Roster-Based Optimisation for Limited Overs Cricket

by

Ankit K. Patel

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Master of Science
in Statistics and Operations Research.

Victoria University of Wellington
2016

Abstract

The objective of this research was to develop a roster-based optimisation system for limited overs cricket by deriving a meaningful, overall team rating using a combination of individual ratings from a playing eleven. The research hypothesis was that an adaptive rating system accounting for individual player abilities, outperforms systems that only consider macro variables such as home advantage, opposition strength and past team performances. The assessment of performance is observed through the prediction accuracy of future match outcomes. The expectation is that in elite sport, better teams are expected to win more often. To test the hypothesis, an adaptive rating system was developed. This framework was a combination of an optimisation system and an individual rating system. The adaptive rating system was selected due to its ability to update player and team ratings based on past performances.

A Binary Integer Programming model was the optimisation method of choice, while a modified product weighted measure (PWM) with an embedded exponentially weighted moving average (EWMA) functionality was the adopted individual rating system. The weights for this system were created using a combination of a Random Forest and Analytical Hierarchical Process. The model constraints were objectively obtained by identifying the player's role and performance outcomes a limited over cricket team must obtain in order to increase their chances of winning. Utilising a random forest technique, it was found that players with strong scoring consistency, scoring efficiency, runs restricting abilities and wicket-taking efficiency are preferred for limited over cricket due to the positive impact those performance metrics have on a team's chance of winning.

To define pertinent individual player ratings, performance metrics that significantly affect match outcomes were identified. Random Forests proved to be an effective means of optimal variable selection. The important performance metrics were derived in terms of contribution to winning, and were input into the modified PWM and EWMA method to generate a player rating.

The underlying framework of this system was validated by demonstrating an increase in the accuracy of predicted match outcomes compared to other established rating methods for cricket teams. Applying the Bradley-Terry method to the team ratings, generated through the adaptive system, we calculated the probability of $team_i$ beating $team_j$.

The adaptive rating system was applied to the Caribbean Premier League 2015 and the Cricket World Cup 2015, and the systems predictive accuracy was benchmarked against the New Zealand T.A.B (Totalisator Agency Board) and the CricHQ algorithm. The results revealed that the developed rating system outperformed the T.A.B by 9% and the commercial algorithm by 6% for the Cricket World Cup (2015), respectively, and outperformed the T.A.B and CricHQ algorithm by 25% and 12%, for the Caribbean Premier League (2015), respectively. These results demonstrate that cricket team ratings based on the aggregation of individual player ratings are superior to ratings based on summaries of team performances and match outcomes; validating the research hypothesis. The insights derived from this research also inform interested parties of the key attributes to win limited over cricket matches and can be used for team selection.

Acknowledgements

I would like to thank my supervisor, Dr. Paul Bracewell, for his patient guidance, encouragement and advice. I am indebted to him for many stimulating discussions about this work and inspiring my interest in the field of Sports Analytics.

I would like to thank DOT loves data for providing the funding for the uptake of this research and providing an excellent working environment.

Finally, I must express my gratitude to my parents for their continued support and encouragement.

Contents

Acknowledgements iii

1 Introduction to Sport Analytics 1

1.1 Overview of Sport Rating Systems 3

1.2 Analytics in Cricket 6

1.3 Formats of Cricket 7

1.4 Intent of Research 7

1.5 Structure of Thesis 8

2 Literature Review 9

2.1 Team Rating Systems for Non-Cricket Sports 9

2.2 Individual Rating Systems for Non-Cricket Sports 12

2.3 Team Rating Systems for Cricket 15

2.4 Individual Rating Systems for Cricket 19

2.5 Literature Review Findings 25

3 Research Objectives and Methodology **27**

3.1 Research Objectives 27

3.2 Research Methodology 29

3.3 Previous Research 31

3.4 Software and Hardware 32

4 Data Extraction and Processing **33**

4.1 Data Manipulation 34

4.2 Data Limitations 36

5 Exploratory Data Analysis and Regression Diagnostics **39**

5.1 Summary Statistics 39

5.2 Multicollinearity and Interrelationships 40

5.2.1 Variance Inflation Factors (VIF) 40

5.2.2 Scatterplots and Correlation matrix 41

5.3 Regression Assumptions 41

5.3.1 Independence of Errors 42

5.3.2 Normality of Residuals 42

5.3.3 Constant Variance 43

5.3.4 Residual Outliers 45

5.4 Chapter Remarks 45

6 Establishing Significant Performance Metrics **47**

6.1 Classical Parametric Techniques 48

6.1.1 Principal Component Analysis 48

6.1.2 Linear Discriminant Analysis 49

6.1.3 Stepwise Regression 51

6.1.4 Hierarchical Clustering Trees 52

6.2 Non-Parametric Techniques 54

6.2.1 Regression Trees 54

6.2.2 Random Forest 57

6.3 Dimension Reduction Application 58

6.3.1 Principal Component Analysis 58

6.3.2 Linear Discriminant Analysis 59

6.3.3 Stepwise Regression 61

6.3.4 Hierarchical Cluster Analysis 62

6.3.5 Regression Trees 64

6.3.6 Random Forest 65

6.4 Summary of Dimension Reduction results 67

6.5 Performance Metric Validation: Lorenz Curve and Linear Discriminant Analysis 68

6.6 Chapter remarks 69

7 Optimisation System **71**

7.1 Mathematical Formulation . . . 71
7.2 Determining the optimisation system . . . 72
7.3 Binary Integer Programming . . . 73
7.4 Branch and Bound Algorithm . . . 74
7.4.1 Binary Integer Programming Framework . . . 75
7.4.2 Model Constraints . . . 76
7.4.3 Establishing Model Constraints . . . 77
7.4.4 Optimisation Formulation . . . 80
7.5 Chapter Remarks . . . 81

8 Evaluating Individual Rating Systems 83

8.1 Optimal team rating using individual player ratings . . . 83
8.2 Establish the optimal team and optimal team rating . . . 84
8.2.1 Bradley-Terry Model . . . 85
8.3 Analytical Hierarchy Process . . . 86
8.3.1 TOPSIS . . . 89
8.3.2 COPRAS . . . 91
8.4 Product Weighted Measure . . . 93
8.4.1 Batsmen Ratings . . . 93
8.4.2 Bowler Ratings . . . 94
8.4.3 All-rounder Ratings . . . 95

