Cricket Analytics

by

Gamage Harsha Perera

M.Sc., Simon Fraser University, Canada, 2011
B.Sc.(Hons.), University of Peradeniya, 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 2015

Approval

Name:	**Gamage Harsha Perera**
Degree:	**Doctor of Philosophy (Statistics)**
Title:	***Cricket Analytics***
Examining Committee:	**Chair:** Yi Lu Associate Professor

Tim Swartz
Senior Supervisor
Professor _____

Paramjit Gill
Supervisor
Associate Professor
Barber School of Arts and Sciences
University of British Columbia
Okanagan _____

Brian Naicker
Internal Examiner
Director
Centre for Online and Distance
Education _____

David Stephens
External Examiner
Professor
Department of Mathematics and
Statistics
McGill University _____

Date Defended: 16 December 2015

Abstract

This thesis consists of a compilation of three research papers and a non-statistical essay.

Chapter 2 considers the decision problem of when to declare during the third innings of a test cricket match. There are various factors that affect the decision of the declaring team including the target score, the number of overs remaining, the relative desire to win versus draw, and the scoring characteristics of the particular match. Decision rules are developed and these are assessed against historical matches. We observe that there are discrepancies between the optimal time to declare and what takes place in practice.

Chapter 3 considers the determination of optimal team lineups in Twenty20 cricket where a lineup consists of three components: team selection, batting order and bowling order. Via match simulation, we estimate the expected runs scored minus the expected runs allowed for a given lineup. The lineup is then optimized over a vast combinatorial space via simulated annealing. We observe that the composition of an optimal Twenty20 lineup sometimes results in nontraditional roles for players. As a by-product of the methodology, we obtain an "all-star" lineup selected from international Twenty20 cricket.

Chapter 4 is a first attempt to investigate the importance of fielding in cricket. We introduce the metric of expected runs saved due to fielding which is both interpretable and is directly relevant to winning matches. The metric is assigned to individual players and is based on a textual analysis of match commentaries using random forest methodology. We observe that the best fielders save on average 1.2 runs per match compared to a typical fielder.

Chapter 5 is a non-statistical essay of two cricketing greats from Sri Lanka who established numerous world records and recently retired from the game. Though their record-breaking performances are now part of cricketing statistics, this chapter is not a contribution which adds to the statistical literature, and should not be regarded as a component of the thesis in terms of analytics.

Keywords: cricket; decision rules; Gibbs sampling; parameter estimation; Random forests; Relative value statistics; Simulated annealing; Simulation; Textual analysis; Twenty20 cricket

Dedication

To my loving family in Sri Lanka and Kavinda Ranil Bibile.....

Acknowledgements

First and foremost I want to express my sincere thanks and gratitude to my senior supervisor Dr. Tim Swartz for his guidance, support, and encouragement over the years. If not for his urging I would not even have enrolled in the Ph.D program and would have gone on to seek a career in industry after completing my Masters. In retrospect that would have been a mistake as I would never again have had the opportunity to pursue the coveted Ph.D. Thank you Tim.

I came to SFU initially because of Dr. Tom Loughin and Dr. Derek Bingham. I am most indebted to them for reposing their trust and confidence in me which enabled me to begin this great educational experience at SFU.

To my examining committee consisting of Dr. Tim Swartz, Dr. David Stephens, Dr. Paramjit Gill and Brian Naicker - a special word of thanks for a patient hearing and for the valuable inputs which I have incorporated into my thesis.

I also want to thank all the faculty members of the Department of Statistics and Actuarial Science who taught me during these last six years, especially Dr. Carl Schwarz, Dr. Richard Lockhart, Dr. Boxin Tang, Dr. Joan Hu, Dr. Rick Routledge, Dr. Brad McNeney, Dr. Jiguo Cao, Dr. Charmaine Dean, Dr.Tom Loughin, and Dr. Tim Swartz.

I am grateful for all the financial support provided by the Department of Statistics and Actuarial Science as well as to the donor of the Randy Sitter Annual Graduate Scholarship in Statistics and Actuarial Science. Special thanks to Statistics Workshop Manager Robin Insley for his support and guidance during my time at SFU. My sincere gratitude to Sadika, Kelly, and Charlene for their kind assistance and backup.

I want to thank my research colleague Jack Davis who helped me a great deal, especially in the preliminary stages of the projects in the area of data collection, which was a real challenge, and for his advice and help with coding. Thanks also to my graduate student colleagues for their friendship and camaraderie and for the fun times we had together. To Ross Churchley, I want to say thanks for helping me with LaTex.

I'm also grateful to my SFU family, my dear circle of Sri Lankan friends, Rajitha, Lasantha, Gaya, Bhagya, Pulindu, Chamara, Nadheera, Nethangi, Shinelle, Sujani and Lalangi for making my journey a fun filled one, and for all the support all of you have given me on various occasions.

I am fortunate to have a great family in Sri Lanka. I would not have come this far without their encouragement.

Last but not least I want to thank Dr. Ranil Waliwitiya for introducing me to SFU, and Kavinda Ranil Bibile for his invaluable support and guidance in so many ways throughout my life in Vancouver.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 SPORTS ANALYTICS

Sports analytics play a major role in various problems associated with sport. Some of these problems are the ranking of individual players and their specialized skills, the composition of teams with an optimal balance of specialized skills, the ranking of teams, the negotiation of contracts, the evaluation of sports businesses and their potential revenue streams, the planning of both physical and mental training, the development of strategies for winning games and tournaments, assessing the effectiveness of coaches and referees, the medical and actuarial aspects of sports injuries (health and insurance), the analysis of existing rules and the need for improving such rules, the improvement of equipment and technology, the determination of awards, the keeping of historical records and the generation of odds for gambling activities. Related to all of the above is the coherent statistical presentation of both raw data and its inferences to the decision makers to facilitate successful planning and implementation. Furthermore, the media and the public have a great appetite for well visualized statistics.

North American sports such as baseball, football, basketball and ice-hockey generate vast sums of money for players, teams, media, advertisers, sponsors, facilities providers and others. The financial imperatives of these activities have given rise to intense statistical analyses.

New opportunities for sports analytics have arisen due to the advent and availability of detailed and high quality data. For example, in Major League Baseball (MLB), the Pitchf/x and FIELDf/x systems have provided comprehensive data on pitching and fielding. These systems record every play while also tracking the exact movements of all players on the field. Using these data sources, [25] used spatial statistics to assess fielding contributions whereas [45] developed methods to assess pitching quality.

In the National Basketball Association (NBA), the SportVU player tracking technology provides coordinates for each player and the ball at a rate of 25 hertz for every NBA regular

season game. The technology has been in place since 2013 and has led to investigation of problems that had not been previously considered [19].

In the National Hockey League (NHL), the website nhl.com provides detailed data on the events that occur in every regular season game. They use a tracking system whereby imbedded chips inside pucks and jerseys provide movement data. Using this data, [44] developed statistics for assessing goaltending.

In golf, the Professional Golfers Association (PGA) has made use of ShotLink data which provides information on every shot played in most PGA tournaments. The data is used in realtime for television telecasts but has also been used to analyze components of the game such as putting [50].

Today's level of sports analytics has evolved where both the technology which provides data, and the statistical methodologies which provide the tools for analysing data, improved very rapidly.

One of the pioneers of modern sports analytics is Bill James who began writing statistically based articles on baseball while working in a factory warehouse. However, many of his articles did not get published because editors felt his style would not have widespread appeal. He then published a series of books titled "Baseball Abstract" beginning in 1977. Although he discontinued writing his self-published books in 1988, he has continued to write books on baseball history. His work was revolutionary because of the use of innovative statistics and his use of play-by-play accounts of games which today is considered absolutely essential for statistical analysis [22].

James initiated a movement today known as sabermetrics, a word coined from SABR (Society for American Baseball Research). SABR is an organization that boasts over 6500 members, produces baseball research and hosts annual meetings. Sabermetric methods have been used to investigate many problems of interest in baseball including the prediction of the outcomes of matches.

Billy Beane, a former professional baseball player who later became the general manager of the Oakland Athletics (a MLB team), further employed sabermetric methods. He was concerned with the evaluation of individual players so as to try and make statistically informed decisions on player selection, rather than depending on traditional scouting methods which involved a certain degree of instinct. Beane's approach provided more objective methods for player selection.

Beane's use of sabermetrics was perhaps the first time that statistics and data were employed to make decisions about player selections in professional sports teams. He believed that a high on-base percentage improved run scoring and resulted in winning more games. He then drafted players who fit that mold rather than considering the outward physical characteristics of individual players. This made the Oakland Athletics a highly successful team and this transformed player selection in baseball as well as other professional sports.

Lewis [30] wrote about Billy Beane and his use of analytics and sabermetrics. Beane put together winning teams with limited budgets by selecting undervalued players whose potential was assessed via sabermetrics. This highly successful book titled "Moneyball: The Art of Winning an Unfair Game" was later made into a movie called "Moneyball" starring Brad Pitt and Jonah Hill which was nominated for six Academy Awards.

The successful results obtained by the use of statistics and the competitive advantage enjoyed by teams that use these techniques has resulted in a shift towards analytics in major North American sports. All of the major professional sports associations have statistics on their websites. Examples include NBA.com, MLB.com, NHL.com, and NFL.com.

Due to the onset of sports analytics and the availability of big data there has been a recent increase in analytics research papers ([32], [31], [21], [11]). Some of the sports focused journals where these appear include the Journal of Quantitative Analysis of Sports (JQAS), the Journal of Sports Analytics (JSA), the International Journal of Computer Science in Sports (IJCSS), the Journal of Sports Science and Medicine (JSSM) and the International Journal of Sports Science and Engineering. There are also numerous international sports analytics conferences which give a platform to disseminate sports analytics. For example, the annual MIT Sloan Sports Analytics Conference provides a forum for professionals, academics, researchers, students, media and other interested parties to discuss sports analytics and learn about the latest innovations.

Though sports analytics has been rapidly developing, it has not been the case with cricket. Due to historical reasons where cricket was perceived as a leisurely gentleman's game played without remuneration to players (until recently), cricket was not subject to large financial transactions.

This has changed in the last few years with the introduction of shorter formats of the game. The shortest and newest format, known as T20, which is explained below, generates intense interest and vast sums of money, especially in the Indian sub-continent. The demand for cricket analytics has increased accordingly and the main website for cricket information and data is cricinfo.com.

The number of research papers on cricket analytics is limited when compared to North American sports. This dearth of research has motivated my own research in this field. This is apart from my avid interest in the game of cricket; as a researcher, as a fan, and as a player. Three of my research papers on cricket form part of this thesis. Due to this exclusive focus on cricket, I believe that my thesis is unique.

1.2 INTRODUCTION TO CRICKET

Cricket is a sport that originated in England in the 16th century and later spread to her colonies. The first international game however did not feature England but was played between Canada and the United States in 1844 at the grounds of the St George's Cricket Club in New York. In time, in both of these countries, cricket took a back seat to other, faster sports like ice-hockey, basketball, and baseball. International cricket is played today by a number of British Commonwealth countries; the main ones being Australia, Bangladesh, England, India, New Zealand, Pakistan, South Africa, Sri Lanka, West Indies and Zimbabwe. These teams are members of the International Cricket Council (ICC). A second rung of international teams called Associates includes numerous countries including Canada.

Cricket is played on an oval-shaped playing field and, apart from baseball, is the only major international sport that does not define an exact size for the playing field. The main action takes place on a rectangular 22 yard area called the pitch in the middle of the large playing field.

Cricket is a game played between two teams of 11 players each, where the two teams alternate scoring (batting) and defending (fielding). A player (bowler) from the fielding team delivers a ball to a player (batsman) from the batting team, who should strike it with a bat in order to score while the rest of the fielding team (fielders) defend the scoring. Furthermore, though it is a team sport, the bowler and batsman in particular, and fielders to some extent, act on their own, each carrying out certain solitary actions independently. A similar sport with respect to individual duties is baseball. The simplicity of these actions (relative to sports such as hockey, soccer and basketball) facilitate statistical modelling. In the process of batting, batsmen can get dismissed (get out) due to a variety of lapses on their part. When all the batsmen from the batting team have been dismissed, or the batting side has faced their allotted number of overs, (each over normally consists of six balls), that team's turn (called innings) is concluded. Their score (number of runs) is recorded. The teams then change places and the fielding team now gets to wield the bat and try to overtake the score of the team that batted first. At the end of one such set of innings (in the shorter versions of cricket) and two such sets of innings (in the longer version of cricket) the winner is selected on the basis of the most runs scored. This is a very simplified explanation of a very complex game, and there are many variables and constraints that come into play. For more details on cricket; see http://www.icc-cricket.com/cricket-rules-and-regulations/.

When international cricket matured, the standard format was a match that could last up to five whole days. This format is called a test match. But even after five days of play the match could end in a draw which means that there is no winner. This was fine in a more leisurely age when both players and spectators had more time, when playing the game was more important than winning, and when most cricketers were amateur players. But as lifestyles became faster, spectators became ever more reluctant or unable to spend five days

watching one match (sometimes with no result). Meanwhile, other faster sports became crowd pullers and earned much more in the way of ticket sales and TV rights. As a result, cricketers became the poor relatives in the sports world.

In the 1960s a shorter version of cricket was developed called 'one-day cricket' with each batting side given 65 overs, and later 50 overs in which to score runs. When this format was used in international matches, they became known as one-day Internationals, or ODIs. This version of cricket was much more exciting to watch, as the batsmen had to wield the bat aggressively. Compared to the five-day long test matches, the advent of the 50-over format was a dramatic improvement in terms of spectator entertainment. However, even a 50-over match lasted about 8 hours and could not compete with the two to three hour match times and attention spans of the fans of ice-hockey, football, baseball and basketball. As competition increased for the sports fans' dollar, and TV advertising income was linked directly to the number of viewers, it was inevitable that a shorter format for cricket would emerge. With declining ticket sales and dwindling sponsorships, the England and Wales Cricket Board (ECB) discussed the options for a shorter and more entertaining game limited to twenty overs per side and the first official game in this format was played on June 13, 2003. This version became known as Twenty20. Since then, Twenty20 cricket has exploded in popularity with the label "Twenty20" being shortened to T20. In 2008 the Indian Premier League (IPL) was inaugurated using the T20 format. The participation of dozens of international players in the IPL tournament, and its exciting format, changed the financial aspects of cricket. The Board of Control for Cricket in India (BCCI) was already the richest cricket administration in the world but the T20 format brought the IPL almost up to the NBA level in terms of team salaries on a pro-rata basis. In 2014, the IPL brand value was estimated at US$ 3.2 billion. Cricket had finally found its Eldorado in T20 (http://www.iplcricketlive.com/indian-premier-league-news/ipl-brand-value-at-4-13-billion/).

Nevertheless, test cricket retains a strong following. Many purists claim it is the only authentic version though it is not financially viable.

1.3 ORGANIZATION OF THE THESIS

The main body of the thesis consists of three projects and a non-statistical essay which are all related to cricket. Chapters 2, 3, and 4 are copies of the papers [36], [35], and [34], respectively, Chapter 5 is an essay of two Sri Lankan cricketing greats, and Chapter 6 provides some concluding remarks.

One of the overriding themes of all three projects is the reliance on computation. All of the work is based on extensive data collection where match commentaries have been parsed. Match commentaries involve the conversations between announcers where the conversations are free flowing with a nonstandard format. We parse these conversations to provide detailed

ball-by-ball data. The use of detailed data of this form has permitted us to go deeper into cricket analytics than otherwise possible. As for computation, we have used a variety of modern tools such as random forests, simulated annealing, empirical Bayes methods, etc.

Chapter 2 considers the decision problem of when to declare during the third innings of a test cricket match. There are various factors that affect the decision of the declaring team including the target score, the number of overs remaining, the relative desire to win versus the desire for a draw, and the scoring characteristics of the particular match. Several decision rules are developed and these are assessed against the decisions made in historical matches. We observed that there are discrepancies between the optimal time to declare and what takes place in practice. One of the main contributions of this chapter is that we provide a table that teams can use in making declaration decisions. This chapter (paper) was published in 2014 in the Journal of Quantitative Analysis in Sports.

Chapter 3 considers the determination of optimal team lineups in T20 cricket where a lineup consists of three components: team composition, batting order, and bowling order. We used the most advanced and sophisticated T20 match simulator to date which is discussed in my research colleague Jack Davis's PhD thesis and the research paper [11]. The simulator generates outcomes for each ball that is bowled where the outcome probabilities depend on the batsman, the bowler, the over, wickets lost, home field advantage, the innings, and the target. Previous cricket simulators by [26], [15], [3], [4] and others did not consider some of these important factors. Using the simulator, this chapter determines optimal lineups using a compelling criterion. Specifically, the expected runs scored minus the expected runs allowed for a given lineup is the basis for assessing lineups. Clearly, run differential is highly correlated with winning matches. Optimal lineups were then determined by optimizing over a vast combinatorial space of lineups via simulated annealing. We observed that the composition of an optimal lineup sometimes resulted in nonstandard lineup. As a by-product of the methodology, we obtained an "all-star" lineup selected from international T20 cricketers. This chapter (paper) might soon be able to say "has been published" by the Journal of Statistical Computation and Simulation.

Chapter 4 is a first attempt to investigate the importance of fielding in cricket. We introduced the metric of expected runs saved due to fielding which is both interpretable and is directly relevant to winning matches. The metric was assigned to individual players and is based on a textual analysis of match commentaries using random forest methodology. We observed that the best fielders save on average 1.2 runs per match compared to a typical fielder. This chapter (paper) has been submitted for publication and is currently under review.

Chapter 5 is a non-statistical essay of two cricketing greats from Sri Lanka (Murali and Sangakkara) who established numerous world records and recently retired from the game. Though their record-breaking performances are now part of cricketing lore, this chapter is not a contribution which adds to the statistical literature in sport, and should not be

regarded as a component of the thesis in terms of analytics. However, during the course of my PhD studies, my supervisor Professor Tim Swartz and I were contacted by Dr. Joel Nathan Rosen, Associate Professor of Sociology at Moravian College, Pennsylvania. He invited us to write a chapter for a scholarly volume which will be published in 2016 by the University Press of Mississippi. This article occupied about two terms of my Ph.D studies and is included to demonstrate my non-statistical writing skills. Again, although Chapter 5 contains nothing statistical, I feel that this contribution rounds out my thesis and reflects my love for the game of cricket.

Chapter 2

Declaration Guidelines in Test Cricket

2.1 INTRODUCTION

Although various versions of the sport of cricket have been played since the 16th century, test cricket is considered by most to be the traditional form of cricket played at the highest level. Test cricket matches between two teams may take up to five days to complete where these teams are full member nations of the International Cricket Council (ICC). Currently, there are 10 full member nations of the ICC and they are Australia, Bangladesh, England, India, New Zealand, Pakistan, South Africa, Sri Lanka, West Indies and Zimbabwe.

The laws (rules) of cricket are extensive, and are maintained by the Marylebone Cricket Club (MCC); see http://www.lords.org/laws-and-spirit/laws-of-cricket/. We now provide a very basic description of test cricket relevant to the problem examined in this paper.

A test match begins with Team A batting, and typically, Team A accumulates *runs* until 10 of their *batsmen* have been *dismissed*. This concludes the first *innings*. The runs that are accumulated during batting are the consequence of batsmen facing balls bowled by *bowlers*. A group of 6 balls bowled by the same bowler is referred to as an *over*. Team B then comes to bat during the second innings and likewise accumulates runs until 10 of their batsmen have been dismissed. Assuming that there is no *follow-on*[1], Team A bats again in the third innings. During the third innings, we are interested in the case where Team A's cumulative runs exceed Team B's runs, and this establishes a *target* score for Team B to achieve in the fourth and final innings. The question that continually faces Team A during the third innings is whether they should voluntarily terminate their innings. This is known as a *declaration*, and the decision to declare is the subject of this paper.

[1]At the completion of the second innings of a scheduled five-day match, if the team batting first leads by at least 200 runs, then it has the option of forcing the second innings batting team to bat again in the third innings. The sequence of batting innings according to Team A, Team B, Team B, and Team A (if required) is the result of team A enforcing the follow-on.

To understand the motivation behind declaration, it is necessary to understand how matches terminate. If during the fourth innings, 10 batsmen of Team B are dismissed with Team B not reaching the target, then Team A is the winner of the match. Alternatively, if at some point during the fourth innings, Team B exceeds the target, then Team B is the winner. However, a common occurrence during the fourth innings is that the five-day time limit for the match is attained prior to Team B exceeding the target or losing its 10 wickets. In this case, the match is recorded as a draw. Therefore, during the third innings, Team A is trying to assess the merits of two subtle alternatives:

- to continue batting which leads to a more imposing target and reduces the chance of losing while simultaneously providing less time to dismiss Team B in the fourth innings and hence increasing the probability of a draw
- to declare, which concludes the third innings, establishes a target and leaves Team A more vulnerable to losing while simultaneously providing more time to dismiss Team B in the fourth innings

Although the declaration problem is fundamental to test cricket strategy, it has surprisingly received little attention from a quantitative perspective. To our knowledge, there are only two academic papers on the problem, [42], and an enhanced approach given by [41]. We refer to the two papers collectively as SETAL (Scarf et al.). Although we later make a comparison of our results with those of SETAL, we note there are substantial differences in the two approaches. In particular, SETAL fit a multinomial regression model for match outcomes with covariates selected to improve fit. SETAL then provide estimated probabilities of match outcomes given the end of the first, second and third innings for various states of matches. SETAL also analyse the follow-on decision. Alternatively, we propose a probabilistic criterion to assess whether a team should declare in the third innings. The criterion essentially addresses whether a team is better off declaring immediately or waiting for another ball to declare. Although we do not utilize covariates, we provide a range of declaration guidelines to suit various types of matches. Another difference concerns the datasets. SETAL used a dataset of 391 test matches to estimate batting characteristics. On the other hand, we use situational datasets and ball-by-ball data. For example, to study third innings batting prior to a declaration, we consider a restricted dataset of 134 test matches where declarations occurred in the third innings. The intuition (which we confirm in section 2.4.1) is that immediately prior to declaring in the third innings, the batting team becomes increasingly aggressive in an attempt to pile up runs without regard to dismissals. For an overview of other decision problems related to test cricket, the reader is referred to [9].

Section 2.2 begins with an exploratory data analysis. We make the observation that declaring teams rarely lose. Section 2.3 provides a formalization of the decision problem in terms of expected match outcomes. The solution to the problem provides a yes/no

answer as to whether a team should declare at any given stage of the third innings. Various parameters are introduced, and this provides flexibility in the decision making according to the circumstances of a particular match. Parameter estimation is discussed in section 2.4. For the casual reader, section 2.4 may be skipped without jeopardizing comprehension of the main results. In contrast to our estimation procedures, [43] fit negative binomial distributions to the runs scored in innings and in partnerships. They also discuss strategy in the context of third innings batting. In section 2.5, we provide guidelines (Table 2.4) as to when declaration is optimal. In some circumstances, the guidelines deviate substantially from traditional practice. Table 2.4 is the major contribution of the paper. In section 2.6, we assess the proposed decision rules. We first provide a comparison of our win-draw-loss probabilities to the probabilities obtained by SETAL. We observe reasonable agreement in many of the scenarios. We also examine historical matches and investigate how the matches may have turned out had our decision rules been implemented. We discover that our decision rules are helpful. We conclude with a brief discussion in section 2.7.

