods in the major sports and describes challenges and problems confronting statistical research in sports.

The discipline of Statistics has recognized that sport is rich in data, and that interesting statistical problems arise in sport. Two of the early sports papers in prominent Statistics journals that are familiar to us are by Elderton (1945) and Wood (1945). Both of these papers concern the distribution of running scoring in Test cricket. As another early example, Mosteller (1952) considered the problem of estimating the probability that the better team wins the World Series in MLB. With the understanding of a need to foster the development of statistics and its applications in sports, the American Statistical Association (ASA) initiated a separate section for “Statistics in Sports”(SIS) in 1992. It promotes publications devoted to statistical theory and methodology and their application to statistics in sports. The SIS section also promotes meetings devoted to sports analytics, provides career guidance, data resources and online sports statistics forums.

With the rapid evolution in the field of sports analytics, some research journals have been created that are totally devoted to statistics in sports. Two main sports analytics journals with a concentration on Statistics are The Journal of Quantitative Analysis in Sports which is an official journal of the American Statistical Association and Journal of Sports Analytics. The Journal of Sports Analytics is a more recent journal which has a focus on practical applications that serve team owners, general managers, coaches, fans and academics. There are many other sports science journals which are not specifically statistical. A very abbreviated list of these journals include Journal of Sports Economics, Journal of Sports Sciences, American Journal of Sports Medicine, International Journal of Computer Science in Sport, Journal of Sports Science and Medicine and International Journal of Sports Science and Engineering.

In addition to journals and books, sports analysts publish their work on blog sites as well. Now, it has become a trend to provide day-by-day analyses of games on blogs with less effort than publishing (e.g. `www.soccermetrics.net`, `www.hockeyanalysis.com`). Sports analytics conferences are also a platform for professionals, researchers and students to discuss related topics in sports. The MIT Sloan Sports Analytics Conference is one of the highly recognized annual conferences. Other regular sports conferences include the New England Symposium on Statistics in Sport, MathSport International and the Australasian Conference on Mathematics and Computers in Sport. Other recent and nearby sports analytics conferences have been the Ottawa Hockey Analytic Conference (January, 2016), the Vancouver Hockey Analytics Conference (April, 2016), the Sports Analytics Innovation

Summit in San Francisco, (August, 2016) and the Cascadia Symposium on Statistics in Sports in Vancouver (September, 2016).

Banding together as societies, online forums, or professional groups is beneficial to both the sport and to the researchers involved. The Society for American Baseball Research (SABR) was established in 1971 with the idea of giving its members opportunities to publish their findings in its journal "Baseball Research Journal" which includes histories, biographies, statistics, personalities, book reviews, and other aspects of the game. SABR has around 6,000 members including major and minor league baseball officials, broadcasters and writers, and numerous former players (`www.sabr.org`). At Simon Fraser University (SFU), a group of SFU faculty, coaches and students who have a passion for sports and analytics formed the Sports Analytic Group (SAG) in 2015, with the goal of creating new knowledge about sports and the underlying analytics that help the decision makers in sports organization. They organize regular talks given by coaches and practitioners in their respective field to understand real world situations in sports and to conduct effective research (`www.sfu.ca/sportsanalytics.html`). Through the organization of two conferences, the Vancouver Hockey Analytics Conference and the Cascadia Symposium on Statistics in Sports (CASSIS), SAG was able to bring sports analytics experts to showcase the power of analytics in sports.

Modern technology has made it easier to visualize the statistical information to public. The SportVU camera system is used in the National Basketball Association (NBA) to track the real-time positions of players and the ball 25 times per second. In the 2014 FIFA World Cup, a system called "Matrics" which was built by an Italian firm called Deltatre, gathered data to deliver the real-time statistics. The rich datasets can be utilized by sports researchers to provide insights on soccer. During matches, commentators were able to explain their opinions with the help of this visual data. Technology is also beneficial to sports management in the area of ticketing analytics, sports betting and fan engagement.

Now, I briefly discuss six of the most popular sports in the world and the impact of statistical analytics on these sports. What all of these team sports have in common is money. Analytics is expensive and requires salaries for team analysts and the purchase of hardware, software and data access.

1.2 Soccer

Using any familiar measure, soccer is clearly the most popular sport in the world. It is, however, behind in analytics compared to various North American sports such as baseball and basketball. Resistance to the use of new technology and a lack of standardization and access to data are ongoing problems. Prozone sports data company is among the pioneers in using player tracking data for technical and tactical statistical analysis. Until recently, player performance has been primarily based on events such as passes and shots on goal.

With the aid of tracking data, new metrics are being developed to measure the impact a player has on subtle processes such as creating space for his teammates, applying defensive pressure and reducing passing options. Although soccer is growing in North America via international soccer and Major League Soccer (MLS), soccer is very much a world sport with the biggest leagues in Europe.

1.3 Basketball

Basketball has the second greatest number of professional leagues worldwide. Some of the useful metrics in basketball analytics are field goal attempts, field goal percentage, free throw attempts, free throw percentage, defensive rebound rate, and adjusted plus/minus. Whereas the above metrics and others have been useful for many years, a revolution has occurred in NBA basketball due to the proliferation of player tracking data made available from the company SportVU. Utilizing tracking data, many detailed aspects of basketball are now being investigated such as the measurement of defensive contributions (Franks et al. 2015).

1.4 Cricket

ICC Cricket World Cup 2015 hosted in Australia and New Zealand has been ranked third from over 80 sporting events by Sportcal's Global Sports Event Index, after the FIFA Women's World Cup 2015 and the Rugby World Cup 2015.

Though the game is more popular in the countries that were once colonized by England, the latest format of cricket known as "Twenty20" is now gaining the attention of other countries. However, cricket lags behind the other major professional sports in terms of advanced analytics.

The availability of cricket data, as provided by the Cricinfo website is the primary source for publicly available cricket data. It contains information on matches going back to the 1770's (`www.espncricinfo.com`). An overview of cricket analytics is provided by Tim Swartz in the chapter "Research Directions in Cricket" in the previously mentioned "Handbook of Statistical Methods and Analyses in Sports".

1.5 Ice Hockey

Ice hockey is the most popular winter sport in the world where it is played in snow and cold countries like Canada, Russia, USA, and some Scandinavian countries. It is the dominant sport in Canada. Though the sport is not widely played throughout the world, hockey analytics are gaining attention, especially in the National Hockey League (NHL). Similar to basketball, there have been many common statistics that have been developed over the years in hockey that have gained widespread usage.

However, one of the most notable recent advances in hockey analytics is the availability of the NHL Real Time Scoring System database. Thomas and Ventura (2014) have created an R package *nhlscrapr* that provides detailed event information from the database in a format that can be handled by analysts. And in terms of hockey analytics, even more data will be on the horizon as player tracking cameras have now been installed in NHL arenas. It is expected that the type of data available in the NBA will soon be available for the NHL. One of the companies leading the way in NHL player tracking technology is Sportlogiq.

1.6 Baseball

Baseball is one of the major spectator sports in both the USA and Japan, and MLB is the oldest of the four major professional North American sports leagues - MLB, NBA, NHL and NFL. Baseball analytics became known to the public in the 1980s, with the publication of Bill James' Baseball Abstract. James was the co-founder of a research based movement called Sabermetrics, named after the Society for American Baseball Research (SABR). Sabermetrics is concerned with the mathematical and statistical study of baseball. We believe that nearly every MLB team now has an analytics staff.

With the introduction of cameras in MLB ballparks, and the implementation of PITCHf/x and FIELDf/x technologies, baseball is undergoing an analytics renaissance. For example PITCHf/x provides over 70 variables of information on every pitch thrown in every MLB game. Some of the new metrics in baseball analytics are true average (TAv), base running runs (BRR), special aggregate fielding evaluation (SAFE), and offensive/defensive efficiency rating (OER/DER).

1.7 American Football

The National Football League (NFL), formed in 1920, is one of the four professional leagues in North America. The NFL official website `www.nfl.com` is a primary source for football data. In addition, the website `www.advancedfootballanalytics.com` appears to be a well-organized source for football analytics. It categorizes analyses into four main areas: team analysis, player analysis, game analysis, and game probabilities. Although the NFL appears secretive about the analytics that are carried out and there is a dearth of research publications related to football, football may be prime for a growth in analytics. In particular, Sam Ventura and his colleagues at Carnegie Mellon University have created an open-source R package *nflscrapR* which allows easy access to detailed NFL data from 2009-2016.

1.8 Organization of the thesis

In my thesis, there are four chapters that follow. What these chapters have in common is sports analytics. My work attempts to investigate some interesting sports problems from a statistical perspective and gain insight on these problems. Another common theme in the four chapters is data. Although there is disagreement these days about what constitutes big data, none of the data sets used in this thesis are trivial. Each data set requires some sort of scraping procedure to secure detailed data. And it is the detailed level of the data (as opposed to aggregrate data which people often see) that has allowed me to study the topics under consideration.

Another common theme in my work is computation. Each of the following four chapters use programs written either in the R programming language or in the WinBUGS programming language. WinBUGS is an especially valuable language when carrying out Bayesian analyses.

Before describing the chapters in more detail, a final general remark about the chapters is that the focus is problem based. Sometimes, the problems call for more sophisticated methodology and sometimes less. However, the focus has always been on solving/investigating the problem.

Chapter 2 is a project related to the powerplay in one-day cricket. This chapter is a copy of the published paper by Silva, Manage and Swartz (2015) which has appeared in the European Journal of Operational Research. In one-day cricket, powerplay rules have changed frequently over the years. Here we investigate the various powerplay rules in terms of run production and the taking of wickets. It is important for the ICC (International Cricket Council) to have a good understanding of the effect of new rules. In our opinion, rules to major sports should not be changed regularly as this affects record keeping and affects the integrity of the game. The main idea in the chapter (paper) is the replacement of powerplay overs with overs that resemble non-powerplay overs. This leads to a dataset of paired matches, actual matches and parallel matches. This chapter (paper) appears to be the first scientific study of the powerplay in one-day cricket

Chapter 3 considers an analysis of substitutions in soccer. This chapter is a copy of the discussed paper by Silva and Swartz (2016) which has appeared in the Journal of Quantitative Analysis in Sports. The project is a response to an earlier paper by Myers (2012). Myers (2012) introduced a decision rule for soccer substitutions to improve goal differential if a team is trailing. Myers' rule has received considerable attention in the media due to its apparent simplicity and for the alleged benefits from following the rule. However, our intuition suggested that there was something problematic with the rule and this was the motivation for our investigation. In addition to a careful analysis of Myers' rule, we introduced a statistical model which took into account covariates that Myers (2012) had not considered. Our primary finding was that there is no goal scoring benefit due to various substitution patterns. In essence, we found a contrary result to what was advocated by Myers (2012). One of our other observations was that when teams are of equal strength, the trailing team is more likely to score the next goal. This secondary finding has importance for the game of soccer where either by coaching instruction or by psychology, the leading team goes into a defensive shell. We have demonstrated that this is detrimental to the team that is leading. We suggest that Jose Mourinho's "parking of the bus" is not the way that teams should play when they are leading.

Chapter 4 considers tactics in the sport of Twenty20 cricket. This chapter is a copy of the paper by Silva, Perera, Davis and Swartz (2016) which has appeared in the South African Statistical Journal. Twenty20 is the most recent format of cricket that has a huge following. This is an extremely practical chapter (paper) where we explain how teams may improve their chances of winning based on some simple strategies. Since Twenty20 is so young, many of its tactics have been borrowed from the sport of one-day cricket. However, Twenty20 and one-day cricket are different games and we have attempted to exploit inefficiencies in the way that Twenty20 is currently played. We believe that this paper has enormous potential for changing the way that Twenty20 is played.

Chapter 5 explores new definitions for pace of play in ice hockey. We will attempt to either blog this chapter or publish it in a journal. The proposed definitions borrow on notions from soccer and involve calculating distance that the puck moves forward with possession. To our great surprise, we observed that pace does not correlate with expected quantities such as goals scored and shots taken. Therefore this chapter may be seen as a negative result.

Chapter 2

A Study of the Powerplay in One-Day Cricket

2.1 Introduction

In the major sports of the world, rule changes are typically considered with great care. For example, FIFA (Fédération Internationale de Football Association) has made very few significant rule changes in soccer over the last 44 years (www.fifa.com). In 1992, legislation was introduced whereby goalkeepers were henceforth forbidden from handling back-passes. The only other significant rule change in soccer during the period concerned the offside rule. The offside rule has been twice liberalized (1995 and 2005) whereby offsides are now less common. Similarly, baseball is a sport steeped in tradition where there is a reluctance to alter the way that the game is played. In Major League Baseball (MLB), one may point to the introduction of the designated hitter in 1973 as the most recent significant rule change (www.baseball-almanac.com/rulechng.shtml). Wright (2014) provides a survey of the analysis of sporting rules from the perspective of operational research (OR).

In contrast to the stability of rules (laws) involving many of the major sports, one-day cricket has tinkered continuously with its powerplay rule. One-day cricket was introduced in the 1960s as an alternative to traditional forms of cricket that can take up to five days to complete. With more aggressive batting, colorful uniforms and fewer matches ending in draws, one-day cricket has become very popular. In the early days of one-day cricket, fielding restrictions were introduced as an additional strategy for making the game more exciting and popular. In simple terms, the powerplay imposes fielding restrictions that encourages aggressive batting and the scoring of runs. More specifically, fielding restrictions on the bowling team are in place during the full 50 overs of an innings. During powerplay overs, the level of fielding restrictions is increased whereby there are fewer fielders allowed in the outfield which may encourage the batting team to play more attacking type shots.

Although fielding restrictions have existed in one-day cricket since the 1996 World Cup, the term “powerplay” was introduced by the International Cricket Council (ICC) in 2005. And since 2005, there have been four distinct implementations of the powerplay rule. This paper investigates the four versions with a specific focus on whether powerplays really do increase run production. Although it may appear self-evident that run scoring increases during the powerplay, it is conceivable that aggressive batting leads to more wickets which in turn results in fewer runs. This is the line of reasoning which has initiated our investigation.

There are various practical questions associated with our investigation. For example, is the run scoring and wicket taking properties associated with the powerplay in line with the desires of the ICC? Also, in-game wagering has become extremely popular with online sportsbooks (`http://bleacherreport.com/articles/54254`). Accordingly, are in-game wagering odds properly reflected by the onset of the powerplay? Other questions involve strategic implications of the powerplay. For example, in what over should a team invoke the powerplay? Moreover, is an individual’s level of batting aggressiveness appropriate during the powerplay?

To our knowledge, there have not been any previous investigations on the effect of the powerplay. However, there are many data analytic studies concerning one-day cricket that have an OR focus. To get a sense of the variety of problems that have been addressed in one-day cricket, we mention a few recent papers. Most notably, Duckworth and Lewis (2004) developed the standard approach for the resetting of targets in rain interrupted matches. The approach known as the “Duckworth-Lewis method” has been adopted by all prominent cricketing boards and is based on the concept of resources which is a function of overs remaining and wickets taken. Following the seminal work of Duckworth and Lewis (2004), there have been various modifications and proposals for the resetting of targets (e.g. McHale and Asif 2013). Various authors including Allsopp and Clarke (2004) and Fernando, Manage and Scariano (2013) have investigated the effect of the home team advantage in one-day cricket. This is obviously important for match prediction. A topic of interest in every sport is player evaluation. Whereas in some sports, the measurement is straightforward, cricket performance involves a combination of batting, bowling and fielding contributions. In limited overs cricket, van Staden (2009) developed some simple and intuitive graphical displays to investigate batting and bowling performances. Valero and Swartz (2012) dispelled the myth that there are synergies in opening partnerships. It is argued that batsmen are not affected by the performance of their partners. Team selection is a problem of real interest to cricketing sides. Lemmer (2013) considered integer optimization methods for team selection. Swartz, Gill, Beaudoin and de Silva (2006) extended the problem to the determination of optimal batting orders using a simulated annealing algorithm. Norton and Phatarfod (2008) used dynamic programming to produce an optimal run scoring strategy for the batting team in both the first and second innings.

In section 2.2, the data are introduced and the four historical versions of the powerplay are described. Section 2.3 is concerned with the construction of hypothetical parallel matches. We take a “what if” approach where powerplay outcomes are substituted with what might have happened had there been no powerplay. This leads to a paired comparisons setting consisting of actual matches and parallel matches where outcomes are imputed during the powerplay period. Section 2.4 carries out the powerplay analyses by comparing the actual matches with the parallel matches. We investigate the difference in run production and the number of wickets taken with respect to the various powerplay rules. We also investigate the difference in run production with respect to the over where the powerplay was initiated. Section 2.5 provides a Bayesian analysis of individual batsmen and their ability to take advantage of the powerplay. We then do likewise for bowlers. We conclude with a short discussion in section 2.6.

2.2 Data and History of the Powerplay

For the analysis, we considered all ODI (one-day international) matches that took place from July 7, 2005 until the end of 2013 which involved full member nations of the International Cricket Council (ICC). Currently, the 10 full members of the ICC are Australia, Bangladesh, England, India, New Zealand, Pakistan, South Africa, Sri Lanka, West Indies and Zimbabwe. Details from these matches can be found via the Archive link at the CricInfo website (`www.espncricinfo.com`).

For these matches, only first innings data were considered. The rationale is that we want to study the powerplay under baseline circumstances. A team’s batting behaviour (aggressive versus passive) in the second innings depends largely on the target score that was established in the first innings. We excluded matches that were discontinued or were shortened to less than 50 overs. We also excluded 197 matches where we were unsure about the starting and ending points of the powerplay. In total, we were left with 597 matches involving reliable full first innings data.

For the imputation methods of section 2.3, we require detailed batting results, at the level of balls bowled. This information does not appear to be generally available in a convenient format. Hence, a proprietary R-script was developed and used to parse and extract ball-by-ball information from the Match Commentaries. For each first innings, we have two rows of data with 300 columns. In the jth column of the first row, we record the number of runs scored on the jth ball bowled (with extras included). In the jth column of the second row, we record either 0/1 according to whether a wicket was taken on the the jth ball bowled. Some additional columns were also recorded such as the match identifier, the batting team, the bowling team and the beginning and ending over for the batting powerplay. This results in a large dataset with $2(597) = 1194$ rows and 305 columns.

We now review the various historical implementations of the powerplay during the period of study, July 7, 2005 through 2013. We sometimes found it difficult to pin down details regarding the history of the powerplay. Some of our information was obtained from the following web sources:

- http://news.bbc.co.uk/sport2/hi/cricket/rules_and_equipment/4180026.stm
- http://voices.yahoo.com/cricket-power-play-rules-one-day-internationals-4720834.html
- http://www.espncricinfo.com/natwestchallenge/content/story/213010.html
- http://www.itsonlycricket/entry/106/

A: July 7/2005 - September 6/2008 - We have 239 observed matches where the match identifiers range from 2259 through 2762. During this period, there were three blocks of powerplays which imposed stricter fielding restrictions compared to the rest of the match. The first 10 overs of the innings imposed fielding restrictions which allowed only two fielders outside the 30-yard circle and two fielders within 15 yards of the on-strike batsman. This was known as the mandatory powerplay. The mandatory powerplay was followed by a five-over block known as powerplay 2 and a subsequent five-over block known as powerplay 3. The initiation of the two non-fixed powerplays were determined at the discretion of the bowling team. In both powerplays, the fielding restrictions allowed only three fielders outside of the 30-yard circle. If no powerplay had been initiated, then overs 41 through 50 automatically became powerplays. If only one powerplay had been initiated, then overs 46 through 50 automatically became powerplay 2.