8.4.4 Wicket-Keepers Ratings . . . 95

8.5 Principal Component Analysis . . . 95

8.6 Application of Individual Rating Systems . . . 96

8.6.1 Analytical Hierarchy Process . . . 96

8.6.2 Principal Component Analysis . . . 97

8.6.3 Product Weighted Measure . . . 98

8.6.4 Application Results . . . 100

8.7 Optimal team vs. Playing team . . . 100

8.8 Adaptive Rating System flaws . . . 101

8.8.1 PCA ranking flaws . . . 101

8.8.2 AHP-TOPSIS and AHP-COPRAS flaws . . . 101

8.8.3 Product Weighted Measure flaws . . . 101

8.9 Forecasting Methods . . . 102

8.9.1 Exponentially Weighted Moving Average . . . 103

8.10 Chapter Remarks . . . 103

9 Future Research, Discussion and Conclusion 105

9.1 Further Research . . . 105

9.2 Discussion . . . 106

9.3 Conclusion . . . 108

Appendices **111**

A Performance Metric Definitions **113**

B Scorecard Data Structure **114**

C Season Performance Metrics **115**

D Ball-by-Ball Data Structure **117**

E Team Level Dataset **118**

F AHP Pairwise Comparison Matrix **119**

G Player-Type Ratings by Team **120**

Chapter 1

Introduction to Sport Analytics

The growth of sport analytics and the need for meaningful sport related statistics has emerged in recent decades due to the large volume of monetary resources that is increasingly being invested in a single player or team. The rise in player salaries and salary caps over the last 25 years provide ample evidence of the growth of sport analytics, with investors, franchises, clubs and other stakeholders wanting to determine the true value of their investment. For example, in the National Football League (NFL) there has been an increase of approximately 950% in player salaries since the 1980's, and an increase of 288% in salary cap since 1994 [77]. With global sports revenue estimated to grow by US$145.3 billion over the 2010-2015 period [27] and winning teams earning significantly larger revenue than that of losing teams, there is a strong incentive for managers and coaching staff of sport teams to succeed. Given the large investment of resources and the stakes involved, coaches and managerial staff cannot solely rely on subjective views and personal beliefs to make team and player selection decisions. Solutions must be augmented with objective approaches by implementing analytical techniques.

Purpose of Research

The explosion in the sporting industry in terms of popularity and revenue is evident in cricket. Cricket has seen a huge global growth in revenue in recent years, and transformed into a sporting juggernaut due to the advent of T20 cricket. This is a relatively new short form format,

where teams each face up to 20 overs. A match typically concludes in three hours, which increases spectator appeal. The Economist reported that global cricket will generate total revenues of approximately $2.5 billion between 2014-2022 [2]. Moreover, the Indian Premier League (i.e. India's domestic T20 Competition) has a brand valuation of $3 Billion [39], while viewership of the Big Bash T20 league, Australia's domestic T20 competition, increased by 17% from 2015 to 2016 [1]. This rapid growth within the sport, and the accessibility to unique datasets and commercially-sensitive models motivated this research.

Formally, sport analytics is defined as "the management of structured historical data, the application of predictive analytic models that utilize such data, and the information systems used to inform decision makers and enable them to help their organizations in gaining a competitive advantage on the field of play" [7, p.1]. It is important to distinguish sport analytics from collecting quantitative data. *Quantitative data collection*, in sport, is the measurement and storage of the behaviours or actions of a team or a player, while *analytics* is the use of data to inform decision makers [11]. An early example of data collection within sports dates back to the 1850's with the publication of cricket averages in magazines. Although the collection and recording of numerical data within sports has been conducted for quite some time, the application of quantitative and statistical methods to this data is still in its infancy. Much of the early work on sports analytics revolved around sports that are popular in the United States, especially American Football and Baseball [11]. However it was not until the 1960's that the practice of sports analytics emerged, initially within baseball in the United States [11].

Given the myriad of numerical data generated by sports it is paramount that meaningful information is extracted from the data. The results generated from applying statistical techniques to sport related data are called sport statistics, which differs from sport analytics in the sense that sports statistics are the outcomes generated from the analytical techniques applied to the data. According to [13] sports statistics fall into two categories: 1. statistics that can be directly observed from a score sheet, known as performance indicators, and 2. statistics that are not directly observable from a score sheet, known as performance outputs. Sport statistics are utilised to make player selection decisions, develop training regimes and determine optimal strategies. There is a breath of academic literature applying various statistical techniques to myriad sports, for example discriminant analysis was utilised in [34] to identify performance

metrics that significantly distinguish between winning, losing and drawing team in the Europe Champions League. In [9] it was claimed that traditional win/loss and points scored ranking models applied to American Football fail to produce satisfactory rankings. The study therefore developed a hybrid paired comparison model which outperformed competitor models, producing robust results under model misspecification. Further, a modified least squares ranking procedure was developed in [44] to rank division 1 American men's college basketball teams using game outcomes. The results showed that the predictive accuracy of the modified least squares (76.3%) method outperformed that of the basic least squares (74.2%).

Due to the nature of human contest, sport lends itself to fluctuations and discrepancies in game outcomes, this in turn generates spectator interest. This outcome volatility is predominately due to variation in performance between individual players and teams. Therefore coaches, managers, fans, media and other interested parties utilise analytical approaches to understand the root of this variation, and handle and reduce its effect in order to produce 'better', more consistent results. Moreover these analytical techniques allow the user to rank and rate player and team performances. In general, sport rating systems provide an objective evaluation of a team or individual based on prior performances, and are implemented for player comparisons, improving the player/ team selection process and betting purposes. A *rank* refers to ordinal placement of ratings, while "*ratings* come from a continuous scale such that the relative strength of a team or individual is directly reflected in the value of its rating" [59, p.2].

1.1 Overview of Sport Rating Systems

Formally, a sports rating system "assigns each team a single numerical value to represent that team's strength relative to the rest of the league on some predetermined scale" [59, p.2]. Sport rating systems have had a long history dating back to the early 1930's. However, these systems were primarily implemented within popular sports in the United states. Despite the long history, sport rating systems have recently experienced a tremendous explosion; predominately due to the increasing volume of monetary resources injected into sports and the rising popularity of sports betting. Global sports revenue is estimated to grow by US$145.3 billion, over the 2010-2015 period, at an annual compound growth rate of 3.7% [62]. Moreover "the regulated sports

betting market is forecasted to reach $70 billion in 2016, representing a 20% increase from 2012" [41, p.5]. These claims were reinforced in [51, p.1] stating that "in the case of growing popularity of online sports betting, the analysis and forecasting of competitive sports has been receiving increasing interest".

Sport ratings are beneficial to numerous parties, especially athletes, coaches and managers who utilise such systems to track form, progress and use the ratings as a motivational tool. "These ratings are typically derived by suitably aggregating the competitors' previous performances and providing predictive power in forecasting tasks" [51, p.3]. Using a common framework, a survey of major world sports rating systems was presented in [74]. The study stated that sport rating systems have three steps:

1. Weigh the observed results to provide competition points - this is the most important factor in determining points for competitor i for a given competition.

2. Combine the competition points to produce seasonal value.

3. Aggregate the seasonal value to produce a rating.

Types of sport rating systems

According to [71] sports rating systems, in general, fall into two categories:

1. *Earned Ranking*

   Earned Ranking systems utilise past performances to provide a suitable method for selecting either a winner or a set of teams that should participate in a playoff [71]. The *earned* ratings are assigned an ordinal rank to produce team rankings. The majority of international sports adopt earned ranking system.