2.2 EXPLORATORY DATA ANALYSIS

For this section we collected data on 134 test matches where a declaration occurred in the third innings followed by batting in the fourth innings. These matches involved the 10 ICC teams during the period March 1, 2001 through March 31, 2012. The data are presented in various ways to provide insight on the declaration problem.

In Table 2.1, we record the number of times that the declaring team won, lost and drew matches. We observe that there are considerable differences amongst the 10 ICC nations. For example, some of the teams (e.g. Bangladesh, West Indies and Zimbabwe) rarely declare. This may not be a reflection of their cautiousness. Rather, these are weaker teams and they are rarely in a position to declare. We also observe that Australia declares often and has a higher percentage of wins in the declared matches than the other ICC teams. This suggests that Australia is bolder with respect to declaring. Although Australia appears to more readily declare, they do so without increasing their loss percentage. Generally, it seems that declaring rarely results in a loss, and overall, declaring teams win at roughly the same rate at which they draw.

In Figure 2.1, we provide a scatterplot of the target set by the declaring team versus the remaining overs in the match at the time of declaration. The remaining overs were calculated by noting that test cricket allows up to 90 overs in a day. As anticipated, the target is greater when more overs remain. We again observe that declaring teams lose very few matches (triangle symbol) which suggests that teams are often cautious in the sense that they declare late in matches. The scatterplot also suggests that with more than 150 overs remaining, the declaring team usually wins.

Team	N	Declared Matches	Wins	Losses	Draws
Australia	62	29	22 (76%)	1 (3%)	6 (21%)
Bangladesh	45	3	1 (33%)	0 (0%)	2 (67%)
England	62	24	12 (50%)	1 (4%)	11 (46%)
India	46	16	6 (38%)	0 (0%)	10 (63%)
New Zealand	38	6	4 (67%)	0 (0%)	2 (33%)
Pakistan	32	11	7 (64%)	0 (0%)	4 (36%)
South Africa	47	22	7 (32%)	1 (5%)	14 (64%)
Sri Lanka	47	16	11 (69%)	0 (0%)	5 (31%)
West Indies	35	3	1 (33%)	0 (0%)	2 (67%)
Zimbabwe	21	4	2 (50%)	0 (0%)	2 (50%)
Overall	435	134	73 (51%)	3 (2%)	67 (47%)

Table 2.1: Match outcomes for teams that declared in the third innings. The column N is the total number of matches over the period where the team batted in the first and third innings.

Figure 2.2 is a scatterplot of the fourth innings runs versus the overs consumed in the fourth innings. We observe that the number of runs scored is roughly proportional to the number of overs. This may be a little surprising since better batsmen tend to bat earlier in batting lineups. We also observe that there does not appear to be any systematic difference in the number of runs scored between batting teams that are all out (plus symbol) and those that are not all out (circle symbol). The variability of the number of runs scored increases with the numbers of overs consumed. All of these observations are relevant to the modelling assumptions made in sections 2.4.2 and 2.4.3.

In Figure 2.3, we provide a more comprehensive scatterplot of runs scored versus overs used. The scatterplot involves all of the 1849 innings arising from test matches involving ICC nations between March 1, 2001 and March 31, 2012. Whereas this data is used for parameter estimation in section 2.4, the key feature at this stage is that the scatterplot is similar to the scatterplot in Figure 2.2. This suggests that the distribution of runs scored in the fourth innings following a third innings declaration does not differ from the distribution of runs scored in general.

2.3 THE DECISION PROBLEM

In formulating the decision problem, we consider Team A batting in the third innings, pondering whether they should declare. We denote a win and a draw for Team A as W and D respectively, and let y be the number of *overs* remaining in the match. Again, the value y can be obtained at any point during a match since test cricket allows 90 overs in a day. Since we are interested in how the situation changes for Team A over time, we let y' be the number of overs y remaining in the match less one ball bowled. To complete the notation,

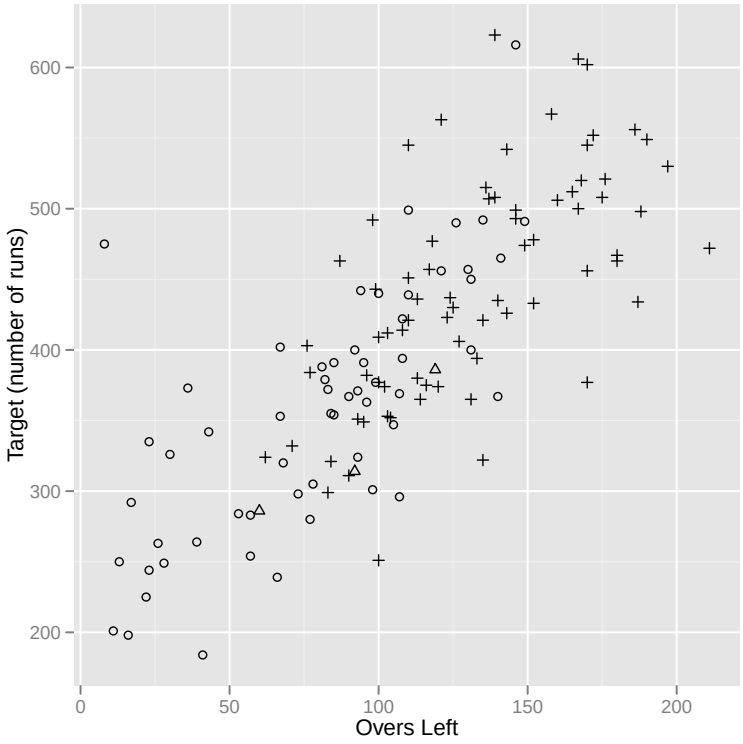

Figure 2.1: Scatterplot of the target versus the remaining overs for matches with declarations in the third innings. Plus/circle/triangle symbols indicate that the declaring team has gone on to win/draw/lose the match.

we let Q_y correspond to the decision that Team A declares with y overs remaining and we let T_y be the corresponding target score with y overs remaining.

In most sports (soccer notably excluded), teams receive two points for winning a match, one point for a draw and zero points for a loss. This is an intuitive scoring system where a draw is assigned half the value of a win. In this case, there is a total of two points assigned to the two teams under all possible match outcomes. Although points are not explicitly awarded in a test cricket series, we note that the above point system corresponds to how cricket series are decided. For example, a five-match series is considered drawn should either a 2-1-2 result or a 1-3-1 result occur according to Win-Draw-Loss. In these two cases, both results yield five points. Accordingly, we assert that Team A should declare when the bowling of an additional ball causes their expected number of match points to decrease. Using the notation introduced above, Team A should declare with y overs remaining if

$$\begin{aligned} & 2 \cdot \text{Prob}(W \mid T_y \ ; \ Q_y) + 1 \cdot \text{Prob}(D \mid T_y \ ; \ Q_y) \\ > \ & 2 \cdot \text{Prob}(W \mid T_y \ ; \ Q_{y'}) + 1 \cdot \text{Prob}(D \mid T_y \ ; \ Q_{y'}) \ . \end{aligned} \tag{2.1}$$

Now, there may be circumstances where winning is imperative. That is, for Team A, a draw is no better than a loss. An example of this may occur in a four-match test cricket

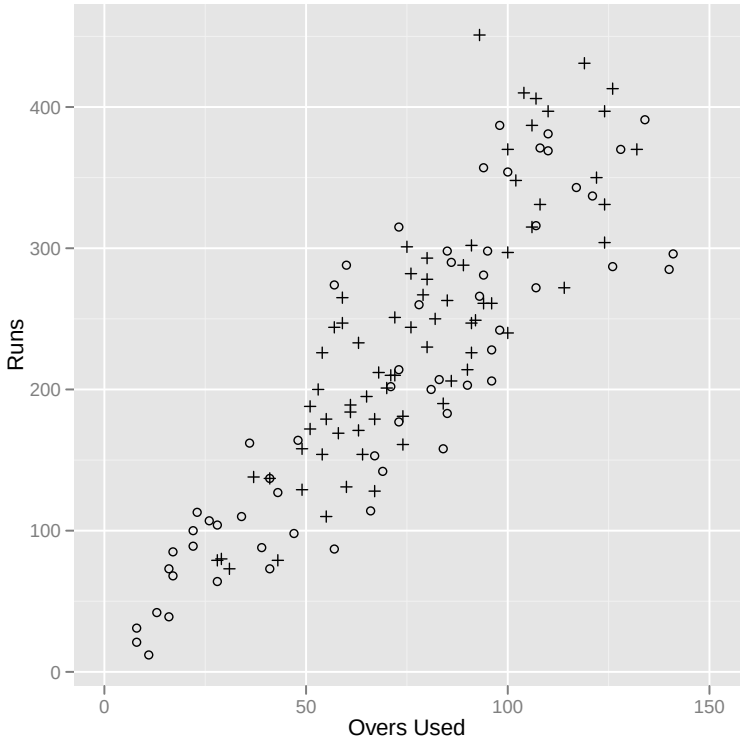

Figure 2.2: Scatterplot of fourth innings runs versus overs consumed for matches with declarations in the third innings. Plus (circle) symbols indicate that the batting team was all out (not all out) at the end of the fourth innings.

series where after three matches, Team A has one win and two losses. For Team A, a win in the fourth and final match will result in a drawn series whereas anything less than a win in the fourth match results in a lost series. In this case, we modify (2.1) and assert that Team A should declare with y overs remaining if

$$\text{Prob}(W \mid T_y \ ; \ Q_y) \ > \ \text{Prob}(W \mid T_y \ ; \ Q_{y'}) \ . \tag{2.2}$$

For a discussion of the implications of awarding two versus three points for a win in soccer, see [7].

It may also be possible that the objective of Team A is to avoid losing a match at all costs. For example, Team A may be leading a five match series with one win and three draws. Either a win or a draw in the fifth match will allow them to win the series. However, a loss in the fifth match will result in a drawn series. When Team A is batting in the third innings, their probability of losing the match decreases as they continue to bat. This is because they are simultaneously accumulating runs and providing less time for Team B to score runs. Therefore, if the objective is strictly not to lose, Team A should never declare. In this scenario, there is no need for a quantitative analysis.

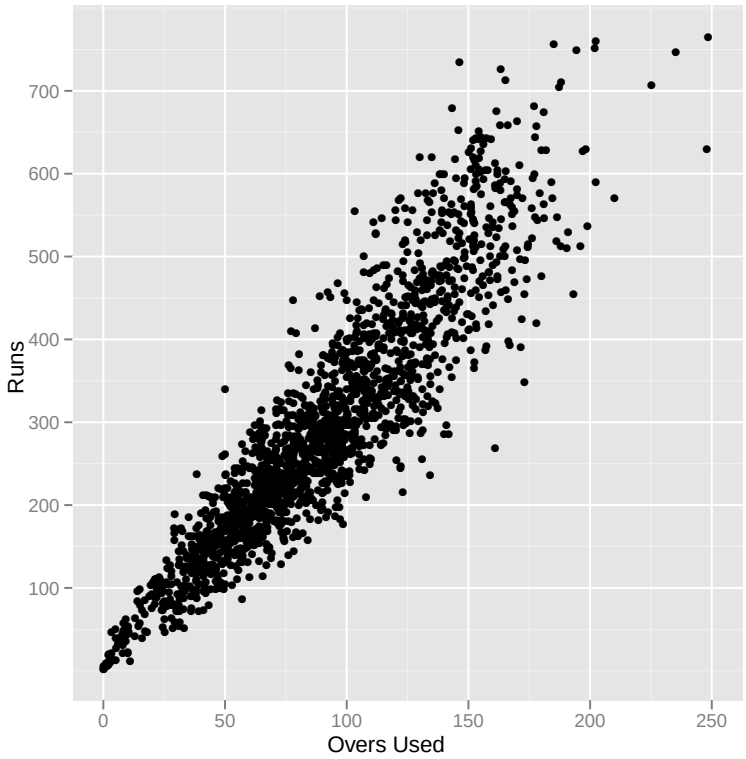

Figure 2.3: Scatterplot of runs scored (in an innings) versus overs used in test matches involving ICC nations between March 1, 2001 and March 31, 2012.

The evaluation of (2.1) and (2.2) can be simplified by invoking the total law of probability. Accordingly, consider the possible outcomes that can occur on the next ball bowled:

$$\begin{aligned} B_{-1} &\equiv \text{dismissal} \\ B_i &\equiv i \text{ runs scored}, \quad i = 0, \ldots, 6 \end{aligned}$$

where we exclude the possibility of extra runs that can be accumulated via wide-balls and no-balls. In fact, we do account for extra runs but avoid the presentation here to maintain the simplified notation. It follows that

$$\begin{aligned} \text{Prob}(W \mid T_y \,;\, Q_{y'}) &= \text{Prob}(WB_{-1} \cup WB_1 \cup \cdots \cup WB_6 \mid T_y \,;\, Q_{y'}) \\ &= \sum_{i=-1}^{6} \text{Prob}(WB_i \mid T_y \,;\, Q_{y'}) \\ &= \sum_{i=-1}^{6} p_i \, \text{Prob}(W \mid T_{y'(i)} \,;\, Q_{y'}) \end{aligned} \tag{2.3}$$

where $T_{y'(i)} = T_y + i\delta(i \neq -1)$, δ is the indicator function and

$$p_i = \text{Prob}(B_i) \tag{2.4}$$

is the probability of B_i, $i = -1, \ldots, 6$, where Team A is batting and is on the verge of declaring. Similarly,

$$\text{Prob}(D \mid T_y \ ; \ Q_{y'}) \ = \ \sum_{i=-1}^{6} p_i \ \text{Prob}(D \mid T_{y'(i)} \ ; \ Q_{y'}) \ . \tag{2.5}$$

We therefore observe that the terms in (2.3) have the same form as $\text{Prob}(W \mid T_y \ ; \ Q_y)$ and that the terms in (2.5) have the same form as $\text{Prob}(D \mid T_y \ ; \ Q_y)$. This simplifies the evaluation of the inequalities (2.1) and (2.2).

At this stage, we provide a partial recap of what needs to be done in order to decide whether Team A should declare in the third innings with y overs remaining and a target of T_y. The captain must first decide whether a win is essential. If so, then the simpler inequality (2.2) provides the appropriate criterion; otherwise inequality (2.1) is the appropriate criterion. To evaluate (2.1) and (2.2), we need to first estimate the batting characteristics of Team A in the third innings given by the parameters $p_{-1}, \ldots, p_6$ in (2.4). Secondly, we need to estimate $\text{Prob}(W \mid T_y \ ; \ Q_y)$ which is dependent on the batting performance of Team B in the fourth innings. Thirdly, we may need to estimate $\text{Prob}(D \mid T_y \ ; \ Q_y)$ which is also dependent on the batting performance of Team B in the fourth innings. We elaborate on all three estimation procedures in section 2.4.

2.4 PARAMETER ESTIMATION

Data were collected from the Cricinfo website http://www.espncricinfo.com/ where matches were filtered using Statsguru. In cases where ball-by-ball data were required, we accessed matches from the Cricinfo Archive where Commentary logs were downloaded. The Commentary logs appear to be the only comprehensive source of ball-by-ball data. An R script was then used to parse the commentary logs and obtain ball-by-ball data in a convenient format.

2.4.1 Estimation of $p_{-1}, \ldots, p_6$

The parameters $p_{-1}, \ldots, p_6$ given in (2.4) describe the batting characteristics of Team A when they are on the verge of declaring. With ball-by-ball data, we are able to obtain the proportions corresponding to the batting events $B_{-1}, \ldots, B_6$. We collected data on 134 test matches for which a declaration occurred in the third innings. These matches involved the 10 ICC teams during the period March 1, 2001 through March 31, 2012.

In Table 2.2, we present the results for varying overs remaining in the third innings. We observe that teams that are on the verge of declaring (i.e. 5 and 10 overs remaining) bat more aggressively than they generally do throughout the third innings. For example, the proportion of 6's is more than three times as large in the second and third rows of Table 2.2 than in the first row. This is sensible since teams on the verge of declaring are trying to pile up runs without regard to dismissals. We also note that there is greater stability

in the proportions $\hat{p}_1, \ldots, \hat{p}_6$ than in $\hat{p}_{-1}$ and $\hat{p}_0$ when comparing the results in the second and third rows of Table 2.2. This is good news since only $p_1, \ldots, p_6$ contribute to the target score $T_{y'(i)}$ in (2.3) and (2.5). For our analyses in section 2.5, we use the estimates from the third row of Table 2.2 to describe Team A's batting characteristics when they are on the verge of declaring in the third innings.

Overs Remaining	Balls Bowled	$\hat{p}_{-1}$	$\hat{p}_0$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$	$\hat{p}_5$	$\hat{p}_6$
All	62948	0.012	0.664	0.206	0.040	0.009	0.065	0.000	0.005
10	7496	0.027	0.519	0.293	0.060	0.007	0.080	0.000	0.014
5	3659	0.038	0.490	0.306	0.063	0.008	0.079	0.000	0.016

Table 2.2: Sample proportions corresponding to the parameters $p_{-1}, \ldots, p_6$ in (2.4). The proportions are obtained for varying overs remaining in the third innings.

Now it is possible that the estimates from the third row in Table 2.2 vary according to the quality of Team A. We therefore repeated the above exercise and obtained team specific estimates over the last five years. We discovered that there are only minor differences in the characteristics of the 10 teams. These differences also proved inconsequential with respect to the optimal time to declare.

2.4.2 Estimation of $\text{Prob}(W \mid T_y \ ; \ Q_y)$

Recall that $\text{Prob}(W \mid T_y \ ; \ Q_y)$ is the probability that Team A wins the match given that they declare with y overs remaining in the match and the target score is T_y. The event W occurs if Team B scores fewer than T_y runs and is dismissed in y overs or less. Therefore

$$\begin{aligned} \text{Prob}(W \mid T_y \ ; \ Q_y) &= \text{Prob}(R_y < T_y \ \cap \ O_1 + \cdots + O_{10} \leq y) \\ &= \sum_{i=0}^{y} q_i \ \text{Prob}(R_y < T_y \mid O_1 + \cdots + O_{10} = i) \end{aligned} \tag{2.6}$$

where R_y is the number of runs scored by Team B in the fourth innings with y overs available, O_i is the number of overs that it takes for the ith wicket to fall and

$$q_i = \text{Prob}(O_1 + \cdots + O_{10} = i) \tag{2.7}$$

for $i = 0, \ldots, y$.

The estimation of q_i in (2.7) is carried out by modelling the O_j's as independent geometric variables. Each O_j has a single parameter that can be estimated from the batting records in the fourth innings following a third innings declaration. When the geometric parameters are obtained, then q_i can be calculated via the convolution formula. The convolution formula is useful when the distribution of a sum of random variables is required.

There is a catch in the estimation of the geometric parameters, and this involves censored data. Consider matches where we wish to estimate the geometric parameter r_j correspond-

ing to O_j. The probability mass function for O_j is

$$\text{Prob}(O_j = k) = r_j(1-r_j)^{k-1} \qquad k = 1, 2, \ldots \tag{2.8}$$

From the n_j fourth innings matches where there were at least $j-1$ dismissals, we have data $O_j^{(1)}, \ldots, O_j^{(l_j)}, O_j^{(l_j+1)}, \ldots, O_j^{(n_j)}$ where the first l_j observations are not censored (i.e. these are the number of overs that it took for the jth wicket to fall) and the remaining $n_j - l_j$ observations are censored (i.e. these are the number of overs from wicket $j-1$ until the end of the match). The likelihood corresponding to $O_j^{(1)}, \ldots, O_j^{(l_j)}, O_j^{(l_j+1)}, \ldots, O_j^{(n_j)}$ is therefore given by

$$\prod_{k=1}^{l_j} r_j(1-r_j)^{O_j^{(k)}-1} \prod_{k=l_j+1}^{n_j} (1-r_j)^{O_j^{(k)}} = \left(\frac{r_j}{1-r_j}\right)^{l_j} (1-r)^{\sum_{k=1}^{n_j} O_j^{(k)}}$$

from which we obtain the maximum likelihood estimator $\hat{r}_j = l_j / \sum_{k=1}^{n_j} O_j^{(k)}$. For our geometric model, the only assumption is that the outcomes of balls (dismissal or not) consist of independent and identically distributed Bernoulli trials.

In section 2.5 we present optimal declaration guidelines, and this is done in an average sense. However, we also want to take into account both weak and strong teams. For the estimation of $r_1, \ldots, r_{10}$, we define a notional weak/strong team as one whose parameter estimates lie one standard deviation above/below the maximum likelihood estimates. For these calculations, we use a Fisher information approximation to the asymptotic variance of the maximum likelihood estimator.

Table 2.3 provides the maximum likelihood estimates of the geometric parameters $r_1, \ldots, r_{10}$ for an average team using the fourth innings data. We also provide estimates for relatively weak and relatively strong teams. As expected, we observe that the geometric parameters are generally increasing as the number of wickets fall. This implies that batsmen further down the batting order are more likely to be dismissed. The interesting exception is the opening partnership where dismissals occur at a higher rate than some of the other early partnerships. This may be due to the fact that there is an adjustment period at the beginning of an innings where batsmen need to acquire comfort with a new ball, changed field conditions, new bowlers, etc.

Team B	$\hat{r}_1$	$\hat{r}_2$	$\hat{r}_3$	$\hat{r}_4$	$\hat{r}_5$	$\hat{r}_6$	$\hat{r}_7$	$\hat{r}_8$	$\hat{r}_9$	$\hat{r}_{10}$
Weak	0.117	0.085	0.082	0.087	0.083	0.102	0.142	0.179	0.193	0.228
Average	0.108	0.079	0.076	0.081	0.077	0.094	0.131	0.166	0.179	0.212
Strong	0.099	0.073	0.070	0.075	0.071	0.086	0.120	0.153	0.165	0.196

Table 2.3: Estimates of the geometric parameters r_j as defined in (2.8) for weak, average and strong teams.

The final estimation procedure in this subsection concerns the conditional distribution of R_y given $O_1 + \cdots + O_{10} = i$ for the probability specified in (2.6). We first postulate that

the number of runs R_y scored by Team B in the fourth innings with y overs available given that they are all out in i overs, $i = 1, \ldots, y$ can be approximated via

$$[R_y \mid O_1 + \cdots + O_{10} = i] \sim \text{Normal}(\mu \cdot i, \sigma^2 \cdot i) \ . \tag{2.9}$$

The proposed distribution (2.9) only requires the estimation of the two parameters μ and σ. The rationale of the normal distribution is based on a Central Limit Theorem effect whereby runs are accumulated as the sum over many balls. By considering the runs scored and the associated number of overs used in the fourth innings data, we obtained maximum likelihood estimates $\hat{\mu} = 3.02$ and $\hat{\sigma} = 5.95$.