B: October 9/2008 - September 20/2011 - We have 224 observed matches where the match identifiers range from 2763 through 3197. Rule B is the same as Rule A except that the start of one of the discretionary powerplays became the decision of the batting team. Hence the nomenclature for the two discretionary powerplays became the "bowling powerplay" and the "batting powerplay" accordingly. Although it is technically possible for the batting powerplay to precede the bowling powerplay, this did not occur in any of the 224 matches. The rationale for the introduction of Rule B was based on the observation that under Rule A, the bowling team often employed powerplays 2 and 3 as soon as possible (i.e. in overs 11-15 and 16-20, respectively). With the decision to start one of the powerplays given to the batting team, the hope was to spread the powerplays throughout the innings.

C: October 13/2011 - September 5/2012 - We have 51 observed matches where the match identifiers range from 3198 through 3304. Rule C is similar to Rule B except that the bowling and batting powerplays were not allowed to take place during overs 11 through 14 nor during overs 41 through 50.

D: November 4/2012 - December 25/2013 - We have 83 observed matches where the match identifiers range from 3305 through 3448. Rule D is the current rule and again requires that the mandatory powerplay takes place during the first 10 overs with the same fielding restrictions that allow only two fielders outside the 30-yard circle. Under Rule D, the bowling powerplay has been dropped. The batting powerplay (a five-over block) must be completed by the 40th over. For the batting powerplay, the fielding restrictions allow only three fielders outside the 30-yard circle.

2.3 Construction of Parallel Matches

The construction of hypothetical parallel matches is based on a "what if" approach where powerplay outcomes are substituted with what might have happened had there been no powerplay. Clearly, we want the imputations to be as realistic as possible, and this should take into account the nuances of the actual matches. For example, in a particular match, it is possible that pitch conditions are poor and batting is subsequently difficult. In this case, we impute overs that reflect the difficulty of scoring runs.

The study focuses on powerplay 3 under Rule A and the batting powerplays (for Rules B, C and D). Consequently, we construct parallel matches that only involve the imputation of powerplay 3 and the batting powerplays. The batting powerplays are of particular interest since most of the rule changes have involved the batting powerplay. Not only are we interested in the effect of the batting powerplay, but we are also interested in the effect due to the over where the batting powerplay is initiated.

We now describe the imputation procedures. The procedures depend on what aspects of the parallel match need imputation and on the information that is available from the corresponding actual match. Table 2.1 provides a summary broken down according to the various powerplay rules and imputation procedures. Whereas Rules A, C and D mostly use imputation method (a), we observe that Rule B uses imputation method (b) roughly 50% of the time. This is because under Rule B, the batting team frequently chose its powerplay in the final overs of the match.

Imputation Method (a): The simplest imputation procedure occurs when the actual batting powerplay is surrounded by 2.5 non-powerplay overs preceding the powerplay and 2.5 non-powerplay overs following the powerplay. For the parallel match, we substitute the powerplay results (both runs and wickets) with what happened during the surrounding overs. The surrounding overs are intended to be a fair representation of how the match would have proceeded had there been no powerplay. For the portion of the parallel match prior to the powerplay, we simply substitute what happened during the actual match. For the portion of the parallel match following the powerplay, we also substitute what happened during the actual match until the parallel innings terminate (i.e. 10 wickets lost or 50 overs consumed). Under these most basic conditions, Figure 2.1 provides a pictorial aid of the imputation procedure. However, if the actual match terminates earlier than the parallel match, there is no corresponding match history for imputation. In this case (which is rare - 68 out of 597 observed matches), the parallel match has $w < 10$ wickets taken at the point in time where the actual match terminated (i.e. the 10th wicket occurred). We then replicate the results from wicket w in the actual match until the end of the parallel match. We make sure that we skip (do not impute from) the powerplay; if the powerplay has x wickets, then our imputation begins from wicket $w - x$ in the actual match. There are only 18 such cases out of the 68.

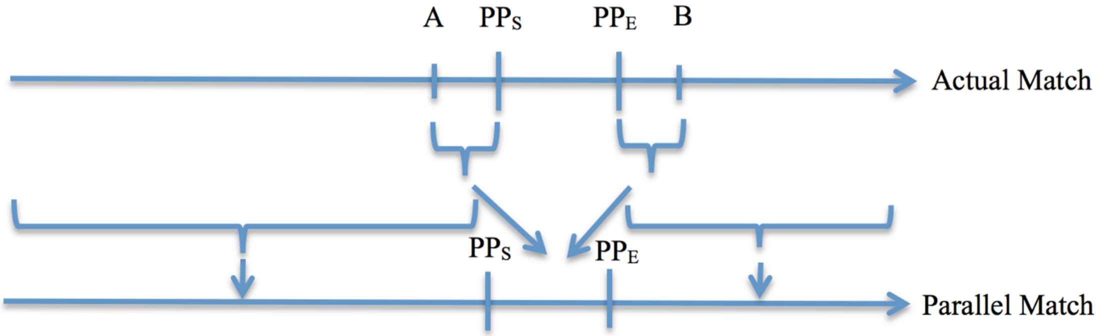

Figure 2.1: Imputation method (a) under the most basic conditions. The symbol A denotes 2.5 overs preceding the start of the powerplay ($\mathrm{PP_S}$) and the symbol B denotes 2.5 overs following the end of the powerplay ($\mathrm{PP_E}$).

Imputation Method (b): When the actual match does not have overs following the powerplay, Imputation Method (a) cannot be used. For example, this occurs when the powerplay takes place during overs 46 through 50. Our approach here is to take the ten overs preceding the powerplay and divide it into two blocks of 5 overs. Let R_1 be the number of runs scored in the first block and let R_2 be the number of runs scored in the second block. Then the ratio R_2/R_1 captures the change in scoring rate as the match progresses over the ten over interval. The number of imputed runs for the 2.5 overs following the powerplay is then set at $R_2(R_2/R_1)$ suggesting that the run scoring rate should change in this period by the same factor. Since we do not want R_2/R_1 to be unrealistically small or large, we do not permit the ratio to be less than 0.67 or greater than 1.5. We carry out the same procedure with the number of wickets. We experimented with values other than (0.67,1.5) and did not observe meaningful differences.

Powerplay Rule	Total Number of Matches	Imputation Method (a)	(b)
A	239	238	1
B	224	110	114
C	51	51	0
D	83	80	3

Table 2.1: Summary of the imputation procedures.

2.4 Powerplay Analyses

With the construction of the hypothetical parallel matches, we have a dataset satisfying a paired comparisons framework. For every match involving a batting powerplay, we have a parallel match where batting outcomes are imputed as though there were no batting powerplay. This facilitates an analysis where we can look at the difference in run production between the actual match and its corresponding parallel match. Let $R_i^{(a)}$ be the number of first innings runs scored in the ith actual match and let $R_i^{(p)}$ be the number of first innings runs scored in the ith parallel match. Then the quantities of interest are the differences

$$D_i = R_i^{(a)} - R_i^{(p)}. \tag{2.1}$$

We also study the difference in the actual number of wickets lost during the powerplay and the number of wickets that were lost during the same window in the parallel match.

2.4.1 Powerplay Rule A

Figure 2.2 provides a histogram of the differences D_i in (2.1) corresponding to powerplay 3 under Rule A. The median and the mean of the dataset are 1.0 and 1.3 runs respectively suggesting that the powerplay had a minor influence on increasing the number of runs scored. A Wilcoxon test of the hypothesis of no effect (i.e. H_0 : median = 0 versus H_1 : median > 0) was carried out and the null hypothesis was not rejected with the p-value = 0.08. Therefore, the effect (i.e. the number of increased runs due to the powerplay) is insignificant.

In a comparison of wickets lost during powerplay 3 versus wickets lost during the corresponding period of the parallel match, a Wilcoxon test was also carried out. The p-value = 0.45 was obtained indicating that there was no increase in the number of wickets taken due to powerplay 3. There were 0.01 more wickets taken on average during powerplay 3.

2.4.2 Powerplay Rules B, C and D

In this subsection, powerplay rules B, C and D are combined since they only differ in terms of when the batting powerplay is allowed to take place. This provides us with a dataset of 358 matches. Having seen that powerplay 3 under Rule A conferred no advantage to the batting team, it is interesting to investigate the revised Rules B, C and D. Recall that Rules

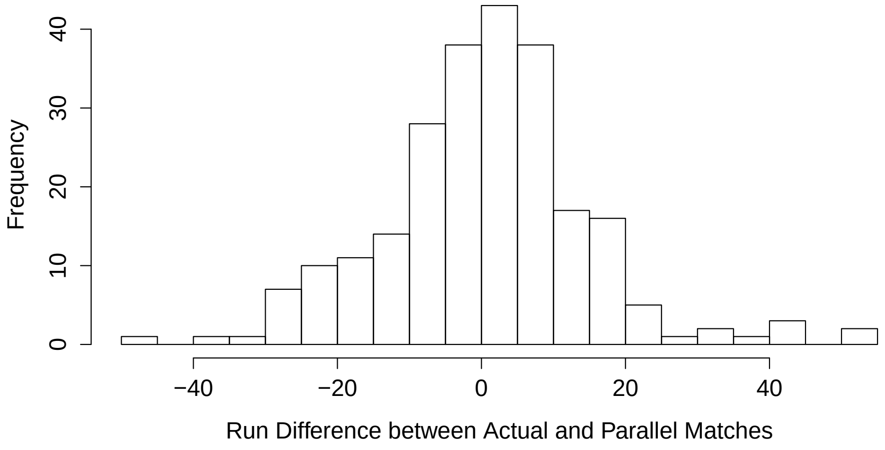

Figure 2.2: Histogram of the 239 differences D_i corresponding to Rule A.

B, C and D put the determination of the batting powerplay into the hands of the batting team. Perhaps the ICC had observed that Rule A was not accomplishing much in terms of increased run production, and that tinkering with the powerplay rule was required.

Figure 2.3 provides a histogram of the differences D_i in (2.1) corresponding to the batting powerplay under Rules B, C and D. The median and the mean of the dataset are 6.0 and 6.5 runs respectively suggesting that the powerplay provides increased run production. A Wilcoxon test of the hypothesis of no effect (i.e. H_0 : median $= 0$ versus H_1 : median > 0) was carried out and the null hypothesis was strongly rejected with the p-value $= 6.7 * 10^{-12}$. So it appears that the powerplay changes implemented by the ICC had the effect of increased run production.

In a comparison of wickets lost during the batting powerplay versus wickets lost during the corresponding period of the parallel match, a Wilcoxon test was also carried out. The p-value $= 0.007$ was obtained indicating that the number of wickets taken during the battng powerplay was greater than had there been no powerplay. There were 0.17 more wickets taken on average during the batting powerplay.

If we compare Figure 2.2 with Figure 2.3, we see that the run difference under Rules B, C and D is greater but is also more variable than under Rule A. This implies that the newer powerplay rules create more runs but also introduce greater uncertainty in the match. The new versions of the powerplay provide more runs for the batting team if they can avoid losing wickets. However, if their aggressiveness during the powerplay leads to an increased number of lost wickets, then the powerplay is detrimental to the batting team.

Recall that the imputation procedure for a parallel match was based on the assumption that had there not been a batting powerplay, then the results during these overs would resemble the surrounding overs. Although apparently reasonable, it is good to check the

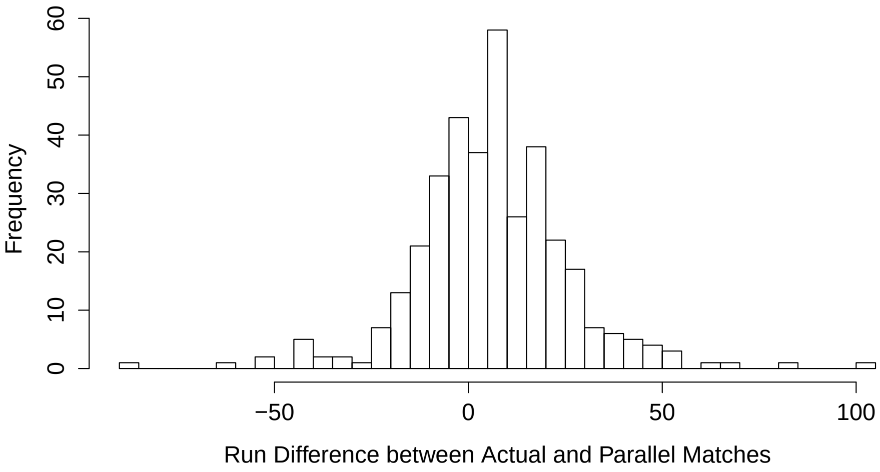

Figure 2.3: Histogram of the 358 differences D_i corresponding to Rules B, C and D.

sensitivity of the assumption. In the imputation procedure, the surrounding overs of a 5-over batting powerplay were defined as the 2.5 overs (i.e. 15 balls) preceding the powerplay and the 2.5 overs (i.e. 15 balls) following the powerplay. We now modify the assumption and instead consider the surrounding 8 balls preceding the powerplay and the 7 balls following the powerplay. Then the number of runs observed during these $8 + 7 = 15$ balls are scaled and imputed where the powerplay occurred. This provides us with a comparison set of 358 parallel matches. Under this alternative imputation procedure, we observed a mean difference of 6.3 more runs during the actual matches than the parallel matches. This is comparable to the 6.5 mean run difference under the original imputation procedure. This suggests a robustness of the proposed imputation procedure.

It is also worth asking whether what happens in the surrounding overs is affected by them being just before or just after the powerplay. To investigate this to some extent, we obtained the mean number of runs scored in the third over prior to the powerplay (4.9), the second over prior to the powerplay (5.6) and the over immediately prior to the powerplay (5.3). The run scoring pattern provides no indication of a change in tactics prior to the powerplay.

We now investigate the effect of the powerplay with respect to the over where the powerplay was initiated. We aggregate all 597 matches. The corresponding plot is given in Figure 2.4 along with a smoothed lowess curve to assess general features. We observe that most of the powerplays in our dataset were initiated in the vicinity of three time points: the 16th over, the 36th over and the 46th over. We also observe that the batting powerplay under Rule B was invoked near the end of the innings in the majority of matches.

Further investigation of Figure 2.4 reveals that under Rule A, powerplay 3 was typically initiated early in the innings. Specifically, powerplay 3 was initiated in the 16th over in 193

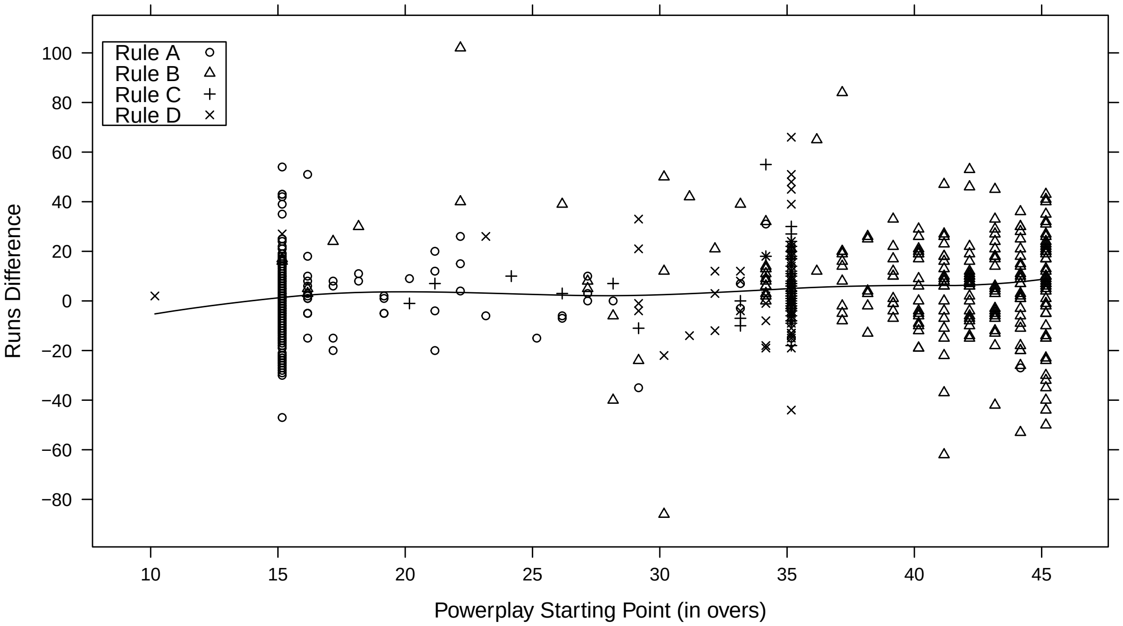

Figure 2.4: The run differences D_i for all rules A, B, C and D plotted against the over where the powerplay was initiated. A lowess curve with parameters span=0.5 and degree=2.0 is superimposed.

of the 239 matches. We hypothesize that the reason why the powerplay was ineffective at this stage was because it was early in the match and the batting team did not want to take the risk of batting aggressively and losing wickets. Recall that the timing of powerplay 3 was at the discretion of the bowling team.

Under Rules B, C and D, the initiation of the batting powerplay was at the discretion of the batting team, and this appears to have made a positive difference in run production. A very close inspection of the smoothed curve suggests that it may not have been advantageous to initiate the powerplay late in the match, say beyond the 41st over. Our intuition here is that when the batting powerplay begins late, the batsmen involved in the powerplay are typically weaker batsmen in the lineup and are unable to take advantage of the situation. Of course, under the current Rule D, it is no longer possible to initiate the batting powerplay beyond the 36th over.

In Figure 2.4, there is also some evidence that initiating the powerplay near the 36th over may be optimal. This stage of the match provides a compromise; it is sufficiently late in the match that the batsmen are free to bat aggressively, and it is sufficiently early in the match so that good batsmen (i.e. mid-order batsmen) are typically available for batting. The smoothed curve also suggests that it may also be beneficial to begin the batting powerplay around the 21st over. However, we are not convinced of this due to the sparsity of data at this stage of the match.

2.5 Individual Batsmen and Bowlers

We have seen that the current implementation of the powerplay contributes on average 6.5 additional runs of scoring had there not been a powerplay. It is interesting to investigate whether some batsmen take a greater advantage of the powerplay conditions than other batsmen.

We therefore considered 45 batsmen during the study period (i.e. rules B, C and D) who faced a minimum of 600 balls. Details on these batsmen are provided in Table 2.2. The number of balls faced during the powerplay varies greatly among batsmen where Kumar Sangakkara of Sri Lanka faced 416 balls and Alastair Cook of England faced 53 balls. All of the batsmen faced many more balls (typically about ten times as many) during non-powerplay conditions.

With many balls faced, we appeal to the Central Limit Theorem and define

$$\begin{aligned} X_i^{(1)} \equiv &\ \text{run rate per over for the } i\text{th batsmen during PP} \sim \text{Normal}(\mu_i^{(1)}, \sigma^2/n_{1i}) \\ X_i^{(2)} \equiv &\ \text{run rate per over for the } i\text{th batsmen during non-PP} \sim \text{Normal}(\mu_i^{(2)}, \sigma^2/n_{2i}) \end{aligned}$$

where n_{1i} and n_{2i} are the number of balls faced by the ith batsman during powerplay conditions and non-powerplay conditions respectively. In a Bayesian analysis, we further define prior distributions

$$\begin{aligned} \mu_i^{(1)} &\sim \text{Normal}(\mu_0^{(1)}, \sigma_\mu^2) \\ \mu_i^{(2)} &\sim \text{Normal}(\mu_0^{(2)}, \sigma_\mu^2) \\ \sigma^2 &\sim \text{Inverse Gamma}(1,1) \\ \sigma_\mu^2 &\sim \text{Inverse Gamma}(1,1) \end{aligned}$$

where $\mu_0^{(1)}$ and $\mu_0^{(2)}$ are set according to the sample means of the dataset (i.e. an empirical Bayes approach). The hyperparameters of the Inverse Gamma distributions provide standard reference priors. The non-constant values n_{1i} and n_{2i} provide a twist that prevents a straightforward classical analysis.