2. *Predictive Ranking*.

   Predictive Ranking systems utilise past performance to "provide the best prediction of the outcome of future games between two teams" [71, p.1].

It was stated in [72] that sport rating systems can be separated into three distinctive types depending on how new ratings are calculated for each rating system: *1. Adjustive systems 2. Accumulative systems and 3. subjective systems.*

Adjustive systems

Adjustive systems, also known as adaptive systems, "provide the best predictors for future performances because each adjustment follows from a predictor-corrector action in which a rating for team i can increase, decrease or stay the same, as each new result is compared to each prediction based on information available prior to the competition" [72, p.8]. Such systems cause ratings to fluctuate, depending on performances, and account for *Leapfrogging*[1]. Adaptive systems are adopted by sports such as golf, cricket, chess, football and rugby. According to [72] an adjustive system for competitor i has the following form:

$$r_i^n = r_i^{n-1} + K[w_i^n - P(r_i^{n-1}, r_j^{n-1}, W, O^{n-1})], \tag{1.1}$$

where r_i^n represents rating for competitor i after competition (i.e game) n, derived by adjusting the previous rating, r_i^{n-1}, for competition i, by a multiple K. The adjustment, K, depends on, w_i^n, which represents the difference between the actual performance of competitor i in competition n, (i.e. w_i^n), and the predicted performance $P(...)$ which is based on competitor $i's$ previous ratings. Competitor $i's$ and opponent $j's$ previous rating is affected by W and O^{n-1}, defined as weightings and other factors present in competition $n-1$, respectively.

Accumulative systems

Accumulative systems are *"running sums"* rating methods that are non-decreasing over a defined time-frame. These systems are predominately adopted by athletic sports such as gymnastics, power-lifting and cycling. According to [72] an accumulative system for

[1] A situation in which a player who can not participate in certain matches, due to injury, is exposed to being 'over-taken' by team mates who can play more games, and therefore have the opportunity to earn more points.

competitor i has the following form:

$$r_i^n = \sum_{k=1}^{n} f_i[w_i^k, W, A, O^k] \tag{1.2}$$

where r_i^n represents competitor $i's$ rating after competition n, based on past performances. "The function f_i for competitor i operates on w_i^k which is the performance of i in competition k, using W, which is a weighting procedure used to convert performances to points" [72, p.7]. The performance points are adjusted by an 'ageing' factor, A, and O^k represents *other* results in competition k, used to adjust $i's$ point score. The factors W and A are dependent on the sport to which the system is applied.

Subjective systems

Subjective systems consist of a panel of *experts* (i.e. judges) who rank the competitors and then combine the individual ratings to produce the overall ranking. Subjective systems are formally adopted by sports such as kick-boxing, mixed martial arts and boxing.

1.2 Analytics in Cricket

One sport which has recently seen an exponential rise in the use of statistics to make informed and strategic decisions regarding player and team performance is cricket. The very core of the sport is entwined with numerical values that translate ultimately to a match result. Cricket data has been recently explored using data mining and knowledge management tools with some success [66]. Data collection and data analysis has been conducted on cricket since the 1850's and 1960's, respectively. Given the rich sports data environment and its increase in popularity over the past decade, cricket has recently seen an increase in analytical literature and the adoption of predictive methodologies at the professional level. It was noted by [53, p.1] that "during the past decade a large number of papers have been published on cricket performance measures and prediction methods". Player performance has been analysed with the help of simple statistical measures, for example using augmented scatterplots it was found, in [49], that medium and slow (spin) bowlers, and fast bowlers tended to appear in different regions on the graph. The author illustrated that 'good' fast bowlers tended to have a low number of balls per wicket (i.e.

low strike rate) and a high number of runs per ball (i.e. high bowling average). While 'good' medium and slow pace bowlers (i.e. spinners) tended to have a low number of runs per wicket (i.e. low bowling average) and a high number of balls per wicket strike rate (i.e. high strike rate). Applying the method to the Indian Premier League (2008) bowling data the scatterplot enabled the author to rank various bowler-types.

1.3 Formats of Cricket

Cricket is a sport consisting of 11 players per team, the role of each player is either a batsmen, bowler, all-rounder or wicket keeper (i.e. keeper). International Cricket has three distinct formats 1. test matches, 2. one dayers and 3. Twenty-twenty (T20). The latter two formats are regarded as limited overs cricket due to restrictions imposed on the number of overs allotted to the batting and bowling side, and the number of overs an individual may bowl during an innings. In one day cricket each batting team is allotted 50 overs, while in T20 cricket each batting team is allotted 20 overs. Additionally restrictions are imposed on the number of fielders that may reside in particular areas of the cricket ground at any given time during an innings. Tests matches are regarded as the purist form of cricket with the longest format. Matches are typically scheduled for 5 days. Unlike limited overs format test matches do not limit the amount of overs allotted to each side, nor does the format impose fielding or bowling restrictions.

1.4 Intent of Research

The objective of this research is to develop a roster-based optimisation system for limited overs cricket by deriving a meaningful, overall team rating using a combination of individual ratings from a playing eleven. The research hypothesis is that a team rating system accounting for individual player abilities, outperforms systems that only consider macro variables such as home advantage, opposition strength and past team performances. The assessment of system performance is observed through the prediction accuracy of future match outcomes. This is based on the expectation that in elite sport, better teams are expected to win more often.

1.5 Structure of Thesis

Given the growth of online sports betting, and the analysis and forecasting of competitive sports, the following chapter discusses various sports ratings systems that were identified in the academic literature, derived using mathematical and statistical techniques, at both the individual and team level. The literature review will be followed by Research objectives and methodology which formally define the research questions and describes the adopted methodology. Subsequently, data extraction and processing procedures are described. Ensuing chapters are dedicated to data and statistical analysis. The final chapter discusses the model results and concludes with the optimal team rating system.

Chapter 2

Literature Review

This chapter provides a review of academic literature outlining the application of statistically derived rating systems for various sports , at both the individual and team level. The chapter has been partitioned into four segments: *1. Team rating systems for non-cricket sports, 2. Team rating systems for cricket, 3. Individual rating systems for cricket and 4. Individual rating systems for non-cricket sports.*

2.1 Team Rating Systems for Non-Cricket Sports

In [78] linear modelling techniques were applied to [American] college football data (2004-2006) to develop a predictive model for the outcome of 'bowl' football matches. Regressing on six predictors (scoring margin, offensive yards per game, defensive yards per game, strength of schedule, defensive touch-downs per game and turnover margin) the authors found all predictors to be practically and statistically significant for match outcome, with the model explaining 22% of variation. Team ratings were calculated by building a predictive model using previous season data and 'bowl' game outcomes. The amount of points a team received was based on the 95% confidence interval ($c.i.$) for the expected outcome for a single game. A team would receive 1 point if the $c.i.$ included 0, 2 points if the $c.i.$ included values > 0 and 0 points if the $c.i.$ included values < 0. A teams ratings would then be generated by aggregating these points across all games. Applying this method to the Bowl College Series competition a correlation

of 0.60 was found between the predicted and actual (end-of-season) ratings.