Whereas (2.9) provides an appealing rough-and-ready distribution for runs scored, we want to ensure that it is realistic. Some doubt concerning the adequacy of (2.9) may be seen by considering the case of $i = 50$ overs. Here, $\hat{\mu} \cdot i \pm 2\hat{\sigma}\sqrt{i}$ provides the interval $(116.6, 185.4)$ which seems too narrow for the number of runs scored when compared to Figure 2.2.

Alternatively, we consider a more data-driven approach with minimal assumptions on the mean-variance structure. Specifically, we propose

$$[R_y \mid O_1 + \cdots + O_{10} = i] \sim \text{Normal}(\mu_i, \sigma_i^2) \tag{2.10}$$

where both the means μ_i and the variances σ_i^2 are assumed to be increasing relative to the number of overs i. By introducing reference priors $[\mu_i] \propto 1$ and $[\sigma_i^2] \propto 1/\sigma_i^2$ subject to the order constraints, we have a high dimensional Bayesian model. Let $R_{i1}, \ldots, R_{in_i}$ be the number of runs scored in the n_i fourth innings that lasted exactly i overs. Then the posterior means of the μ_i and σ_i are readily obtained through a Gibbs sampling algorithm where full conditional distributions are given by

$$\begin{array}{lcll} [\mu_i \mid \cdot\,] & \sim & \text{Normal}(\bar{R}_i, \sigma_i^2/n_i) & \mu_{i-1} < \mu_i < \mu_{i+1} \\ [\sigma_i^2 \mid \cdot\,] & \sim & \text{Inverse Gamma}(n_i/2, 2/(\sum_{j=1}^{n_i}(R_{ij} - \mu_i)^2) & \sigma_{i-1} < \sigma_i < \sigma_{i+1},\ n_i > 1 \\ [\sigma_i^2 \mid \cdot\,] & \propto & 1/\sigma_i^2 & \sigma_{i-1} < \sigma_i < \sigma_{i+1},\ n_i = 1 \ . \end{array}$$

The generation of variates from constrained distributions is facilitated using inversion. When a random variable X has a cumulative distribution function F and is further constrained to the interval (a, b), inversion proceeds by obtaining $x = F^{-1}(F(a) + u(F(b) - F(a)))$ where $u \sim \text{uniform}(0, 1)$.

2.4.3 Estimation of $\text{Prob}(D \mid T_y \ ; \ Q_y)$

Recall that $\text{Prob}(D \mid T_y \ ; \ Q_y)$ is the probability that Team A draws the match given that they declare with y overs remaining and a target score T_y. The event occurs if Team B

scores no more than T_y runs and is not dismissed in the y overs available. Therefore

$$\begin{aligned} \text{Prob}(D \mid T_y \; ; \; Q_y) &= \text{Prob}(R_y \leq T_y \; \cap \; O_1 + \cdots + O_{10} > y) \\ &= \text{Prob}(R_y \leq T_y \mid O_1 + \cdots + O_{10} > y) \; \text{Prob}(O_1 + \cdots + O_{10} > y) \end{aligned} \tag{2.11}$$

where R_y is the number of runs scored by Team B in the fourth innings with y overs available, O_i is the number of overs that it takes for the ith wicket to fall and

$$\text{Prob}(O_1 + \cdots + O_{10} > y) \;\; = \;\; 1 - \sum_{i=0}^{y} \; q_i \tag{2.12}$$

with q_i defined in (2.7).

The estimation of (2.12) is the same as that described in subsection 2.4.2. As for the estimation of the first term in (2.11), we propose the distribution

$$[R_y \mid O_1 + \cdots + O_{10} > y] \sim \text{Normal}(\mu_y, \sigma_y^2) \; . \tag{2.13}$$

When comparing (2.10) to (2.13), we argue that it is the number of overs used which is paramount in the run distribution. In fact, the scatterplot of runs versus overs used (when all-out) does not differ materially from the scatterplot of runs versus overs used (when not all-out). The synthesis of the two scatterplots appears in Figure 2.3.

For the estimation of the constrained parameters μ_i and σ_i, we considered the 1849 innings corresponding to matches between ICC nations from March 1, 2001 through March 31, 2012. Using the Gibbs sampling algorithm described above, examples of estimates that we obtained are $\hat{\mu}_{50} = 169.6$, $\hat{\mu}_{100} = 316.9$, $\hat{\mu}_{150} = 508.6$, $\hat{\sigma}_{50} = 35.3$, $\hat{\sigma}_{100} = 55.1$ and $\hat{\sigma}_{150} = 75.1$. These estimates appear sensible when compared to both Figure 2.2 and Figure 2.3. To gain some further intuition for the estimates, consider the probability of scoring at least 400 runs in 100 overs of batting. The probability is given by $\text{Prob}(Z \geq (400 - 316.9)/55.1) = 0.07$ where Z is a standard normal random variable.

Posterior standard deviations of the parameters are also readily available from the Gibbs sampling algorithm and they provide us with a measure of confidence in the estimates. For example, we obtained $\text{sd}(\mu_{50}) = 3.7$, $\text{sd}(\mu_{100}) = 5.5$, $\text{sd}(\mu_{150}) = 7.0$, $\text{sd}(\sigma_{50}) = 1.2$, $\text{sd}(\sigma_{100}) = 3.0$ and $\text{sd}(\sigma_{150}) = 5.8$. We observe that the posterior standard deviations increase for increasing numbers of overs. This is somewhat expected as data becomes more scarce for increasing numbers of overs. We therefore did not produce estimates beyond $i = 180$ overs. The upper value of $i = 180$ overs corresponds to two full days of batting.

Again, we also want to account for relatively strong and weak teams. We define a relatively strong (weak) Team B as one whose expected number of runs $\hat{\mu}_i$ in i overs of batting is 20% above (below) the posterior mean. In the above example, we obtain $\hat{\mu}_{100} =$ 380.3 (253.5) runs for a relatively strong (weak) Team B. Of course, nothing prevents the selection of differentials other than 20%.

2.5 RESULTS

At this stage, we provide some practical declaration guidelines. First, a decision maker needs to determine how badly Team A wants to win a given match. If winning the match is of the utmost importance, then criterion (2.2) is used for decision making; otherwise criterion (2.1) is used. In the midst of a match, it may be inconvenient for a decision maker to access a computer, input parameter estimates and evaluate the relevant inequality. Instead, we have produced Table 2.4 which summarizes optimal declaration targets for specified overs remaining in a match. The table takes into account whether Team B is relatively weak, average or strong compared to Team A. For a particular situation, Table 2.4 was constructed by systematically increasing the target for a fixed number of overs until the relevant inequality was satisfied. Table 2.4 is the major contribution of the paper, and it is our hope that it may be utilized for better decision making in test cricket.

A stunning feature of Table 2.4 is that the decision to declare depends critically on how badly Team A wants to win. Under criterion (2.2) where Team A is desperate to win, they should declare with a much smaller target than under criterion (2.1). For example, with 90 overs remaining in a match, a desperate Team A (criterion (2.2)) should declare with a target of 296 runs against an average opponent. This is contrasted with a target of 359 runs using criterion (2.1) against an average opponent. The difference sensibly dissipates with a large number of overs remaining since Team B will most likely be made out and there is little chance that the match will end in a draw. For example, with 180 overs remaining in a match, a desperate Team A should declare with a target of 629 runs against an average opponent. This is contrasted with a comparable target of 652 runs using criterion (2.1). We also observe that the quality of Team B has an impact on the declaration guidelines where weaker/stronger opponents naturally require a lower/higher target. For example, under criterion (2.1) with 120 overs remaining in the match, Team A should declare with a target of 389/467/548 runs against weak/average/strong opponents. We also observe that the recommendations in Table 2.4 can differ substantially from common practice. For example, Figure 2.1 suggests that with 50 overs remaining, teams declare when the target is roughly 300 runs. In contrast, Table 2.4 requires far fewer runs than that, a target of 254 runs against an average team using criterion (2.1). With large numbers of overs remaining, the opposite phenomenon takes place. That is, Table 2.4 indicates that teams should declare with a higher target than is common practice. For example, when 150 overs are remaining, common practice suggests a declaration target of roughly 500 runs whereas Table 2.4 provides a target of 592 runs against an average team using criterion (2.1). The target guidelines for large numbers of overs may first appear to be in conflict with Figure 2.1. In Figure 2.1, we observe that the declaring team (Team A) nearly always wins when there is a large number of overs remaining, and this is suggestive that they could declare earlier. However, there were only 21 matches where teams declared with more than 150

overs remaining. Although Team B was dismissed and did not reach its target in this small sample of matches, Figure 2.3 suggests that teams have the run scoring capability of doing so.

Overs	Criterion (2.1) - Win or Draw			Criterion (2.2) - Must Win		
Remaining	Weak	Average	Strong	Weak	Average	Strong
50	214	254	295	131	161	192
60	229	274	320	158	193	228
70	253	304	356	189	228	269
80	274	331	389	219	263	309
90	298	359	422	247	296	346
100	331	399	469	277	331	387
110	361	436	513	311	371	433
120	389	467	548	343	409	477
130	423	507	595	377	448	522
140	457	548	643	413	491	571
150	494	592	679	450	535	623
160	527	612	704	486	576	648
170	556	631	728	521	605	674
180	585	652	752	552	629	699

Table 2.4: Optimal target scores for declaration. The targets are calculated for a specified number of overs remaining in a match under both criterion (2.1) and criterion (2.2) and according to whether Team B is a weak, average or strong scoring team relative to Team A.

We note that the quality of Team B in Table 2.4 is a characterization of its ability to score runs in the fourth innings. Clearly, the capacity for Team B to score runs also depends on the bowling and fielding standard of Team A. Therefore, a decision ought to take Team A into account when referring to Table 2.4. For example, if Team A is a strong bowling team and Team B is an average batting team, then one may want to refer to the "Weak" category in Table 2.4. A decision maker may alternatively choose intermediate values between the three categories of weak, average and strong.

2.6 ASSESSING THE DECISION RULES

Associated with our decision rules in Table 2.4 are win, loss and draw probabilities corresponding to Team A where Team B is assumed to be a relatively average team. In Table 2.5, we provide these estimated probabilities for declarations in the third innings, and we compare these probabilities against the probabilities reported by SETAL. Specifically, we use the SETAL results from Table 2.4 of their 2011 paper which represent standard conditions.

From Table 2.5, we first observe that our probabilities generally correspond to intuition. For example, for a given target score, it becomes more probable for Team A to lose a match as the number of remaining overs increase. The same is true with the SETAL probabilities.

We also observe general agreement between our probabilities and the SETAL probabilities which is reassuring since the two methods of estimation are dramatically different. In the case of substantial differences, there does not appear to be any obvious pattern. One of the largest differences occurs in the cell with 60 overs remaining and a target score of 200 runs. Referring to Figure 2.3, although it is plausible to score at least 200 runs from 60 overs, it seems that the probability of the event may be less than 0.5. Recall that the horizontal axis in Figure 2.3 is overs used, and frequently, Team B is dismissed prior to 60 overs. This suggests that the SETAL L (loss) probability of 0.49 may be too large. Perhaps the true probability of L lies somewhere between 0.29 and 0.49. We also comment on some of the other substantial differences between our probabilities and the SETAL probabilities. These larger differences tend to occur in the cells with 100-160 overs remaining and target scores between 250 and 350 runs. Referring to Figure 2.1, these cells correspond to scenarios where declarations rarely occur. In these instances, we provide larger L probabilities than SETAL. Although we do not have great intuition for these scenarios, the reason teams do not declare in these cases is because they believe that there is considerable probability that they may lose.

		Overs Remaining					
Target		60	80	100	120	140	160
	W	0.09(0.19)	0.19(0.23)	0.20(0.25)	0.20(0.25)	0.20(0.25)	0.20(0.24)
200	D	0.62(0.33)	0.10(0.14)	0.00(0.05)	0.00(0.02)	0.00(0.01)	0.00(0.00)
	L	0.29(0.49)	0.71(0.63)	0.80(0.70)	0.80(0.73)	0.80(0.75)	0.80(0.76)
	W	0.11(0.23)	0.30(0.37)	0.39(0.45)	0.40(0.49)	0.40(0.50)	0.40(0.50)
250	D	0.87(0.57)	0.38(0.32)	0.03(0.14)	0.00(0.05)	0.00(0.02)	0.00(0.01)
	L	0.02(0.20)	0.32(0.32)	0.58(0.41)	0.60(0.46)	0.60(0.48)	0.60(0.50)
	W	0.11(0.21)	0.34(0.40)	0.54(0.58)	0.57(0.69)	0.58(0.74)	0.56(0.75)
300	D	0.89(0.73)	0.61(0.49)	0.13(0.25)	0.02(0.10)	0.00(0.04)	0.00(0.01)
	L	0.00(0.06)	0.05(0.11)	0.33(0.17)	0.41(0.21)	0.42(0.23)	0.44(0.24)
	W	0.11(0.17)	0.35(0.36)	0.60(0.59)	0.71(0.77)	0.73(0.85)	0.72(0.89)
350	D	0.89(0.82)	0.65(0.61)	0.29(0.36)	0.08(0.16)	0.01(0.06)	0.00(0.02)
	L	0.00(0.01)	0.00(0.03)	0.11(0.05)	0.21(0.07)	0.26(0.08)	0.28(0.09)
	W	0.11(0.13)	0.35(0.29)	0.61(0.53)	0.79(0.75)	0.85(0.88)	0.86(0.94)
400	D	0.89(0.87)	0.65(0.70)	0.37(0.45)	0.14(0.22)	0.03(0.09)	0.01(0.03)
	L	0.00(0.00)	0.00(0.01)	0.02(0.02)	0.07(0.02)	0.12(0.03)	0.13(0.03)
	W	0.11(0.09)	0.35(0.23)	0.62(0.46)	0.81(0.70)	0.90(0.87)	0.91(0.94)
450	D	0.89(0.91)	0.65(0.77)	0.38(0.54)	0.18(0.29)	0.06(0.13)	0.04(0.05)
	L	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.01(0.01)	0.04(0.01)	0.05(0.01)

Table 2.5: The estimated win (W), draw (D) and loss (L) probabilities when Team A has declared in the third innings with a specified target and a specified number of overs remaining. The SETAL probabilities are given in parentheses.

However, we stress that the focus of the paper is not whether our probabilities are preferable to the SETAL probabilities. The two sets of probabilities are somewhat comparable,

and this is comforting. The more important issue is whether our decision rules regarding declaration lead to more favourable outcomes than what is observed in current practice.

We now evaluate our decision rules given in Table 2.4 via a comparison with actual data. We take a validation approach whereby we apply our results to matches that were not used in the estimation of parameters. We consider the 49 test matches between ICC teams that were played between April 1, 2012 and April 30, 2013. In these matches, we examine the third innings scorecard, and upon the completion of each over, we ask whether our declaration rule ought to have been invoked. There were 13 matches where actual declarations took place in the third innings. Of the remaining 36 matches where there were no actual third innings declarations, our decision rules also did not invoke a declaration. This suggests that teams are making good decisions with respect to declaration although they may not be doing so at optimal times. From the 13 matches where declarations occurred in the third innings, two of these matches were rained out before the fourth innings began. This reduces our study to 11 matches of interest. In the 11 matches of interest, we simulated match results according to our declaration rule. If our declaration preceded the actual declaration, we simulated fourth innings batting using Table 2.6 from our point of declaration. If our declaration did not precede the actual declaration, we simulated third innings batting using Table 2.2 from the actual point of declaration until our rule prescribed a declaration (if at all), and then from that juncture, simulated fourth innings data. In summary, we compare how the matches actually turned out with the way they would have turned out (via simulation) had our decision rules been followed.

$\hat{p}_{-1}$	$\hat{p}_0$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$	$\hat{p}_5$	$\hat{p}_6$
0.016	0.743	0.128	0.035	0.010	0.065	0.000	0.003

Table 2.6: Sample proportions corresponding to the parameters $p_{-1}, \ldots, p_6$ for fourth innings batting using the original 134 matches.

In Table 2.7 and Table 2.8, we provide a summary of the 11 match results based on 1000 simulations per match. The simulation in Table 2.7 used the standard criterion (2.1) where wins and draws are both valued. The simulation in Table 2.8 used criterion (2.2) where Team A is desperate and has a win at all costs perspective. In both tables, we provide simulation results where Team B is considered to be a relatively average team in comparison to Team A. We also consider the case where Team B is considered to be either relatively weak or relatively strong in comparison to Team A. The determination of weak or strong is obtained from the ICC Test rankings www.cricketworld4u.com/team-ranking.php.

From Table 2.7, we first observe that in 6 of the 7 matches, our declaration took place slightly earlier than the actual declaration. In the April 15 match where Australia actually declared earlier than what our decision rules stipulate, we note that Australia's declaration is premature compared to the historical declaration decisions in Figure 2.1. As expected, we also observe that when Team B is a weaker opponent, then the declaration occurs earlier

Actual Results						Simulated Results					
Date	TA	TB	Target	OR	Result	TB	Target	OR	P(W)	P(D)	P(L)
Apr 15	Aus	WI	214	59	D	weak	233	55	0.10	0.90	0.00
						average	268	49	0.02	0.98	0.00
Jun 22	SL	Pak	509	195	W	average	-	-	-	-	-
Jun 30	Pak	SL	260	37	D	weak	218	43	0.02	0.98	0.00
						average	239	40	0.01	0.99	0.00
Jul 08	Pak	SL	269	71	D	weak	261	74	0.36	0.64	0.00
						average	290	65	0.11	0.89	0.00
Aug 02	SA	Eng	252	39	D	average	241	40	0.01	0.99	0.00
Nov 25	NZ	SL	363	106	W	average	358	107	0.27	0.39	0.34
Feb 01	SA	Pak	480	255	W	weak	-	-	-	-	-
Mar 08	SL	Ban	268	22	D	weak	-	-	-	-	-
Mar 22	NZ	Eng	481	143	D	average	472	145	0.44	0.51	0.05
Apr 17	Zim	Ban	483	236	W	average	-	-	-	-	-
Apr 25	Ban	Zim	401	134	W	average	398	135	0.32	0.08	0.60

Table 2.7: Comparison of actual results versus simulated results using our declaration rules and according to the standard criterion (2.1). Abbreviated column headings are TA (Team A), TB (Team B) and OR (Overs Remaining). Team B is characterized as relatively average, weak or strong in comparison to Team A. The matches from June 22/12, Feb 1/13, Mar8/13 and Apr 17/13 were not simulated because the optimal number of overs exceeded 180.

than if Team B is an average opponent. When looking at the simulation probabilities for wins, draws and losses, it appears that the outcomes using our declaration rules are not in conflict with the actual outcomes.

From Table 2.8, we observe that Team A declares earlier than in Table 2.7. This is anticipated since in Table 2.8, Team A is desperate to win. Consequently, the win and loss probabilities in Table 2.8 are larger than those in Table 2.7.

2.7 DISCUSSION

The determination of when to declare in test cricket is an important problem which affects the outcomes of matches. In this paper, we provide specific guidelines (Table 2.4) whether teams should declare at various stages of matches.

Relating our optimality results to current practice, we have the following summary guidelines. Teams that declare with less than 90 overs remaining (which is an entire day of batting), are behaving cautiously and should actually declare earlier. In particular, teams wait far too long when it is essential that they win the match. Although waiting longer always reduces the chance of losing, it may also decrease the chance of winning. SETAL also remark that teams are generally cautious when declaring. However, for optimal decision making, teams that declare earlier (say, with more than 135 overs or 1.5 days remaining), should actually wait a little longer to declare.

Actual Results						Simulated Results					
Date	TA	TB	Target	OR	Result	TB	Target	OR	P(W)	P(D)	P(L)
Apr 15	Aus	WI	214	59	D	weak	173	68	0.23	0.51	0.26
						average	194	64	0.08	0.46	0.46
Jun 22	SL	Pak	509	195	W	average	465	210	0.93	0.00	0.07
Jun 30	Pak	SL	260	37	D	weak	160	55	0.10	0.85	0.05
						average	165	53	0.03	0.59	0.38
Jul 08	Pak	SL	269	71	D	weak	212	80	0.41	0.47	0.12
						average	243	77	0.19	0.51	0.30
Aug 02	SA	Eng	252	39	D	average	165	58	0.03	0.29	0.68
Nov 25	NZ	SL	363	106	W	average	339	114	0.31	0.28	0.41
Feb 01	SA	Pak	480	255	W	weak	-	-	-	-	-
Mar 08	SL	Ban	268	22	D	weak	-	-	-	-	-
Mar 22	NZ	Eng	481	143	D	average	459	148	0.48	0.40	0.12
Apr 17	Zim	Ban	483	236	W	average	-	-	-	-	-
Apr 25	Ban	Zim	401	134	W	average	379	140	0.36	0.03	0.61

Table 2.8: Comparison of actual results versus simulated results using our declaration rules and according to the "must win" criterion (2.2). Abbreviated column headings are TA (Team A), TB (Team B) and OR (Overs Remaining). Team B is characterized as relatively average, weak or strong in comparison to Team A. The matches from June 22/12, Feb 1/13, Mar8/13 and Apr 17/13 were not simulated because the optimal number of overs exceeded 180.

We add a caveat with respect to our results. The run scoring characteristics of Team B in the fourth innings are based on standard batting behaviour. It is possible that teams become cautious in situations where they believe that winning is impossible, and they instead play for a draw. However, our Table 2.4 provides declaration guidelines at the tipping point where it turns advantageous for Team A to declare. These are not extreme situations, and we suggest that Team B is likely to employ standard batting behaviour in these circumstances.

In sport, teams that adopt improved strategies are able to gain a competitive edge (see Lewis, 2003). It is hoped that papers such as this will help promote the adoption of analytics in cricket. At the present time, the use of cricket analytics appears to trail many of the popular professional sports.

Finally, we note that in recent years, there appears to be an increase in run rates in Test cricket. This phenomenon should be taken into account when updating Table 2.4.

Chapter 3

Optimal Lineups in Twenty20 Cricket

3.1 INTRODUCTION

Twenty20 cricket (or T20 cricket) is a form of limited overs cricket which has gained popularity worldwide. Twenty20 cricket was showcased in 2003 and involved matches between English and Welsh domestic sides. The rationale behind the introduction of T20 was to provide an exciting version of cricket with matches concluding in three hours duration or less. There are now various professional T20 competitions where the Indian Premier League (IPL) is regarded as the most prestigious. Even in Canada (not exactly known as a cricketing country), every game of the IPL is telecast on live television.

Except for some subtle differences (e.g. fielding restrictions, limits on the number of overs per bowler, powerplays, etc.), Twenty20 cricket shares many of the features of one-day cricket. One-day cricket was introduced in the 1960s, and like T20 cricket, is a version of cricket based on limited overs. The main difference between T20 cricket and one-day cricket is that each batting side in T20 is allotted 20 overs compared to 50 overs in one-day cricket.