We implemented the model using WinBUGS software (Spiegelhalter, Thomas and Best 2003) where our primary interest concerns the parameters $\mu_i^{(1)}$ and $\mu_i^{(2)}$ for $i = 1, \ldots, 45$. A WinBUGS implementation is straightforward as the generation of parameters from the posterior distribution is done in the background. A user is only required to specify the statistical distributions. We ran 15,000 iterations of the Markov chain where 5,000 iterations were used as burn-in. Standard diagnostics provide evidence that the Markov chain has adequately converged.

In Figure 2.5, we provide boxplots of the differences $\mu_i^{(1)} - \mu_i^{(2)}$ generated from the Markov chain where the boxplots are sorted in ascending order of mean difference. We observe that nearly all of the mean differences exceed zero which implies that batsmen score runs at a higher rate during the powerplay. The only exceptions to this are T.

Iqbal of Bangladesh where $E(\mu_i^{(1)} - \mu_i^{(2)}) = -0.26$ and M. Guptill of New Zealand where $E(\mu_i^{(1)} - \mu_i^{(2)}) = -0.05$. At the other end of the plot, the greatest value of $E(\mu_i^{(1)} - \mu_i^{(2)})$ is attributed to M. Mahmudullah of Bangladesh for whom $E(\mu_i^{(1)} - \mu_i^{(2)}) = 2.35$. This difference translates to $5(2.35) = 11.75$ extra runs during the five-over powerplay block than during non-powerplay conditions. Also notable among the exceptional powerplay batsmen are M. Hussey of Australia where $E(\mu_i^{(1)} - \mu_i^{(2)}) = 2.28$ and S. Marsh of Australia where $E(\mu_i^{(1)} - \mu_i^{(2)}) = 2.19$. The batsman with the highest mean value of $\mu^{(1)}$ is V. Sehwag of India where $E(\mu_i^{(1)}) = 7.96$. Not only is Sehwag great during the powerplay but he is great at all times with $E(\mu_i^{(2)}) = 7.19$ and therefore his mean difference is only $E(\mu_i^{(1)} - \mu_i^{(2)}) = 0.77$.

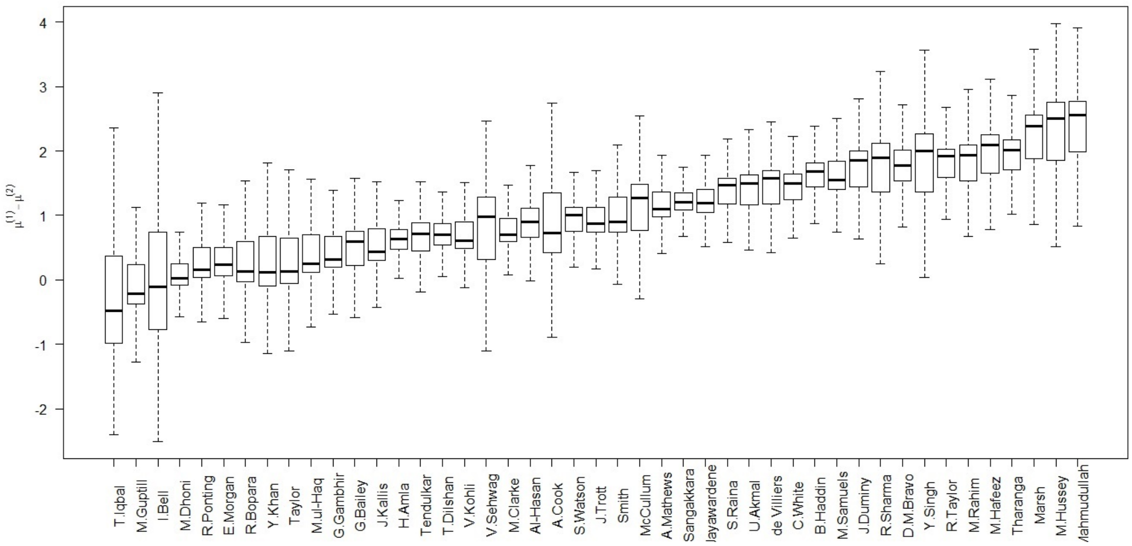

Figure 2.5: Boxplots of the differences $\mu_i^{(1)} - \mu_i^{(2)}$ for the 45 batsmen based on output from the Markov chain corresponding to the hierarchical model of section 2.5.

For bowlers, we carry out a similar analysis to investigate individual powerplay performances. Here we have obtained data on 28 bowlers during the study period (i.e. rules B, C and D) who have bowled at least 1000 balls. Details on these bowlers are provided in Table 2.3. The inferential quantities of interest for the ith bowler are the mean run rate per over during the powerplay $\mu_i^{(1)}$ and the mean run rate per over $\mu_i^{(2)}$ during non-powerplay conditions. In cricket parlance, $\mu_i^{(1)}$ and $\mu_i^{(2)}$ are referred to as the mean economy rates. The economy rate is often regarded as more important than both the bowling strike rate and the bowling average in one-day cricket. The bowling strike rate is defined as the average number of balls bowled per wicket and the bowling average is defined as the average number of runs conceded per wicket.

In Figure 2.6, we provide boxplots of the differences $\mu_i^{(1)} - \mu_i^{(2)}$ where the boxplots are sorted in ascending order of mean difference. In this case, it is bowlers on the left side of the plot who have performed exceptionally during the powerplay. We observe that the only bowler with a negative value of $\mu_i^{(1)} - \mu_i^{(2)}$ is A. Mathews of Sri Lanka with $\mu_i^{(1)} - \mu_i^{(2)} = -0.07$. This implies that he bowls better during the powerplay than during non-powerplay overs. Of course, this may simply be a case of small sample size as Matthews has bowled only 116 powerplay balls. The median value of the $\mu_i^{(1)} - \mu_i^{(2)}$ values amongst the 28 bowlers is 1.15 which says that the median bowler allows 5(1.15) = 5.75 more runs on average during the five-over powerplay block than during non-powerplay overs. At the right end of the plot is P. Kumar of India for whom $\mu_i^{(1)} - \mu_i^{(2)} = 2.03$. This implies that Kumar allows on average 5(2.03) = 10.15 more runs during the five-over powerplay block than during non-powerplay overs.

In Table 2.3, we have distinguished the bowlers as either fast or spin bowlers. A cursory inspection of Figure 2.6 indicates that fast bowlers tend to be situated in the rightmost boxplots. This suggests that spin bowlers adjust better to the powerplay overs than do fast bowlers. To test this formally, we divide the 28 bowlers according to 10 spinners and 18 fast bowlers. We then carry out a two-sample t-test on the null hypothesis of no difference between the two types of bowlers where our observations are the posterior means of $\mu_i^{(1)} - \mu_i^{(2)}$. With the p-value = 0.045, we reject the hypothesis using a two-tailed test. This shows that there is a significant difference between the economy rates of the fast bowlers and the spin bowlers during the powerplay.

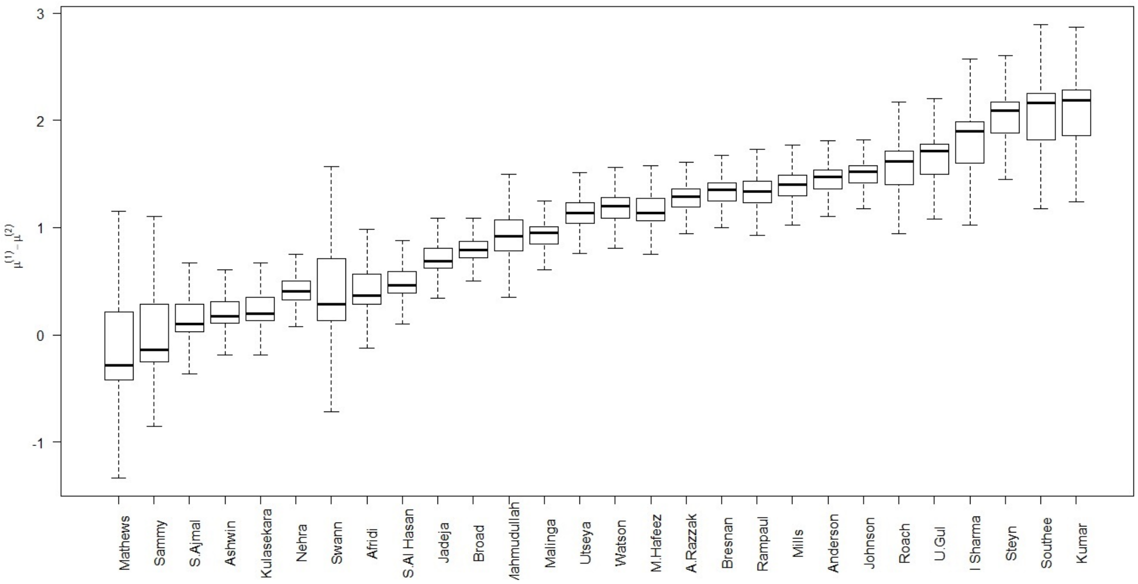

Figure 2.6: Boxplots of the differences $\mu_i^{(1)} - \mu_i^{(2)}$ for the 28 bowlers based on output from the Markov chain corresponding to the hierarchical model of section 2.5.

2.6 Discussion

This paper appears to be the first quantitative investigation on the effect of the powerplay in one-day cricket. The main result is that recent versions of the powerplay rule contribute an average of 6.5 additional runs. However, the contribution of increased runs is countered by an increase in the number of lost wickets which adds uncertainty to the match. Furthermore, the choice of over where the powerplay is initiated has some effect on the number of runs scored. It appears that initiating the batting powerplay in the 36th over is roughly optimal.

Based on the above findings, there are possible implications for the game:

1. The ICC may want to again revisit the powerplay with a focus on the intention of the powerplay. If the intention is to create more runs, we have now quantified the average number of increased runs. Is 6.5 runs adequate? Altering the fielding restrictions may further modify run scoring. Of course, the ICC may be wary of changing the powerplay rule once again. If a goal of the batting powerplay is to introduce more uncertainty into the game, then this appears to have been accomplished since the average number of wickets lost during the powerplay is greater than had there been no powerplay. When the batting team loses large numbers of wickets during the powerplay, then their run production decreases.

2. There may be strategic implications for the powerplay. Although invoking the powerplay around the 36th over appears to be roughly optimal, teams may want to consider their batting style (i.e. aggressive versus passive) during the powerplay. They may be able to invoke the powerplay earlier if they tone down their level of aggressiveness. The advantage of initiating the powerplay earlier is that early-order batsmen may be able to take better advantage of the powerplay opportunity.

3. One might ask “what are the implications of this study for Twenty20 cricket?” In Twenty20 cricket, the powerplay is mandated to take place during the first six overs of each innings when the best batsmen are typically batting. At the end of the 6th over, there are 14 overs remaining since Twenty20 matches are allotted 20 overs. In one-day cricket, the optimal completion of the powerplay occurs roughly at the end of the 40th over (i.e. 10 overs remaining). Also, in Twenty20, losing wickets is less of a concern for the batting side than in one-day cricket. Therefore, there is some suggestion that the timing of the powerplay in Twenty20 may be optimal in terms of creating additional runs.

Name	Runs(non-PP)	Balls Faced(non-PP)	Runs(PP)	Balls Faced(PP)
K. Sangakkara (SL)	1358	1804	395	416
Misbah-ul-Haq(PAK)	1172	1631	257	347
M. Clarke (AUS)	1373	1737	310	345
R.Ponting (AUS)	1039	1222	277	321
A.B. de Villiers (SA)	1705	1782	384	311
H. Amla (SA)	1793	1868	325	304
M.S. Dhoni(IND)	1795	1800	297	298
R. Taylor (NZ)	1035	1274	311	272
A. Mathews (SL)	940	1295	235	260
S. Raina (IND)	1044	1184	286	252
J.P. Duminy (SA)	992	1127	301	249
M.Hussey (AUS)	1181	1314	325	239
S. Watson (AUS)	1338	1420	261	234
T.M. Dilshan (SL)	1482	1607	239	230
V. Kohli (IND)	1005	1233	204	225
Y. Khan (PAK)	909	1227	161	222
J. Trott (ENG)	1012	1260	208	221
M. Jayawardene (SL)	834	1097	210	220
M. Guptill (NZ)	1128	1238	186	218
B. Haddin (AUS)	637	857	223	217
M. Hafeez (PAK)	1268	1559	237	200
Y. Singh (IND)	801	804	268	195
G. Gambhir (IND)	890	1010	179	194
M. Mahmudullah (BAN)	662	864	237	193
B. Taylor (ZIM)	709	893	146	186
J. Kallis (SA)	587	693	164	181
R. Bopara (ENG)	433	504	153	178
U. Akmal (PAK)	712	838	184	166
V. Sehwag (IND)	1121	903	241	165
S.E. Marsh (AUS)	561	762	191	165
U. Tharanga (SL)	736	1045	170	162
C. White (AUS)	850	1065	161	153
M. Samuels (WI)	511	833	132	153
S. Tendulkar (IND)	609	587	171	147
E. Morgan (ENG)	680	665	143	135
G. Smith (SA)	670	913	117	135
B.B. McCullum (NZ)	728	716	155	123
T. Iqbal (BAN)	812	989	72	111
G. Bailey (AUS)	733	669	133	110
Shakib-Al-Hasan (BAN)	732	821	100	96
M. Rahim (BAN)	823	1094	99	91
D.M. Bravo (WI)	421	639	85	89
I. Bell (ENG)	635	802	44	68
R.G. Sharma (IND)	433	550	68	60
A. Cook (ENG)	618	831	43	53

Table 2.2: Summary data on the 45 batsmen considered in section 2.5.

Name	Style	Runs	Balls Bowled	Runs	Balls Bowled
		(non-PP)	(non-PP)	(PP)	(PP)
M. Johnson (AUS)	Fast	829	1129	496	501
R. Ashwin (IND)	Spin	1075	1344	408	496
S. Ajmal (PAK)	Spin	1160	1592	360	489
S. Broad (ENG)	Fast	827	978	446	457
L. Malinga (SL)	Fast	1057	1175	423	399
Shakib-Al-Hasan (BAN)	Spin	738	963	328	391
T. Bresnan (ENG)	Fast	886	1097	391	377
J. Anderson (ENG)	Fast	1085	1414	326	321
A. Nehra (IND)	Fast	830	845	336	320
U. Gul (PAK)	Fast	828	953	368	317
R. Jadeja (IND)	Spin	1526	2012	272	313
A. Razzak (BAN)	Spin	787	1009	311	312
S. Watson (AUS)	Fast	610	774	299	302
N. Kulasekara (SL)	Fast	1131	1307	252	282
D. Steyn (SA)	Fast	706	984	276	257
S. Afridi (PAK)	Spin	1566	2023	210	254
I. Sharma (IND)	Fast	674	773	298	248
P. Utseya (ZIM)	Spin	958	1209	240	244
P. Kumar (IND)	Fast	1219	1522	280	237
K. Mills (NZ)	Fast	769	1093	222	237
T. Southee (NZ)	Fast	925	1164	270	231
M. Hafeez (PAK)	Spin	969	1399	197	224
R. Rampaul (WI)	Fast	692	952	210	221
D. Sammy (WI)	Fast	773	1022	133	186
K. Roach (WI)	Fast	664	853	168	159
A. Mathews (SL)	Fast	954	1161	87	116
M. Mahmudullah (BAN)	Spin	838	1032	55	57
G. Swann (ENG)	Spin	948	1206	46	57

Table 2.3: Summary data on the 28 bowlers considered in section 2.5.

Chapter 3

Analysis of Substitution Times in Soccer

3.1 Introduction

In the game of soccer (known as football outside of North America), teams are allowed three player substitutions in a match. The timing of the substitutions is strategic. For example, if a team is losing, the manager (coach) may want to replace a player with a more attacking player. On the other hand, teams should be wary of early substitutions. Once a team has made their three substitutions, a subsequent injury on the field may force the team to play the remainder of the match with 10 players instead of 11.

Myers (2012) proposed a substitution scheme based on regression tree methodology that analyzed data from the top four soccer leagues in the world: the 2009/2010 seasons of the English Premier League (EPL), the German Bundesliga, the Spanish La Liga and the Italian Serie A. In addition, data were analyzed from the 2010 season of North America's Major League Soccer (MLS) and from the 2010 FIFA World Cup. The decision rule for substitutions (page 11 of Myers, 2012) was succinctly stated as follows:

- if losing:
  - make the 1st substitution prior to the 58th minute
  - make the 2nd substitution prior to the 73rd minute
  - make the 3rd substitution prior to the 79th minute (3.1)
- if tied or winning:
  - make substitutions at will

The subsequent analysis in Myers (2012) demonstrated that teams that followed the decision rule improved their goal differential 42.27 percent of the time. For teams that did not follow the decision rule, they improved their goal differential only 20.52 percent of the time.

The decision rule (3.1) is attractive both in its apparent simplicity and also due to the benefits from following the rule. Consequently, the decision rule has received considerable attention in the mainstream media. For example, a quick Google search reveals YouTube interviews, blogs and newspaper articles concerning the study, many of which marvel at the findings (e.g. Diamond 2011 and Cholst 2013). Chapter 9 of Anderson and Sally (2013) endorses the results in Myers (2012). They argue that by the time managers observe that a player is tired, it is already too late. The substitution of the player ought to have occurred earlier. They suggest that the substitution rule proposed by Myers (2012) is an analytics-based approach that provides prescience beyond what managers are able to ascertain.

In this paper, we provide both a review of Myers (2012) and an alternative analysis of the soccer substitution problem. At a surface level, the results appear contradictory as our analysis indicates that there is no discernible time during the second half when there is a benefit due to substitution. However, as we discuss, the two approaches are not directly comparable as they use different statistical methodologies, different response variables and different explanatory variables. Our analysis also indicates that with evenly matched teams, the trailing team is more likely to score the next goal during the second half. This observation has implications for the game of soccer. Teams that are leading may be “parking the bus” or failing to send attackers forward in sufficient numbers. These tentative reactions or strategies are seemingly detrimental.

In Section 3.2, we carefully review the paper by Myers (2012). We begin by providing two examples where there are subtleties associated with the decision rule. In the first example, we note that the proposed substitution scheme is not entirely practical as it provides substitution directives that refer to earlier stages of a match. The two examples lead to a formal characterization of the decision rule. We then discuss various aspects of the analysis in Myers (2012). In Section 3.3, we present an alternative analysis that is based on Bayesian logistic regression where team strength is considered and subjective priors are utilized. The prior specification facilitates the smoothing of temporal parameters. We conclude with a short discussion in Section 3.4.