A common practice in American College Football is the use of computer models to produce team rankings, however these computational models often receive considerable criticism due to their tendency to heavily weigh margin of victory. To counter this weighing issue a penalised maximum likelihood approach was proposed in [60]. The result was a ranking process that attempted to reflect the opinion of human pollsters. The author began by assuming a normal distribution with mean θ_i and variance $\frac{1}{2}$, for the day-to-day variation in the intrinsic performance level of each team i. "Treating the performance level as random is consistent with the fact that even good teams can be 'upset' by weaker teams" [60, p.243]. The model assumed that a team's intrinsic performance level is independent of its opponent's level and that the team with the greater performance level, on the day, wins. Therefore given these assumptions the probability that team X defeats team Y was $\Phi(\theta_x - \theta_y)$, where parameter θ_i is the mean performance level, for team i. The developed likelihood approach reproduced the thought process of human pollsters, penalised undefeated team's by ensuring the MLE for an undefeated team was finite, producing rankings which agreed with human pollsters. Moreover the model accounted for all game outcomes in which teams were from different divisions since ignoring such events could lead to controversy if a teams only loss was to a team from a lower division. Applying the model to 1998 American College Football data and comparing the proposed model outcomes to computer-based outcomes, it was found that the penalised maximum likelihood approach outperformed two of the three [computer-based] models adopted by American College football.

In [36] a method of predicting the distribution of scores in international soccer matches was developed. The author treated each team's goals scored as independent Poisson variables dependent on the FIFA team ratings and the match venue. This was achieved by using a Poisson regression model based on two assumptions: 1. The number of goals scored by a team in a soccer match is Poisson distributed and 2. It is independent of the number of goals scored by the opposing team [36]. The Poisson regression implemented current FIFA ratings, opponent's FIFA rating and a parameter which changed according to venue (i.e. home, away and neutral) as predictors. The author calculated the expected number of goals scored per team, and using these as means the marginal probabilities for each team's Poisson distribution of goals scored was calculated. Using independence, the mean and marginal probabilities were multiplied to

produce the probability of each individual match result [36]. Using the latest FIFA ratings to calculate the expected number of goals estimated through the regression analysis, it was possible to generate two Poisson random variables for every game, and run a simulation for the entire tournament. After running the simulation, the author aggregated the probabilities for each of the World Cup matches to calculate the expected number of wins, draws and losses for each team. From output it was established that the raw FIFA ratings were slightly poorer predictors than the adjusted Poisson ratings generated through simulation, concluding that a Poisson assumption for goals scored was sufficient.

In [16] a generic rating system was developed, generating outputs known as 'Team Lodeings'. The output "measures the relative performance of sports teams and the competitive balance of competition" [16, p.4]. The lodeings framework enabled the authors to measure a teams performance relative to the opponents within the same division. This allowed for meaningful team comparisons. Applying the framework to the 2004 New Zealand National Provincial Rugby Championship revealed that the ratings engine produced suitable comparisons of team performance across divisions. The authors showed that the standard deviation of the ratings provided good representation of the competitiveness of a given sports league. Moreover, it was found that a competitive league results in teams having similar winning percentages, and therefore a smaller standard deviation. The method was externally validated by comparing the standard deviation of team ratings to the standard deviation of winning percentage, a strong positive correlation of 0.81 was found between the two variables. Applying the ratings engine to 23 domestic competitions across 7 different sports, it was found that soccer was the most competitive sport, followed by Basketball and American football, while Rugby was found to be the least competitive with 4 out of the bottom 5 least competitive leagues.

Implementing the lodeings algorithm developed in [16] and re-calibrating the results, a method to measure the relative performance of team's across divisions was developed in [47]. The author applied the methodology to the NBA, as it naturally splits into two divisions (i.e. Western and Eastern). Simulations were run on 3 groups. The first group contained matches played between the Eastern conference teams; the second contained matches played between Western conference teams; while the third group contained matches played between teams across conferences (i.e. interaction group). Applying the lodeings simulation to the first and second

group produced 'within conference' ratings, while simulations on group 3 produced 'between conference' ratings. Additionally 'overall NBA' ratings were produced by running the simulation when the groups were not defined. The author established strong links between team lodeings, winning percentage and final standings. Correlations > 0.80 existed between 'within' and 'between' conference lodeings' and final winning percentage, and 'overall NBA' lodeings. Next, a method to develop a recalibration equation was established by adopting Generalised Linear Models. This was established by regressing 'overall NBA' team lodeings (Y) on team lodeings across the three groups ($X_i's$). Using the regression equation to re-weight the 'within' conference and 'between' conference lodeings the author was able to re-calibrate the 'overall NBA' lodeings, allowing for meaningful comparisons of team performances across divisions.

2.2 Individual Rating Systems for Non-Cricket Sports

In [24] multiple linear regression was applied to rate tennis players using results from an Australian domestic doubles competition. Using indicator variables to tag the individual players the author fitted a regression model to 'games-up per set played' as a linear function of the two players involved and found the model to be statistically significant with an R^2 of 0.074. Next, percentage of games won by opposition and 'set weakness' were added to the regression model. The model produced was practically and statistically significant with an R^2 of 0.26. However given the large amount of unexplained variation in the model, an analysis considering the ability of individual opponents was conducted. Using separate player ratings a larger regression model was considered, incorporating a constant for home advantage. The home advantage coefficient of 0.51 was significant with a p-value of 0.026. The two sets [of ratings] had an almost perfect linear relationship suggesting that the method of calculating ratings using only the data available, to clubs, provide reasonable estimates of a players' relative ability. The author then assessed the difficulty of playing in certain positions (i.e. 1 or 2) by summing the ratings of the actual players who played in those positions. An exponential smoothing method was implemented to estimate a players rating at the end of the season. A correlation of 0.85 between the exponential smoothed ratings and regression ratings indicated that the smoothing method produced reasonable results. Additionally, given that the regression ratings were a single rating for the entire season's performance, and the smoothed ratings were an estimate of the

players ratings at the end of the season, indicated that the smoothing method was able to give reasonable ratings. "A comparison of the ratings of the pair of participating players gives the expected set-margin" [24, p.1389], "if the players do better than predicted, their ratings goes up; worse than predicted and their ratings goes down" [24, p.1389]. Therefore:

$$Predicted\,set\,margin = Player A\,rating + Player B\,rating - Opponent A\,rating - Opponent B\,rating \quad (2.1)$$

$$Updated\,player\,rating = previous\,player\,rating + \alpha(actual\,set\,margin - predicted\,set\,margin), \quad (2.2)$$

where α was a smoothing constant between 0 and 1, optimised such that the best fit to the predicted set of results was produced. Each refinement in the method showed an increase in the correlation of the [end of season] exponentially smoothed ratings and the least squares regression ratings [24], reinforcing the use of exponentially smoothed ratings to rank tennis players.