Consequently, many of the strategies used in one-day cricket have trickled down to Twenty20 cricket. This paper investigates the determination of T20 team lineups (i.e. team selection, batting orders and bowling orders). It is desirable that teams field their strongest sides, and doing so, requires a judgment on the relative value of batting versus bowling.

The question of team selection in cricket has been investigated by various authors. [5] advocated the use of a nonlinear combination of the strike rate (runs scored per 100 balls faced) with the batting average (runs scored per dismissal) to select batsmen in one-day international (ODI) cricket. We note that the proposed metric involves an arbitrary weight α and that individual player selection does not account conditionally for players already selected. [46] obtained optimal batting orders in ODI cricket using batting characteristics from multinomial regression. However, the paper failed to look at the effect of bowling

and did not address the initial team selection problem. [6] and [29] considered integer optimization methods for selecting players in fantasy league cricket and in limited overs cricket. Their methodology is based on performance measures computed from summary statistics such as the batting average, strike rate, etc. Integer programming constraints include various desiderata such as the inclusion of a single wicket-keeper and specified numbers of batsmen, all-rounders and bowlers. Two major drawbacks of the approach are the ad hoc choice of performance statistics (the objective function) and the lack of validation against "optimal" team selections.

In this paper, we investigate the problem of determining team lineups in T20 cricket from the point of view of *relative value statistics*. Relative value statistics have become prominent in the sporting literature as they attempt to quantify what is really important in terms of winning and losing matches. For example, in Major League Baseball (MLB), the VORP (value over replacement player) statistic has been developed to measure the impact of player performance. For a batter, VORP measures how much a player contributes offensively in comparison to a replacement-level player [49]. A replacement-level player is a player who can be readily enlisted from the minor leagues. As another example, the National Hockey League (NHL) reports plus-minus statistics. The statistic is calculated as the goals scored by a player's team minus the goals scored against the player's team while he is on the ice. More sophisticated versions of the plus-minus statistic have been developed by [44] and [21].

In cricket, a team wins a match when the runs scored while batting exceed the runs conceded while bowling. Therefore, it is the run differential that is the holy grail of performance measures in cricket. An individual player can be evaluated by considering his team's run differential based on his inclusion and exclusion in a lineup. Clearly, run differential cannot be calculated from actual match results in a straightforward way because there is variability in match results and conditions change from match to match. Our approach in assessing performance is based on simulation methodology where matches are replicated. Through simulation, we can obtain long run properties (i.e. expectations) involving run differential. By concentrating on what is really important (i.e. run differential), we believe that our approach addresses the core problem of interest in the determination of T20 team lineups.

In Section 3.2, we begin with some exploratory data analyses which investigate batting and bowling characteristics of T20 cricketers. We observe that player characteristics are not clustered according to position. Rather, player skills appear to vary on the continuum. We then provide an overview of the simulator developed by [11] which is the backbone of our analysis and is used in the estimation of expected run differential.

In Section 3.3, a simulated annealing algorithm is proposed to obtain optimal team lineups (i.e. the joint determination of team selection, batting order and bowling order). The algorithm searches over the vast combinatorial space of team lineups to produce an optimal lineup with the greatest expected run differential. This is a more complex problem

than is typically considered in the literature where only team selection is discussed. We remark that we have not seen any previous work that addresses bowling orders. In the search for an optimal team lineup, the objective function is the run differential which is the quantity that is relevant to winning and losing matches. The simulated annealing algorithm requires fine tuning in order to effectively search the space and converge to the optimal lineup.

In Section 3.4, we provide two applications of the simulated annealing algorithm. First, we determine an optimal lineup for both India and South Africa in T20 cricket and we compare these lineups to actual lineups that have been used in the recent past. Some comments are then made on the optimal composition of teams. Second, the simulated annealing algorithm is applied to an international pool of players to identify an "all-star" lineup. We compare the resulting team selection with some common beliefs concerning "star" players. We conclude with a short discussion in Section 3.5.

3.2 PRELIMINARIES

To initiate discussion on T20 team composition, we begin with some exploratory data analyses. We consider T20 matches involving 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. The matches considered were those that took place from the first official match in 2005 until May 21, 2014. Details from these matches can be found in the Archive section of the CricInfo website (www.espncricinfo.com). In total, we obtained data from 282 matches.

For batting, we use familiar quantities. We let BA denote the batting average (runs scored per dismissal) and we let SR denote the batting strike rate (runs scored per 100 balls). Good batting is characterized by large values of both BA and SR. We define similar quantities for the bowling average BA (runs allowed per dismissal) and the bowling economy rate ER (runs allowed per over).

In Figure 3.1, we produce a scatterplot of BA versus SR for the 40 batsmen in our dataset who have faced at least 500 balls. As expected, players who are known as bowlers are not as proficient at batting (lower left section of the plot). We observe that batsmen have different styles. For example, V. Kohli of India has the best batting average but his strike rate is only slightly above average. On the other hand, Y. Singh of India scores many runs (i.e. has a high strike rate) yet his batting average is not exceptional. A. Hales of England and K. Pietersen of England are plotted deep in the upper right quadrant, and are the ideal combination of reliability (i.e. batting average) and performance (i.e. strike rate). We also observe a continuum in batting abilities along both axes with no apparent clustering. This is an important observation for the determination of optimal team selections. Consequently,

players should be selected on their own merits rather than filling quotas of pure batsmen, all-rounders and bowlers.

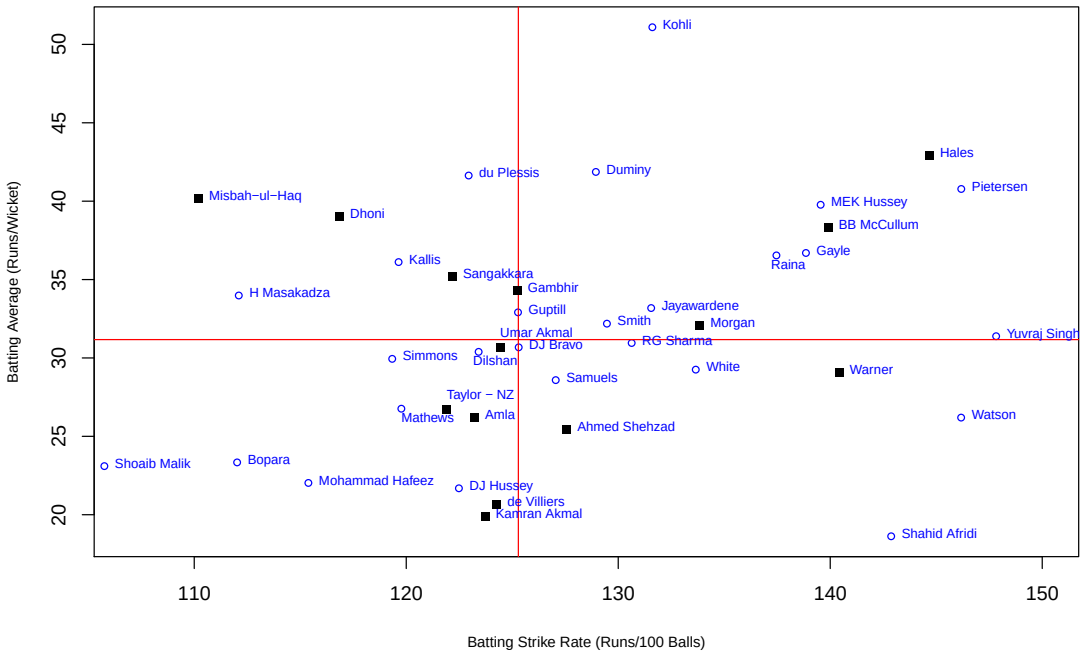

Figure 3.1: Scatterplot of batting average versus batting strike rate. Pure batsmen (i.e. those who never bowl) are indicated by black squares.

In Figure 3.2, we produce a scatterplot of BA versus ER for the 33 bowlers in our dataset who have bowled at least 500 balls. Again, we observe that bowlers have different characteristics. For example D. Vettori of New Zealand concedes very few runs (i.e. low economy rate) but he is mediocre in taking wickets as highlighted by his middling bowling average. We also observe a continuum in bowling abilities with no apparent clustering.

We now provide an overview of the simulator developed by [11] which we use for the estimation of expected run differential. Ignoring extras (sundries) that arise via wide-balls and no-balls, there are 8 broadly defined outcomes that can occur when a batsman faces a bowled ball. These batting outcomes are listed below:

$$\begin{aligned}
\text{outcome } j = 0 &\equiv 0 \text{ runs scored} \\
\text{outcome } j = 1 &\equiv 1 \text{ runs scored} \\
\text{outcome } j = 2 &\equiv 2 \text{ runs scored} \\
\text{outcome } j = 3 &\equiv 3 \text{ runs scored} \\
\text{outcome } j = 4 &\equiv 4 \text{ runs scored} \\
\text{outcome } j = 5 &\equiv 5 \text{ runs scored} \\
\text{outcome } j = 6 &\equiv 6 \text{ runs scored} \\
\text{outcome } j = 7 &\equiv \text{dismissal}
\end{aligned} \tag{3.1}$$

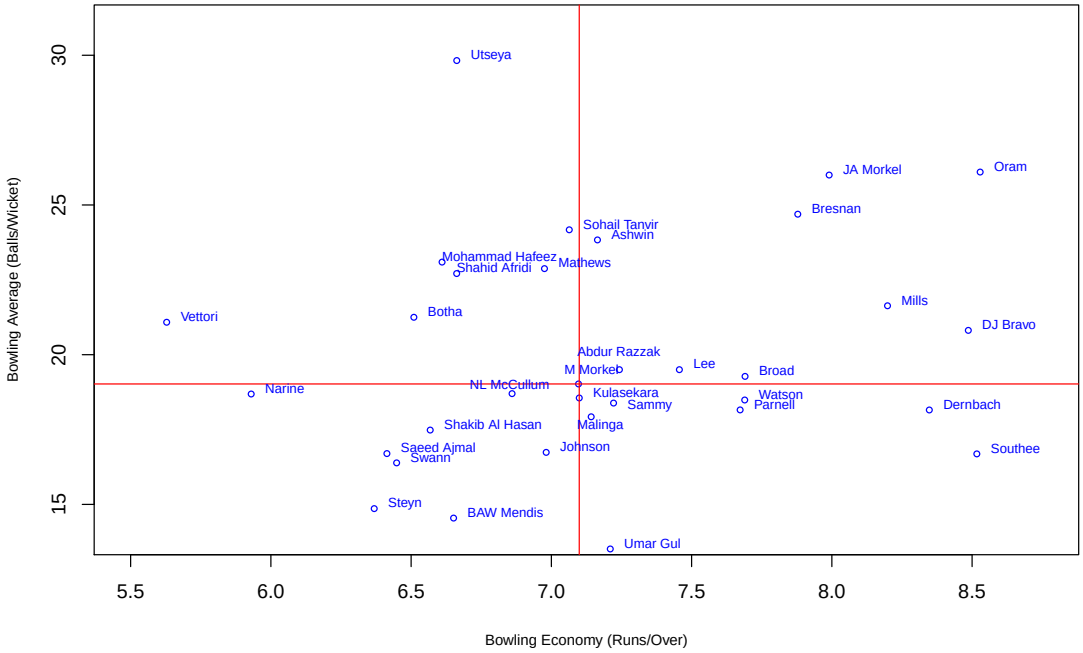

Figure 3.2: Scatterplot of bowling average versus bowling economy rate.

In the list (3.1) of possible batting outcomes, we include byes, leg byes and no balls where the resultant number of runs determines one of the outcomes $j = 0, \ldots, 7$. We note that the outcome $j = 5$ is rare but is retained to facilitate straightforward notation.

According to the enumeration of the batting outcomes in (3.1), [11] suggested a statistical model for the number of runs scored by the ith batsman:

$$(X_{iow0}, \ldots, X_{iow7}) \quad \sim \quad \text{multinomial}(m_{iow};\ p_{iow0}, \ldots, p_{iow7}) \tag{3.2}$$

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 (3.2), 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 is "special" in the sense that it consists of detailed ball-by-ball data. Typically, researchers study aggregate match data. The data were obtained using a proprietary parser which was applied to the commentary logs of matches listed on the CricInfo website.

The estimation of the multinomial parameters p_{iowj} in (3.2) 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, [11] introduced the simplification

$$p_{iowj} = \frac{\tau_{owj}\ p_{i70j}}{\sum_j \tau_{owj}\ p_{i70j}}\ . \tag{3.3}$$

In (3.3), the parameter p_{i70j} represents the baseline characteristic for batsman i with respect to batting outcome j. The characteristic p_{i70j} is the probability of outcome j associated with the ith batsman at the juncture of the match immediately following the powerplay (i.e. the 7th over) when no wickets have been taken. The multiplicative parameter τ_{owj} scales the baseline performance characteristic p_{i70j} to the stage of the match corresponding to the oth over with w wickets taken. The denominator in (3.3) ensures that the relevant probabilities sum to unity. There is an implicit assumption in (3.3) that although batsmen are unique, their batting characteristics change by the same multiplicative factor which is essentially an indicator of aggression. For example, when aggressiveness increases relative to the baseline state, one would expect $\tau_{ow4} > 1$ and $\tau_{ow6} > 1$ since bolder batting leads to more 4's and 6's.

Given the estimation of the parameters in (3.3) (see [11]), a straightforward algorithm for simulating first innings runs against an average bowler is available. One simply generates multinomial batting outcomes in (3.1) according to the laws of cricket. For example, when either 10 wickets are accumulated or the number of overs reaches 20, the first innings is terminated. [11] also provide modifications for batsmen facing specific bowlers (instead of average bowlers), they account for the home field advantage and they provide adjustments for second innings simulation.

In summary, with such a simulator, we are able to replicate matches and estimate the expected run differential when Team A (lineup specified) plays against Team B (lineup specified). In [11], the simulator is demonstrated to provide realistic realizations for Twenty20 cricket.

3.3 OPTIMAL LINEUPS

Consider the problem of selecting 11 players from a pool of players for a Twenty20 cricket team and then determining the batting order and the bowling order for the selected players. We refer to this overall specification as a "team lineup". The objective is to select a team lineup that produces the greatest expected difference between runs scored R_s and runs allowed R_a

$$\begin{aligned} \mathrm{E}(D) &= \mathrm{E}(R_s - R_a) \\ &= \mathrm{E}(R_s) - \mathrm{E}(R_a) \end{aligned} \tag{3.4}$$

against an average opponent.

For a particular team lineup, the calculation of E(D) in (3.4) depends on the batting and bowling characteristics of the selected players. Clearly, this is not something that we can obtain analytically. Therefore, we simulate the batting innings and the bowling innings for a particular lineup and then take the average of the run difference D over many hypothetical matches. In calculating D, we simulate first innings runs for both teams. This seems reasonable since in practice, second innings batting terminates if second innings runs exceed the target. Since the performance measure E(D) is recorded in runs, it is desirable that our estimates are accurate to within a single run. We have found that simulating $N = 25000$ pairs of innings provides a standard error of less than 0.2 for the expected run difference E(D).

We now formalize the solution space of team lineups. Although the set of possible team lineups is discrete and finite, it is a vast combinatorial space. Let M denote the number of players who are available in the team selection pool. Then the first step in obtaining an optimal lineup is the specification of the 11 active players. There are $\binom{M}{11}$ potential selections. Once team selection is determined, there are $11! \approx 40$ million potential batting orders alone. And with a potentially different bowler for each over in a bowling innings, there is an upper bound of $(11)^{20}$ bowling orders. The number of bowling orders is an upper bound since the rules of T20 cricket prohibit a bowler for bowling more than four overs. Therefore a simple upper bound for the cardinality of the solution space is given by

$$\binom{M}{11}(11!)(11)^{20} = \frac{M!\ (11)^{20}}{(M-11)!}\ . \tag{3.5}$$

Our problem is therefore computationally demanding. We need to optimize team lineups over an enormous combinatorial space where the objective function E(D) itself requires many simulations of innings. For example, if $M = 15$, the upper bound (3.5) yields 3.67×10^{31}. In Appendix A, we provide a more nuanced description of the solution space taking into account some detailed aspects of the game.

To carry out the optimization, we employed a simulated annealing algorithm [27]. Simulated annealing is a probabilistic search algorithm that explores the combinatorial space, spending more time in regions corresponding to promising team lineups. Successful implementations of simulated annealing typically require careful tuning with respect to the application of interest.

For our problem, simulated annealing proceeds as follows: Denote the current team lineup at the beginning of step i of the algorithm as lineup c_{i-1} where $i = 1, 2, \ldots$ The algorithm has a prescribed starting lineup c_0 where a typical T20 lineup is straightforward to specify and is recommended. During each step of the algorithm, a candidate lineup c^* is generated (as described in the fine tuning of the algorithm - Section 3.3.1). Then N_i matches are generated with the lineup c^* which provides the estimated run differential

$\hat{\mathrm{E}}(D_{c^*})$. The candidate lineup is accepted as the current lineup, i.e. we set $c_i = c^*$ if

$$\hat{\mathrm{E}}(D_{c^*}) > \hat{\mathrm{E}}(D_{c_{i-1}}) \ . \tag{3.6}$$

The candidate lineup is also accepted, i.e. we set $c_i = c^*$ if both

$$\hat{\mathrm{E}}(D_{c^*}) \leq \hat{\mathrm{E}}(D_{c_{i-1}}) \tag{3.7}$$

and if a randomly generated $u \sim \mathrm{Uniform}(0, 1)$ satisfies

$$u < \exp\left\{ \frac{\hat{\mathrm{E}}(D_{c^*}) - \hat{\mathrm{E}}(D_{c_{i-1}})}{t_i} \right\} \tag{3.8}$$

where t_i is a specified parameter referred to as the "temperature". If the proposed lineup c^* is not accepted, then the current lineup is carried forward, i.e. we set $c_i = c_{i-1}$. The temperature is subject to a non-increasing "cooling schedule" $t_i \to 0$. We therefore observe that the simulated annealing algorithm goes "up the hill" (3.6) to states corresponding to preferred lineups but also occasionally goes "down the hill" (3.7) allowing the algorithm to escape regions of local maxima in the search for a global maximum. Intuitively, condition (3.8) says that as the temperature cools (i.e. gets closer to zero) and the system is more stable, then it becomes more difficult (probabilistically) to escape a state (i.e. a lineup) for a state with a lower expected run differential. Under suitable conditions, the simulated annealing algorithm converges to an optimal lineup. Practically, the algorithm terminates after a fixed number of iterations or when there are infrequent state changes. The asymptotic results suggest that the final state will be nearly optimal.

The fine tuning of the algorithm corresponds to the cooling schedule and the mechanism for generation of candidate lineups. A necessary condition of the asymptotic theory is that all team lineups are "connected". In other words, it must be possible for every team lineup to be reached from any other team lineup in a finite number of steps. A guiding principle in the development of our simulated annealing algorithm is that we enable large transitions (moving to "distant" team lineups) during the early phases of the algorithm and we restrict transitions to neighbouring (i.e. close) lineups during the latter phases of the algorithm.

With the combinatorial space described above, we note that our problem has similarities to the longstanding travelling salesman's problem (TSP) where the potential routes of a salesman consist of permutations of the order of visited cities. In the TSP, given n cities, there are n! orderings in which the cities may be visited. As this is likewise a large discrete space for even moderate n, there are some useful heuristics in proposing candidate routes. For example, total distance travelled is unlikely to vary greatly if the order of the visited cities is only slightly changed. This introduces a concept of closeness where total distance travelled is close if there are only small changes to the route. This corresponds to small changes in the permutations such as interchanging the order of two adjacent cities. Also,

in simulated annealing we want to explore the solution space widely in the initial phase of the search so that with high probablity, we are eventually in the neighbourhood of the global optimum. In the TSP, exploring wide neighbourhoods corresponds to more extreme changes in the permutations of cities visited. Although our problem is more complex than the TSP, we borrow ideas from [2] and [46] where simulated annealing has been successfully utilized to address optimization problems for related discrete spaces.

We note that there is considerable flexibility in the proposed procedure. For example, suppose that the captain wants to try out a new player in the 4th position in the batting order. In this case, the proposal distribution of the simulated annealing algorithm can be hard-coded such that the new player is forced into the 4th position, and the remaining lineup is optimized according to this constraint. The introduction of such constraints provides a systematic approach for experimentation with lineups, and is particularly useful with new players who do not have much of a batting/bowling history.

3.3.1 Fine Tuning of the Algorithm

Recall that the determination of a team lineup has three components:

(a) the selection of 11 players from a roster of available players

(b) the specification of the batting order for the selected 11 players

(c) the specification of the bowling order for the selected 11 players

Our mechanism for the generation of candidate lineups takes these three components into account. In each step of simulated annealing, we generate a random variate $u \sim \text{Uniform}(0, 1)$. If $0 \leq u < 1/3$, we address issue (a). If $1/3 \leq u < 2/3$, we address issue (b). If $2/3 \leq u < 1$, we address issue (c). Specifically, we generate candidate lineups as follows:

(a) Team Selection - Denote the roster of M potential players as $\{i_1, \ldots, i_M\}$ where $i_1, \ldots, i_{11}$ are the players in the current lineup. We randomly choose a player from $i_1, \ldots, i_{11}$ and swap this player with a randomly chosen player from $i_{12}, \ldots, i_M$. The swapped-in player replaces the swapped-out player in both the batting order and the bowling order. If the swapped-out player is a bowler who is active in the bowling order and the swapped-in player is a pure batsman, then the excess bowling overs are randomly distributed to the current bowlers in the lineup. We do not allow swaps that result in lineups with fewer than five players who can bowl nor do we allow swaps that result in lineups without a wicketkeeper. The resultant lineup is the candidate lineup.

(b) Batting Orders - Denote the batting order corresponding to the current lineup by $i_1, \ldots, i_{11}$ where i_j corresponds to the batsman in the jth batting position. We randomly choose a batsman i_j from $i_1, \ldots, i_{10}$ and then interchange i_j with i_{j+1}. The resultant lineup is the candidate lineup.

(c) Bowling Orders - Suppose that we have M_B potential bowlers in the current lineup. The generation of a candidate bowling order begins with an ordered list of $4M_B$ bowler symbols where each bowler appears in the list four times. The setup therefore enforces one of the T20 rules that no bowler may bowl more than four overs in a match. The first 20 entries in the list define the current bowling order where i_j corresponds to the bowler who bowls during over j. We randomly choose a bowler from $i_1, \ldots, i_{20}$ and then swap him with a randomly chosen player from one of $i_1, \ldots, i_{4M_B}$ According to T20 rules, bowlers are not permitted to bowl in consecutive overs, and we do not allow such a swap. The resultant lineup is the candidate lineup.