There are at least two other papers in the literature that have addressed substitution issues in soccer. Hirotsu and Wright (2002) use hypothetical soccer results to demonstrate the estimation of a four-state Markov process model. With such a model (which requires the estimation of player specific parameters), optimal substitution times may be obtained, optimal in the sense of maximizing league points. In Del Corral, Barros and Prieto-Rodriguez (2008), the substitution patterns from the 2004-2005 Spanish First Division are studied. They determine that the score of the match is the most important factor affecting substitutions. In addition, they find that defensive substitutions occur later in a match than offensive substitutions.

In our analysis, we consider the probability that the trailing team scores the next goal. However, scoring intensity is also relevant to soccer. It is well known that scoring intensity increases throughout a match (Morris 1981). For example, Ridder, Cramer and Hopstaken (1994) provided the total goals scored during the six 15-minute segments in a 90 minute match corresponding to the 340 matches played during the 1992 season in the two professional Dutch soccer divisions. Based on 952 goals, the percentages in the six segments were 13.4, 14.7, 15.4, 17.8, 17.9 and 20.8. They also demonstrated that after a red card is issued, the scoring intensity of the 11-man team increased by a factor of 1.88 whereas the scoring intensity of the 10-man team decreased only slightly by a factor of 0.95. Increased scoring intensity towards the end of matches was corroborated by Armatas, Yiannakos and Sileloglou (2007) who studied the 1998, 2002 and 2006 World Cups.

3.2 The Original Decision Rule

To gain a better understanding of the decision rule (3.1) proposed by Myers (2012), we consider two illustrative examples.

Example 3.1: Team A scores in the 50th minute. Team B substitutes in the 45th minute, substitutes in the 70th minute and then scores in the 75th minute.
Discussion: In this match, the conditions for use of the decision rule are applicable. The reason is that Team B is losing at the critical 73rd minute. Therefore, we see that the rule is not prospective - based on the score in the 73rd minute, it tells us how we should have substituted previously in the match. From a management perspective, it would be preferable to have a rule that provides decision guidelines at any point in time. We also see that the simple formulation (3.1) is not entirely clear in defining an instance of "when" a team is losing. In this example, Team B followed the decision rule and improved their goal differential.

Example 3.2: In an actual match (March 10, 2009) between Burnley and Birmingham in the English Premier League, the home team Burnley scored goals in the 53rd and 62nd minutes. Birmingham substituted in the 45th minute, the 45th minute, the 67th minute and then scored in the 90th minute.
Discussion: Here, Birmingham falls behind in the 53rd minute and remains behind for the entire match. Birmingham substitutes in accordance with the decision rule. The question arises as to whether Birmingham improved their goal differential. The final score of 2-1 (for Burnley) represents no change in differential from the 53rd minute (the time of Burnley's first goal to make the score 1-0). However, from the time of Birmingham's third substitution in the 67th minute when the score was 2-0 for Burnley, there is a positive change in differential by the end of the match. In a personal communication with Myers, he indicates that indeed Birmingham should be credited with an improved goal differential.

Therefore, the decision rule is more complex in its implementation than as simply specified by (3.1). Given that the rule has gained some traction in soccer, it is useful to have an unambiguous specification of the rule. We consider a formulation which is unfortunately more complicated than (3.1) but facilitates statistical analysis.

Accordingly, observe the first time t_0 that a team has fallen behind in a match and let $j(t_0)$ be the number of substitutions that the team has made prior to t_0. We define s_i as the time of the ith substitution and let SL(t) be true (false) if the team is losing (no longer losing) at time t. We further define the next substitution time $s_n = s_{1+j(t_0)}$ and the next critical time

$$t^* = \begin{cases} 58 & \text{if} \quad t_0 \leq 58 \\ 73 & \text{if} \quad 58 < t_0 \leq 73 \\ 79 & \text{if} \quad 73 < t_0 \leq 79 \end{cases}$$

Table 3.1 provides a breakdown of the 9 situations where the decision rule is applicable and the corresponding substitution patterns under which the decision rule is followed. When following the decision rule, a success in reducing the goal differential is defined by observing the change in goal differential between s_n and the 90th minute. When not following the decision rule, a success in reducing the goal differential is defined by observing the change in goal differential between t^* and the 90th minute.

Situations DR Applicable			Substitution Pattern Required to Follow DR		
$t_0 \leq 58$,	$j(t_0) = 0$,	$\text{SL}(s_1) = T$	$s_1 \leq 58$,	$s_2 \leq 73$ (if SL(73)=T),	$s_3 \leq 79$ (if SL(79)=T)
$t_0 \leq 58$,	$j(t_0) = 1$,	$\text{SL}(s_2) = T$	$s_2 \leq 73$,	$s_3 \leq 79$ (if SL(79)=T)	
$t_0 \leq 58$,	$j(t_0) = 2$,	$\text{SL}(s_3) = T$	$s_3 \leq 79$		
$58 < t_0 \leq 73$,	$j(t_0) = 0$,	$\text{SL}(s_2) = T$	$s_2 \leq 73$,	$s_3 \leq 79$ (if SL(79)=T)	
$58 < t_0 \leq 73$,	$j(t_0) = 1$,	$\text{SL}(s_2) = T$	$s_2 \leq 73$,	$s_3 \leq 79$ (if SL(79)=T)	
$58 < t_0 \leq 73$,	$j(t_0) = 2$,	$\text{SL}(s_3) = T$	$s_3 \leq 79$		
$73 < t_0 \leq 79$,	$j(t_0) = 0$,	$\text{SL}(s_3) = T$	$s_3 \leq 79$		
$73 < t_0 \leq 79$,	$j(t_0) = 1$,	$\text{SL}(s_3) = T$	$s_3 \leq 79$		
$73 < t_0 \leq 79$,	$j(t_0) = 2$,	$\text{SL}(s_3) = T$	$s_3 \leq 79$		

Table 3.1: The 9 situations under which the decision rule (DR) is applicable and the corresponding conditions under which the DR is followed.

3.2.1 Examination of the Original Decision Rule

In this subsection, we provide a discussion of various aspects of the analysis related to Myers (2012).

Recall, we have re-formulated the original decision rule (3.1) proposed by Myers (2012) with the description provided in Table 3.1. To check our characterization, we attempted to replicate the analysis in Myers (2012) using the formulation in Table 3.1. We aggregated results over the same six competitions as Myers (2012); namely the English Premier League 2009-2010 season, the German Bundesliga 2009-2010 season, the Spanish La Liga 2009-2010 season, the Italian Serie A 2009-2010 season, North America's Major League Soccer 2010 season and the 2010 World Cup held in South Africa. We obtained an improved goal differential 40.07 percent of the time when following the decision rule and 17.90 percent of the time when not following the decision rule. These results are very close to the values

42.27 percent and 20.52 percent reported by Myers (2012). Because of our limited data sources, we excluded matches with red cards and matches where substitutions occurred in the first half. These decisions likely account for the small discrepancies in the two analyses. Our replicated analysis was based on 292 occasions where teams followed the decision rule and 620 occasions where teams did not follow the decision rule.

We were concerned with sample size inadequacies in the above analysis, especially the 292 instances where the decision rule was followed. We therefore augmented the dataset by including three more English Premier League seasons (2010-2011, 2011-2012 and 2012-2013). This provided a total of 446 occasions where the decision rule was followed and 1,118 occasions where the decision rule was not followed. With the larger dataset, improved goal differential was achieved 39.01 percent of the time when following the decision rule and 20.48 percent of the time when not following the decision rule. We therefore observe that the difference between following the rule and not following the rule is slightly less than previously reported. In Section 3.3, an alternative analysis is presented which is based on a much larger dataset.

One of the assumptions of analyses based on regression trees is that observations are statistically independent. According to the formulation of the decision rule in Table 3.1, it is possible that both teams in a match may be subject to the decision rule. In this case, the two situations are not statistically independent. For example, if one team improves its goal differential, it is less likely that the opponent will improve its goal differential. The lack of independence is not taken into account in the analysis by Myers (2012). We note that the analysis presented in Section 3.3 does not have such issues.

In the analysis presented in Myers (2012), the decision rule is based on whether a team follows the 58-73-79 substitution pattern. It seems to us that any possible advantage due to a team's substitution pattern should also depend on their opponent's substitution pattern. The analysis in Myers (2012) does not take the opponent's substitution pattern into account. However, we note that the opponent's substitution pattern is considered in the analysis presented in Section 3.3.

A nuanced consideration of Myers (2012) is that the analysis is based on a comparison of following the 58-73-79 rule versus not following the 58-73-79 rule. There are many ways that teams can fail to follow the decision rule. For example, a team could follow a 60-73-79 rule but fail to follow the 58-73-79 rule. However, it is doubtful that there would be much difference in team performance between the recommended 58-73-79 rule and a 60-73-79 rule. When the 58-73-79 rule is compared against all other substitution patterns, it is possible that the rule is compared against some "bad" substitution patterns. Therefore, it would be preferable if substitutions could be compared at different points in time. The analysis presented in Section 3.3 provides such comparisons.

There is an aspect of the substitution analysis in Myers (2012) that is nonstandard and is highlighted in the following example.

Example 3.3: Team A scores in the 50th minute and Team B scores in the 56th minute. **Discussion:** We consider the substitution problem from Team B's perspective. We therefore have $t_0 = 50$, $j(t_0) = 0$, $s_n = s_1$ and $t^* = 58$. Following Table 3.1, if Team B substitutes in the 54th minute, we refer to the first row and note that the decision rule is applicable since $\mathrm{SL}(s_1 = 54) = T$. However, if Team B substitutes in the 57th minute, then the decision rule is not applicable since $\mathrm{SL}(s_1 = 57) = F$. What makes the analysis nonstandard is that the substitution protocol determines whether the match is a case in question.

3.2.2 Accounting for Team Strength

A final discussion point concerning Myers (2012) relates to the well-known fact that the assessment of cause and effect is best investigated using randomized experiments. However, in the soccer dataset, the decisions to follow the 58-73-79 rule were not randomized. It is possible that some confounding factor could have been involved, a factor that is related to the success of the decision rule.

When studying the decision rule, it is apparent that teams essentially follow the decision rule when they make their substitutions early, and we hypothesize that strong teams are more likely to substitute early. Strong teams tend to have "deeper" benches and are better able to replace players with quality players. Obviously, stronger teams are more able to improve goal differential.

To investigate the hypothesis, we define a variable that describes a team's relative strength in a given match. When determining the team's strength, we also account for home team advantage. Here we consider a balanced schedule where each team in a league plays every other team the same number of times, both home and away. For a given league in a given season, let HTA denote the league-wide home team advantage calculated as

$$HTA = \frac{\text{total home goals} - \text{total away goals}}{\text{total matches}} \ .$$

For Team j, define its average goal differential during a season by

$$D_j = \frac{\text{Team } j\text{'s total goals scored} - \text{Team } j\text{'s total goals allowed}}{\text{total matches by Team } j} \ .$$

Then, if Team j is playing Team k, we define the relative strength of Team j as

$$z = \begin{cases} D_j - D_k + HTA & \text{if } j\text{'s home field} \\ D_j - D_k - HTA & \text{if } k\text{'s home field} \end{cases} \tag{3.2}$$

where a positive (negative) value of z suggests that Team j (k) is favored to win the match.

The value z in (3.2) has a straightforward interpretation as the number of goals by which Team j is expected to defeat Team k. This interpretation is useful for the subjective priors that are developed in Section 3.1. Alternative measures of team strength have been developed for soccer including latent variable probit models (Koning 2000), extended dynamic models (Knorr-Held 2000) and various Poisson-type models (Karlis and Ntzoufras 2003). There are also alternative measures of the home team advantage. For example, Clarke and Norman (1995) use regression methods to obtain team specific measures for English soccer. Issues surrounding the use of team specific measures versus a single league-wide measure is discussed in Swartz and Arce (2014).

Having developed the team strength parameter z, we now return to the question of whether team strength is confounded with success of the decision rule. We use the dataset from Myers (2012) but exclude the 2010 World Cup results where the strength parameter z is unavailable. When teams are stronger, they follow the decision rule 37 percent of the time (105 times out of 283 opportunities). When teams are weaker, they follow the decision rule 30 percent of the time (177 times out of 589 opportunities). Moreover, stronger teams that followed the decision rule improved their goal differential in 56.19 percent of the cases (59 out of 105 times). This is a much higher value than the previously reported 40.07 percent success rate for following the decision rule.

It therefore appears that team strength is relevant to the success of the decision rule. Although team strength was not considered by Myers (2012), the analysis in Section 3.3 takes team strength into account.

3.3 An Alternative Analysis

In Myers (2012), regression trees were used to search over potential substitution times to determine an optimal substitution rule. Recall that optimality was based on improving goal differential. We consider a related approach that considers whether the trailing team scores the next goal. Therefore, the response variables are different in the new analyses. In addition, we use more data, we take into account the relative strength of the trailing team and we also consider the time of the match. Our analysis is based on Bayesian logistic regression using informative prior distributions.

We consider goals scored during all matches in the dataset where a team was trailing prior to the goal being scored. Recall that Myers (2012) considered the change in goal differential for which a team could have at most one observation per game. Accordingly, let $Y_i = 1(0)$ denote that the ith goal was scored by the trailing (leading) team where $i = 1, \ldots, n$. Then $Y_i \sim \text{Bernoulli}(p_i)$. Therefore, we do not consider goals that occur when the score is tied. Our focus is on the behavior of the trailing team.

Following (3.2), we let z_i denote the strength parameter of the trailing team which takes into account the home team advantage. We introduce the substitution variable s_i where the underlying assumption is that extra substitutions refresh or infuse energy to a team in the same way across all teams. Corresponding to the ith goal, we define

$$s_i = \begin{cases} 1 & \text{trailing team has made more substitutions than the leading team} \\ \text{-1} & \text{trailing team has made fewer substitutions than the leading team} \\ 0 & \text{trailing team has made the same number of substitutions as the leading team} \ . \end{cases}$$

This leads to the logistic model

$$\log\left(\frac{p_i}{1-p_i}\right) = \lambda z_i + \beta_{0t} + \beta_{1t} s_i \ . \tag{3.3}$$

In (3.3), we have attempted to incorporate the relevant factors that affect the probability of a goal being scored by the trailing team. The relative strength of the trailing team including the home team advantage is expressed through λz_i. It is also well-known that trailing teams become more desperate to score as the match progresses. We therefore see that the term β_{0t} includes a subscript for time where the number of minutes played is given by $t = 1, \ldots, 90$. The substitution parameter β_{1t} also includes a time subscript where our intention is to assess the most beneficial times for substitution.

Again, our dataset corresponds to all of the matches considered in Myers (2012) except for the World Cup matches for which the strength variable z_i is not available. In addition, we supplement the dataset with English Premier League matches from three additional seasons, 2010-2011, 2011-2012 and 2012-2013. This leads to a dataset with $n = 4,226$ observations.

A first attempt in fitting model (3.3) is straightforward logistic regression. In Figure 3.1, we have plotted the estimates $\hat{\beta}_{0t} + \hat{\beta}_{1t} s$ with respect to the time index t for $s = -1, 0, 1$. The plots correspond to the log-odds of the probability that the trailing team scores the next goal when teams are equally matched (i.e. $z = 0$). We have plotted the values for the second half only (i.e. $t \geq 46$) as this is the most interesting part of the match. We note that prior to halftime, substitutions are typically made only when there is an injury. In all three plots, we observe that the estimates are mostly positive which implies that the trailing team has a greater chance of scoring next. This suggests that the common strategy of playing defensively given the lead is counter-productive. Conversely, teams that fall behind are more likely to play more aggressively, and this behaviour appears to have merit. A value $\beta_{0t} = 0.2$ which appears typical from Figure 3.1 translates to $p = 0.55$. This implies that the next goal will be scored by the trailing team 55 percent of the time compared to 45 percent of the time by the leading team. We also observe that the substitution covariate s does not appear to have much impact on which team scores the next goal.

A main purpose in displaying Figure 3.1 is to observe the variability of the estimates. We would like to reduce the variability by taking into account prior knowledge. For example, we know that there should be only a small difference in the parameters β_{0t} at adjacent times t and $t + 1$. To improve the smoothness in the estimates with respect to time, we

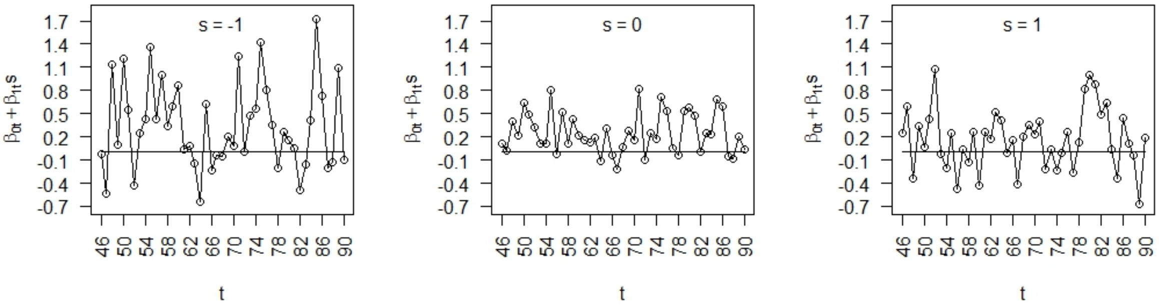

Figure 3.1: Estimates of the parameters $\beta_{0t}+\beta_{1t}s$ based on logistic regression for the second half of play. The three plots correspond to the the substitution covariate $s = -1, 0, 1$. The lines $\beta_{0t} + \beta_{1t}s = 0$ are superimposed.

next consider a Bayesian approach where parameters borrow information from one another. The variability in Figure 3.1 obscures potential trends with respect to time. For example, it may be possible that there is a decreasing trend in the final minutes of a match. This may be due to increased risk taking by the trailing team which is now more exposed to goals on the counter-attack. It is also possible that β_{0t} has larger values for times slightly greater than $t = 45$. This may be due to inspirational instruction at halftime by the manager.

3.3.1 The Prior Distribution

We take a Bayesian approach and require the specification of the prior distribution for the parameters in (3.3). Although many Bayesian statisticians advocate a subjective formulation of prior opinions (Goldstein 2006, Lindley 2000), most practitioners avoid the challenge involved in the elicitation of prior opinions. In many applications, priors of convenience are chosen which are often diffuse and improper.

One of the advantages in sports analytics is that researchers typically have good instincts. For example, the known objectives for winning, the rules of the game and the limited durations of matches give sport a simplicity when compared to the investigation of more complex phenomena. When processes are well understood, this facilitates the use of subjective priors. We consider subjective priors for the parameters in model (3.3). Subjective priors are particularly important for logistic regression; it is well known that diffuse default priors on the coefficients in logistic regression induce probability distributions on p that are convex and are typically inappropriate (Baskurt and Evans 2015).

Referring to the logistic model in (3.3), the parameters are λ, β_{0t} and β_{1t} for $t = 1, \ldots, 90$. To reduce parameter specification to situations of interest, we restrict the time variable to $t = 46, \ldots, 90$. This leaves us with 91 primary parameters. During this timeframe, 2,989 observations were recorded which provides a ratio of 2,989/91 $\approx$ 32.8 observations per parameter. The time restriction also improves the speed of computation.