In [25] various heuristic ranking methods were assessed to produce player rankings in round robin tournaments. The aim was to find the optimal heuristic method to determine the most appropriate player ranking. The optimal method was evaluated on computational time and number of violations. Under the conventional tournament framework, each match is assumed to result in a decision. A tournament T is represented by an $n \times n$ matrix $A(T) = (a_{ij})$, where

$$a_{ij} = \begin{cases} 1, & \text{if player } i \text{ defeats player } j \\ 0, & \text{otherwise} \end{cases} \quad (2.3)$$

The study began by evaluating Kendall heuristics whereby players are ranked according to the number of opponents each defeats. However a disadvantage associated with this method is its

failure to break ties among players, meaning that a complete ranking can not be obtained. Next, the study evaluated the Iterated Kendall (IK) method which is an algorithmic extension of the Kendall scores. The method begins by computing the Kendall scores by finding the row sums of $A(T)$, and then ordering the players according to these scores. "If no ties are produced, the resulting ranking will produce no violations and the procedure terminates" [25, p.138]. If they exist, ties are broken by considering the sub-matrix of $A(T)$ containing only rows and columns of the k tied players,and perform the ranking as stated above [25]. Next, the study assessed the P-connectivity heuristic method which claims to remedy the unbreakable tie problem. The method "assumes that the 1-connectivity matrix is $A = A(T)$ and the p-connectivity matrix is p_{ij}, the $(p+1)$-connectivity matrix, $((p+1)_{ij})$, is derived from" [25, p.137]:

$$(p+1)_{ij} = \left(\sum_{k=1}^{n} p_{ik} + \sum_{k=1}^{n} p_{jk} \right) \times a_{ij}$$

"Each time a new matrix is derived, the row sums are computed. At any stage where there are no ties among these sums, the procedure terminates and the players are ranked according to these row sums. If any two row sums are tied, p is increased by 1 and the next p-connectivity matrix is computed" [25, p.137]. However the p-connectivity method produces a violation relative to the actual outcomes of the tournaments, whereby player i defeated player j, yet player j ranks above player i according to the heuristic. To overcome these violations incurred by the $p - connectivity$ method the Hamiltonian Path method was introduced. The Hamiltonian algorithm was applied to transform non-Hamiltonian rankings into a Hamiltonian path, which produces a ranking, R, such that if for any i ($i = 1, 2, ..., n$) the player ranked in the ith position has defeated the player ranked in the $i + 1$ position. However it was found that the Hamiltonian path algorithm only determines if two adjacent players can be switched [25]. Finally the Generalised Iterated Kendall (GIK) method was evaluated. The method applies a re-scoring procedure which takes the 'new' sub-tournament (only includes all tied players) and attempts to resolve the tie by to finding a player who defeated the last player put in the ranking. "If such a player exists, it is advantageous to put that player next, and then immediately change his place with the previously last ranked player" [25, p.140]. However if the tie is not resolved than an

attempt is made to rank the players such that there is a reduction in the number of violations. This ensures a reduction in the overall number of violations of 1. Each heuristic method was tested across 3 groups, $(n = 10, 40, 100)$, where n = number of players, and it was found that the GIK method was the optimal heuristic technique, producing the least amount of violations and lowest computational times for 2 out of the 3 groups ($n = 40$ and 100).

2.3 Team Rating Systems for Cricket

In [15] it was realised that the method in which limited overs cricket results (i.e. margin of victory) are recorded complicate the ability to generate meaningful team ratings, for example "if the team that batted first wins the margin of victory is expressed in terms of the differential of runs scored. However, if the team batting second wins, than the margin of victory is expressed as the number of remaining wickets for the second innings batting team" [15, p.1]. As there was no meaningful mapping function between these two forms of margin of victory, team ratings in cricket are based on win/loss records. To resolve this issue the study developed a method for creating meaningful performance based team ratings for cricket utilising a margin of victory that was solely runs based. This was achieved by developing a method for calculating the margin of victory for when the team batting second wins. The method estimated the number of runs that would have been scored had the team batting second continued until their resources were exhausted. The underlying philosophy was the same as that used by Duckworth & Lewis (1998). Using this framework a score projection was carried out if both resources had been exhausted using $T_2 = \frac{C_2}{R_2}$, where C_2 is team two's actual score and R_2 is the (Duckworth-Lewis) resources remaining. It was found that the score projections did not produce margins of victory that were significantly different from those produced when the team batting first wins. Logarithmically transformed score ratios (i.e $\frac{T_1}{T_1+T_2}$, where T_1 is the total score for team 1) were used in creating team ratings which were regressed against the winning percentages in order to deduce a linear transformation that would increase the spread of the ratings between 0 and 1. These score ratios were then input into the team lodeings algorithm developed to quantify relative performance. The performance of the ratings was validated by drawing comparisons with the ratings produced by the International Cricketing Council (ICC), a correlation of 0.91 indicated that the team ratings generated by the proposed performance based rating system was

valid.

In [23] a dynamic programming model was applied to one day cricket to calculate, at any stage of an innings, the optimal scoring rate, an estimate of the total number of runs to be scored in the first innings and an estimate of the probability of winning in the second innings. Given that the objective of the team batting first is to score as many runs as possible off their allotted resources (i.e. wickets and balls), at any stage of the game the aim is to maximise the expected score with the remaining resource. The author produced the following first innings formulation:

$$f_n(i) = \max_R [p_d * f_{n-1}(i-1) + \frac{R}{6} + (1-p_d) * f_{n-1}(i)],$$

where $f_n(i)$ represents the maximum expected score in the remaining n balls and i wickets in hand, p_d denotes the probability of dismissals[1], and $R = 6r$ denotes the runs rate per over. The author calculated the average number of balls faced per dismissal, and showed that "a team should try to score slightly faster than they expect their average rate for the rest of the innings to be, and if wickets are lost, slow up, rather than the current practice of scoring slower than average and speeding up if wickets are not lost" [23, p.333].

Since the team batting second knows the total scored by the first innings batting team, the aim for the second innings batting team is to maximise the probability of achieving a certain score, therefore the author introduced a variable, S, denoting the *number of runs* to go. "Each ball, a batsman either goes out with probability p_d and the team still has S runs to score with one less wicket in hand and one less ball, or scores X runs with probability p_x and so the team has $S-x$ runs to score with one less ball to go and the same number of wickets in hand" [23, p.334]. The author produced the following second innings formulation:

$$p_n(s,i) = \max_R p_d * p_{n-1}(s, i-1) + \sum_{0 \leq x \leq 6} p_x * p_{n-1}(S-x, i),$$

where $p_n(S,i)$ is the probability of scoring at least another S runs with n balls and i wickets remaining.