In the early stages of the algorithm ($i = 1, \ldots, 500$), we permit double swaps instead of single swaps in steps (a), (b) and (c) to facilitate larger transitions to escape local neighbourhoods. To complete the description of the algorithm, we use an exponential cooling schedule defined by a sequence of temperature plateaux

$$t_i = 0.5(0.9)^{\lfloor i/100 \rfloor} \ . \tag{3.9}$$

We introduce a provision to the cooling schedule (3.9) whereby we move to the next temperature plateau if there are more than 20 state changes at any temperature. The rationale is that we do not want to "waste" time at temperatures where we are regularly accepting the candidate lineup (i.e. moving both up and down the hill). At such temperatures, we would not be moving in the direction of an optimal lineup.

As the algorithm proceeds, we also want to make sure that the objective function $\mathrm{E}(D)$ is estimated accurately. We therefore increase the number of innings simulations N_i as the algorithm proceeds. Specifically, we set

$$N_i = \min(25000, 1000 + 16i) \ .$$

For example, $N_1 = 1016$, and when our algorithm has typically converged, $N_{1500} = 25000$.

3.4 APPLICATIONS

We have run the simulated annealing algorithm for various teams and have obtained sensible yet provocative results in each case. In our testing, we have found that 1500 iterations are sufficient for convergence and this requires approximately 24 hours of computation on a laptop computer.

A large fraction of the computational cost comes from the simulation of innings. Since the number of simulations increases as the algorithm proceeds, so does the cost of each successive iteration. We eliminate unnecessary simulations wherever possible. The expected run differential and the lineup are stored for each proposal that is accepted, so only the newly proposed lineup needs to be simulated during any given iteration. Furthermore, if the batting order is unchanged between the current lineup and the proposed lineup in

an iteration, then only the bowling innings is simulated and the current lineup's batting average is reused. Likewise, if the bowling assignment is unchanged in a proposal, then only the batting innings needs to be simulated.

Since the algorithm simulates a large number of independent innings during each iteration, the simulation step is embarrassingly parallelizable, meaning it requires $1/n$ as much time to complete using n identical CPU cores as it would on one such core. We use the Snow package [47] and the Snowfall package [28] in R to parallelize simulation on four cores using an Intel i7 processor. Other steps in the algorithm such as creating candidate lineups and checking them for validity are not parallelizable, and there is a computational cost associated with work assignment and aggregation of output between cores.

In Figure 3.3, we illustrate a sample path taken by the simulated annealing algorithm in the search for the optimal Indian lineup described in more detail in Section 3.4.1. We observe that the starting lineup is far from optimal (i.e. $\mathrm{E}(D_1) \approx 0$) but very quickly iterates to good lineups (i.e. $\mathrm{E}(D_i) > 10$) for larger values of i. Recall that our computational strategy was to not spend much time exploring unsuitable lineups. Hence we see that confidence intervals (based on Monte Carlo simulations) corresponding to the estimates of $\mathrm{E}(D_i)$ are wider in the early stages of the algorithm where N_i is small. We note that we have confidence in the algorithm since different starting values (i.e. lineups) all lead to the same neighbourhood of solutions. We say "neighbourhood of solutions" since we have found that small lineup changes (e.g. exchanging the batting order of the 9th and 10th batsmen) lead to changes in $\mathrm{E}(D)$ that are not meaningul (i.e. less than 0.5 runs). As the temperature decreases, candidate lineups continue to be proposed but fewer are accepted. Also, as the temperature decreases, the magnitude of the dips in Figure 3.3 decrease as it becomes less probable that condition (3.8) is satisfied.

The algorithm has been designed such that all batsmen are drawn from the same pool. This prevents double drawing of players, and also allows for multiple wicketkeepers to be in the lineup. It is therefore possible that a wicketkeep such as A.B. de Villiers could be selected in his non-core role.

3.4.1 Optimal T20 Lineup for India

To investigate our methodology, we considered 17 currently active players for India who have a playing history in T20 ICC matches. The optimal lineup (i.e. team selection, batting order and bowling selection) based on the proposed simulated annealing algorithm is given in the first column of Table 3.1. For comparison, columns 2, 3 and 4 provide actual lineups used by India in the T20 World Cups of 2010, 2012 and 2014, respectively. In general, we see many similarities between the optimal lineup and the lineups used in practice. Also, there seems to be as much variation between the optimal lineup and the actual lineups as between the actual lineups themselves.

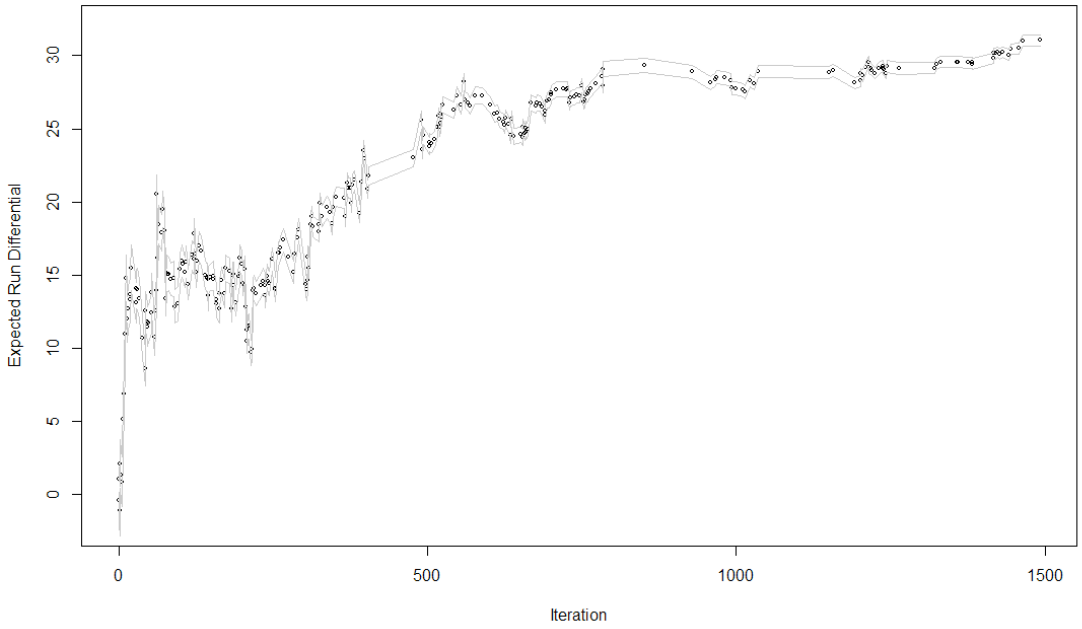

Figure 3.3: A plot of the estimated run differential versus the iteration number in a single run of simulated annealing corresponding to India. Confidence intervals for the estimates are provided.

However, there are a number of interesting discrepancies between the optimal lineup and the actual lineups used by India. For example, we observe that India may not be positioning M.S. Dhoni (captain and wicketkeeper) properly in the batting order. Whereas India places him in the middle batting positions (No 4, 5 and 7), they may be better served to have him bat later, such as in the 9th position as specified in the optimal lineup. It is also interesting to observe that the optimal lineup has only one fast bowler, namely B. Kumar. In each of the actual lineups there are two fast bowlers. In cricket, there is a tradition of mixing the bowling sequence between fast and spin bowlers. However, from an analytics point of view, if your team does not have many good fast bowlers (for example), then maybe the composition should be re-examined to include more spin bowlers. Sometimes it is useful to question tradition and folklore in the presence of hard numbers.

One question of interest is whether the optimal T20 lineup has a different composition in terms of the number of pure batsmen, all-rounders and bowlers who are used in actual T20 matches and actual ODI matches. In analyzing actual matches, we found that there is considerable variability in this regard. The variability is both across matches and across teams. Complicating the analysis is the definition of "all-rounder". An all-rounder is an ambiguous term as it suggests that a player is both good at batting and bowling. But technically, any player who bowls is forced to bat occasionally, and therefore, whether or not he is considered an all-rounder is subject to some debate. Therefore, in terms of team

	Optimal Lineup	May 11/10 Lineup	Oct 02/12 Lineup	Apr 06/14 Lineup
	01. V Kohli (1)	KD Karthik	G Gambhir	RG Sharma
	02. G Gambhir	G Gambhir	V Sehwag	AM Rahane
	03. S Raina (3)	S Raina	V Kohli	V Kohli
	04. Y Singh (4)	MS Dhoni*	RG Sharma (1)	Y Singh
	05. V Sehwag	Y Singh (1)	Y Singh (4)	MS Dhoni*
	06. RG Sharma	YK Pathan (3)	S Raina	S Raina (4)
	07. YK Pathan	RG Sharma	MS Dhoni*	R Ashwin (4)
	08. H Singh (4)	PP Chawla (4)	IK Pathan (3)	RA Jadeja (1)
	09. MS Dhoni*	H Singh (4)	R Ashwin (4)	A Mishra (4)
	10. A Mishra (4)	V Kumar (4)	L Balaji (4)	B Kumar (3)
	11. B Kumar (4)	A Nehra (4)	Z Khan (4)	MM Sharma (2)
	NS Z Khan			
	NS R Ashwin			
	NS R Jadeja			
	NS V Kumar			
	NS S Dhawan			
	NS AM Rahane			
E(D)	32.4	2.9	7.4	1.1

Table 3.1: Optimal lineup for India and three typical lineups that were used on the specified dates. The vertical numbering corresponds to the batting order where the players labelled NS were not selected. In parentheses, we provide the number of overs of bowling and the asterisk denotes the wicketkeeper.

composition, the main take-away point is that teams should play their optimal lineup even though it may sometimes result in nontraditional uses of players. The continuum of abilities suggested in Figure 3.1 and Figure 3.2 lead credence to the idea that there are not always clear delineations between pure batsmen, all-rounders and bowlers.

It is worth commenting that E(D) for the optimal T20 lineup is remarkably larger than for the actual lineups in Table 3.1. While this is a promising reason to consider the proposed methodology, there may be various nuances involved in this observation. First, there are important changes that collectively explain the 30-run gap between the optimal lineup and the ones that were actually used. For example, M.M. Sharma was unknown before the 2014 World Cup match, and his limited international T20 information implies that he is worse than an average bowler. Hence, his inclusion in the April 6/14 lineup reduced E(D). India included Sharma in their roster despite his limited international experience. It was Sharma's stellar performance in other leagues (e.g. IPL) and one-day cricket which earned him the right to bowl in a Twenty20 World Cup final. Other notable inclusions in the optimal lineup who did not play in 2014 are Gambhir, Sehwag, Pathan and Singh. Also, we recall that the timeframe of our dataset used for estimating player characteristics was 2005-2014. It is therefore likely that some of the players who were selected in the actual lineups were experiencing a good run of form, perhaps better than their historical characteristics

which were used in obtaining E(D). Related to this comment, G. Ghambir and V. Sehwag were included in our optimal squad but not in 2014. By 2014, they were likely perceived as aging players with sub-optimal performance.

3.4.2 Optimal T20 Lineup for South Africa

Similarly, we carried out the optimization procedure for South Africa. This time, 15 current players were considered for team selection. The optimal lineup and the actual lineups (used in the three most recent World Cups) are given in Table 3.2. The optimal lineup ranges from 16.8 runs to 21.9 runs superior to the actual lineups.

	Optimal Lineup	May 10/10 Lineup	Oct 02/12 Lineup	Apr 04/14 Lineup
	01. H Amla	HH Gibbs	H Amla	Q de Kock*
	02. AB de Villiers*	GC Smith	JH Kallis (3)	H Amla
	03. JP Duminy	JH Kallis (4)	AB de Villiers*	F du Plessis
	04. F du Plessis (4)	AB de Villiers	F du Plessis (1)	JP Duminy (3)
	05. F Behardien	JP Duminy	JP Duminy (1)	AB de Villiers
	06. D Miller	MV Boucher*	F Behardien	D Miller
	07. JA Morkel	JA Morkel (4)	RJ Peterson (4)	JA Morkel (2)
	08. M Morkel (4)	J Botha (2)	JA Morkel	D Steyn (3)
	09. LL Tsotsobe (4)	RE van der Merwe (2)	J Botha (3)	BE Hendricks (4)
	10. I Tahir (4)	D Steyn (4)	D Steyn (4)	I Tahir (4)
	11. D Steyn (4)	CK Langeveldt (4)	M Morkel (4)	WD Parnell (3)
	NS Q de Kock			
	NS BE Hendricks			
	NS WD Parnell			
	NS AM Phangiso			
E(D)	36.6	14.7	19.8	14.7

Table 3.2: Optimal lineup for South Africa and three typical lineups that were used on the specified dates. The vertical numbering corresponds to the batting order where the players labelled NS were not selected. In parentheses, we provide the number of overs of bowling and the asterisk denotes the wicketkeeper.

A significant difference between Table 3.2 and Table 3.1 is that the South African optimal lineup has three fast bowlers (Morkel, Tsotsobe and Steyn) compared to India's one fast bowler. This implies that the South African fast bowlers are more effective than their Indian counterparts. And, this is again suggestive that teams should consider the best players available for team selection, and not be tied to a tradition of having an even mix of fast and spin bowlers.

Perhaps the largest discrepancy between the optimal lineup and the 2014 actual lineup is that Q. de Kock was not selected to the optimal lineup whereas he is the opening batsman in 2014. We note that de Kock did not play all that well in the World Cup scoring only

64 runs in five innings. Also, the optimal lineup appears to place more value on A.B. de Villiers as he is selected as an opener compared to No 5 in the 2014 lineup.

3.4.3 All-Star Lineup

To our knowledge, the determination of an "all-star" team which attracts interest in many team sports (e.g. the National Basketball Association, the National Hockey League, etc.) is not something that is common to cricket. We have therefore carried out this novel exercise by including a larger pool of 25 widely recognized and current T20 players for team selection. This provided a further test of the algorithm to see that the size of the player candidate pool did not impose an impediment to obtaining an optimal lineup.

The optimal all-star lineup is shown in Table 3.3. Although there is no way of assessing whether the optimal lineup is "correct", our cricketing intuition is that the resultant all-star team is very strong. It is interesting to note that four of the selected players are from the West Indies, of whom three are all-rounders and are well-known power hitters. The West Indies have been a strong T20 team in recent years, winning the 2012 World Cup and reaching the semi-finals in the 2014 World Cup.

The fact that M.S. Dhoni was not selected to the all-star team reinforces the observation that he was placed down at No 9 in the optimal Indian batting order. We note that although V. Kohli is thought by some to be the best current Indian cricketer, he was not selected to the all-star team, yet his teammate, S. Raina was selected. It was also interesting to observe that B.B. McCullum was selected as the wicketkeeper instead of A.B. de Villiers who is known as a great power hitter. The all-star team had an impressive expected run differential of $\mathrm{E}(D) = 61.6$.

3.5 DISCUSSION

This paper provides a powerful methodology for the joint problem of determining team selection, the optimal batting order and the optimal bowling order in Twenty20 cricket. No such work has been previously carried out on this signficant problem. Moreover, the approach is based strictly on analytics, avoiding opinion, folklore and tradition.

One of the features of the proposed approach is that teams are free to modify player characteristics. Perhaps a player is in particularly good form and it is believed that his probability of dismissal is reduced. It is a simple matter of lowering his value of p_{i707} in (3.3) and observe what optimal lineup is obtained. Alternatively, one can hard-code a subset of players into the lineup and build the remainder of the roster around them.

We note that optimal lineups may not be greatly different from some alternative lineups in terms of expected run differential. The simulator provides a means for investigating the difference.

Optimal Lineup	Pure Batsmen not Selected	Bowlers & All-Rounders & Wicketkeepers not Selected
01. AJ Finch, Aus	G Bailey, Aus	S Afridi, Pak
02. BB McCullum, NZ*	MJ Guptill, NZ	S Ajmal, Pak
03. NLTC Perera, SL (2)	AD Hales, Eng	K Akmal, Pak*
04. G Maxwell, Aus (2)	R Levi, SA	C Anderson, NZ
05. C Gayle, WI	D Miller, SA	DJ Bravo, WI
06. K Pollard, WI	A Shehzad, Pak	SCJ Broad, Eng
07. D Sammy, WI (4)	R Taylor, NZ	AB de Villiers, SA*
08. S Raina, Ind		MS Dhoni, Ind*
09. F du Plessis, SA (4)		JP Duminy, SA
10. S Badree, WI (4)		U Gul, Pak
11. D Steyn, SA (4)		M Johnson, Aus
		V Kohli, Ind
		L Malinga, SL
		S Narine, WI
		D Ramdin, WI*
		KC Sangakkara, SL*
		S Watson, Aus

Table 3.3: All-Star lineup including players not selected. In parentheses, we provide the number of overs of bowling and asterisks denote wicketkeepers.

It is our hope that papers like this will help promote the adoption of analytics in cricket. Big money is now being spent in leagues such as the IPL, and it is only sensible to make use of the knowledge contained in data.

Chapter 4

Assessing the Impact of Fielding in Twenty20 Cricket

4.1 INTRODUCTION

The three major components that lead to success in cricket are batting, bowling and fielding. Whereas expertise in batting and bowling is readily quantifiable with familiar measures such as batting average, bowling average, strike rate, etc., there are no popular measures that exist for fielding.

In Twenty20 cricket, [10] introduced simulation procedures that estimate the expected number of runs that players contribute to their teams in terms of batting and bowling when compared against average players. In cricket, a player's expected run differential (in comparison to an average player) is an unequivocal performance measure of interest as it leads to directly to wins and losses. In [10], it is seen that the best Twenty20 players contribute nearly 10 additional runs when compared to average players. Yet, there have been no similar investigations on the effects of fielding. An exceptional fielding play in cricket may be described as "fantastic" but there is no sense of the scale of the contribution in terms of runs. This paper is a first attempt to quantify the impact of fielding.

In an ideal world, cricket matches would yield detailed spatial data where fielding contributions could be objectively assessed. For example, one could determine the distance covered by a fielder in making a catch and the time that it takes to cover the distance. In Major League Baseball (MLB), spatial data is facilitated through FIELDf/x technology and papers have been written (e.g. [25]) that assess fielding contributions via spatial statistics. Similarly, in the National Basketball Association (NBA), spatial data is provided through the SportVU player tracking system which records the coordinates of the ball and each player on the court 25 times per second. Detailed data of this sort have provided opportunities to investigate lesser-studied basketball characteristics such as defensive proficiency [19]. As cricket is the second most popular game in the world (following soccer), it seems only a matter of time until spatial data will likewise be available for analysis.

However, for the time being, the most comprehensive cricket data consist of *match commentaries*. A match commentary is a recorded conversation between announcers that provides a description of what has occurred on the field. Our approach first involves the parsing of match commentaries to obtain ball-by-ball data. This represents a considerable improvement over the use of *match summaries* which only report aggregate data. With match commentaries, we then carry out a textual analysis using random forest methodology to quantify the impact of fielding.

Broadly speaking, this paper introduces "moneyball" concepts which strike at the core of what is important in Twenty20 cricket. The book Moneyball [30] and its ensuing Hollywood movie starring Brad Pitt chronicled the 2002 season of the Oakland Athletics, a small-market MLB team who through advanced analytics recognized and acquired undervalued baseball players. Moneyball analyses often rely on *relative value statistics*. Relative value statistics have become prominent in the sporting literature as they attempt to quantify what is really important in terms of winning and losing matches. For example, in Major League Baseball (MLB), the VORP (value over replacement player) statistic has been developed to measure the impact of player performance. For a batter, VORP measures how much a player contributes offensively in comparison to a replacement-level player [49]. A replacement-level player is a player who can be readily enlisted from the minor leagues. Baseball also has the related WAR (wins above replacement) statistic which is gaining a foothold in advanced analytics (http://bleacherreport.com/articles/1642919). In the National Hockey League (NHL), the plus-minus statistic is prevalent. The statistic is calculated as the goals scored by a player's team minus the goals scored against the player's team while the player is on the ice. More sophisticated versions of the plus-minus statistic have been developed by [44] and [21]. In this paper, we introduce *expected runs saved due to fielding* in Twenty20 cricket. This may be classified as a relative value statistic as it measures the number of runs that a team saves on average due to the fielding performance of a given player when compared against a baseline player.

To our knowledge, the first and only quantitative investigation of fielding was undertaken by [40]. They proposed measures that are based on subjective weights. By contrast with the statistic proposed in this paper, their approach requires a video assessment of every fielding play to provide a measure of fielding proficiency.

In Section 4.2, we provide an overview of the match simulator developed by [11]. The simulator is the backbone for calculating expected runs saved due to fielding. 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 define the proposed metric of expected runs saved due to fielding and describe its calculation via simulation methodology. The calculation relies on a fielding matrix Λ for a given fielder. The estimation of Λ is carried out by a textual analysis using random forest methodology. An innovation in our estimation procedure involves the amalgamation of matches from

international Twenty20 cricket and the Indian Premier League (IPL). In Section 4.4, we calculate expected runs saved due to fielding for players with a sufficient Twenty20 history. Some surprising results are revealed and we are able to put into context the importance of fielding in Twenty20 cricket. We conclude with a short discussion in Section 4.5.

4.2 OVERVIEW OF SIMULATION METHODOLOGY

We now provide an overview of the match simulator developed by [11] which we use for the calculation of expected runs saved due to fielding. There are 8 broadly defined outcomes that can occur when a batsman faces a bowled ball. These batting outcomes are listed below:

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

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 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), [11] 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 is "special" in the sense that it consists 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 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 p_{iowj}, [11] introduced the simplification

$$p_{iowj} = \frac{\tau_{owj}\ p_{i70j}}{\sum_j \tau_{owj}\ p_{i70j}}\ . \tag{4.3}$$

In (4.3), the parameter p_{i70j} represents the baseline characteristic for batsman i with respect to batting outcome j. The characteristic p_{i70j} is the probability of outcome j associated with the ith batsman at the juncture of the match immediately following the *powerplay* (i.e. the 7th over) when no wickets have been taken. The multiplicative parameter τ_{owj} scales the baseline performance characteristic p_{i70j} to the stage of the match corresponding to the oth over with w wickets taken. The denominator in (4.3) ensures that the relevant probabilities sum to unity. There is an implicit assumption in (4.3) that although batsmen are unique, their batting characteristics change with respect to overs and wickets by the same multiplicative factor which is essentially an indicator of aggression. For example, when aggressiveness increases relative to the baseline state, one would expect $\tau_{ow4} > 1$ and $\tau_{ow6} > 1$ since bolder batting leads to more 4's and 6's.

Given the estimation of the parameters in (4.3) (see [11]), 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 accumulated or the number of overs reaches 20, the first innings is terminated. [11] also provide modifications for batsmen facing specific bowlers (instead of average bowlers), they account for the home field advantage and they provide adjustments for second innings batting. In [11], the simulator is demonstrated to provide realistic realizations for Twenty20 cricket.

4.3 THE APPROACH

Recall that the match simulator of Section 4.2 generates batting outcomes according to situational probabilities. Specficially, p_{iowj} is the probability of batting outcome j when batsman i is batting in the oth over having lost w wickets. We first consider an average batting team (i.e. average positional player at each batting position) batting against an average fielding team. Simulating over many first innings (based on the Twenty20 International dataset in [11]), we obtained the expected runs scored $\mathrm{E}(R) = 149.9$.