The parameter λ relates the strength of the trailing team to the probability that the trailing team scores the next goal. We expect that as the strength of the trailing team increases so should their probability of scoring the next goal (i.e. $\lambda > 0$). We therefore prefer a prior distribution for λ that is defined on $\mathcal{R}^+$, and it is also intuitive that the density should be concave. Therefore, we impose the prior $\lambda \sim \text{Gamma}(a_0, b_0)$. The specification of a_0 and b_0 are obtained by referring to gambling websites where soccer markets are thought to be close to efficient (Nyberg 2014). In our dataset, the largest values of the relative team strength covariate z are roughly $z = 1.5$. For most of the soccer matches considered in this analysis, when an exceptionally strong team faces an exceptionally weak team, the handicap in favor of the strong team is roughly 1.5 goals[1] with roughly 2.5 total goals. This implies a scoreline of 2.0-0.5 in favor of the strong team. Consequently, goal scoring in favor of the strong team can be expected to occur in roughly a 4:1 ratio, i.e. with probability 0.80. When $\beta_{0t} = 0$ and $s = 0$ in (3.3), we solve the logit expression, $\log(p/(1-p)) = \log(0.80/0.20) = \lambda z = \lambda(1.5)$, yielding an expected value of $\lambda = 0.92$. We therefore select hyperparameters $a_0 = 10.0$ and $b_0 = 10.9$ where we observe that the specified prior has $\text{E}(\lambda) = 0.92$ and there is sufficient variability surrounding λ to allow for errors in our subjectivity.

Recall that when a goal is scored at time t, the parameter β_{0t} relates the probability that the trailing team scores the goal. It is conceivable that β_{0t} could be either positive or negative. It is also clear that β_{0t} values are dependent in the sense that β_{0t_1} and β_{0t_2} should be comparable when $|t_1 - t_2|$ is small. This suggests that the multivariate distribution

$$\beta_0 = (\beta_{046}, \ldots, \beta_{090})' \sim \text{Normal}(\mu_0, \Sigma) \tag{3.4}$$

provides a sensible subjective prior. When the two teams are evenly matched (i.e. $z = 0$) and when the two teams have made the same number of substitutions (i.e. $s = 0$), we have little intuition as to who will score the next goal. We therefore choose $\mu_0 = (0, \ldots, 0)'$. We then define Σ as a first order autoregressive covariance matrix where the (i, j)th element of Σ is given by $\sigma^2 \rho^{|i-j|}$. The remaining prior specification concerns the variance parameter $\sigma^2 > 0$ and the correlation parameter $\rho \in (0, 1)$. In an evenly contested match (i.e. $z = 0$) when both teams have made the same number of substitutions (i.e. $s = 0$), we cannot imagine the goal ratio for the trailing team at any time t varying beyond 1:2 or 2:1. Therefore $\log(2) - \log(1/2) = (\beta_{0t} + 3\sigma) - (\beta_{0t} - 3\sigma)$ which yields $\sigma = 0.23$. To introduce some variability in σ, we assign $\sigma \sim \text{Gamma}(2.3, 10)$ where $\text{E}(\sigma) = 0.23$. For ρ, we assume that there is no meaningful difference in goal scoring rates at times t and $t + 1$. We express this as imposing the correlation $\rho = 0.97$. We note that at five minute differences t and $t + 5$, this implies a correlation of $\rho^{|t+5-t|} = 0.86$. To introduce some variability in ρ, we assign $\rho \sim \text{Beta}(38, 1)$ where $\text{E}(\rho) = 0.97$. We note that ρ serves as a smoothing parameter where

[1] In gambling circles, a 1.5 handicap means that a wager on the favorite team is successful if the team wins by two or more goals, and the wager is unsuccessful otherwise.

the variability in neighbouring β_{0t} values is reduced as $\rho \to 1$. We note that we experimented with alternative prior specifications and observed qualitatively similar results.

Recall that when a goal is scored at time t, the parameter β_{1t} relates the probability that the trailing team scores the goal when they have made at least one more substitution than the opposition. The arguments advanced in the prior specification of β_0 can be repeated in the case of $\beta_1 = (\beta_{146}, \ldots, \beta_{190})'$. Therefore β_1 will also be assigned a multivariate normal distribution with parameters that have the same hyperparameter specifications as in the case of β_0.

We remark that sometimes statisticians entertain complex models where resulting estimates are subsequently used in secondary analyses. Although sometimes this may be the only viable route, these approaches may be viewed as somewhat ad-hoc where there is a mixing of inferential procedures. For example, in this application, we could have taken the β_{0t} estimates from ordinary logistic regression and simply smoothed the estimates using some sort of procedure such as lowess. Instead, we have proposed a comprehensive model where the smoothing mechanism is facilitated through the prior specification. This strikes us as a more appealing approach for statistical inference.

3.3.2 Results from Bayesian Logistic Regression

We implemented the Bayesian logistic regression model (3.3) via the WinBUGS programming language (Spiegelhalter, Thomas, Best and Lunn 2003). WinBUGS is often convenient for Bayesian analysis as the user only needs to specify the model and provide the data; the associated and sometimes difficult Markov chain Monte Carlo operations are handled in the background by WinBUGS. In our implementation, we carried out 5,000 burn-in iterations followed by 10,000 iterations which were used to estimate posterior characteristics. Standard diagnostic procedures were carried out which suggested practical convergence of the Markov chain.

We first consider the parameter λ which relates the relative strength of the trailing team to the probability that the trailing team scores the next goal. The posterior mean and posterior standard deviation are given by $\mathrm{E}(\lambda \mid y) = 1.00$ and $\mathrm{SD}(\lambda \mid y) = 0.05$. The posterior density of λ is provided in Figure 3.2. We see that the posterior distribution is roughly symmetric. In comparison to the subjective prior distribution for λ which had mean $\mathrm{E}(\lambda) = 0.92$, the posterior distribution is more concentrated and shifted further to the right. The main message involving λ is as expected - with everything else being equal (i.e. $\beta_{0t} = \beta_{1t} = 0$), the stronger team is more likely to score the next goal. Putting this into greater context, imagine that the trailing team is expected to defeat the leading team by one goal (i.e. $z = 1$). Then $\hat{\lambda}z = 1.00$ and the probability that the next goal is scored by the trailing team is $p = \exp(1.00)/(1 + \exp(1.00)) = 0.73$.

We now turn our attention to the parameter $\beta_{0t} + \beta_{1t}s$ which relates the combined effect of the time of the match t and the substitution advantage s to the probability that the

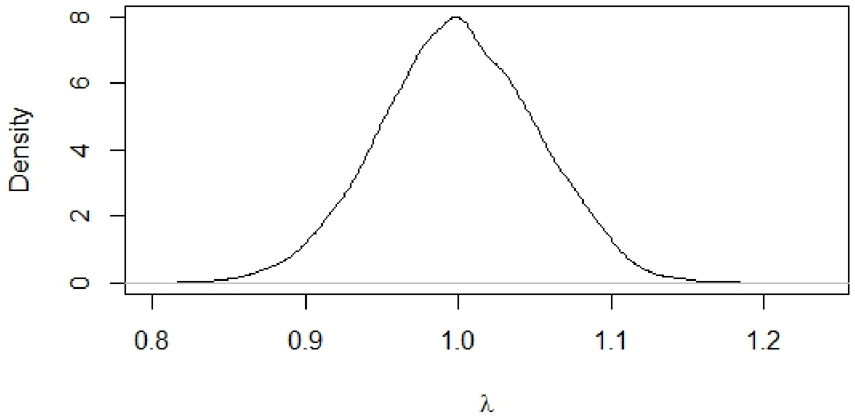

Figure 3.2: The posterior density of λ based on the Bayesian logistic regression model (3.3).

trailing team scores the next goal when teams are equally matched (i.e. $z = 0$). Figure 3.3 provides posterior means of $\beta_{0t} + \beta_{1t}s$ in the second half for each of $s = -1, 0, 1$. In the plot corresponding to $s = 0$, we first observe that the correlation structure introduced in the prior specification (3.4) was successful in smoothing the β_{0t} estimates when compared to the extreme variability observed in Figure 3.1. From a practical point of view, the plots reveal practices and consequences for the game of soccer. The positive estimates in Figure 3.3 suggest that during the second half there is a goal scoring advantage provided to the trailing team. Why might this be? One explanation is tactical. Perhaps managers of teams that are leading instruct players to play cautiously, to stay back, and consequently the leading team is defending more than attacking. In these situations, the trailing team is more likely to score the next goal. Another explanation is psychological. Perhaps teams that are leading are fearful of giving up the lead, and hence play with the cautious characteristics described previously. In any case, the message is clear - teams that are leading should not play as though they are leading. Generally, they should adopt the same style that allowed them to obtain the lead. The tactical and psychological explanations are also relevant to the trailing team. The trailing team may be taking chances, playing fearless and attacking.

From Figure 3.3, we are also able to quantify the scoring effect due to the time of the match and the substitution covariate s. We observe that $\beta_{0t} \approx 0.2$ for most of the second half. With $\beta_{0t} = 0.2$, the probability that the trailing team team scores the next goal is a substantial $p = \exp(0.2)/(1 + \exp(0.2)) = 0.55$. Also, it appears that the plot dips slightly from roughly the 50-minute mark and dips again from roughly the 80-minute mark. A possible explanation is that the manager of the trailing team provides an inspiring talk at halftime, but the motivation begins to wear off beyond $t = 50$. Also, beyond $t = 80$, the

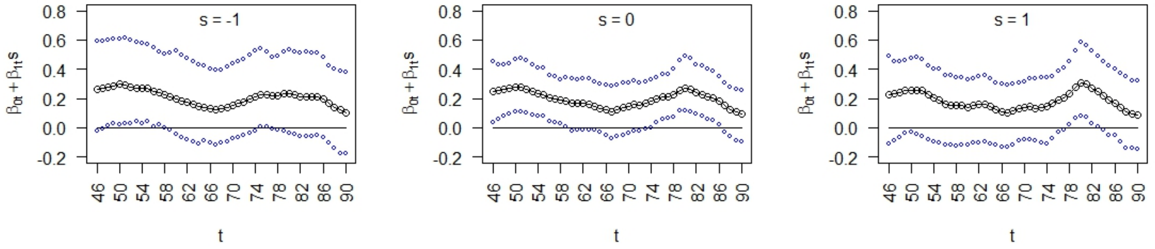

Figure 3.3: Posterior means of the parameters $\beta_{0t} + \beta_{1t}s$ based on the Bayesian logistic regression model (3.3) for the second half of play. The three plots correspond to the the substitution covariate $s = -1, 0, 1$. The lines $\beta_{0t} + \beta_{1t}s = 0$ are superimposed as well as the 95 percent posterior intervals.

aggressive attacking style adopted by the trailing team becomes overly aggressive to the extent that they become more vulnerable to the counter-attack.

We now consider the parameter β_{1t} which was the initial focus of our investigation. We are interested in the relationship between the substitution time t and the probability that the trailing team scores the next goal. The detailed effects due to β_{1t} are not easily assessed from Figure 3.3 as the plots corresponding to $s = -1, 0, 1$ are similar. Posterior means for β_{1t} in the second half of a match are given in Figure 3.4. The noteworthy feature of Figure 3.4 is that the estimates are not discernible from zero when looking at the 95 percent posterior interval bands. That is, at any time t during the second half, if the trailing team has made more substitutions than the leading team, there is no scoring benefit. This finding is in stark contrast to Myers (2012) who claimed there is a strong benefit to the trailing team when they substitute prior to the 58th, 73rd and 79th minutes.

We have observed that the parameters β_{0t} and β_{1t} appear constant with respect to t in the Bayesian analysis. For sake of comparison, we fit two sub-models of model (3.3) in a classical logistical regression context, suppressing the dependence on the time variable t. Under maximum likelihood estimation, we observed $\hat{\beta}_0 = 0.200$ with standard error 0.039, and $\hat{\beta}_1 =$ -0.088 with standard error 0.053. These results are consistent with the magnitude of estimates obtained in the Bayesian analysis as seen in Figure 3.3.

3.4 Discussion

This paper investigates various influences on scoring in soccer by considering a dataset involving 2,989 second half goals when teams were trailing.

An important result that does not seem to be widely recognized is that when teams are of equal strength (i.e. $z = 0$), the trailing team is more likely to score the next goal during the second half. This has implications for strategy. When teams are leading, managers should encourage their teams to play the sort of style that allowed them to obtain the lead.

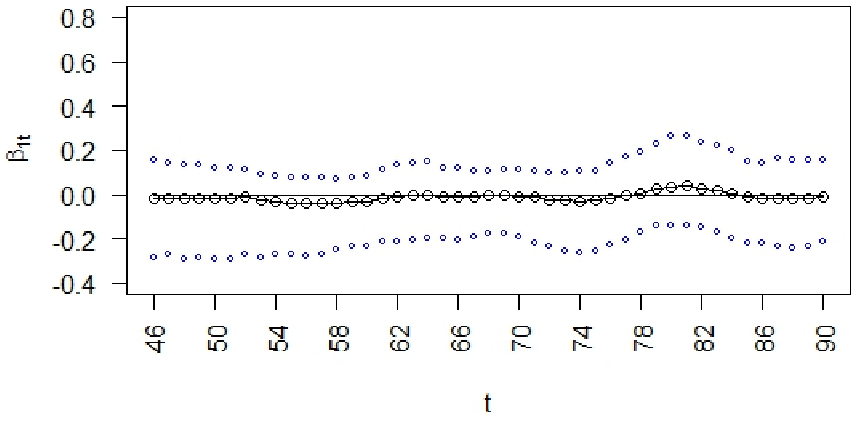

Figure 3.4: Posterior means of the parameters β_{1t} based on the Bayesian logistic regression model (3.3) for the second half of play. The line $\beta_{1t} = 0$ is superimposed as well as the 95 percent posterior intervals.

Going into a defensive shell (whether intentionally or as a psychological consequence) is not optimal. A similar sentiment has been attributed to John Madden in reference to American football: “All a prevent defense does is prevent you from winning.”

Even more surprising than the above result is the impact of substitutions. When the strength of the teams and the time of the match have been considered, there is no discernible benefit for the team that has made more substitutions. This observation needs to be assessed carefully. We are not saying that there is no need to replace players. Instead, we believe that managers are adept at observing player performance. For example, when a player is injured or tired, this is noticed by the manager and they substitute accordingly. Managers are essentially making good decisions, and there are no prolonged periods where teams are significantly weakened. What has happened via substitution is that a quality player has been replaced with another quality player, and there is little distinction. Therefore, in our analysis, there are no times t in Figure 3.4 that appear advantageous with respect to substitution. In fact, one may argue that managers typically put out their best teams at the start of a match, and therefore substitutions are often cases of replacing quality with slightly lower quality. Perhaps this is why we see the trend in Figure 3.4 falling slightly below the line $\beta_{1t} = 0$. In summary, we suggest that managers should substitute, especially when they see a drop in a player’s performance. But there is no reason to tie these substitutions to critical times such as the 58th, 73rd and 79th minutes as in Myers (2012).

We also remark that soccer matches are not randomized experiments where substitutions are made according to some randomization protocol. As is well known, randomization helps deal with the influence of confounding variables.

All soccer fans probably recall occasions when a substitute immediately made an impact on the game, perhaps by scoring a critical goal. Was this managerial brilliance in terms of knowing when to substitute? Perhaps it is simply a case of memory bias and confirmation bias (Schacter 1999). In sports (and in other activities), people tend to remember outstanding events and use these occasions to solidify previously held opinions.

Finally, in the comparison of our approach with Myers (2012), we note that different response variables were used and that our approach introduced new covariates. Therefore, although both analyses address the substitution problem, they do so in different ways and the results are not directly comparable. In terms of practice, Myers (2012) states that managers should substitute according to the 58-73-79 minute rule. On the other hand, our analysis suggests that there is no discernible time during the second half where there is a clear benefit due to substitution. What then is a manager to do? We leave this as a bit of a conundrum that may be considered in future research.

3.5 Commentary

The following is the commentary, directly copied from the article written by Bret R. Myers, (Myers 2016) on our JQAS article entitled “Analysis of Substitution Times in Soccer.”

3.5.1 Analysis of substitution times in soccer (Silva and Swartz)

Overall, I applaud the effort of Silva and Swartz in extending research on soccer substitutions on their JQAS article entitled “Analysis of Substitution Times in Soccer”. The problem is still very relevant in soccer today as FIFA’s substitution rules remain unchanged and the results of matches continue to be impacted significantly by the roles of players coming off of the bench. My original 2012 JQAS paper entitled “A Proposed Decision Rule for the Timing of Soccer Substitutions” advanced an important conversation concerning a critical problem that soccer managers face in matched. As the paper has drawn much attention, it is a natural progression for the likes of Silva and Swartz to evaluate the rule and also create and improved approach that produces a more informative conclusion.

It is reassuring that when trying to replicate my study across the same data sets, Silva and Swartz found similar results with just minor variation. In my analysis, there was 42.27% improvement when following the proposed decision rule and 20.52% improvement when not following the rule (for a difference of 21.75%). When replicating the study through their interpretation of the decision rule, Silva and Swartz obtained 40.07% improvement when following the rule and 17.90% improvement when not following the rule (for a difference of 22.17%). For me, this both validates the results of my study, and also, confirms the idea that substitution patterns significantly impact a team’s likelihood improving goal differential when behind in a match.

While my proposed (58-73-79) decision rule was shaped using classification and regression tree (CART) analysis, Silva and Swartz use Bayesian logistic regression and also account for the strength of each team and home/away. Another difference is in the target variables of the two studies. Concerning my proposed decision rule, the target variable is a categorical of whether or not the trailing team improves their goal differential. However, Silva and Swartz create a categorical variable of whether or not the trailing team scores the next goal. The authors also expand on the original data set of my study by also including three additional English Premier League Seasons.

Based on the results of the two studies, there are rather contradictory conclusions. While I would argue towards a general policy that early substitutions help to improve goal differential when behind, Silva and Swartz would argue that the timing of substitutions do not impact the trailing team's likelihood of scoring the next goal. With the contrasting approaches of the two studies, it is difficult to directly compare the results and to suggest that one paper is right and the other one wrong. Therefore, readers should not view the work of Silva and Swartz to be a replacement for my original article in terms of the information provided, but rather, an alternative way to view the problem.

One key advantage of my proposed decision rule is the ease of interpretability. This benefit stems from the CART approach with binary outcomes of substituting before/after critical times in a soccer match. Although Silva and Swartz do a very thorough and rigorous job with their analysis, their model construct and corresponding results may be more difficult for a practitioner to interpret. In order for sports research to be truly valuable, the information needs to be able to connect with key decision makers in sports organizations. My primary concern at the conclusion of this paper is that a practitioner (or even an academic) may end up being confused, rather than informed about the timing of soccer substitutions.

My original study also looked at the impact of the number of substitutions and a team's ability to successfully comeback in matches. The evidence was clear that a team was much better off using all three substitutions as opposed to using only two. It was unclear in the Silva and Swartz article how quantity of substitutions impacted the trailing team's ability to score.

In conclusion, I encourage readers of the Silva and Swartz article to recognize both the similarities and differences of their approach to that of my original work. I am also hoping that more researchers will be eager to expand on the analyses of this problem, just as Silva and Swartz have done. While I'm inclined to defend my original work and offer perspective on how it compares to this new article, I truly admire Silva and Swartz's novel approach and well-executed analysis in their paper.

3.6 Rejoinder to Myers (2016)

3.6.1 Introduction

By most measures, soccer is the most popular game in the world, and the substitution rule in soccer is unique to the sport. Bret Myers has introduced a substitution guideline for managers which has caught the attention of soccer insiders, fans and academics (Myers 2012). We are thankful for Myers' original contribution and we are also thankful for his comments on our paper (Silva and Swartz 2016) which offers an alternative analysis of substitutions.