[1] Even though p_d depends on batting style, skill state of the bat, the bowler, run rate etc., all factors apart from run rate were ignored in the dynamic program.

The first-innings formulation allowed the author to develop an 'optimal scoring table' outlining a team's optimal scoring rate (i.e. runs per over) to obtain a given expected total, with i wickets in hand and n balls remaining. The second innings formulation allowed the author to develop a 'probability scoring table' outlining the probability of the second innings batting teams scoring, S runs with i wickets in hand and 300 balls remaining.

In [8] multiple linear regression techniques were applied to determine the relative batting, bowling strength and common home advantage, for each team, for the first innings of an international test match. A teams batting and bowling ratings were utilised to produce an overall team rating. The authors focused on the first innings as teams attempt to optimise their performance to establish a substantial first innings lead which provides the opportunity to control the match [8]. A teams first innings score in a test match played between the batting team i and the bowling team j on ground k, was modelled as:

$$S_{ijk} = A + \alpha_i - b_j + h_{ijk} + \epsilon_{ijk} \tag{2.4}$$

"The response variable S_{ijk} signifies a team's score, the intercept A represents the expected score between average teams on neutral ground, h_{ijk} represents home advantage and α_i and b_j signify the batting and bowling ratings of team i and j, respectively" [8, p.659]. S_{ijk} was logarithmically transformed and, applying an inverse transformation to the resultant parameters, produced estimates for the batting and bowling ratings for which the product equalled 1. A multiplicative combination of these individual ratings provided a measure of a team's overall strength. "A multinomial logistic regression model was adopted to determine which factors associated with a teams first innings performance significantly affected match outcomes" [8, p.664]. The predictors were common home advantage, result of the toss, the teams relative batting superiority over the oppositions bowling (i.e. $\frac{\alpha_i}{\beta_j}$) in the third innings and the team's relative batting superiority over the oppositions bowling in the fourth innings. The results illustrated three key findings: 1. the tendency of the home team to win, 2. the winning advantage that is gained by establishing a first-innings lead, and 3. a large advantage in batting second. There was also a clear indication that the probability of winning a test match is highly influenced by the relative difference in strength of the team in the fourth innings, but not the third innings. The overall strength of a team was treated as a combined effect of its first-innings batting and

bowling attributes. Analysing the effect of establishing a first-innings lead, it was found that the average lead that the home team needed to establish to have a better than even chance of winning was 157 runs. Conversely, if the team batting first was the away team it was found that the away team needed a lead of 292 runs to have a 50% chance of winning, confirming a clear home advantage.

A fair method for resetting the target for interrupted overs in one-day cricket was developed in [35]. The authors recognised that the batting side has two resources at their disposal from which to score runs: wickets in hand and overs remaining, and the number of runs that maybe scored from any position depends on the combination of these resources. The target score set for team 2, when an interruption has occurred during the second innings, is reflective of the relative resources remaining compared with team 1. A two-factor relation was established between the proportion of total runs which maybe scored and the two resources: $Z(u,w) = Z_0(w)[1 - \exp b(w)u]$[2], where $b(w)$ is the exponential decay constant and $Z_0(w)$ is the asymptotic average total score from the last 10 wickets. The average proportion of the runs still to be scored in an innings, with u overs bowled and w wickets down, was then calculated via: $P(u,w) = \frac{Z(u,w)}{Z(N,0)}$. This calculated the proportion of the combined scoring resources remaining in an innings, when u overs are left and w wickets down, enabling the authors to produce a table of proportions from which the correction to an interruption maybe made for any target score. Applying the method to hypothetical and real world examples it was found that the framework produced sensible and fair targets for all interruptions.

In [20] a Markov chain approach was developed to evaluate the expected performance of a cricket teams batting order in an innings. Realising that the interaction between bowler and batsmen is the main factor dictating the dynamics of run production, the author modelled the game as a sequence of one-on-one interactions. The authors created a multidimensional matrix, M, with entries (b, r, w, b_1, b_2), representing the number of balls, runs scored, wickets down, the striking and non-striking batsmen, respectively. It was claimed that, in general, only one of seven states can occur to which a given situation will commonly transition on a single ball bowled: (1) the probability a batsman is dismissed and zero runs are scored, p_d, (2) the probability that zero runs are scored, p_0, (3) the probability that one run is scored, p_1, (4) the

[2]Commercial confidentiality prevents the disclosure of the mathematical definitions of these functions.

probability that two runs are scored, p_2, (5) the probability that three runs are scored, p_3, (6) the probability that four runs are scored, p_4, (7) the probability that six runs are scored, p_6[3]. Given these states "it was possible to compute the probability of being in a given situation, for each number of balls bowled, by multiplying the matrix M, representing the set of probabilities after the $b-1$ balls, by the probability of each of the events" [20, p.497]. For example, at the beginning of a match the matrix, M, entries are $(0,0,0,1,2) = 1$. This represents the state in which there have been 0 balls bowled, 0 runs scored, 0 wickets down, batsman number 1 at the strikers-end and batsman number 2 at the non-strikers. All other entries of M $(0,r,w,b_1,b_2)$ are zero. "These values were obtained in the general case by multiplying each non-zero entry of M $(b-1,r,w,b_1,b_2)$ by each of the state probabilities and placing the result in the appropriate location in the matrix M" [20, p.497]. Simulation results in a 'runs distribution' table (i.e. estimates for p_0,..,p_d) representing the probability of any given number of runs having been scored. "Summing the product of each possible number of runs and its probability of being the result for the game gives the expected number of runs for the batting order considered" [20, p.497]. Applying this method to the 2005 Australian national team a 'runs distribution' table was generated for each player. The players were then ranked in terms of batting ability by determining a teams expected runs if the team was made up of only 1 player occupying all 11 positions in the batting line-up. Using these rankings the author computed the expected team total with the goal of determining the 'optimal' batting line-up (i.e. batting line-up producing the highest expected runs). The results indicated that the optimal batting line-up had a minimum and maximum expected number of runs of approximately 219 and 235, respectively. Moreover the best batting order produced at least 70 runs more than the worst batting order.