We now contemplate the introduction of a given fielder to the bowling side. With the introduction of such a fielder, the average batsmen at the relevant stage of the innings with batting characteristics p_{iowj} now bats according to p^*_{iowj} where the updated characteristics are due to the impact of the presence of the fielder. We then simulate first innings according to p^* and obtain the expected runs scored $\mathrm{E}(R^*)$. For the particular fielder, this leads to

our proposed metric

$$\mathrm{E}(RSF) \quad = \quad \mathrm{E}(R) - \mathrm{E}(R^*) \tag{4.4}$$

which is the expected runs saved due to fielding. The larger the value of E(RSF) in (4.4), the better the fielder. We note that E(RSF) is an appealing statistic as it is directly interpretable in terms of runs.

In order to carry out the simulations required in (4.4), it is necessary to obtain the batting characteristics p^*_{iowj} which have been modified from p_{iowj} due to the presence of the fielder of interest. For ease of notation, we temporarily suppress the subscripts iow and express

$$p^*_k = \lambda_{0k}p_0 + \lambda_{1k}p_1 + \cdots + \lambda_{7k}p_7 \ . \tag{4.5}$$

In (4.5), the fielding characteristic λ_{jk} represents the conditional probability that the fielder converts the batting outcome j to the batting outcome k. For example, if $\lambda_{21} = 0.05$, this denotes that 5% of the time, the fielder can alter the outcome of a typical 2-run scoring play to a single run due to exceptional fielding. We then define the column vectors $p = (p_0, p_1, \ldots, p_7)^T$ and $p^* = (p^*_0, p^*_1, \ldots, p^*_7)^T$ which describe the probability of batting outcomes based on an average fielder and the fielder of interest, respectively. We also define the matrix of fielding characteristics

$$\Lambda = (\lambda_{jk}) \tag{4.6}$$

for the fielder of interest. The matrix Λ describes the impact of the fielder in transforming typical batting outcomes to other batting outcomes due to his fielding. From (4.5), we then have

$$p^* = \Lambda^T p \tag{4.7}$$

where probability restrictions require that $0 \leq \lambda_{jk} \leq 1$ for all j, k and that the rows of Λ sum to unity. Given the rarity of exceptional fielding plays, we expect the diagonal elements λ_{kk} of Λ to be large (i.e. close to unity).

To summarize, we first consider a typical batting lineup against a typical bowling lineup based on the batting characteristics p. Via simulation, this typical team scores $\mathrm{E}(R) = 149.9$ first innings runs on average. With the inclusion of a fielder of interest, the batting characteristics are modified from p to p^* via (4.7) and the expected runs scored with the presence of the fielder can be simulated to obtain $\mathrm{E}(R^*)$. Therefore, the fielder's contribution in terms of expected runs saved due to fielding is given by $\mathrm{E}(RSF) = 149.9 - \mathrm{E}(R^*)$. In the next section, we use textual analysis and random forests to estimate the fielding matrix Λ.

4.3.1 Parameter Estimation

We now outline the estimation of the fielding matrix Λ. Recall that λ_{jk} is the conditional probability that given a batting outcome j, the fielder is able to alter the batting outcome to k.

In the estimation of Λ, we have used data from both international Twenty20 cricket and the IPL. Specifically, we have 286 international Twenty20 matches involving the 10 full-member nations of the ICC from the period February 17, 2005 through April 3, 2014. For the IPL, we have 324 matches taken from the 2009 through 2015 seasons. Whereas the batting and bowling standards of the two data sources may not be exactly the same, we believe that the assessment of good and bad fielding does not depend greatly on the level of the competition. For example, a good or bad fielding play does not depend on the quality of the batsman nor on the quality of the bowler. Our combination of the two data sources appears to be novel and helps provide more reliable estimation of the fielding matrix Λ.

The data consist of match commentary logs which are available from the CricInfo website (www.espncricinfo.com). Each line of commentary provides us with information on a particular ball that was bowled. The main idea is that we pay particular attention to events where a fielder's name is mentioned in the commentary logs. When a fielder has done something either good or bad with respect to fielding, invariably his name will be mentioned. For example, consider the following two excerpts from commentary logs where a fielder's name has been mentioned:

- Apr 18/09, Bangalore vs Rajasthan (IPL), 2nd innings, 11.2 over - Kumble to Jadeja, OUT, That should seal it for Bangalore, gave it air this time unlike the previous one where he bowled it flat and short, Jadeja got down on one knee and tried to slog-sweep it over deep midwicket, got a lot of elevation and though it was struck well, didn't get the distance he desired, Kohli ran well to his right to take a neat catch well inside the ropes.

- Sep 27/12, Sri Lanka vs New Zealand (Twenty20 International), 1st innings, 6.2 over - Mathews to Guptill, FOUR, short of a length outside off, Guptill slashes at it and gets a thick outside edge towards third man, Malinga is on the boundary and he runs to his right and misses the ball. The ball spun past him and he was clutching at air. Horribly dozy effort.

The first example highlights an example corresponding to good fielding whereas the second example corresponds to poor fielding. As can be seen, there is a diversity of language in the match commentaries to describe events. This is suggestive of the use of machine learning tools to reveal patterns in the text.

Our approach partitions the 149,764 balls in the dataset (i.e. lines of commentary) into A: the 137,107 cases where no fielders' names are mentioned and B: the remaining 12,657

cases where fielders' names are mentioned. We designated 66 keywords (see Table 4.1) in dataset A as features that provide predictive information on the batting outcome (i.e. the observed dependent variable (4.1)). The 66 keywords were the most frequently occurring words in the training set A after removing *stop* words (e.g. prepositions, pronouns), words that directly restate the outcome (e.g. "run", "six", "catch", etc.) and player names. The most common keyword was "leg" which appeared 17,297 times in dataset A. The least common keyword was "brilliant" which appeared 146 times in dataset A. We observe the contextual nature of the words in Table 4.1 which provide insight on the outcome from a bowled ball.

across	brilliant	edge	great	pace	straight
air	excellent	hard	pads	strike	almost
charge	fine	high	power	stump	angle
chipped	firm	hit	pull	super	arms
clear	flash	knee	reverse	sweep	back
clip	flick	leg	running	swing	backward
cut	foot	length	short	swung	
bad	deep	forward	long	shot	
banged	delivery	front	low	side	
bat	drill	gap	middle	slog	
beat	drive	gloves	midwicket	slower	
boundary	easy	good	misses	square	

Table 4.1: Contextual words referring to batting used in the random forest to predict batting outcome probabilities.

A random forest [23] was then grown using dataset A which provides predictive probability distributions for the outcomes in (4.1) for any line of commentary. More specifically, we used the `randomForest` function from the `randomForest` package in R. The random forest procedure has various tuning parameters to optimize predictive performance. For the application discussed here, we trained the random forest on a simple random sample of 100,000 commentary lines from dataset A. We then used the remaining observations as a validation set for choosing the tuning parameters. The optimal predictive performance was found using 2,500 trees, with each leaf including at least five observations. At each node, the best split was found by searching over a random size 15 subset of the total covariates. The covariates are binary variables indicating the presence or absence of the keywords in Table 4.1. This random subsetting causes the individual regression trees in the random forest to be less correlated and allows the model to identify subtle effects of keywords that might otherwise be missed by individual regression trees.

We now describe how the random forest is used to estimate the fielding parameters λ_{jk} for a specified fielder. For a specified fielder, suppose that he has been on the field for n balls and suppose that his name has been mentioned in the commentary lines of dataset B for m balls. Designate his noteworthy plays with the index $i = 1, \ldots, m$ such that the random

forest produces predicted outcome probabilities $q_{i0}, \ldots, q_{i7}$. These are the probabilities corresponding to what would have happened had the fielder not made a noteworthy fielding play. However, the ith fielding play did produce a realized batting outcome and we denote this outcome by O_i. This leads to the following estimators

$$\hat{\lambda}_{kk} = \frac{n-m}{n} + \frac{m}{n}\left(\frac{\sum_{i=1}^{m} I(O_i = k) q_{ik}}{\sum_{i=1}^{m} q_{ik}}\right) \tag{4.8}$$

and

$$\hat{\lambda}_{jk} = \frac{m}{n}\left(\frac{\sum_{i=1}^{m} I(O_i = k) q_{ij}}{\sum_{i=1}^{m} q_{ij}}\right) \tag{4.9}$$

for $j \neq k$ where I is the indicator function. A detailed construction of (4.8) and (4.9) is provided in Appendix B.

We note that equations (4.8) and (4.9) make sense theoretically. If a player's name is never mentioned in dataset B, then $\hat{\lambda}_{kk} = 1$. This implies that a player never makes mistakes and never makes an exceptional fielding play. Consequently $p^* = p$ for the player of interest, and through simulation, $\mathrm{E}(RSF) = 0$. Later, in our data analysis, we see that it is easier for players to make mistakes than to make exceptional fielding plays. Therefore $\mathrm{E}(RSF) = 0$ actually corresponds to an elevated standard of play.

A slight difficulty with (4.8) and (4.9) is that machine learning algorithms do not take into account the physical aspects of the game. In cricket, there are various scenarios which cannot occur. For example, it would be impossible for a fielder to convert what would have been zero runs to six runs no matter how poor his fielding. In terms of the fielding matrix, this implies the conditional probability $\lambda_{06} = 0$. This also holds true for other scenarios and we therefore introduce the constraints $\lambda_{06} = \lambda_{16} = \lambda_{26} = \lambda_{46} = \lambda_{60} = 0$. When we estimate the Λ matrix using (4.8) and (4.9), we make a final adjustment by scaling the non-zero λ's in a given row proportionally so that each row sums to unity.

Another adjustment that we make to improve estimation is based on the commentaries in dataset B. The content of each commentary line indicates whether a fielder has made a "good", "bad" or "neutral" fielding play. A good play means that the fielder did something that improved the outcome compared to what an average fielder would have done. Therefore, we adjust the prediction distribution obtained from running the random forest on the commentary line of interest. We do this by assigning zero probability to outcomes that are better than the actual outcome. We then scale the remaining prediction probabilities to sum to unity. Likewise for a bad play, we assign zero probability to outcomes that are worse than the actual outcome and we scale the remaining prediction probabilities to sum to unity. For commentaries that are neutral with respect to a fielder's performance, we reassign these cases to dataset A.

In the next section, we see that MS Dhoni, India's legendary wicketkeeper, has been identified by our methodology as a weaker fielder according to the $\mathrm{E}(RSF)$ metric. Table

4.2 provides the estimated fielding matrix for Dhoni. Note that we have combined outcomes $j = 2$ and $j = 3$, and we have combined outcomes $j = 4$ and $j = 5$. This was done since 3's and 5's are rare. One thing we observe from Table 4.2 is that λ_{70} is relatively large. This means that Dhoni drops potential catches and misses stumping opportunities more often than other wicketkeepers. One of the other things that we observe from Table 4.2 is that the relative sizes and magnitudes of the λ_{jk} conform to common sense. For example, it seems that λ_{01} should exceed λ_{02}. The reason is that the consequences of fielding errors from lightly hit balls are easily contained.

		k					
		0	1	2, 3	4, 5	6	7
	0	0.979	0.008	0.001	0.004	0.000	0.007
	1	0.012	0.977	0.003	0.004	0.000	0.005
j	2, 3	0.004	0.017	0.970	0.005	0.000	0.004
	4, 5	0.006	0.012	0.003	0.974	0.000	0.005
	6	0.000	0.006	0.005	0.013	0.974	0.003
	7	0.021	0.002	0.001	0.003	0.000	0.973

Table 4.2: Estimated fielding matrix $\Lambda = (\lambda_{jk})$ for MS Dhoni.

4.4 PLAYER ANALYSIS

We now consider the analysis of specific players. We restrict our attention to the 157 players who have been on the field for at least 2000 balls bowled. Therefore our analysis consists of well-established Twenty20 players.

In Table 4.3, we consider the E(RSF) metric for wicketkeepers. We have included all 13 wicketkeepers in our dataset. Wicket-keepers are considered separately as their positioning on the field allows them greater opportunity to make fielding plays. We observe that Mushfiqur Rahim of Bangladesh is the best fielding wicketkeeper followed closely by Quinton de Kock of South Africa. At the bottom of the list, we were suprised to find the current Australian wicketkeeper Brad Haddin since Australia has traditionally been a strong Twenty20 side.

Another observation concerning Table 4.3 is that there is a negative skewness in the E(RSF) statistic. This can be explained by noting that fielders may make mistakes in a myriad of ways. For example, they can drop balls, they can make throwing errors, they can trip, they can fail to stop balls, etc. On the other hand, there are limited opportunites for a fielder to make an exceptional play. For example, if a batted ball is not within reach, nothing can be done.

In Table 4.4, we list the top 10 fielders who are non-wicketkeepers. It would have been nice to differentiate those players who are positioned in the infield and are more often in

Name	Team	n	m	E(RSF)
M Rahim	BAN	2995	18	-0.55
Q de Kock	SA	2899	36	-0.65
K Akmal	PAK	6698	138	-1.07
T Taibu	ZIM	2017	24	-1.08
D Ramdin	WI	4979	115	-1.36
KC Sangakkara	SL	9879	128	-1.43
AC Gilchrist	AUS	5710	105	-1.76
BB McCullum	NZ	6128	119	-2.07
MV Boucher	SA	5289	137	-2.25
JC Buttler	ENG	3033	57	-2.79
MS Wade	AUS	2347	55	-3.29
MS Dhoni	IND	13077	395	-3.59
BJ Haddin	AUS	4950	133	-4.51

Table 4.3: Expected runs saved due to fielding by wicketkeepers. The variable n refers to the fielder's total number of fielding opportunities and m is the number of notable plays by the fielder such that his name appeared in dataset B.

positions to make fielding plays. However, the game is fluid and players frequently change positions on the field according to the game conditions. We note that there is a single exceptional fielder Rob Nicol of New Zealand. This is surprising as a cursory internet search reveals nothing concerning his fielding prowess. In the media, there seems to be a belief that AB de Villiers of South Africa is a remarkable fielder. According to our methodology, he ranks 22nd of the 144 non-wicketkeepers with E(RSF) $= -0.51$. We note that many of the players in Table 4.4 have small values of m. This means that while fielding, their names rarely appear. In fact, this is a good thing. Such players are not mentioned because they make few mistakes. As discussed previously, there are more opportunities to make mistakes than to make exceptional fielding plays. The average E(RSF) value amongst the 144 fielders is -1.19 and the lowest value is -4.61 (Ryan Sidebottom of England). Therefore the top fielders who have an E(RSF) metric near zero save approximately 1.2 runs compared to a typical fielder.

In Figure 4.1, we provide a plot of m/n versus E(RSF). What is interesting is the decreasing trend. The trend implies that players who are infrequently mentioned relative to their fielding opportunities are doing a good job of fielding. In other words, if a player is seldom identified for his fielding, then he is likely not making mistakes and is contributing to his team. Again, in a cricket match, there are more opportunities to make mistakes than to do something exceptional with respect to fielding during a match. We also note that wicketkeepers tend to be mentioned more frequently.

Name	Team	n	m	E(RSF)
RJ Nicol	NZ	2313	13	1.80
JE Taylor	WI	2021	6	0.22
HDRL Thirimanne	SL	2111	7	0.05
CK Kapugedera	SL	3105	11	0.00
P Kumar	IND	5853	7	-0.01
MM Sharma	IND	3447	7	-0.02
Tamim Iqbal	BD	2510	2	-0.02
B Kumar	IND	4058	4	-0.05
CL White	AUS	8266	58	-0.08
E Chigumbura	ZIM	2492	8	-0.18

Table 4.4: Expected runs saved due to fielding by the top 10 non-wicketkeepers. The variable n refers to the fielder's total number of fielding opportunities and m is the number of notable plays by the fielder such that his name appeared in dataset B.

4.5 DISCUSSION

The home for comprehensive cricket data is the Statsguru search engine at espn.cricinfo.com. Currently, the only information collected on fielding via Statsguru is aggregate data which involves catches and stumps for individual fielders. While such data may be of some value, it does not consider fielding contributions due to run-outs. In addition, Statsguru does not differentiate between difficult and easy fielding plays, nor does it consider exceptional fielding plays that reduce the number of runs scored.

On the other hand, we have developed methods that potentially account for all fielding contributions. The approach is based on the textual analysis of ball-by-ball match commentaries using random forests.

Whereas the best individual batters and bowlers contribute roughly 10 runs per match on average in Twenty20 cricket [10], we have seen that a really good fielder may only save 1.2 expected runs for his team. Whereas this may seem small, it is possible to have multiple good fielders on a team, and therefore the impact of fielding becomes more meaningful.

Now that the impact of fielding can be described in terms of an easily understood quantity (runs), it is possible that better decision making can be made with respect to team selection and salaries. Furthermore, there seems to be a divergence of popular opinion concerning the very best fielders; our Tables 4.3 and 4.4 can shed some quantitative light on this topic.

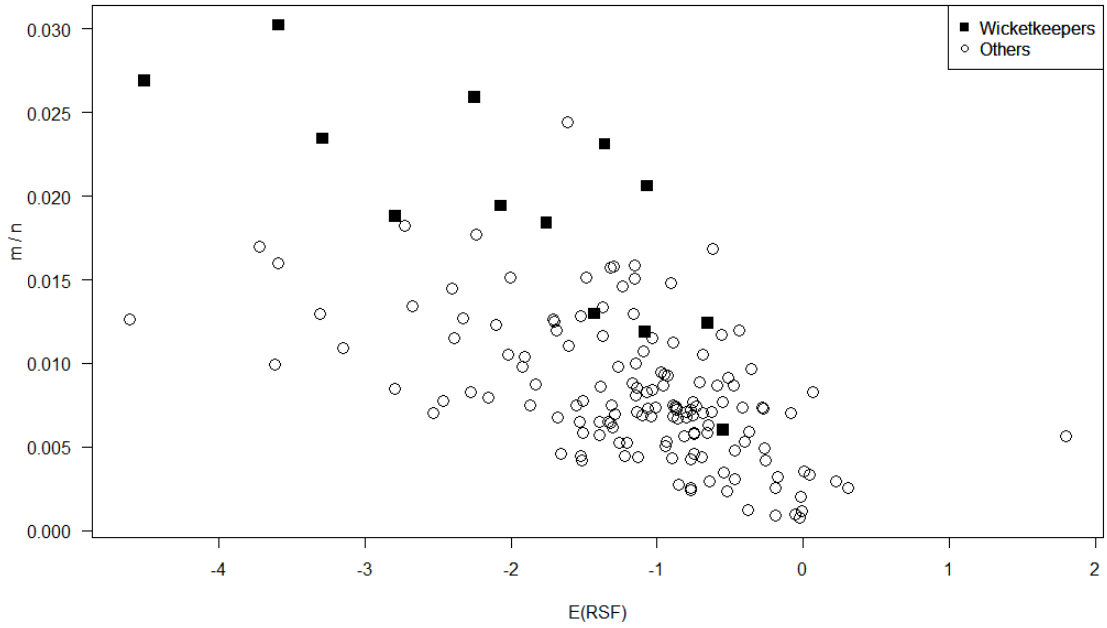

Figure 4.1: Scatterplot of the ratio of mentions to fielding opportunities (m/n) versus E(RSF).

Chapter 5

Muralitharan and Sangakkara: Forging Identity and Pride through Cricket in a Small Island Nation

5.1 INTRODUCTION

It was the seminal moment in Sri Lankan cricketing history. A nation's heart was aflutter. Nearly every eye was glued to the television. For those without means, every ear was tuned to the radio. It was Sri Lanka versus powerhouse Australia in the 1996 One-Day International (ODI) World Cup final, only the sixth ODI World Cup to be contested.

It was also a time of uncertainty in Sri Lanka. The Tamil Tigers had recently bombed the Central Bank in Colombo, killing 91 people and injuring 1400 others. With India, Pakistan, and Sri Lanka jointly hosting the World Cup, there were security concerns over the Sri Lankan venues with two of the Group Stage matches abandoned. Yet, in this troubling time, on the world stage, tiny Sri Lanka had made it to the final. They were a decisive underdog. In the openings innings of the match, Australia was held to 241 runs, largely due to the bowling expertise of Muttiah Muralitharan ("Murali"), then a 23 year old newcomer to international cricket. He was the most outstanding bowler of the finals, limiting Australia to 31 runs through 10 overs. In the second innings, Sri Lanka scored 245 runs, winning by 7 wickets with 22 balls remaining.

At the end of the match, wild celebrations erupted in Sri Lanka, and amongst the Sri Lankan diaspora throughout the world, including Australia. Murali had become a national icon. Hardly an imposing figure at 5 feet 7 inches tall and of Tamil background, the Sinhalese majority worshipped him. Back in Kandy, a young schoolboy of 18 years, Kumar Sangakkara ("Sanga"), had been watching the final and had been duly inspired.

This chapter tells the story of Murali and Sanga, two Sri Lankan cricketers whose backgrounds could not be more different. Yet together, they have made a deep impression on the game of cricket, forging identity and pride in the small island nation of Sri Lanka. Moreover, this chapter speaks to the excellence of these two cricketers in the face of challenging

politics and ethnic turmoil in a country that is relatively new to the game of international cricket. Our chapter begins by examining the roots of cricket in Sri Lanka beginning from colonial times. We then describe the paths taken by Murali and Sanga from childhood to their prominence in world cricket, to their status today as ambassadors and icons for Sri Lanka.

5.2 SRI LANKAN CRICKET IN CONTEXT

To understand the context of cricket in Sri Lanka, we first need to delve into the history of Sri Lanka in general and the history of cricket in Sri Lanka in particular. Sri Lanka is an island with a diverse population of nearly 21 million people in a land area of 25,000 square miles - about the size of West Virginia. It is located just off the southeast coast of India, which is the second most populous country in the world.

Originally the island was inhabited by Veddah tribes who were the equivalent of Canada's First Nations, i.e. aboriginal people. Migration first began with Prince Vijayaand his followers (ca 543 BC) who arrived from what is now West Bengal in India, and originated the Sinhalese ethnicity. They were followed by Dravidians (of Tamil ethnicity) from South India.

Traders and sailors from Arab nations followed - the Moors in particular - who in turn were followed by colonists from Europe (Portuguese, Dutch and British) lured by the rich trade in spices. In 1815, Ceylon (later named Sri Lanka) became a British Crown Colony. During the British Raj, numerous ethnicities settled in Sri Lanka. A second wave of Tamils, brought by the British, worked as labourers in the tea plantations in the 1860s, including Murali's grandfather.

All of these groups had deep influences on aspects of Sri Lankan life in terms of politics, ethnicity, language, religion, art, architecture, music, food, dress, etc., all of which remain evident in the early 21st century. The Portuguese and the Dutch, many of whom intermarried with the local population (unlike the British) and whose descendants are called *Burghers*, left their mark in place names, family names, food, and religion. The island has been known by many names by many peoples, including the British 'Ceylon' the name that continued to be used after independence in 1948, until the island was renamed Sri Lanka in 1972 [13].