Since the decision criteria, the covariates and the methods in Myers (2012) differ from those in Silva and Swartz (2016), one would not expect the conclusions to be exactly the same. However, the two papers pose a conundrum in that the results are so strikingly different. Distilling the two papers to their essence,

(A) Myers (2012) claims that there is a large competitive advantage for trailing teams that follow his substitution guidelines.

(B) Silva and Swartz (2016) argue that there are no special substitution times or periods of a match that yield a competitive advantage for trailing teams.

In this rejoinder, we first encourage the interested reader to carefully examine the discussion provided in Section 2 of Silva and Swartz (2016). The takeaway message is that more faith ought to be placed on (B) than on (A). Second, we now expand on the analysis of Silva and Swartz (2016) by extending their model.

3.6.2 Extension of the Silva and Swartz (2016) Model

Daniel Stenz (former Director of Analytics and Scouting for the Vancouver Whitecaps) suggested that goal differential may also have an impact on whether the trailing team scores the next goal. Accordingly, we have expanded model (3) in Silva and Swartz (2016) as follows

$$\log\left(\frac{p_i}{1-p_i}\right) = \lambda z_i + \beta_{0t} + \beta_{1t} s_i + \tau_2 w_2 + \tau_3 w_3 \tag{3.5}$$

where the new covariate $w_2 = 1$ corresponds to a two-goal deficit by the trailing team and $w_2 = 0$ otherwise. The other new covariate $w_3 = 1$ corresponds to a large deficit by the trailing team (three or more goals) and $w_3 = 0$ otherwise. We assume independent Normal$(0, 1)$ priors for τ_2 and τ_3.

Upon fitting the model, we see no qualitative changes in the posterior means of $\lambda \approx 1.00$, $\beta_{0t} \approx 0.20$ and $\beta_{1t} \approx 0.00$. For τ_2, we obtained posterior means and standard deviations 0.08 and 0.08 respectively. For τ_3, we obtained posterior means and standard deviations

0.12 and 0.11 respectively. The impact is that in the case of all things being equal (i.e. $z = 0$),

- Prob(trailing team scores next when down by 1 goal) $\approx \text{logit}^{-1}(0.20) = 0.55$
- Prob(trailing team scores next when down by 2 goals) $\approx \text{logit}^{-1}(0.20 + 0.08) = 0.57$
- Prob(trailing team scores next when down by $\geq$ 3 goals) $\approx \text{logit}^{-1}$(0.20+0.12) $= 0.58$

Chapter 4

Tactics for Twenty20 Cricket

4.1 Introduction

Twenty20 cricket is the most recent format of cricket. It was introduced in 2003, and gained widespread acceptance with the first World Cup in 2007 and with the introduction of the Indian Premier League in 2008. The main difference between Twenty20 cricket and the more established format of limited *overs* cricket known as one-day cricket is that the former is based on a maximum of 20 overs of batting whereas the latter is restricted to a maximum of 50 overs of batting. Consequently, Twenty20 cricket has a shorter duration of play than one-day cricket, and this is appealing to those with limited time to follow sport. Because the two formats of cricket are so similar, it appears that many of the practices of one-day cricket have transferred to Twenty20 cricket. For example, although there are critics (Perera and Swartz 2013), the Duckworth-Lewis method for resetting targets in interrupted one-day cricket matches is also used in Twenty20 cricket. As another example, it is often the case that a nation's Twenty20 side will resemble its one-day side even though there are different skill sets required in the two formats of cricket.

Since Twenty20 cricket is a relatively new sport, it may be the case that optimal strategies have not yet been fully developed, and instead, Twenty20 cricket is played in much the same way as one-day cricket. This paper explores two avenues for the modification of tactics in Twenty20 cricket which may provide competitive advantages to teams. Of course, with the universal adoption of strategies by all teams, advantages cease to exist. This is one of the themes discussed in the novel "The Blind Side: Evolution of the Game" (Lewis 2006) which was later popularized as a motion picture starring Sandra Bullock.

The first avenue for improving tactics in Twenty20 cricket is based on the realization that *wickets* are of less importance in Twenty20 cricket than in other formats of cricket (e.g. one-day cricket and Test cricket). A consequence is that batting sides in Twenty20 cricket should place more emphasis on scoring runs and less emphasis on avoiding wickets falling. On the flip side, fielding sides should place more emphasis on preventing runs and less emphasis on taking wickets. To justify the claim that wickets are of less importance in Twenty20 cricket than in one-day cricket, Table 4.1 provides a wicket comparison between Twenty20 cricket ($n = 243$ matches) and one-day cricket ($n = 835$ matches) based on international matches involving full member nations of the ICC (International Cricket Council). The matches were played during the period of February 17/05 through December 25/13. We see in Table 4.1 that batting reaches the 8th batsman (i.e. 6 or more wickets taken) 84% of the time in one-day cricket but only 65% of the time in Twenty20 cricket. Since the 8th, 9th, 10th and 11th batsmen tend to be weaker batsmen, we observe that weak batsmen are batting less often and that teams rarely (10% of the time) expend all of their wickets in Twenty20 cricket. Since we are less concerned with wickets, it follows that a potential strategy for Twenty20 batting is to ensure that batsmen with high *strike rates* bat early in the batting lineup. Conversely, it may make sense for the bowling team to prevent runs by introducing bowlers with low *economy rates* early in the bowling order.

	Proportion of first innings with x or more wickets taken when the innings terminate					
	$x = 5$	$x = 6$	$x = 7$	$x = 8$	$x = 9$	$x = 10$
Twenty20	0.84	0.65	0.45	0.27	0.17	0.10
One-Day	0.94	0.84	0.73	0.58	0.44	0.29

Table 4.1: Proportion of first innings with x or more wickets taken when the innings terminate, $x = 5, 6, \ldots, 10$.

To emphasize the distinction between Twenty20 cricket and one-day cricket involving wicket usage, Table 4.2 considers the same time frame as Table 4.1 and shows the distribution of wickets taken after 90% of the overs are used. In Table 4.2, all Twenty20 first innings were considered that reached the end of the 18th over (i.e. 90% of the maximum number of overs). For one-day cricket, we considered all first innings that reached the end of the 45th over (i.e. 90% of the maximum number of overs). From these stages of a match, we again see that late-order batsmen bat less often in Twenty20 cricket than in one-day cricket.

	Proportion of first innings with x or more wickets taken when 90% of the overs are completed					
	$x = 5$	$x = 6$	$x = 7$	$x = 8$	$x = 9$	$x = 10$
Twenty20	0.66	0.37	0.20	0.09	0.04	0.01
One-Day	0.76	0.54	0.35	0.24	0.14	0.09

Table 4.2: Proportion of first innings with x or more wickets taken at the time when 90% of the overs are completed, $x = 5, 6, \ldots, 10$.

The second avenue for improving tactics is motivated by Figure 4.1 which plots the distribution of the amount by which Team A defeats Team B. This is a general density plot that is applicable to many sports where "amount" could represent runs, goals, points, time, etc. We have made the distribution symmetric although this is not required. We have also created the plot so that Team A is much stronger than Team B, and on average, Team A will win by a considerable amount under standard tactics. The probability that Team B wins corresponds to the area under the density curve to the left of zero. There is a second distribution displayed in Figure 4.1 where Team B has modified its tactics so as to increase variance of the response variable. It is possible that this change of tactics will result in Team B losing on average by an even greater amount (i.e. the mean of the distribution is shifted to the right). However, our emphasis is on left tail probabilities corresponding to negative values. These are the cases in which Team B wins. What we see in Figure 4.1 is that Team B wins more often under modified tactics with increased variance than under standard tactics. In this paper, we explore tactics with inflated variance which may allow a weaker team in Twenty20 cricket to win more often.

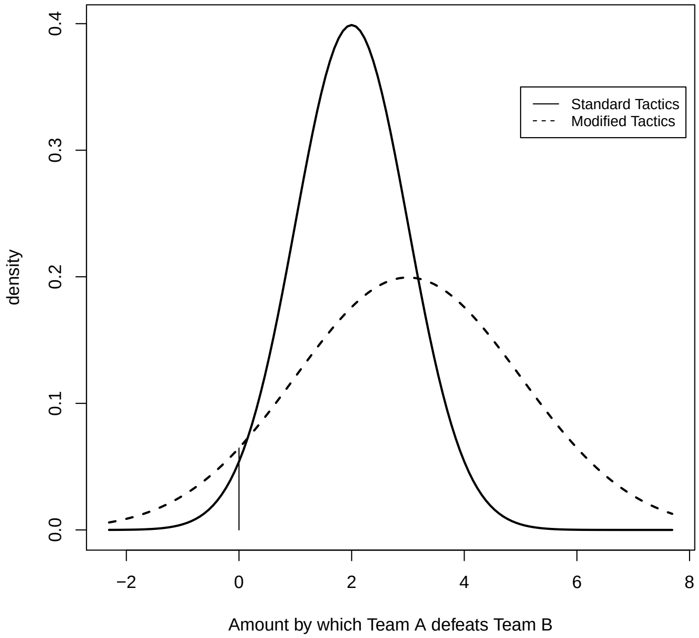

Figure 4.1: Probability density functions of the amount by which Team A (the stronger team) defeats Team B (the weaker team). The tail regions to the left of zero correspond to matches where Team B wins.

In Twenty20 cricket, the quantity of interest that leads directly to wins and losses is run differential. When a team scores more runs than its opposition, they win the match. To investigate run differential, the study of historical matches between two teams is of little value. The composition of the teams change from match to match, and there is rarely a sufficient match history between two teams from which to draw reliable inferences. In addition, matches from the distant past are irrelevant in predicting the future. We therefore use simulation techniques under altered tactics to investigate the distribution of run differential.

In Section 4.2, we provide an overview of the match simulator developed by Davis, Perera and Swartz (2015). The simulator is the backbone for investigating run differential. For the casual reader, this section can be skimmed, as it is only important to know that methodology has been developed for realistically simulating Twenty20 matches. In Section 4.3, we consider modified batting orders. The proposal is to load the batting order so that batsmen with higher strike rates bat earlier in the batting order. This idea aligns with the theme that wickets are less important in Twenty20 cricket than in one-day cricket. In Section 4.4, we consider modified bowling orders. The proposal is that bowlers with low economy rates should bowl early in the bowling lineup. This idea also aligns with the theme that wickets are less important in Twenty20 cricket than in one-day cricket. Here, our focus is to suppress runs rather than be concerned with taking wickets. In Section 4.5, we increase the aggressiveness of batsmen. This has the dual effect of increasing run scoring while simultaneously increasing the rate of wickets falling. This is clearly a variance inflation technique. In Section 4.6, we consider a more comprehensive strategy involving modified batting and bowling orders. Here we use a simulated annealing algorithm over the vast combinatorial space of lineups (i.e. team selection, batting order and bowling order) so as to maximize win percentage. This approach is based on ideas from Perera, Davis and Swartz (2016). We conclude with a short discussion in Section 4.7.

The exploration of tactics appears to be a novel exercise for cricket generally, and Twenty20 cricket in particular. Clarke (1998) recommends that teams should score more quickly in the first innings in one-day cricket than is the current practice. Swartz (2016) provides a survey of cricket analytics with some discussion devoted to tactics and strategy.

4.2 Overview of Simulation Methodology

We now provide an overview of the match simulator developed by Davis, Perera and Swartz (2015) which we use for the estimation of the run distribution for a given team. In cricket, there are 8 broadly defined outcomes that can occur when a batsman faces a bowled ball. These batting outcomes are listed below:

$$\begin{array}{lcl}
\text{outcome } j = 0 & \equiv & 0 \text{ runs scored} \\
\text{outcome } j = 1 & \equiv & 1 \text{ runs scored} \\
\text{outcome } j = 2 & \equiv & 2 \text{ runs scored} \\
\text{outcome } j = 3 & \equiv & 3 \text{ runs scored} \\
\text{outcome } j = 4 & \equiv & 4 \text{ runs scored} \\
\text{outcome } j = 5 & \equiv & 5 \text{ runs scored} \\
\text{outcome } j = 6 & \equiv & 6 \text{ runs scored} \\
\text{outcome } j = 7 & \equiv & \text{dismissal}
\end{array} \tag{4.1}$$

In the list (4.1) of possible batting outcomes, *extras* such as *byes*, *leg byes*, *wide-balls* and *no balls* are excluded. In the simulation, extras are introduced by generating occurrences at the appropriate rates. Extras occur at the rate of 5.1% in Twenty20 cricket. The outcomes $j = 3$ and $j = 5$ are rare but are retained to facilitate straightforward notation.

According to the enumeration of the batting outcomes in (4.1), Davis, Perera and Swartz (2015) suggested the statistical model:

$$(X_{iow0}, \ldots, X_{iow7}) \quad \sim \quad \text{multinomial}(m_{iow};\ p_{iow0}, \ldots, p_{iow7}) \tag{4.2}$$

where X_{iowj} is the number of occurrences of outcome j by the ith batsman during the oth over when w wickets have been taken. In (4.2), m_{iow} is the number of balls that batsman i has faced in the dataset corresponding to the oth over when w wickets have been taken. The dataset was special in the sense that it consisted of detailed ball-by-ball data. The data were obtained using a proprietary parser which was applied to the commentary logs of matches listed on the CricInfo website (`www.espncricinfo.com`).

The estimation of the multinomial parameters p_{iowj} in (4.2) is a high-dimensional and complex problem. The complexity is partly due to the sparsity of the data; there are many match situations (i.e. combinations of overs and wickets) where batsmen do not have batting outcomes. For example, bowlers typically bat near the end of the batting order and do not face situations when zero wickets have been taken.

To facilitate the estimation of the multinomial parameters, Davis, Perera and Swartz (2015) introduced parametric simplifications and a hybrid estimation scheme using Markov chain Monte Carlo in an empirical Bayes setup. A key idea of their estimation procedure was a bridging framework where the multinomial probabilities in a given situation (i.e. over and wickets lost) could be estimated reliably from a "nearby" situation.

Given the estimation of the parameters in (4.2) (see Davis, Perera and Swartz 2015), first innings runs can be simulated for a specified batting lineup facing an average team. This is done by generating multinomial batting outcomes in (4.1) according to the laws of cricket. For example, when either 10 wickets are taken or 20 overs are bowled, the first innings is terminated. Davis, Perera and Swartz (2015) also provided modifications for batsmen facing specific bowlers (instead of average bowlers), accounted for the home field advantage and provided adjustments for second innings batting.

4.3 Modified Batting Orders

In Twenty20 cricket, the objective is to score more runs than your opponent. To maximize runs scored, it is important to carefully consider team selection, and once a team is selected, to determine a good batting order (Perera, Davis and Swartz 2016). The criterion "good" is not straightforward as the consensus opinion is that you want batsmen at the beginning of the batting lineup who both score runs at a high rate but are dismissed at a low rate. Recall that batting in the first innings of a Twenty20 match concludes when either 20 overs have been completed or when 10 wickets have been lost.

However, we have argued that wickets are of less importance in Twenty20 cricket than in the more established one-day format. We therefore consider an extremely simple idea of altering the batting order such that batsmen with high strike rates (average runs per 100 balls) bat early in the batting lineup.

At the time of writing, India may be regarded as one of the stronger Twenty20 sides and we consider their batting order as given in Table 4.3. This was the batting order used in their January 31/16 match versus Australia where India won by 7 wickets with 0 balls remaining. As an opponent, we consider Bangladesh which is well-known to be a weaker side. In the 2016 Twenty20 World Cup, Bangladesh were placed in the group stage consisting of eight teams, from which two teams advanced to the Super 10 stage. We consider Bangladesh's Twenty20 batting lineup from January 17/16 where they defeated Zimbabwe by 42 runs. Based on repeated match simulations with these lineups, we see from Table 4.3 that Bangladesh is expected to defeat India only 21% of the time. The simulated matches were carried out in a simple way; we generated first inning runs for both India and Bangladesh, and then calculated the run differential to determine the match winner.

India (Jan 31/16)	Bangladesh (Jan 17/16)	Bangladesh (alternative)
01. RG Sharma	T Iqbal	S Al Hasan (132.6)
02. S Dhawan	S Sarkar	S Sarkar (130.6)
03. V Kohli	S Rahman	S Rahman (119.0)
04. SK Raina	M Mahmudullah Riyad	T Iqbal (117.1)
05. Y Singh	M Rahim	M Rahim (115.9)
06. MS Dhoni	S Al Hasan	M Mahmudullah Riyad (107.3)
07. HH Pandya	S Hom	M Mortaza (104.6)
08. RA Jadeja	N Hasan	N Hasan
09. R Ashwin	M Mortaza	S Hom
10. JJ Bumrah	A-A Hossian	A-A Hossian
11. A Nehra	M Rahman	M Rahman
	Win Pct = 21%	Win Pct = 37%
	Mean(Run Diff) = -22.1	Mean(Run Diff) = -10.0
	StdErr(Run Diff) = 28.6	StdErr(Run Diff) = 30.0

Table 4.3: Batting orders used in the match simulator for India versus Bangladesh lineups. The career Twenty20 strike rates for the Bangladesh batsmen are given in parentheses Summary statistics regarding the simulation are given at the bottom.

What we further observe in Table 4.3 are the strike rates corresponding to the Bangladesh batsmen (we ignore the four pure bowlers). We therefore consider an alternative batting order that Bangladesh has never utilized in practice. In the alternative lineup, we place the Bangladesh batsmen in decreasing order according to their career strike rates based on international and IPL data up to October 25/15. The biggest changes involves Shakib Al Hasan who moves from batting position #6 to position #1, and Tamin Iqbal who moves from position #1 to #4. With these radical changes, we observe a huge improvement for Bangladesh who now win 37% of the time via the simulation procedure. We note that Al Hasan is an explosive batsmen and the Jan 17/16 lineup does not take advantage of his run scoring capability. In Twenty20 cricket, it is sometimes the case that the 6th batsman in an order may not have the opportunity to bat. We also note that Iqbal is an experienced player, and perhaps his longstanding tenure and reputation plays a role in his batting position with Bangladesh. In Table 4.3, we also observe that the standard lineup used by Bangladesh would have 22.1 fewer average first innings runs than India. When the Bangladesh lineup is altered with the highest strike rate batsmen at the beginning of the batting order, the mean run differential is reduced to 10.0 runs.

The results in Table 4.3 are stunning, and this is particularly due to the batting placement of Al Hasan. Other teams may not be able to have such dramatic improvements. It depends on whether or not their standard lineups use high strike rate batsmen near the beginning of the batting order. Also, we have used career strike rate as a criterion for batting order. This may not be optimal as we note that a batsman's batting position on his team impacts how freely he can bat which in turn affects his strike rate.

4.4 Modified Bowling Orders

From the bowling perspective, we now consider how a fielding team can suppress runs. We again use the sample case from Section 4.3 involving a hypothetical match between Bangladesh and India, and we consider bowling from the perspective of Bangladesh.

In Table 4.4, we provide the bowling order that was used by Bangladesh in their recent January 17/16 match against Zimbabwe. We observe that they used six bowlers in the match. If this bowling order is used against the India lineup listed in Table 4.3, we recall from the simulation procedure that Bangladesh wins only 21% of the time and has an average deficit in run differential of 22.1 runs.