2.4 Individual Rating Systems for Cricket

In [50] a Poisson Hidden Markov Model (HMM), in conjunction with reliability analysis, was utilised to evaluate individual batting performances of one day cricketers. The number of runs scored by a batsmen, X_n, for game $n = 0,1,2,..$, followed a conditional Poisson distribution with a mean that depended on the underlying batting performance, Y_n. The model parameters were the means, and the transition matrix $\underline{\mathbf{P}} = P_{i,j}$, where $P_{i,j} = P(Y_n = i|Y_{n-1} = i)$, repre-

[3]Rare events such as fives and run-outs on run scoring balls were ignored.

sents the one-step transition probability from state i to j. A forward-backward EM algorithm was implemented to compute the data-likelihood. Implementing a uniform prior a Bayesian approach in combination with MCMC sampling enabled samples and inferences to be drawn from the posterior distribution. The appropriate number of performance states, k, for each player was identified by deducing the posterior distribution using a MCMC parallel sampling technique. Next, batting performance was evaluated via reliability analysis using the Poisson HMM and batting average as a performance measure. Each batsmen was treated as a type of system that was expected to perform adequately. Applying the method to the top 20 ODI batsmen (2014) it was found that the posterior estimates of the batting average provided a more meaningful [summary] measure of batting performance compared to the traditional batting average, as the HMM accommodated not-out scores, over-dispersion and serial dependence in the data.

In [17] a mixed distribution, called the Ducks 'n' Runs distribution was proposed. The distribution consisted of a beta distribution to model zero scores (i.e. ducks) and a geometric distribution to model non-zero scores (i.e. runs). It utilised runs scored and contribution to evaluate individual batsman in an innings. The suitability of the probability distribution was demonstrated, at a macro level (i.e. similar batsmen based on batting position) and micro level (i.e. individual batsmen), using data from New Zealand first class batsmen over a four year period (1994-1998). At the macro level scores were grouped into 20 bins $(0, 1-2, 3-6, 7-10, 11, 20, 21-30, ..., 151-200, > 200)$ and the observed proportion of scores were compared with expected probabilities. A Q-Q plot illustrated strong linear relationship between the observed and expected instances of scores indicating the the 'Duck n Runs' model was a good approximation for batting scores. At the micro level the probability distribution model for individual scores was fitted to all individuals and was used to calculate the proportion of Ducks, numbers of 50's and number of 100's an individual was expected to score. The results showed that all experiment-wise $p - values$ were less than 5% across all three measures, indicating that the 'Duck'n'Runs' distribution adequately models individual batting scores. Control charts based on quartiles of individual batting scores, were developed, to monitor an individuals batting performance. It was found that the control charts were able to detect significant changes in batting performance which suggested a change in an individuals *'form'*.

In [30] a Bayesian simulation and Stochastic Dominance, a technique used to analyse securities and portfolios, approach was applied to investigate the contribution of individual batsmen to overall team performance. Using a Bayesian approach, the author was able to replace the 'not-out' scores with a conditional average, representing an optimal estimate of the score the batsmen would have obtained had the 'not-out' innings been completed. "In every instance of 'not out', the batsman's score in that innings is replaced by the Bayesian estimate" [30, p.506]. The adjusted data was then analysed using Stochastic Dominance technique[4]. The utility function of a batsmen was characterised according to first-order Stochastic Dominance rules- "The first derivative of the utility function, with respect to runs scored, of an ODI batsmen was assumed to be positive" [30, p.503], indicating that more runs are preferred to less. The author then adjusted the individual batting averages due to the bias introduced by the conventional batting average formula. Graphically representing the cumulative probability of individual batting performances for 5 cricketers (4 batsmen and a bowler) revealed that the specialist batsmen curves dominated the curves for the specialist bowlers, indicating that batsmen have a higher probability of scoring a particular number of runs than bowlers.

In [37] time series clustering analysis was used to map the test career progression of Australian cricketing legend Sir Don Bradman, acknowledged as the greatest Batsman of all time with an unparalleled career batting average of 99.94, from 80 innings. However part of his career was interrupted when all international cricket was suspended due to World War II. Given this 'disruption' in his test career the authors utilised time series clustering to characterise Bradman's test career and compared him to other 'great' batsmen to test whether or not Bradman was denied his prime. The selected clustering method was based on global characteristics measures "as it does not require many conditions to be true before it can be utilised, relative to other clustering techniques" [37, p.3]. Additionally the approach clusters global features extracted from individual time series and can be applied on different length time series. The performance measure used to compare batsman was average 'contribution' per innings. A [scaled] average contribution was then modelled using weighted least squares regression. This smoothed standardised data was then fitted to a polynomial function, for each batsman, and the parameters of the model were used to generate meaningful clusters. The results showed that Bradman's ca-

[4]The problem of portfolio choice is that of selecting a portfolio that maximizes the utility of the investor.

reer progression was most similar to West Indian legend Brian Lara, indicating that Bradman's peak performance would have occurred in the 12th to 14th years of his career (1939-1941), coinciding with World War II. Imputing Bradman's likely performances (i.e. batting average) for 1939-1945 the authors estimated his batting average to be 105.41, which was significantly higher at the 5% significance level than Bradman's actual [career] average of 99.94. The authors concluded that Bradman was indeed denied his prime.

In [5] a multinomial logistic regression model was fitted to session by session test match data to calculate match outcome probabilities. These probabilities were used to measure the overall contribution of each player to match outcome based on their individual contribution during each session. The model assumed a multinomial distribution: $Y \sim MN(p_1, p_0, p_{-1}, \sum p_i = 1)$ where p_1, p_0 and p_{-1} represent the probability of a win, draw and a loss, respectively. The fitted predictors were lead, ground effect and total wickets lost for each team (W_1 and W_2). Using multinomial regression models the authors were able to predict match outcome probabilities given the match position at the end of each session t, ($t = 1, 2, 3, ..., 15$). Next a hypothetical position at the end of session t was defined, in which the batsmen had scored no runs, and match outcome probabilities were generated. Additionally, a hypothetical position at the end-of session t was defined, in which *bowlers* had not taken any wickets, and match outcome probabilities were generated. A players overall contribution during a given session was assessed by using the difference between the hypothetical match outcome probabilities and the actual match probabilities. The batting probability differences were observed with respect to 'not losing' and bowlers with respect to winning [5]. "These probability differences were then distributed to batsman according to their share of the runs scored in the session, and to bowlers according to their share of wickets taken in the session" [5, p.687]. An individual$i's$ batting contribution in session t was evaluated via:

$$C_{i,t,bat} = C_{t,bat} \times \frac{r_{i,t}}{r_t},$$

where $r_{i,t}$ is the runs scored by player i in session t and r_t is the total runs scored by his team in session t. An individual, i, bowling contribution in session t was evaluated via:

$$C_{i,t,bowl} = C_{t,bowl} \times \frac{\sum_{j=1}^{n} Z_{itj}\alpha_j}{Z_t},$$

where Z_{itj} represents the total number of wicket taken by player i during session t for wicket-taking contribution j, $j = \{1, 2, 3\}$, where $j = 1$ corresponds to a wicket taken by the bowler with no fielder involvement, $j = 2$ corresponds to catches taken by a fielder and $j = 3$ corresponds to run-outs. The α_j represents the share of points for a wicket awarded to the fielder. The net contribution of player i in the match is then the sum of contributions from all sessions. However it wad found that the contributions rating system took little account of contribution after a point when the win or draw probability of any team is close to unity. To overcome this problem the author used the contributions as one component of a weighted average rating system, while the other was raw runs and wickets in the match. Points gained were placed on a 'runs-like' scale by multiplying the net player contribution by the average runs per match based on test matches from 1877-2007. Team ratings for each nation were calculated by combining the individual player ratings and the final summed value represented the nation's overall team rating.