With such a history, one can readily appreciate the current multi-ethnic and multicultural society in Sri Lanka where practices include Buddhism, Hinduism, Islam, and Christianity. Looking forward, it is difficult to imagine how a simple game (cricket) came to dominate the consciousness of this diverse country. Under British control of the maritime provinces in 1802 and the whole island in 1815, the British instituted many customs, including cricket, which was chief among the British recreational activities. The game itself dates back to England circa 1600. In Sri Lanka, however, the Colombo Cricket Club, the first

such club, was set up in 1832, though they did not played their first official match against a British regiment until the next calendar year. The first international match between a Sri Lankan team and an English team was held in Colombo in 1882 and was facilitated by the opening of the Suez Canal in 1869. The match was a consequence of a travel break taken by the English team enroute to Australia when the ship stopped for fuel and provisions in Colombo. These matches were what are known as *limb-looseners* for the traveling cricketers. The matches provided an opportunity for local players to gain 'first-class' experience and were awaited with bated breath by the residents of Colombo. With the popularisation of air travel in the mid-1960s, ships carrying cricketers no longer called in Colombo, and as a result, few international cricketers came to Sri Lanka. Many of the founding Sri Lankan cricketing clubs have ethnically based names such as the Sinhalese Sports Club (SSC), the Tamil Union, the Moor's Sports Club, and the Burgher Recreation Club (BRC), dating back to the 1890s. In the early 21st century, club teams are no longer ethnically based, although the club names have not changed.

Significantly, Murali and Sanga hail from the two main ethnic groups in Sri Lanka, the Tamil and Sinhalese communities respectively. Although these two communities have had a troubling relationship, cricket has been a unifying factor in Sri Lanka, and with respect to cricket, ethnicity has never been an issue. In fact, the current Sri Lankan cricket team includes almost all of the different ethnicities and religious persuasions prevalent on the island. This speaks volumes for the passions generated by the game to the extent that ethnic loyalties can be drowned by an overriding national loyalty for the Sri Lanka cricket team. Between 1953 and 1975, the only serious international matches involving Sri Lankan teams were between the Ceylon Cricket Association and the Madras team of India. These were first-class matches played for the Gopalan Trophy with the venue alternating between Colombo and Madras (now Chennai). During this period, cricket was commonly played by boys in village fields, backyards, beaches, and even in fallow rice paddies using discarded tennis balls and bats, when such were available, or with improvised equipment such as balls made of 'kaduru' a type of large round hard seed, and bats made of 'polpiti' (the leaf stalk of coconut fronds). In contrast, real cricket with proper equipment was reserved for boys at elite schools that had the necessary resources and cricket grounds. These were the youngsters who had an opportunity to play for the local cricket clubs that held first-class tournaments.

To be sure, cricket is the most popular sport in Sri Lanka whether it be at international, club or school levels. However, school cricket has the longest history and also perhaps the most partisan supporters. Some of the schools have played each other for well over a century. The 'Big Matches' between schools get far more publicity and press coverage than any first class club matches. The associated parades and carnivals create a festival atmosphere and is now embedded in the local culture. The original 'Big Match' between Royal College and St. Thomas' College, two elite boys' schools in Colombo, dates back to 1879. By the

beginning of the First World War in 1914, there were three 'Big Matches' on the island, one of which was played between the two elite schools in Kandy, St. Anthony's College and Trinity College, which respectively were to be the crucibles for schoolboys Murali (born in 1972) and Sanga (born in 1977). Since school cricket teams are age denoted, Murali and Sanga did not actually face each other on the field in any of the 'Trinity-Anthonian' Big Matches.

Murali and Sanga began their cricketing careers at a time when Sri Lanka cricket was still in its infancy. Sri Lanka had not yet attained international 'Test' status and even the opportunity to view high level cricket was limited. Although future talent was being developed through the school system, the game was raw, coaching was undeveloped, and the odds of rising to international star status seemed insurmountable.

Sri Lanka attained Test status, the highest designation in international cricket, in 1981 after years of canvassing with the Marylebone Cricket Club (MCC). The MCC is England's original cricket club established in 1787 with its home grounds being the hallowed Lord's Cricket Ground in St John's Wood, London. The MCC evolved over time to be the governing body for international cricket until the establishment of the International Cricket Council (ICC) in the early 1990s. Until 1981, the only countries enjoying Test status were England (1877), Australia (1877), South Africa (1889), West Indies (1928), New Zealand (1930), India (1932), and Pakistan (1952). For a while it seemed like only these seven countries would ever play Test cricket. However, after a hiatus of thirty years, Sri Lanka managed to gain entry as the 8th country by displaying their skills in both cricket and diplomacy. Subsequent to Sri Lanka's ascension to Test status in 1981, two other countries have been admitted to the Test cricket club: Zimbabwe in 1992 and Bangladesh in 2000 [1].

Notably, all of the ICC's Test playing nations were once part of the British Empire through varying degrees of colonization. Understandably, there has always been particular joy for Sri Lankans (and other former British colonies) when the English are defeated in cricket. Sri Lanka also revels when it defeats its neighbouring superpower India.

The story of how Sri Lanka achieved Test status from its humble beginnings in cricket is a consequence of diplomacy, determination, and skill. Sri Lankan's inclusion as a Test country in 1981 can be mainly attributed to Gamini Dissanayake, a senior Cabinet Minister in the Sri who delivered a watershed speech to the MCC. Dissanayake was a lawyer by profession as well as a great orator with a dynamic personality. At the time, there was a feeling amongst MCC committee members that Sri Lanka should wait several more years before being admitted to the fold. However, Dissanayake was persuasive as he unleashed his charm, political savvy, oratorical skills, and legal mind with a promise to deliver on various infrastructure commitments to the MCC. In 2001, journalist David Hopps wrote, "The cricket world will be forever grateful that nearly 30 years ago the mission to grant Sri Lanka Test status was successful." [24]

Sri Lanka played their first Test match, fittingly against England, in Colombo in February, 1982. After attaining Test status, the government established funding for proper cricket to be played all over the island. Consequently, as the selection pool of talent increased, Sri Lanka began to achieve greater success at the international level. By the time Sri Lanka achieved Test match status in 1981, Murali was 9 years old and Sanga was 4 years old. For them, the impetus and the opportunities were beginning to open. However, relentless effort and dedication were required to get to the top of the world.

The popularity and excitement of cricket has grown dramatically in the last two decades, especially due to the shorter formats of the game such as one-day cricket and Twenty20 (or T20) cricket, a roughly three-hour game limited to 20 overs per side. In whatever format, the drama in the sport has historically been known as the "glorious uncertainty of cricket". In The *Wisden Cricket Almanack*, Rowland Ryder writes:

"Among the myriad delights of cricket, not least is the glorious uncertainty of the game. Nothing is certain in cricket except its uncertainty. It is not likely that a batsman will hit every ball of an over for six; that a last wicket stand will add three hundred runs to the score; that a wicket-keeper will take off his pads and do the hat trick: none of these things are anything more than remotely possible, yet all of them have happened; and improbable events, their duration in time varying from a split second to a long drawn out week, interesting, exhilarating, something unbearably exciting, are happening every year that cricket is played." [39]

5.3 MURALI: SPORT TRIUMPING OVER ETHNIC DIVISIONS

As mentioned earlier, Murali's grandfather emigrated from southern India and came to Ceylon, as the country was then called, to work in the tea plantations run by the British. Not much is known about the grandfather as he eventually went back to India. However, the grandfather's sons stayed on and established themselves in the central hill capital of Kandy, which is an area adjacent to the high mountains where the best tea plantations flourish.

Murali's opportunity to succeed at cricket from a young age shows how age-old antagonisms can be overcome by the egalitarian nature of modern sports, particularly spectator sports that engulf whole nations in fervent national pride. It is true to say that when important cricket matches involving the Sri Lanka team are being played, especially the shorter versions of ODI and T20, the nation comes to a halt, with most people watching the progress of play on TV in their homes or even their offices, with very little work being done. A palpable cricket above all else attitude is omnipresent, much to the detriment of

office productivity, though on the positive social side, it brings a sense of national unity where it is sorely needed.

Murali was disadvantaged, both economically in terms of his family's vocation and socially, as part of a minority ethnic group. Murali rarely gives interviews, but there is a 2010 piece where he discusses his ethnicity and childhood in an interview with Peter Roebuck in Melbourne, Australia newspaper *The Age* on 5th November 2010:

Peter Roebuck (PR): Let's talk about your background. Tell me about your ancestors.

Muttiah Muralitharan (MM): They come from India. I still have [the] right to live there. My grandfather came to Sri Lanka to work on a tea plantation. Afterwards he went back, but my father and his brothers stayed, and they built a biscuit factory in Kandy in the 1950s. All sorts of biscuits. Still we have that.

PR: Growing up as a Tamil in Sri Lanka wasn't an easy thing in your early days?

MM: There were riots, but after 1983, it was normal. Remember: I was staying at hostel in school for seven years and living with many Sinhalese and Tamils in the same dormitories, so it was not that difficult.

PR: But in the early days a lot of harm was done to the Tamils. Do you have any memories of that?

MM: Our factory and our house were burnt down in 1977, and that was painful for a time. We were saved by Sinhalese. They came and stopped the crazy people before they killed us. We never forgot that. We rebuilt them and moved on. That was our family way. We are businessmen not politicians. My father kept things as simple as possible.

PR: Do you think that these troubles and growing up in a mixed community helped to give you strength of character? The Tamils had a hard time.

MM: The Sinhalese as well. They had hard times when the communist party came. [T]hey were targeted and a lot of people were killed.

PR: You've never spoken up on political issues. You've been a unifying figure. Is that how you see yourself?

MM: Our lives in Kandy were mostly fine. I could not talk about problems I had not seen. [38]

The 'hostel in school' to which Murali refers is the hostel of St. Anthony's College, Kandy, a private school run by Benedictine monks. It is interesting to note that in Sri Lanka's multicultural, multi-ethnic, and multi religious society, such private schools foster a sense of egalitarian citizenship. Here was Murali, a Tamil follower of the Hindu religion, enrolled in a school run by Catholic Benedictine monks, some of whom came from Italy. Similarly, as we discuss later, Sanga, a Sinhalese and a follower of Buddhism, was put into

a school run by the British Anglican Church through the Ceylon Missionary Society, and that school was Trinity College, the other big private school in Kandy.

Ethnically, Murali is Tamil. He grew up in a Sri Lanka where ethnic fissures were a by-product of many factors including land hungerin an overpopulated island, 'official language' issues, job shortages, and terrorism; a conflict that have been fought between the majority Sinhalese and the minority Tamils. Though the origins of the conflict date back more than two thousand years when there were repeated invasions from southern India, there were intervening centuries of peace due to various geographical and political factors, including 450 years of colonisation by European powers [14].

Murali began playing school cricket at eight years of age and went on to become a schoolboy cricketing prodigy with a haul of over 100 wickets in a 14-game season in 1990-91 and winning the coveted 'Bata Schoolboy Cricketer of the Year', an award which considers all schools in Sri Lanka [12]. Following his schoolboy career, Murali was recruited to play for the Tamil Union Club while being included in the Sri Lanka 'A' team that toured England in 1991. Though that tour was not very successful, by 1992, he joined the Sri Lanka Test team and played his first Test match in August against the Australians when he was just 20 years old. During this otherwise unremarkable match, however, Murali took one wicket that put his extraordinary skills on display and focused the spotlight on him with the promise of greatness to come. Murali pitched a ball to Tom Moody that had the most unusual movement in the sense that it bounced in a completely unanticipated direction.To use Moody's own words: "I can clearly remember the ball he got me out with. It almost pitched off the strip and spun back five feet to bowl me middle-and-off while I was padding up. We thought he was a leg-spinner; his action was that unusual."

Murali's form continued to improve, and he played a pivotal role in most of Sri Lanka's matches, even though the team's wins were few and far between. Arjuna Ranatunga, the Sri Lanka captain, as well as a Sinhalese Buddhist, had absolute faith in Murali despite the traditional ethnic divide, and believed that Murali's presence would herald a new age for Sri Lanka's struggling, and thus far, short, Test history. Murali exceeded the high expectations. He continued to mesmerise batsmen from opposing teams, both in Sri Lanka and overseas, even when Sri Lanka as a team failed miserably.

By the last quarter of 1995, Murali had played in 22 Test matches against Australia, England, New Zealand, South Africa, India, and Pakistan while taking 80 wickets, though his accompanying bowling average allowing 32.7 runs per 100 balls was not that flattering. In the midst of this remarkable run came the fateful Boxing Day Test against Australia in Melbourne in 1995, a match that was destined to change some fundamental rules in relation to bowling and umpiring.

Murali was bowling against Australia in the 2nd Test in front of a huge crowd of 55,000 spectators at the hallowed MCG (Melbourne Cricket Ground) on Boxing Day in 1995. Suddenly, Australian umpire Darrell Hair called Murali for 'throwing' which in cricketing

parlance means bending one's arm at the elbow and straightening it whilst making the delivery, the sort of action that is characteristic of a toss in baseball. Throwing is illegal in cricket and is given as a 'no-ball' which means that the batsman cannot be called out, a run is added to the batting side's total, and the bowler has to bowl an extra ball. Hair continued to no-ball Murali seven times in three overs. Murali had by then already bowled three overs in this Test and had also gone through 22 Tests previously without incident, using the same action, and was therefore completely perplexed.

As the controversy continued, there was much discussion involving Sri Lankan captain Ranatunga and others in team management. After a subsequent no-ball ruling by Hair, it was decided to make Murali bowl from the other end of the pitch to avoid Hair's judgment, and he did so for another 12 overs that day without further no-balling. What was more puzzling was that Hair, being at the bowler's end, did not have the necessary square-on (90°) view of the bowling action, yet he assigned to himself the authority normally given to the square leg umpire who is in a much better position to judge the action. To add to the controversy, Murali bowled many more overs during the Test with Hair as the square leg umpire without any more protests from either umpire. This led to various allegations including that the Australians had been acting in a racist fashion against Murali. There was a sense that Murali was too good and,therefore, must have been cheating. The claims of racial discrimination were ironic since such claims would be expected to come from the majority Sinhalese dominated Sri Lanka Cricket Board and not from some other source. Murali would later address these events in his interview with Peter Roebuck:

PR: You had a great time in Melbourne yesterday, but in 1995 it was not so good. On Boxing Day you were called for throwing from the bowler's end. What was your reaction?

MM: I was shocked. Darrell Hair had umpired me so many times before. Before the match I had bowled 10 overs in Sydney in a one-day game. So I was very surprised when he said I was illegal next match.

PR: What was it like to be called in front of 55,000 people on the first day of a series?

MM: I was so upset. The team was behind me, and I was able to change ends, but that's not real cricket. He had made up his mind what he wanted to do. That should not happen to another bowler. It's very embarrassing. A single umpire cannot decide on the career of a bowler. If you are narrow-minded, then you will see it that way.

PR: Don Bradman said it was the worst umpiring decision he had seen, and that you were obviously not throwing. Not every Australian was on your case. How did you feel that night?

MM: It was terrible because I didn't know what to do or what was going to happen to my career.

PR: Alone among modern bowlers, you put your arm in a splint, went live on television and bowled all your variations. You went to England with Michael Slater and Mark Nicholas

in charge, both [sceptics] at the time, and bowled with your arm in the splint. Both changed their minds. What made you do that?

MM: Because I always thought I was not doing anything wrong - it's an illusion caused by my wrist and the way my joints and arm are built. To the naked eye it looks like throwing, but when you use technology, it shows I don't throw. I have gone through more tests than any other bowler since 1995 and passed them all. But it wanted to prove it. But still I was being booed in Australia, so a reporter gave me the idea and I thought it might end the talk.

PR: What material was used?

MM: Doctors said plaster of Paris can bend, so we put in steel rods. They weighed two pounds, which made it harder to bowl. But I bowled [the] same pace.

PR: Nicholas said there was no way a bowler could straighten his arm in that splint. It's a pity more Australians have not seen the footage. They say it's too expensive to buy. These things seem to crop up only in Australia. Why is that?

MM: Hard to say. Maybe the two umpires [were] premeditated. Maybe someone [was] behind it. I don't know.

In 1996, Murali's action was subjected to scientific biomechanical analysis by both the University of Western Australia in Perth and the Hong Kong University of Science & Technology, who concluded that his action was legitimate [20]. To be sure, Darrell Hair was no stranger to controversy both before and after the Murali incident. There was the 1992 Adelaide Test between Australia and India where Hair was umpiring, and which Australia won. *Wisden Cricketers' Almanack* said that the match was "marred ... by controversy on lbw (Leg Before Wicket) decisions - eight times Indians were given out, while all but two of their own appeals were rejected." Subsequently, Hair was involved in litigation with the ICC on other matters.

The Murali fiasco was one that, amongst others, acted as a catalyst for ICC to eventually bring in rules appointing umpires from non-participating countries for all international matches. It also led to the further development and refinement of biomechanical analysis of bowlers' actions and to the revision of the ICC rules governing 'throwing' where the permissible allowed elbow extension or straightening was changed from a range of 5° to 10° to the current 15° as it was found that 99% of bowlers examined exceeded the existing elbow flexion limits [17]. This incident brought Murali to international attention. The resolution concerning his arm action was critical as it established that he was a truly great bowler rather than the possibility that he was simply a cheater. Furthermore, his impact upon the game in terms of rule changes is something that can be said about few (if any) cricketers.

Murali went on to become the greatest Test bowler of all time, ending his Test cricket career in spectacular fashion by taking his 800th and last Test wicket with the last ball of his last over in his last Test match in 2010 [48]. There is no other bowler on the horizon

who appears likely to match this phenomenal record. Furthermore, Murali has amassed 534 wickets in ODI matches. He held the number one spot in the ICC's player rankings for Test bowlers for a record period of 1,711 days (roughly four and half years).

5.4 SANGA: PERSUING SPORT VERSUS FAMILIAL EXPECTATIONS

Unlike Murali, whose grandfather was an immigrant labourer brought to work in the tea fields by the British, Sanga was raised in Kandy by a privileged family with an aristocratic lineage coming from the era of the Kandyan Kings. As with many of the Kandyan privileged classes, all of whom were Sinhalese Buddhists, the children were, paradoxically, sent to study at the elite Anglican Christian school called Trinity College. Perhaps such parents saw the benefits of a Western liberal education that emphasized all round development and egalitarianism over privilege as a means to success in the wider world.

Soon after entering Trinity, Sanga was seen not only playing the violin, but he participated in most of the games that the school provided in the junior section (ages 5-12), i.e. badminton, tennis, table tennis, swimming, and cricket. A natural at sports, he won junior national colours for badminton and tennis. However, the principal of the school wanted him to concentrate and excel in one activity, and persuaded Sanga's mother that the boy should take up cricket. Thus began a career that was to take Sanga to the top of the cricket world.

As he grew up, Sanga played for his school's cricket teams at all age levels beginning with teams in the under-13 circuit. In schoolboy cricket, Sanga was ultimately awarded the "Trinity Lion" the most prestigious yearly sporting prize at Trinity College given to a member of the senior team. Sanga earned the award due to his remarkable batting and wicket-keeping performances in the 1996 school season.

Trinity College has always tried to foster the 'complete man' who is an all-rounder in academics, in sport, and in other extracurricular activities. Sanga was the definition of the complete man and was accordingly awarded the prestigious Ryde Gold Medal in 1996. Awarded each year to the "best all-round boy" at Trinity, the Ryde Gold Medal is the highest honour that the school can bestow, and the recipient is decided by a secret ballot conducted among the senior boys, the staff, and the school's principal. Historical records show that such a prize has been awarded as early as 1894 at Trinity. Every parent of a Trinity student wishes that his or her son will win this coveted award and therefore tries to foster a balance between academics and sport. Sanga's father, a well-known lawyer in their hometown of Kandy, was no different. As Sanga remarked during an interview on Cricinfo, "âĂę his father, Kshema, would throw him balls and instruct him on technique in the backyard of their beautiful hillside home in Kandy." The article goes on to explain that "Sanga had designs to follow his father into the legal profession before cricket called him properly, part-way through law school, at 22." Maybe the father, whilst promoting

Sanga's cricketing ambitions, had hoped that the son would follow him as a lawyer. In any event, once Sanga achieved the rank of captain on the Sri Lanka Test team, he had the full support of the father, who continued to be his most ardent and perhaps critical supporter. To quote from the interview again, Sanga's father remarks, "Actually, he has never reached my expectations" [sic]. "You see, now, for example, Don Bradman was one person whose every other match gave him a century .... According to Bradman, it is he who gets out - the bowler can't get him out. So Kumar must perfect first the art of not getting out, and the balance will work for itself." [18]

As mentioned earlier, upon leaving school, Sanga's first choice of career was to be a lawyer like his father, and he entered the Law Faculty of the University of Colombo. However, fate had other plans. While studying for his law degree, Sanga continued to play club cricket. Whereas Murali gravitated to the Tamil Union Club, Sanga played for the Nondescripts Cricket Club (NCC) in Colombo. As Sanga stated in a later interview, he really blossomed at 19 while playing at NCC and "seeing real competition ... in a very competitive surrounding, amongst better known players." [16]

In 2000 Sanga made a dramatic entry into international matches, playing for the Sri Lanka 'A' team and scored an impressive 156 runs against Zimbabwe in only 140 balls, which ensured his selection to the Test team. But it also foretold doom to his university studies when he was on the verge of completion. The demands of Test cricket and the rigorous levels of training and travel precluded any opportunities to attend classes or sit for exams. Nevertheless, years later, Sanga was cited as the inspiration to continue his academic studies by Bangladesh captain Mushfiqur Rahim who went on to receive a Masters degree. Perhaps if Sanga had envisioned beforehand the immense amount of time and effort, the endless hours of practice that was needed to make the national team, leave alone achieve success in the international arena, he might have had second thoughts about abandoning a university education for cricket. Fortunately for the game of cricket, Sanga persevered.

It is worth reflecting on the possibility that Sanga the cricketer could easily have been Sanga the lawyer. Giving up his university education was a risky decision at a young age, compounded by the fact that cricket is full of 'glorious uncertainties' or more likely, inglorious certainties. A lawyer with Sanga's oratorical skills can have a lucrative career and a comfortable life almost anywhere. A cricketer who does not make it to the big leagues and is unable to have sustained success over a long period of time can end up in penury. There are few, if any, active cricketers with university levels of formal education. In a country like Sri Lanka, the level of adulation by fans is such that one rarely sees a press article that is critical of the star players. Young people often overlook the fact that there is more to life than cricket, much to the detriment of their studies. In that respect, perhaps Sanga is not the best example for everyone to follow, for few reach Test level success, and tens of thousands fall by the wayside.