We now consider what would happen if Bangladesh's batting order was left unchanged from January 17/16 but we require that the five bowlers (M Rahman, S Al Hasan, A-A Hossain, M Mortaza and S Rahman) bowl in the order of increasing economy rate. In other words, each would bowl four consecutive overs in the specified order. This idea aligns with the theme that wickets are less important in Twenty20 cricket than in one-day cricket. We note that the proposed bowling order is unrealistic as teams are required to change bowlers between overs and teams strategize concerning the utilization of spin and fast bowlers. However, using the proposed bowling order in our simulation procedure, the Bangladesh win rate increases from 21% to 24% and the average run differential deficit improves from 22.1 runs to 20.1 runs.

Although the results above are not as dramatic as with the modified batting orders in Section 4.3, this may be due to the fact that the Bangleshi bowlers have comparable

Ball	Bowler
0.1-0.6	S Hom
1.1-1.6	S Al Hasan (7.20)
2.1-2.6	A-A Hossain (7.74)
3.1-3.6	M Rahman (6.03)
4.1-4.6	M Mortaza (8.46)
5.1-5.6	M Rahman
6.1-6.6	M Mortaza
7.1-7.6	S Al Hasan
8.1-8.6	S Rahman (8.52)
9.1-9.6	S Hom
10.1-10.6	S Rahman
11.1-11.6	S Al Hasan
12.1-12.6	M Mortaza
13.1-13.6	S Al Hasan
14.1-14.6	M Mortaza
15.1-15.6	A-A Hossain
16.1-16.6	M Rahman
17.1-17.6	A-A Hossain
18.1-18.6	S Hom
19.1-19.5	M Rahman
19.6	S Rahman

Table 4.4: Bowling order used by Bangladesh in their January 17/16 match versus Zimbabwe. Career economy rates are given in parentheses based on international and IPL data up to October 25/15. Shuvagata Hom's economy rate is not listed as this was his first international Twenty20 match where he bowled.

economy rates. For teams with greater disparities in their bowling economy rates, the modification of bowling orders may yield greater improvements. Also, suppose that you had three bowlers with comparable economy rates. You would not need to have them bowl in the order ABCABCABCABC, for example. They could bowl in alternative orders such as CBACBACBACBA.

4.5 Increased Aggressiveness

In this section, we explore the idea of variance inflation by increasing the aggressiveness of batsmen. For implementation of this idea, we recognize that batsmen are more aggressive when fewer wickets have been taken. We therefore define *wicket shift behaviour* (WSB) of -1 as a modification in batting style as if one fewer wicket had been taken. In other words, let the state of the match (o, w) correspond to the oth over when w wickets have been taken. Then wicket shift behaviour of -1 corresponds to

- during $(o, w = 0)$, modify batting behaviour as though the state were $(o, w = 0)$
- during $(o, w = 1)$, modify batting behaviour as though the state were $(o, w = 0)$
- during $(o, w = 2)$, modify batting behaviour as though the state were $(o, w = 1)$
-
-
-
- during $(o, w = 9)$, modify batting behaviour as though the state were $(o, w = 8)$

We similarly define wicket shift behaviours of $-2, -3, \ldots, -9$ which correspond to increasing levels of batting aggressiveness. It is also possible to define non-integer levels of wicket shift behaviour. For example, with respect to a given ball, wicket shift behaviour of -1.2 corresponds to wicket shift behaviour of -1 with probability 0.8 and wicket shift behaviour of -2 with probability 0.2.

The proposed batting schemes are well-suited for analysis using the simulator developed by Davis, Perera and Swartz (2015). In the simulator, every batsman has a baseline state of batting characteristics and these characteristics are modified to provide characteristics p_{iowj} which are applicable to the oth over when w wickets have been taken. We therefore only need to slightly modify the code in order to account for prescribed wicket shift behaviours.

To test the idea of increasing batting aggressiveness, we return to the Bangladesh-India matchup previously discussed, and we alter the batting style of Bangladesh using various wicket shift behaviours. The results are provided in Table 4.5. Again, the results are based on simulating first innings for both Bangladesh and India, and calculating the difference in runs. We first observe that when the wicket shift behaviour is zero (ordinary batting), the win percentage of 21.3% corroborates with the win percentage in Table 4.3 under the standard lineup. More importantly, we observe that the numbers in Table 4.5 coincide with our motivating intuition described in Section 4.1. In particular, we see that the variability (last column) increases as batting aggressiveness (i.e. wicket shift behaviour) increases. Also, in terms of win percentage, we observe that there is an initial benefit to Bangladesh through increased aggressiveness although the benefit decreases when aggressiveness becomes too great. Additional simulations indicate that the maximum benefit occurs for wicket shift behaviour of -0.9. At this value, the win percentage increases to 22.8% from 21.3% under ordinary batting.

WSB	W%	RD	SD(RD)
0	21.3	-22.1	28.6
-1	22.8	-20.9	28.9
-2	22.5	-21.6	29.4
-3	21.2	-23.2	29.7
-4	19.6	-25.2	30.2
-5	17.1	-28.5	30.7

Table 4.5: Investigation of various wicket shift behaviour (WSB) for Bangladesh based on their their January 17/16 lineup in a match versus Zimbabwe. The opposition team is India based on their their January 31/16 lineup in a match versus Australia. The table reports win percentage (W%) for Bangladesh, run differential in favour of Bangladesh (RD) and the standard deviation of RD.

In Table 4.6, we repeat the analysis except this time we consider New Zealand versus India based on New Zealand's lineup on August 16/15 in a match versus South Africa. New Zealand may provide a different perspective than Bangladesh since New Zealand is a strong team. In this matchup, we see the same patterns as with Bangladesh versus India. New Zealand has a 60.2% win percentage under wicket shift behaviour -1.2 which represents an increase from a 59.3% win percentage under ordinary batting behaviour. In this example, because New Zealand is the stronger team (see WSB= 0), the motivation of Section 4.1 does not apply directly. Although the variance of run differential increases with increasing aggressiveness (see the last column of Table 4.6), the maximum win percentage achieved at WSB= -1.2 is due to a shift in the distribution of run differential rather than variance inflation.

WSB	W%	RD	SD(RD)
0	59.3	6.3	27.3
-1	60.2	7.0	27.5
-2	59.8	6.7	27.9
-3	59.0	6.3	28.0
-4	56.6	4.4	28.3
-5	52.7	1.9	28.8

Table 4.6: Investigation of various wicket shift behaviour (WSB) for New Zealand based on their their August 16/15 lineup in a match versus South Africa. The opposition team is India based on their their January 31/16 lineup in a match versus Australia. The table reports win percentage (W%) for New Zealand, run differential in favour of New Zealand (RD) and the standard deviation of RD.

4.6 General Modified Lineups

In this section, we consider the comprehensive strategy of determining an optimal lineup. By lineup, we mean the simultaneous consideration of team selection, batting order and bowling order. This problem was considered in Perera, Davis and Swartz (2016) in the context of maximizing expected run differential. We now consider the problem of maximizing expected win percentage. Optimality is achieved through a stochastic search algorithm over the combinatorial space of lineups where expected win percentage for a particular lineup is obtained via the match simulator.

For illustration, we again consider India based on their January 31/16 lineup. The opposition is New Zealand and their baseline lineup from August 16/15 is given in Table 4.7. Corroborating the results from Table 4.6, we see that New Zealand wins 59% of the simulated matches between these two teams. However, we now optimize the New Zealand lineup and consider team selection from the 15 players which New Zealand named for the 2016 World Cup. We see that the optimal team selection differs considerably from the August 16/15 match where Tom Latham, James Neesham, Nathan McCullum, Adam Milne and Mitchell McClenaghan are replaced by Henry Nicholls, Corey Anderson, Tim Southee, Trent Boult and Mitchell Santner. We also observe that the batting lineups differ, especially in the case of Kane Williamson who moves from the opening partnership to the

6th position and Colin Munro who moves from the 7th position to the opening partnership. We remark that throughout the 2016 World Cup, New Zealand placed Munro in the third batting position which is more in keeping with our optimal batting lineup. However, the takeaway message from Table 4.7 is that New Zealand improved its winning percentage from 59% to 70% against India by using the optimal lineup. In terms of explanation, there may be a number of contributing factors including new players, a changed batting order and a different bowling emphasis.

India (Jan 31/16)	New Zealand (Aug 16/15)	New Zealand (optimal)
01. RG Sharma	MJ Guptill	MJ Guptill
02. S Dhawan	KS Williamson	C Munro
03. V Kohli	TWM Latham	H Nicholls
04. SK Raina	GD Elliott	L Ronchi
05. Y Singh	JDS Neesham	CJ Anderson
06. MS Dhoni	L Ronchi	KS Williamson
07. HH Pandya	C Munro	GD Elliott (4)
08. RA Jadeja	NL McCullum	T Southee (4)
09. R Ashwin	AF Milne	T Boult (4)
10. JJ Bumrah	MJ McClenaghan	MJ Santner (4)
11. A Nehra	IS Sodhi	IS Sodhi (4)
	Win Pct = 59%	Win Pct = 70%
	Mean(Run Diff) = 6.3	Mean(Run Diff) = 14.8
	StdErr(Run Diff) = 27.3	StdErr(Run Diff) = 28.9

Table 4.7: Batting orders used in the match simulator for India versus two New Zealand lineups. The number of overs of bowling in the optimal New Zealand lineup is given in parentheses. Summary statistics regarding the simulation are given at the bottom.

4.7 Discussion

This is an extremely practical paper. We have outlined in simple terms how teams may improve their chances of winning. They may do this through modifying their batting order and by modifying their bowling order. The determination of general optimal lineups as discussed in Section 4.6 requires the specialized software developed by Perera, Davis and Swartz (2016).

The suggestion of modifying aggressiveness in batsmen is not as easy to achieve as the modification of batting and bowling orders. Asking a batsman to be a little more aggressive needs to be communicated and executed in a careful way. Maybe one way of doing this is to ask a batting partnership to try to achieve a specified run rate in a given over. Batting a little more aggressively is something that would require both training (on the part of the batsman) and quantitative expertise (on the part of the team captain or those providing instruction) to specify the correct run rate.

The big issue for us is a desire to see the sport of cricket begin to adopt analytic methods to improve performance. At this stage in time, the sport of cricket appears to lag behind many of the world's major sports.

Chapter 5

The Evaluation of Pace of Play in Hockey

5.1 Introduction

In possession sports, pace of play is a characteristic that influences the style of a match. Generally speaking, when the pace of a game is high, the game is more fluid and there is more opportunity for scoring.

There are different measurements of pace for different sports. For example, in the National Basketball Association (NBA), pace is typically measured by the average number of possessions per game. For example, in the 2015-2016 regular season, the Sacramento Kings lead the NBA with 102.2 possessions per game which is contrasted with the Utah Jazz who ranked last with 93.3 possessions per game (see `www.espn.go.com/nba/hollinger/teamstats`). With more possessions, teams typically score and allow more points. For example, the Sacramento Kings and Utah Jazz ranked 2nd and 30th respectively in the 30-team NBA for total points scored and allowed in the 2015-2016 regular season.

In American football, although there is a clear notion of pace of play, there is no commonly reported statistic that directly measures pace. In the National Football League (NFL), the average number of plays per game is recorded for each team (`www.teamrankings.com/nfl/stat/plays-per-game`). Although this statistic is related to pace, it is obvious that poor offensive teams who rarely make first downs have fewer plays per game. Therefore, in football, plays per game for a team is confounded with offensive strength and is not a pure measure of pace. Pace in football can be increased for a team by using a “hurry-up offense” which affords more plays in a given period of time provided that the team continues to make first downs. Furthermore, teams that frequently pass the ball (as opposed to run) typically use up less of the clock and have more plays from scrimmage.

In both basketball and football, increasing the number of possesions can be seen as a strategy, particularly when a team is losing. In basketball, intentional fouling stops the clock and provides more opportunities to score and overcome a deficit. In football, ensuring that plays are terminated by going “out of bounds” stops the clock and provides more possessions.

In soccer and hockey, there are also notions of pace where a “stretched” game is one that goes from end to end, and is thought to be a game which is played at a high pace. However, in both of these sports, there is again no commonly reported measurement for pace of play.

In this paper, we explore various measures for pace of play in hockey that could also be applied to soccer. In hockey, there is a limited body of literature concerning pace. In a recent investigation, Petbugs (2016) considered the percentage of shot attempts taken by a given team in a game (i.e. the Corsi percentage) and used this as a measure of pace. The idea is that teams that are taking most of the shots are playing at a higher pace. As a measure of pace, an immediate difficulty with the Corsi percentage is that the statistic is associated with the quality of the team. If one team is playing much better, they will be in the offensive zone for a greater period of time and will consequently have a higher Corsi percentage. This however, does not mean that they are playing at a high pace. Hohl (2011) provided a brief discussion on possession metrics where Corsi and the related Fenwick statistics are considered as proxy variables for possession.

What makes this paper unusual is that we essentially report a negative result. In the mathematical sciences, negative results are rarely communicated. For example, if an investigator does not establish a theorem, this does not imply that the theorem is not true. It only means that the investigator was unable to prove the result.

In the experimental sciences, the publication of negative results is also not a widespread practice. Sometimes an experimental result is only seen as significant and publishable if a p-value less than 0.05 is attained (Wasserstein and Lazar 2016). However, there has been an increased calling for the publication of negative results. For example, the reputed multidisciplinary journal PLOS ONE now contains a collection of studies that present inconclusive, null findings or demonstrate failed replications of published work (`www.ploscollections.org/missingpieces`). Without the recognition of negative results, publication biases are introduced, and this affects the validity of meta analyses. In particular, when controversial and important questions of public safety are at stake, it is important to have access to all major studies, either positive or negative. One can think of examples such as the effects due to second hand smoke, the effects of high voltage transmission lines and the effects due to marijuana legislation.

There is another reason why negative results should sometimes be reported. Granqvist (2015) writes, “it causes a huge waste of time and resources, as other scientists considering the same questions may perform the same experiments”. Our investigation may fall under this category. We believe that our measures of pace are intuitive and sensible. With the advent of the availability of detailed NHL event data, we imagine that other researchers may consider similar investigations of pace to what we have attempted. In the context of hockey analytics, Sam Ventura (analytics consultant for the Pittsburgh Penguins) tweeted, “I’ve said this to a large number of colleagues & students recently, so I’m posting it here too: Null results are still interesting results!” (`https://twitter.com/stat_sam/status/717109886430158848`). Ventura then tweets, “Publish all of your results, regardless of how "strong" or "weak" they are. It can only serve to benefit the research community by putting this information out there.”

In Section 5.2, we describe the initial approach that we use in defining pace. We also describe the data which we use to investigate various pace of play statistics. The proposed statistics are based on big data sources that take the form of event data. Consequently, the statistics could not have been computed prior to the advent of modern rink technology and computing. In Section 5.3, we calculate the various pace statistics for the 2015/2016 NHL season. We observe that none of the proposed statistics correlate positively with expected and familiar quantities such as goals scored and shots taken. Consequently, there is no appealing narrative for how pace affects games, how pace should be used as a tactic, etc. We conclude with a brief discussion in Section 5.4.

5.2 Pace Calculation

Our understanding of pace is that the pace of play is fast when teams are rushing from end to end, attacking and retreating. In fast paced games, there is less opportunity to be organized in the defensive zone in terms of the numbers of defensive players and positioning. A team that sends players forward exposes themselves to counter-attacks. When a team has the puck and are moving sideways or passing backwards, then they are behaving cautiously and we would say that they are playing at a slow pace. We now attempt to incorporate these general ideas.

Our initial game pace statistic is evaluated as follows: We consider the consecutive events $E_1, \ldots, E_n$ in a game consisting of n events. For each $i = 1, \ldots, n$, the location of the event E_i is obtained according to the Cartesian coordinates (x_i, y_i) where $(0, 0)$ is centre ice and $(-100, 0)$ is the position directly behind the home team’s goal. Rink sizes in the NHL are standardized with dimensions of 200 feet by 85 feet. It is important to note that teams change ends at the beginning of the second and third periods.

For each $i = 1, \ldots, n$, we determine whether the home team (H) or the road team (R) had possession of the puck immediately following E_i. We then determine which team had possession immediately preceding E_{i+1}. If it is the same team, then there is a pace contribution d_i which is the “attacking distance” travelled and is defined by

$$d_i = \begin{cases} \max\{x_{i+1} - x_i,\ 0\} & \text{if H had possession} \\ \max\{-x_{i+1} + x_i,\ 0\} & \text{if R had possession} \\ 0 & \text{if change of possession} \end{cases} \tag{5.1}$$

We therefore define a pace contribution d_i only in the case of a team moving forward in an attacking direction with possession. For example, "dumping the puck" into the offensive zone with retrieval by the defensive team is not considered as a contribution to game pace. Also, drifting sideways with possession is not considered as a contribution to game pace. The total attacking distance in a game is then defined as

$$D_1 = \sum_{i=1}^{n-1} d_i \tag{5.2}$$

The remaining detail in the calculation of (5.2) is the determination of possession as required in (5.1). Thomas and Ventura (2014) have created an R package *nhlscrapr* that provides detailed event information and processing for NHL games. The scraper retrieves play-by-play game data from the NHL Real Time Scoring System database and stores the data in convenient files that can be handled by the R programming language. The *nhlscrapr* package can access NHL matches back to the 2002-2003 regular season. We note that there are 10 types of events E_i provided by *nhlscrapr* as listed in Table 5.1.

Event Type	Frequency
Line Change	27.2 %
Faceoff	16.8 %
Shot on Goal	15.1 %
Hit/Check	13.7 %
Blocked Shot	7.9 %
Missed Shot	6.4 %
Giveaway	4.5 %
Takeaway	3.7 %
Penalty	2.2 %
Goal	1.5 %

Table 5.1: The 10 types of mutually exclusive events that are recorded using *nhlscraper*. The events are listed in order of their percentage frequency from the 2014-2015 NHL regular season based on 451,919 observed events. There are some unlisted rare events and error codes that comprise the remaining 1%. Note that Line Change corresponds to line changes "on the fly" and not line changes that occur during stoppages. A Missed Shot is a shot that was not a shot on goal.

More can be said about the NHL Real Time Scoring System database and the determination of possession. However, a stumbling block with this freely accessible database is that there are roughly 400 events recorded per match. Over a 60 minute hockey game, this translates to an event every 9 seconds on average. Given the action in hockey, much can transpire over 9 seconds, much more than what is recorded in the database. For example, Figure 5.1 provides a potential path taken during 9 seconds of possession. In this case, the pace contribution d_i according to (5.1) does not reflect the amount of forward progress made by the team in possession. It is even possible for possession to change over a 9 second interval and for this not to be recorded. Consequently, although the NHL Real Time Scoring System database has provided a breakthrough for hockey analytics, it is not detailed enough for our purposes.