The MCC itself has realised this problem and has initiated a program called 'A degree in life, not just cricket' with the goal to prepare cricketers for 'life after the game.' As the journalist George Dobellso eloquently states on the subject:

Professional sport is a seductive beast. It sucks you in with whispered promises of glory and glamour and spits you out with broken dreams and an aching body. For every cricketing career that ends in a raised bat and warm ovation, there are a thousand that end on a physio's treatment table or in an uncomfortable meeting in a director of cricket's office. Many, many more stall well before that level.

And that's where the trouble starts. Young men trained for little other than sport can suddenly find themselves in a world for which they have little training and little preparation. Without status, salary or support, the world can seem an inhospitable place. It is relevant, surely, that the suicide rate of former cricketers is three times the national average.

By the age of 22, Sanga had graduated to Test cricket, debuting against South Africa in July 2000. The following year he achieved his first Test century. In his second Test match, he won his first 'Man of the Match' award, giving further rise to the notion that a new star was on the horizon in Sri Lanka.

Sanga became Captain of the Sri Lanka team for two years, from 2009 to 2011. He then decided that it was not for him, stating, "captaining Sri Lanka is a job that ages you very quickly ... It's rarely a job you will last long in ... I had a two-year stint, and I enjoyed it at times, certainly on the field where our results showed we were one of the top two sides in the world for one-and-a-half years, especially in the shorter form of the game." What was left unsaid, but broadly hinted was the debilitating effect the politics of the game in Sri Lanka was having on him and the other players. A suave diplomat as always, he did not directly attack any politicians but everyone got the message. But as Pathiravithana queried in the *Sunday Times*, "One of the greatest cricketing mysteries of the twenty-first century would be as to why Sanga, after leading his side to a Cricket World Cup final, turned his back on the crown a few hours later and decided to abdicate. Yes, we have listened to reasons touted, but none of them are plausible. The truth is yet to arrive, but we at this end do not want to hear it." [33]

In the midst of his cricket career, Sanga delivered the 2011 MCC 'Spirit of Cricket' Cowdrey Lecture at Lords and became the youngest person and the first current international player to deliver that lecture, which was widely praised by the cricketing community. The one hour long speech appeared in the front pages of almost all of Britain's mainstream newspapers, a first for any speech of this nature. At the Cowdrey Lecture, Sanga spoke reverently of his father's heroism during perhaps the most defining week of Sri Lanka's post-colonial history. When mobs scoured parts of the nation in late July 1983, hounding Tamils, killing them and burning down houses in retaliation for an LTTE (Tamil Tiger terrorist group)

ambush on troops in the north, Kshema and his wife, Kumari, (Sanga's parents) rallied around 35 Tamil neighbours and friends, providing refuge at Engeltine Cottage, (Sanga's family home) at great personal risk.

The journalist Peter Roebuck called Sanga's lecture "the most important speech in the history of cricket", because it grounded the game in Sri Lanka's history and implored administrators to safeguard cricket for its immense social value, if nothing else. Sri Lankan Tamils have not forgotten what brave Sinhalese men and women like Kshema and Kumari did in those dark times [37]. Their story certainly hasn't escaped the denizens of the once-embattled northern city of Jaffna, where Sanga is wildly, unreservedly popular: every kid's idol, every coaches' favourite exemplar. "What his father did for Tamils in 1983" is rarely far from northerners' lips when Sanga comes up in discussion.

Sanga's various records are too numerous to elaborate in detail, but a few are worth noting. Sanga has 38 Test centuries including eleven double centuries (i.e. 200 runs and above) and one triple century with 319 against Bangladesh in February 2014. To put this in perspective, there have been a total of only 28 triple centuries in Test cricket since the first one was recorded in 1930. Sanga's tally of 11 double centuries is second only to legendary Don Bradman's tally of twelve. He is also the first cricketer ever to score 150+ scores in four consecutive Test matches. With teammate Mahela Jayawardene, he holds the world record for the highest partnership in Test cricket - 624 runs against South Africa in 2006, where Sanga scored 287 runs and Mahela 374. As a wicketkeeper, Sanga has the 3rd highest number of dismissals in ODIs, 382 including 81 stumpings, the highest ever for a wicketkeeper in one-day international cricket. Moreover, Sanga was the fastest batsman in terms of innings to reach 8,000, 9,000, and 11,000 runs in Test cricket.

In 2012, Sanga had an unprecedented year with five notable distinctions. He was awarded the Sir Garfield Sobers Trophy for being the 'ICC Cricketer of the Year', the 'Test Cricketer of the Year' award and the 'People's Choice' award. He was also selected to the 'World Test' and the 'World One-day' cricket squads. There have been many more awards throughout Sanga's career.

Sanga retired from the T20 format soon after leading Sri Lanka to victory in the T20 World Cup in March 2014. He retired from ODI cricket after the 2015 Cricket World Cup where he established an all time tournament record of scoring four consecutive centuries. At the age 37, he announced his retirement from Test cricket after the second Test match with India in Colombo (August 20-24, 2015). It was a farewell to his international cricket career in front of his home crowd.

5.5 BEYOND CRICKET

We have made the case that through different and difficult paths, Murali and Sanga have risen to the pinnacles of world cricket. However, as icons and as ambassadors for Sri Lanka, their impact has been felt beyond cricket.

While both Murali and Sanga, whose faces adorn numerous advertising billboards around Colombo and other parts of Sri Lanka, have earned money by endorsing products, they also lend their names and make their presence felt for charities, having established their own charitable trusts. Murali partnered with his Sinhalese manager Kushil Gunasekara to establish the charity 'Foundation of Goodness'. With the support of cricketers and administrators from England and Australia, this charity raises funds to support local needs in Seenigama in the Sinhalese dominated south of the island. The foundation helps children, providing education and training, health care, livelihoods and sporting facilities.

When the Boxing Day tsunami of 2004 devastated many parts of Sri Lanka, including Seenigama, Murali mobilised resources to bring aid to the affected people. Due in Seenigama 20 minutes later for a prize distribution ceremony involving his charity, he narrowly escaped death himself when the tsunami hit. In the subsequent rebuilding efforts where cement was badly needed, Murali signed a barter deal with the Lafarge Cement Company to provide cement in return for his endorsements of their products. Subsequently, with the support of Bryan Adams, the Canadian pop-star, Murali raised funds to build a swimming pool in Seenigama where hundreds died needlessly because of their inability to swim.

With the ending of the war in 2009, Murali was also able to extend his charitable work to the Tamil dominated areas of the north, and is building a sports complex, IT and English training centres, and other facilities in Mankulam. In June 2004, Muralitharan was appointed by the United Nations World Food Program as an ambassador to fight hunger among school children.

Sanga has also been involved in charitable work, giving of his time and money. For some time, he has been an ambassador for the ICC's 'Think Wise' campaign, a partnership with UNAIDS and UNICEF that works to eliminate stigma and discrimination and promote HIV and Aids awareness. The Test playing nations are home to around a third of the world's population living with HIV [8].

In Sri Lanka, Sanga has also been helping to bridge the ethnic gap, often in joint efforts with Murali. Large crowds of Tamil people in the north crowded to meet Sanga and get his autograph when he attended the 2014 Murali Harmony Cup cricket tournament in Jaffna. The Murali Harmony Cup tournament is a reconciliation T20 tournament to promote community-building and friendship in post-war Sri Lanka and is organised by the Foundation of Goodness - the charity that now involves Murali, Sanga, their cricket team mate Mahela Jayawardene and other friends and colleagues. In this tournament, cricket is played between schools and clubs, both men's and women's, from the Sinhalese dominated

south and the Tamil dominated north. The matches are held in the north, which was deprived of sporting activities for decades during the war.

5.6 FINAL THOUGHTS

Murali and Sanga, two all-time greats in the world of cricket, were both born in Sri Lanka but came from two often antagonistic ethnic groups and from two very different social milieus. Playing cricket for the same team but with decidedly opposing skills and facing common adversaries over many years no doubt reinforced their mutual respect for each other and their joint efforts for national reconciliation.

Together, Murali and Sanga helped build Sri Lankan cricket from its infancy to a world cricketing power. For 20 years, they were the face of Sri Lankan cricket, and ambassadors for the game. Today, in the twilight of their careers, they are widely admired, and they represent the hopes and strengths of a country once divided by war that today appears to be on its way to harmony and greater prosperity.

In terms of how they will be remembered, it is worth noting that Sri Lanka has a written history exceeding 2000 years and that Sri Lankans love to evoke memories of their heroes. We believe that Murali and Sanga will be remembered as two of the greatest cricketers of all time. As Sanga remarked to England's *The Guardian* newspaper in 2011, "It is always going to be something like sport that brings people together ... cricket has been the heal-all of social evils, the one thing that held the country together during 30 years of war." Perhaps this is their greater legacy. Through their sheer excellence and accomplishments, Murali and Sanga provided a diversion for people during difficult times. Their role in uniting the country of Sri Lanka should not be underestimated.

Chapter 6

Conclusions

The subject of sports analytics, specifically related to the game of cricket was selected for my PhD thesis due to the rapid increase in the demand for sports analytics across the world, the value that it brings across the spectrum of sports, and, in regards to cricket, the potential demand for analytics fueled by the advent of lucrative T20 cricket.

Furthermore, while every major professional sports league in North America employs experts in sports analytics, cricket teams do not as yet have permanent analytics practitioners. Therefore, there is huge potential for the application of these methodologies in cricket. For example, when it comes to the bidding for star cricketers from around the world in the Indian Premier League (IPL), team owners' bids can be in the millions of dollars. Yet, the "selection" of players, and the bids made for them, are not based on detailed analytics as used in North America, but rather based on players' perceived reputations and a high level of "instinct" or "gut feel" on the part of the team owners, which is all very subjective. For the IPL 2015, 349 players were put up for "auction" of whom 67 players (43 Indians and 24 players from overseas) were "sold". These transactions took place in the heated frenzy of an auction where emotion, and the personal competitiveness and rivalry of team owners hold much sway. If the bidders had the benefit of analytics to guide their bidding, they might make more informed choices. This is where the recent research findings discussed in this thesis can help play a major role.

The first project in the thesis relates to the five-day test match format of cricket. It focuses on "declaration"'which is an important aspect of the game where patience and strategy play major roles. Declaration is always a gamble, taken with the intention of winning, but fraught with the danger of losing. Therefore analytics, which helps to mitigate the danger of the gamble, can play an important part in a team's winning strategy by providing team captains, coaches, and managers with a rational basis for their decisions.

Our second and third projects are particularly relevant to the fast paced T20 format. The second project provides a powerful analytical methodology for determining an objective optimal team line-up in terms of both batting and bowling. Our third project adds to this

by providing a basis for evaluating fielding which plays a crucial role in winning and losing matches.

It is our hope that the methodology presented in this thesis will encourage the cricketing world to adopt analytics as a major decision tool.

Bibliography

[1] *Wisden Cricketers' Almanack*. ESPNcricinfo-Online Archive, 1864–2014.

[2] E. Aarts and J. Korst. *Simulated Annealing and Boltzmann Machines*. New York: Wiley, 1989.

[3] M.J. Bailey and S.R. Clarke. Market inefficiencies in player head to head betting on the 2003 cricket world cup. *Economics, Management and Optimization in Sport*, pages 185–202, 2004.

[4] M.J. Bailey and S.R. Clarke. Predicting the match outcome in one day international cricket matches, while the match is in progress. *Journal of Science and Sports Medicine*, 5:480–487, 2006.

[5] G.D.I. Barr and B.S. Kantor. A criterion for comparing and selecting batsmen in limited overs cricket. *Journal of the Operational Research Society*, 55:1266–1274, 2004.

[6] W. Bretteny. *Integer optimization for the selection of a fantasy league cricket team. MSc Thesis, Nelson Mandela Metropolitan University*. 2010.

[7] J. Bring and M. Thuresson. Three points for a win in soccer: Is it fair? *Chance*, 24:47–53, 2011.

[8] A. Bull. Kumar sangakkara focuses on hiv charity work after test debacle. *The Guardian*, June 02, 2011.

[9] S.R. Clarke. Test statistics. *Statistics in Sport*, pages 83–101, 1998.

[10] J. Davis, H. Perera, and T.B. Swartz. Player evaluation in twenty20 cricket. *Journal of Sports Analytics*, 2015.

[11] J. Davis, H. Perera, and T.B. Swartz. A simulator for twenty20 cricket. *Australian & New Zealand Journal of Statistics*, 57(1):55–71, 2015.

[12] A.C. De Silva. Murali won observer schoolboy cricketer of the year title in 1991. *Sunday Observer*, March 16, 2008.

[13] K.M. De Silva. *A History of Sri Lanka*. Berkeley: University of California Press, 1981.

[14] P. L. De Silva. The growth of tamil paramilitary nationalisms: Sinhala chauvinism and tamil responses. *South Asia: Journal of South Asian Studies*, 20(1):97–118, 1997.

[15] D. Dyte. Constructing a plausible test cricket simulation using available real world data. *Mathematics and Computers in Sport. Queensland, Australia: Bond University*, pages 153–159, 1998.

[16] P. Epasinghe. Kumar sangakkara's long journey to world's leading batsman. *The Island*, September 29, 2010.

[17] R.E.D. Ferdinands and U.G. Kersting. An evaluation of biomechanical measures of bowling action legality in cricket. *Sports Biomechanics*, September, 2007.

[18] A.F. Fernando. My father, my critic. *ESPNcricinfo*, December 20, 2013.

[19] A. Franks, A. Miller, L. Bornn, and K. Goldsberry. Characterizing the spatial structure of defensive skill in professional basketball. *Annals of Applied Statistics*, 9(1):94–121, 2015.

[20] R. Goonetilleke. Biomechanical tests done on muttiah muralitharan at hong kong university of science and technology. January 28, 2008.

[21] R. B. Gramacy, S.T. Jensen, and M. Taddy. Estimating player contribution in hockey with regularized logistic regression. *Journal of Quantitative Analysis in Sports*, 9(1):97–111, 2013.

[22] S. Gray. *The mind of Bill James: How a complete outsider changed baseball*. Three Rivers Press, 2006.

[23] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction, Second Edition*. Springer: New York, 2009.

[24] D. Hopps. The speech that set free sri lanka cricket and glued a troubled nation. *The Guardian*, May 18, 2011.

[25] S.T. Jensen, K.E. Shirley, and A.J. Wyner. Bayesball: A bayesian hierarchical model for evaluating fielding in major league baseball. *The Annals of Applied Statistics*, 3(2):491–520, 2009.

[26] A.C. Kimber and A.R. Hansford. A statistical analysis of batting in cricket. *Journal of the Royal Statistical Society, Series A*, 156:443–455, 1993.

[27] C.D. Kirkpatrick, S.and Gelatt Jr and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[28] J. Knaus. *Easier cluster computing (based on snow), version 1.84-6*. 2013.

[29] H.H. Lemmer. Team selection after a short cricket series. *European Journal of Sport Science*, 13:200–206, 2013.

[30] M. Lewis. *Moneyball: The art of winning an unfair game.* WW Norton & Company, 2003.

[31] I.G. McHale, P.A. Scarf, and D.E. Folker. On the development of a soccer player performance rating system for the english premier league. *Interfaces*, 42(4):339–351, 2012.

[32] D. Oliver. *Basketball on paper: rules and tools for performance analysis.* Potomac Books, Inc., 2004.

[33] S.R. Pathiravithana. Kumar says test cricket is the pinnacle. *The Sunday Times*, September 8, 2011.

[34] H. Perera, J. Davis, and T.B. Swartz. Assessing the impact of fielding in twenty20 cricket. Under review.

[35] H. Perera, J. Davis, and T.B. Swartz. Optimal lineups in twenty20 cricket. *Journal of Computational Statistics and Data Analysis*, To appear.

[36] H. Perera, P. Gill, and T.B. Swartz. Declaration guidelines in test cricket. *Journal of Quantitative Analysis in Sports*, 10(1):15–26, 2014.

[37] P. Roebuck. Sangakkara's challenge to cricket. *ESPNcricinfo*, July 05, 2011.

[38] P. Roebuck. The kandy man. *The Age*, November 5, 2010.

[39] R. Ryder. Nothing is certain in cricket-except its uncertainty, 1974 - the glorious uncertainty. *Wisden Cricketers' Almanack, ESPNcricinfo.*

[40] H. Saikia, D. Bhattacharjee, and H.H. Lemmer. A double weighted tool to measure the fielding performance in cricket. *International Journal of Sports Science and Coaching*, 7(4):Article 6, 2012.

[41] P. Scarf and S. Akhtar. An analysis of strategy in the first three innings in test cricket: declaration and the follow-on. *Journal of the Operational Research Society*, 62:1931–1940, 2011.

[42] P. Scarf and X. Shi. Modelling match outcomes and decision support for setting a final innings target in test cricket. *IMA Journal of Management Mathematics*, 16:161–178, 2005.

[43] P. Scarf, X. Shi, and S. Akhtar. On the distribution of runs scored and batting strategy in test cricket. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2):471–497, 2011.

[44] M.E. Schuckers, D. F. Lock, C. Wells, C.J. Knickerbocker, and R.H. Lock. National hockey league skater ratings based upon all on-ice events: An adjusted minus/plus probability (ampp) approach. *Unpublished manuscript.*

[45] P. Swartz, M. Grosskopf, D.R. Bingham, and T.B. Swartz. Assessing the quality of pitches in major league baseball. 2015.

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

[47] L. Tierney, A.J. Rossini, N. Li, and H. Sevcikova. *snow: Simple Network of Workstations, version 0.3-13.* 2013.

[48] S. Veera. Murali gets 800, sri lanka win by ten wickets. *ESPNcricinfo*, July 22, 2010.

[49] K. Woolner. Understanding and measuring replacement level. *Baseball Prospectus 2002*, pages 55–66, 2002.

[50] K. Yousefi and T.B. Swartz. Advanced putting metrics in golf. *Journal of Quantitative Analysis in Sports*, 10(1):15–26, 2013.

Appendix A

Calculation of Upper Bound for the Cardinality of the Solution Space for the Optimization in Section 3.3

In equation (3.5), we provide a simple upper bound for the cardinality of the solution space for the optimization problem. For various reasons, there are some lineups which we exclude as possibilities from the solution space. We discuss this here and provide an improved upper bound for the cardinality of the solution space.

We expand on our earlier notation and let $M = M_1 + M_2 + M_3$ denote the number of players that are available in the team selection pool where M_1 is the number of wicketkeepers, M_2 is the number of non-wicketkeepers who are pure batsmen and M_3 is the number of non-wicketkeepers who are able to bowl. From these players, an optimal lineup of 11 players is chosen where $m_1 + m_2 + m_3 = 11$ using the obvious notation for m_1, m_2 and m_3.

Although there is no formal rule in cricket that prevents two wicketkeepers from playing at the same time, we assume that $m_1 = 1$. This assumption is in accordance with the way that cricket is played in practice. And since only one wicketkeeper is selected, it follows that this wicketkeeper does not bowl (for if he did, there would be no available wicketkeeper). Also, we recall that no player may bowl more than four overs (i.e. this implies that there must be at least five players who are able to bowl). Therefore, the selection of the 11 active players can be chosen in

$$\sum_{A} \binom{M_1}{1}\binom{M_2}{m_2}\binom{M_3}{m_3}$$

ways where $A = \{\ (m_2, m_3)\ :\ 1 + m_2 + m_3 = 11,\ m_2 \leq M_2,\ 5 \leq m_3 \leq M_3\ \}$.

For batting, there are 11! possible batting orders given the team selection. For bowling, given that m_3 bowlers have been selected, let i_j denote the number of overs bowled by bowler j. Then $i_1, \ldots, i_{m_3}$ are restricted according to the set $B = \{\ i_j \leq 4\ :\ i_1 + \cdots + i_{m_3} = 20\ \}$. Given $i_1, \ldots, i_{m_3}$, there are $\frac{20!}{i_1! \cdots i_{m_3}!}$ indistinguishable orderings of the bowlers. However this term is an upper bound for the number of bowling orders given $i_1, \ldots, i_{m_3}$ because it does not take into account the restriction that no bowler is allowed to bowl in consecutive overs. Unfortunately, we were unable to derive a combinatorial expression for the number of distinct orderings of $i_1 + \cdots + i_{m_3} = 20$ symbols where i_j symbols are of type j and no two symbols may be ordered consecutively.

Putting all three lineup components together, an improved upper bound for the cardinality of the solution space is given by

$$\sum_A \left[\binom{M_1}{1}\binom{M_2}{m_2}\binom{M_3}{m_3}\ (11!) \sum_B \frac{20!}{i_1! \cdots i_{m_3}!} \right] .$$

Appendix B

Construction of the Estimators (4.8) and (4.9) in Section 4.3.1

For clarity, we introduce some additional notation. Let N denote a notable fielding play for a given fielder, one that would appear in dataset B of the match commentary. Let A_k denote that the actual batting outcome is k and let S_j denote that the standard batting outcome is j. A standard batting outcome is the outcome that would have occurred without the impact of a notable fielding play. Then using the rules of conditional probability, the probability of outcome k due to the presence of the fielder is given by

$$\begin{aligned}
p_k^* &= \text{Prob}(A_k) \\
&= \Pr(N \cap A_k) + \Pr(\bar{N} \cap A_k) \\
&= \Pr(N)\Pr(A_k \mid N) + \Pr(\bar{N})\Pr(A_k \mid \bar{N}) \\
&= \Pr(N)\sum_{j=0}^{7} \Pr(A_k \cap S_j \mid N) + \Pr(\bar{N})\Pr(A_k \mid \bar{N}) \\
&= \Pr(N)\sum_{j=0}^{7} \Pr(A_k \mid S_j \cap N)\Pr(S_j \mid N) + \Pr(\bar{N})\Pr(A_k \mid \bar{N}) \\
&= \Pr(N)\sum_{j=0}^{7} \Pr(A_k \mid S_j \cap N)p_j + \Pr(\bar{N})p_k
\end{aligned}$$

where we recall that p_j is the probability of batting outcome j when a typical fielder is present. Referring to (4.5) and matching coefficients, the fielding characteristics are therefore given by

$$\lambda_{kk} = \Pr(\bar{N}) + \Pr(N)\Pr(A_k \mid S_k \cap N)$$

and

$$\lambda_{jk} \quad = \quad \Pr(N)\Pr(A_k \mid S_j \cap N)$$

for $j \neq k$. Finally, the estimators (4.8) and (4.9) are obtained by using the sample proportions $\hat{\Pr}(N) = m/n$ and $\hat{\Pr}(A_k \mid S_j N) = \sum_{i=1}^{m} I(O_i = k) q_{ij} / \sum_{i=1}^{m} q_{ij}$. We recall that n is the observed number of fielding opportunities by the fielder of interest. From these n opportunities, he made m notable fielding plays, O_i is the batting outcome associated with the ith notable fielding play and q_{ij} is the probability obtained from the random forest that the ith notable fielding play would have resulted in batting outcome j had the fielding play not been notable.

We note that although the point estimates for λ_{kk} and λ_{jk} are based on statistical theory, it is desirable to assign a standard error to these estimates. This is a problem which we hope to address in future research.