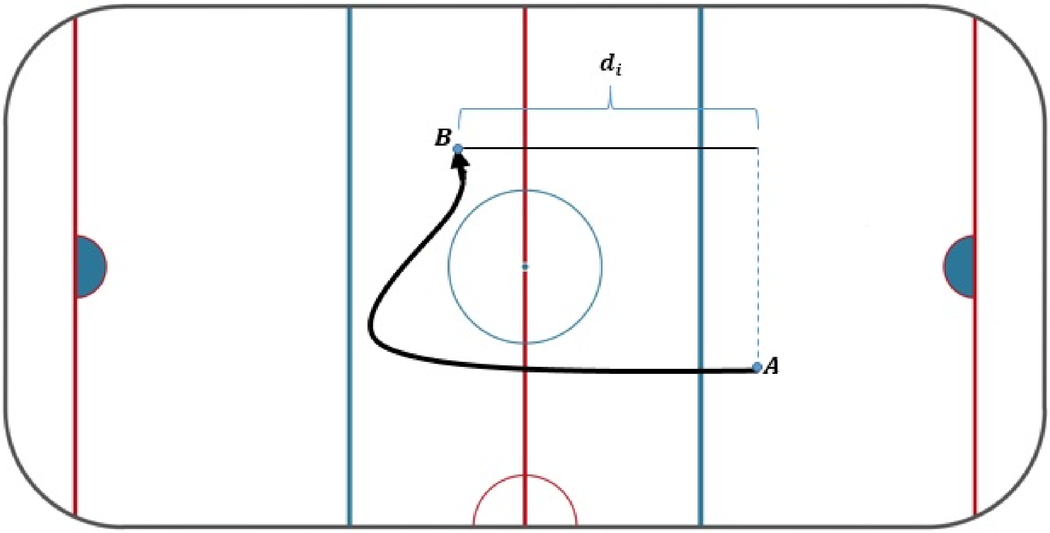

Figure 5.1: Potential path taken by a team during 9 seconds of possession. Given the starting point A and the endpoint B, the pace contribution d_i is shown.

At this point in time, the NHL is moving towards the collection of data via player tracking cameras in every NHL venue. Consequently, there will soon be an explosion of data in the NHL. A similar initiative has already taken place in the NBA where the SportVU system has been in place since the 2013/2014 season. The NBA data has promoted a surge in research activities including previously difficult topics of investigation such as the evaluation of contributions to defense (Franks et al. 2015). In the NHL, the company SPORTLOGiQ has provided us with proprietary data for most games (1140 out of 1230) during the 2015/2016 NHL season. Most importantly for our purposes, there is great detail in the SPORTLOGiQ database with events occurring every 1.2 seconds on average. Although we are not at liberty to discuss aspects of the SPORTLOGiQ database, we can say that the database has an extended number of events compared to those in Table 5.1. Furthermore, possession is easily determined so that the calculations of (5.1) and (5.2) are easily facilitated. In Section 5.3, we describe our investigation of pace using the SPORTLOGiQ database.

5.3 Investigation of Pace

We begin with the distance metric D_1 defined in (5.2) which is the sum of forward attacking distances by both teams in a game measured in feet. We have omitted overtime periods because teams may play differently during overtime. Specifically, since the 2015/2016 season, teams play with three skaters instead of five during overtime periods and this may “open up” the ice and lead to more transitions and greater pace.

To provide an intuitive measure of pace for a game, we define

$$P_1 = D_1/T \tag{5.3}$$

where T is the number of seconds in a match where teams are playing at full strength (i.e. 5v5). We reason that teams may play differently in non-5v5 situations as there is more open ice. Therefore P_1 represents the average forward attacking distance in feet during a game.

In Figure 5.2, we provide scatterplots of the pace variable P_1 versus total goals while full strength and versus total shots while full strength in a game. The plots are based on the 1140 recorded SPORTLOGiQ matches during the 2015/2016 regular season. In neither plot do we see a positive correlation. The correlation coefficients for the two plots are -0.079 (total goals) and -0.281 (total shots). This is surprising as one would think that high paced games would lead to more scoring opportunities. In fact, the correlation for total shots $r = -0.281$ is highly statistically significant (negative) with a t-statistic of $t = r\sqrt{n-2}/\sqrt{1-r^2} = -9.88$. Our intuition is that in high paced games, teams fall out of defensive positions, that there is more open space and consequently more opportunities to score.

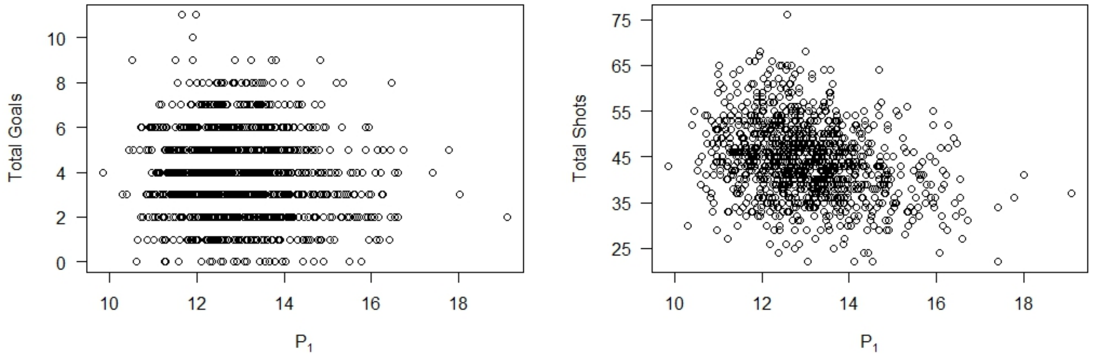

Figure 5.2: Plots of familiar measures (total goals and total shots while full strength) versus P_1 for games during the 2015/2016 NHL regular season.

As a second attempt to investigate pace, we modify the calculation of D_1 to D_2. With D_2, we only consider attacking distances d_i that were traversed at a sufficient speed. The intuition is that teams are not playing at high pace if they are moving slowly. Therefore, we take the attacking distance d_i in (5.1) and obtain the time t_i in seconds that it took the team to travel the distance d_i. The time variable t_i is available from the SPORTLOGiQ database. Then we only include a d_i contribution in D_2 if $d_i/t_i \geq 5.0$ feet per second. This cutoff retains 96.5% of the observations used in calculating P_1. This leads to a second measure of pace in a game given by

$$P_2 = D_2/T \tag{5.4}$$

where T is the number of seconds in a match where teams are playing at full strength.

In Figure 5.3, we provide scatterplots of the pace variable P_2 in (5.4) versus total goals while full strength and versus total shots while full strength in a game. The plots are based on the 1140 recorded SPORTLOGiQ matches during the 2015/2016 regular season. Again, in neither plot do we see a positive correlation. The correlation coefficients for the two plots are -0.091 (total goals) and -0.360 (total shots). Here the correlation for total shots is even more negative than with P_1. We note that we experimented with alternative threshold speeds and observed qualitatively similar results.

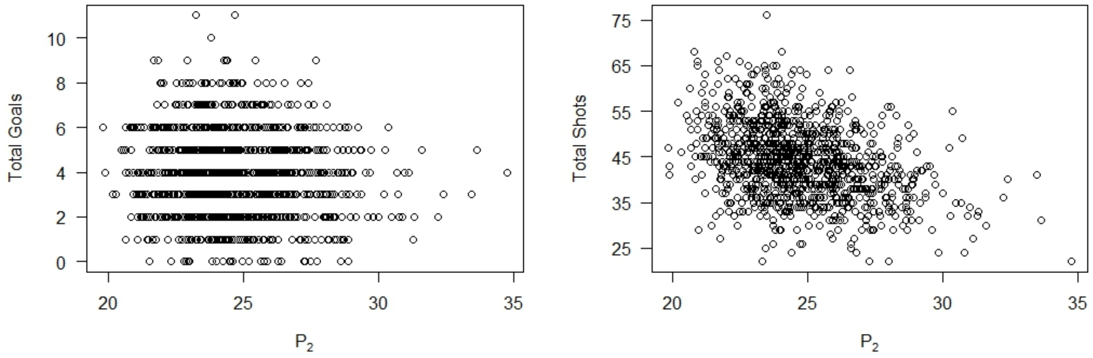

Figure 5.3: Plots of familiar measures (total goals and total shots while full strength) versus P_2 for games during the 2015/2016 NHL regular season.

As a third attempt to investigate pace, we modify the calculation of D_1 to D_3. With D_3, we only consider attacking distances d_i that occurred between the blue lines. The intuition is that frequent transitions between the blue lines (i.e. in the neutral zone) characterize games that have a back and forth quality. Operationally, if we have a distance d_i that begins within a team's own blue line, we truncate it so it begins at the blue line. If a distance d_i ends within the opponent's blue line, then it is truncated to the blue line. This leads to a third measure of pace in a game given by

$$P_3 = D_3/T \tag{5.5}$$

where T is the number of seconds in a match where teams are playing at full strength.

In Figure 5.4, we provide scatterplots of the pace variable P_3 in (5.5) versus total goals while full strength and versus total shots while full strength in a game. The plots are based on the 1140 recorded SPORTLOGiQ matches during the 2015/2016 regular season. Again, in neither plot do we see a positive correlation. The correlation coefficients for the two plots are -0.040 (total goals) and -0.230 (total shots).

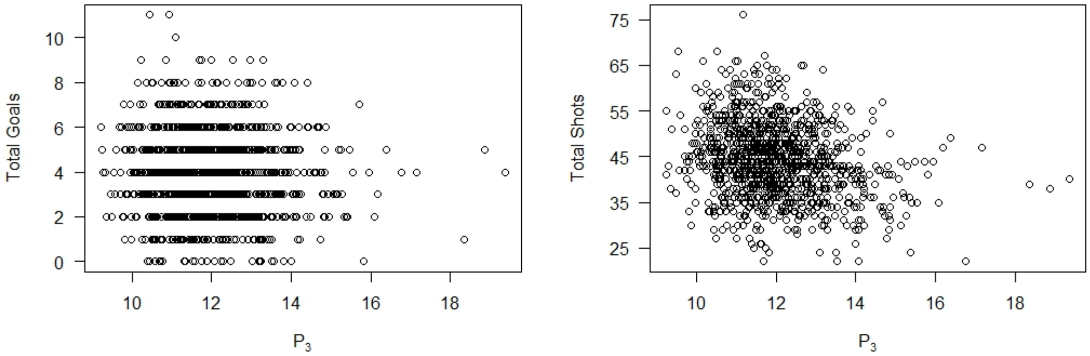

Figure 5.4: Plots of familiar measures (total goals and total shots while full strength) versus P_3 for games during the 2015/2016 NHL regular season.

5.4 Discussion

This paper introduces various measures for pace of play in hockey which are based on the lengthwise distances travelled (skated) by both teams while in possession of the puck during a game. To our great surprise, we found that our definition of pace does not correlate positively with either total goals or shots on goal.

Therefore, our communication may be seen as a negative result. However, since the result is counter-intuitive, we believe that it deserves mention in the hockey analytics community.

Should future refinements to pace provide meaningful correlations, then a host of interesting questions may be addressed. For example, does pace contribute to scoring? Does pace contribute to winning? Which teams are pacey? Has pace changed over seasons? Are there pacey players? Can teams incorporate strategies related to pace and goal scoring? If pace does increase goals at both ends of the ice, then a tradeoff between increasing pace and goal scoring may be similar to the tradeoff between pulling the goaltender earlier and goal scoring (Beaudoin and Swartz 2010).

Bibliography

Alamar, B. and Mehrotra, V. (2011). Beyond moneyball: The rapidly evolving world of sports analytics, Part I. In *Analytics-magazine*, http://analytics-magazine.org/beyond-moneyball-the-rapidly-evolving-world-of-sports-analytics-part-i-2/.

Allsopp, P.E. and Clarke, S.R. (2004). Rating teams and analysing outcomes in one-day and test cricket. Journal of the Royal Statistical Society: Series A, 4, 657-667.

Anderson, C. and Sally, D. (2013). *The Numbers Game: Why Everything you know about Soccer is Wrong*. Penguin Books: New York

Armatas, V., Yiannakos, A. and Sileloglou, P. (2007). Relationship betweeen time and goal scoring in soccer games: analysis of three World Cups. *International Journal of Performance Analysis in Sport*, 7(2), 48-58.

Baskurt, Z. and Evans, M. (2015). Goodness of fit and inference for the logistic regression model. Technical Report, Department of Statistical Sciences, University of Toronto.

Beaudoin, D. and Swartz, T.B. (2010). Strategies for pulling the goalie in hockey. *The American Statistician*, 64(3), 197-204.

Cholst, N. (2013). Research roundup - the best sub strategy, will financial fair play ruin Man City, and why you shouldn't always fire your coach. In *Café Futebol*, http://www.cafefutebol.net /2013/12/20/research-roundup-part-one/.

Clarke, S.R. (1998). "Test statistics" in *Statistics in Sport, editor J. Bennett*, Arnold Applications of Statistics Series, Arnold: London, 83-103.

Clarke, S.R. and Norman, J.M. (1995). Home ground advantage of individual clubs in English soccer. *The Statistician*, 44, 509-521.

Davis, J., Perera, H. and Swartz, T.B. (2015). A simulator for Twenty20 cricket. *Australian and New Zealand Journal of Statistics*, 57, 55-71.

Del Corral, J., Barros, C.P. and Prieto-Rodriguez, J. (2008). The determinants of player substitutions: A survival analysis of the Spanish soccer league. *The Journal of Sports Economics*, 9, 160-172.

Diamond, J. (2011). The science of soccer substitutions. In *The Wall Street Journal*, http://www.wsj.com/articles/SB10001424052748704364004576132203619576930.

Duckworth, F.C. and Lewis, A.J. (2004). A successful operational research intervention in one-day cricket. *Journal of the Operational Research Society*, 55, 749-759.

Elderton, W.E. (1945). Cricket scores and some skew correlation distributions. *Journal of the Royal Statistical Society*, Series A, 108, 1-11.

Fernando, M., Manage, A.B.W. and Scariano, S. (2013). Is the home-field advantage in limited overs one-day international cricket only for day matches? *South African Statistics Journal*, 47, 1-13.

Franks, A., Miller, A., Bornn, L. and Goldsberry, K. (2015). Counterpoints: advanced defensive metrics for NBA basketball. *Sloan Sports Analytics Conference, 2015.*

Goldstein, M. (2006). Subjective Bayesian analysis: principles and practice. *Bayesian Analysis*, 1, 403-420.

Granqvist, E. (2015). Looking at research from a new angle: why science needs to publish negative results. In *Elsevier Publishing Ethics*, Accessed October 4/2016 at https://www.elsevier.com/editors-update/story/publishing-ethics.

Hirotsu, N. and Wright, M. (2002). Using a Markov process model of an Association football match to determine the optimal timing of substitutions and tactical decisions. *Journal of the Operational Research Society*, 53, 88-96.

Hohl, G. (2011). Introduction to hockey analytics Part 4.1: Possession metrics (Corsi/Fenwick). In *SB Nation: Lighthouse Hockey*, www.lighthousehockey.com/2011/8/7/2302188/ an-introduction -to-hockey-analytics-part-4-1-an-introduction-to.

Karlis, D. and Ntzoufras, I. (2003). Analysis of sports data using bivariate poisson models. *The Statistician*, 52, 381-393.

Knorr-Held, L. (2000). Dynamic rating of sports teams. *The Statistician*, 49, 261-276.

Koning, R.H. (2000). Balance in competition in Dutch soccer. *The Statistician*, 49, 419-431.

Lemmer, H.H. (2013). Team selection after a short cricket series. *European Journal of Sport Science*, 13, 200-206.

Lewis, M. (2004). *Moneyball: The Art of Winning an Unfair Game*, W. W. Norton & Company, New York.

Lewis, M. (2006). *The Blind Side: Evolution of a Game*, W. W. Norton & Company, New York.

Lindley, D. (2000). The philosophy of statistics (with discussion). *The Statistician*, 49, 293-337.

McHale, I.G. and Asif, M. (2013). A modified Duckworth-Lewis method for adjusting targets in interrupted limited overs cricket. *European Journal of Operational Research*, 225, 353-362.

Morris, D. (1981). *The Soccer Tribe*, London: Jonathan Cabe.

Mosteller, F. (1952). The World Series competition. *Journal of the American Statistical Association*, 47, 355-380.

Myers, B.R. (2012). A proposed decision rule for the timing of soccer substitutions. *Journal of Quantitative Analysis in Sports*, 8, Article 9.

Myers, B.R. (2016).Commentary, Analysis of substitution times in soccer (Silva and Swartz) *Journal of Quantitative Analysis in Sports*, 12(3), 123-124.

Norton, P. and Phatarfod, R. (2008). Optimal strategies in one-day cricket. *Asia-Pacific Journal of Operational Research*, 25, 495-511.

Nyberg, H. (2014). A multinomial logit-based statistical test of association football betting market efficiency. *Helsinki Center of Economic Research (HECER) Discussion Papers*, No. 380.

Perera, H. and Swartz, T.B. (2013). Resource estimation in Twenty20 cricket. *IMA Journal of Management Mathematics*, 24, 337-347.

Perera, H., Davis, J. and Swartz, T.B. (2016). Optimal lineups in Twenty20 cricket. *Journal of Statistical Computation and Simulation*, 86, 2888-2900.

Petbugs (2016). Run and gun or slow it down: Identifying optimal game pace strategies based on team strengths. *Vancouver Hockey Analytics Conference, Harbour Centre, Simon Fraser University, April 9*, recording at www.stat.sfu.ca/hockey.html.

Ridder, G., Cramer, J.S. and Hopstaken, P. (1994). Down to ten: estimating the effect of a red card in soccer. *Journal of the American Statistical Association*, 89, 1124-1127.

Schacter, D.L. (1999). The seven sins of memory: insights from psychology and cognitive neuroscience. *American Psychologist*, 54, 182-203.

Silva, R.M and Swartz, T.B. (2016). Analysis of substitution times in soccer. *Journal of Quantitative Analysis in Sports*, 12(3), 113-122.

Silva, R.M and Swartz, T.B. (2016). Rejoinder to Myers (2016). *Journal of Quantitative Analysis in Sports*, 12(3), 125-125.

Silva, R.M., Manage, A.B.W. and Swartz, T.B. (2015). A study of the powerplay in one-day cricket. *European Journal of Operational Research*, 244, 931-938.

Silva, R.M., Perera, H., Davis, J. and Swartz, T.B. (2016). Tactics for Twenty20 cricket. *South African Statistical Journal*, 50(2), 261-271

Spiegelhalter, D.J., Thomas, A. and Best, N.G. (2003). WinBUGS (Version 1.4) User Manual. MRC Biostatistics Unit, Cambridge.

Swartz, T.B. (2016). Research directions in cricket. *Handbook of Statistical Methods and Analysis in Sports, editors J.H. Albert, M.E. Glickman, T.B. Swartz and R.H. Koning*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods: Boca Raton, FL.

Swartz, T.B. and Arce, A. (2014). New insights involving the home team advantage. *Ineternational Journal of Sports Science and Coaching*, 9, 681-692.

Swartz, T.B., Gill, P.S., Beaudoin, D. and de Silva, B.M. (2006). Optimal batting orders in one-day cricket. *Computers and Operations Research*, 33, 1939-1950.

Thomas, A.C. and Ventura, S.L. (2014). nhlscrapr: Compiling the NHL Real Time Scoring System Database for easy use in R. *R package version 1.8*, http://CRAN.R-project.org/package=nhlscrapr.

Valero, J. and Swartz, T.B. (2012). An investigation of synergy between batsmen in opening partnerships. *Sri Lankan Journal of Applied Statistics*, 13, 87-98.

van Staden, P.J. (2009). Comparison of cricketers' bowling and batting performances using graphical displays. *Current Science*, 96, 764-766.

Wasserstein, R.L. and Lazar, N.A. (2016). The ASA's statement on p-values: context, process and purpose. *The American Statistician*, 70(2), 129-133.

Wood, G.H. (1945). Cricket scores and geometrical progression. *Journal of the Royal Statistical Society*, Series A, 108, 12-22.

Wright, M. (2014). OR analysis of sporting rules - A survey. *European Journal of Operational Research*, 232, 1-8.