In [54] the limitations of conventional batting and bowling performance measures was recognised. It was claimed that the Duckworth-Lewis methodology could be used to evaluate player contributions for any stage of an innings, and performance metrics producing context based measures were developed. "At any stage of an innings, the worth of a player's contribution, per ball can be evaluated using equation $Z(u, w) = Z_o F(w)(1 - \exp[-bu/F(w)])$" [54, p.806]. This function is interpreted as the proportion of runs accumulated with w wickets lost relative to no wickets lost and, hypothetically, infinitely many overs remaining. $F(w)$ is a positive decreasing step function with $F(0) = 1$. $Z(u, w)$ represents the average further runs obtained in the u remaining overs when w wickets have been lost, and Z_0 and b are positive constants. For example if there are i balls remaining and w wickets have been lost then the expected runs, r_i, from ball i will be either $r_i = Z(i, w) - Z(i - 1, w)$ or $r_i = Z(i, w) - Z(i - 1, w + 1)$ [54], depending on whether the batsmen survives the next ball. If the batsmen scores S_i runs from ball i the batsmen's net contribution, c_i for ball i is either $c_i = S_i - [Z(i, w) - Z(i - 1, w)]$ or $c_i = s_i - [Z(i, w) - Z(i - 1, w + 1)]$. The author then calculated the proportion of resources left with u overs left and w wickets down, depending on whether the batsmen survives the ith ball. The proportion resources consumed on the i_{th} ball is either $p_i = P(i, w) - P(i - 1, w)$

or $p_i = P(i, w) - P(i-1, w+1)$[5]. Next, the batsmen's average run contribution per unit of resources consumed to the team's total was assessed by $\frac{\sum S_i}{p_i}$, while a bowlers average runs contribution per unit resource consumed was measured by $\frac{\sum(S_i+h_i)}{\sum p_i}$, where h_i represents the number of extras conceded by the bowler from ball i. Applying these measures to the 2003 VB series final (Australia vs. England) it was shown that the Duckworth & Lewis based contribution measures were less susceptible to distortions compared to traditional measures.

In [61] a regression tree technique was applied to New Zealand youth test match data (1986-2008) to identify fast bowlers likely to play test cricket, based on New Zealand age-group performances. A regression tree was implemented as a predictive model to account for the multi-collinearity and complex interactions among the performance metrics. The model found balls bowled and strike rate to be practically and statistically significant predictors for a international test career. Results revealed that the regression tree correctly classified 80% of the fast bowlers who went onto represent New Zealand at the test level. Additionally, a Lorenz curve based on the significant metrics showed that within the top 25% of fast bowlers approximately 75% had played international test cricket, illustrating adequate discrimination between successful and unsuccessful [fast] bowlers. A residual logistic regression technique was adopted to rank the bowlers in terms of their probability of success (i.e. playing international test cricket). Applying this technique to New Zealand youth cricket performances (1986-2008) the residual regression tree model correctly ranked and classified 93% of the fast bowlers involved in the study.

[5] $P(i, w)$ is defined as the proportion of resources consumed from the ith ball with w wickets left.

2.5 Literature Review Findings

Through the literature review process the author identified a scarcity in literature surrounding team rating systems, utilising individual ability. This lack of academic depth revealed an inadequacy in understanding, lack of demand and a literature gap. Given the gap in the literature the author established an entry point in the market for this research and attempts to address the literature gap. The primary focus is to develop a novel method to generate the optimal team using individual player ability, while the secondary focus is to identify a method that accurately measures a teams ability to win, given individual player abilities. Given these objectives the research centred on the development of an adaptive-predictive rating system, characterised by utilising past player performances, and accounting for the long and short term variability of a team's performance.

An adaptive method was preferred as the ratings produced by such systems are recalculated whenever new results are obtained. Specifically, adaptive systems update player and team ratings "based on historic performances upon availability of data about current performances" [51, p.3] and can be tailored to incorporate the distinctive features of cricket (i.e. batsmen, bowlers, etc.). Given these findings, the following chapter formally defines the research objectives and methodology adopted to develop an adaptive-predictive rating system. Moreover chapter 3 distinguishes the academic contribution of this research from existing work and attempts to address the scarcity in the literature surrounding team rating systems, utilising individual ability.

Chapter 3

Research Objectives and Methodology

The literature review revealed extensive published research surrounding team and individual rating systems, across various sporting disciplines. However the scarcity of literature surrounding team rating systems, based on individual ability, reflects a historical lack of access to data and computing resources. This has resulted in a gap in the literature. Moreover given the lack of literature applying modelling techniques to predict match outcomes for limited overs cricket, the growing popularity of sports betting within the sport highlights the potential demand for this research. Given this gap in the literature the following research objectives were established:

3.1 Research Objectives

The primary objective of this research was to develop a roster-based optimisation system (i.e. adaptive rating system) for limited overs cricket, using individual player ratings. The goal was to build an adaptive rating system that selects a cricket team (i.e. n = 11 players), based on a set of criteria, from a playing squad (i.e. $n > 15$), such that the optimal team produces the greatest team rating. For example if team A has a 15 'man' squad the optimisation system should select a cricket team which optimises the team's overall rating, using individual ratings of the selected players, across a set of key roles and responsibilities. Consequently, the *optimal team* was defined as *the set of 11 individual players that produce the greatest probability of winning for team i against any given opponent j.* An adaptive rating method was the system

of choice because it updates player and team ratings based on historic performances. Ratings fluctuate according to performance. Additionally it was established that adaptive systems were favoured by object sports, such as rugby, cricket, soccer etc. [73].

The secondary research objective was to ensure that the developed rating system accurately predicted match outcomes (i.e. a system with high predictive power) and could outperform the predictive power of well-established and recognised predictive sporting algorithms. This serves as a validation of the individual player rating system. However applying the adaptive rating system across the two competitions (i.e. CPL and CWC2015), the author encountered a problem: on occasion the 'optimal' team generated by the optimisation model would differ from that selected by coaches and managers; meaning the 'optimal' team rating would not relate to the playing team. To counter this issue, rather than using the optimal team rating, the author simply selected the player ratings of those chosen by coaches, and aggregated the ratings to generate a team rating. Even though this did not represent the *optimal* team rating, it did provide a quantitative indication of the strength for the playing team, and demonstrates the value of the combining individual metrics for a team rating.

The author hypothesised that a team-based [adaptive] rating system, accounting for individual player performances, should outperform rating systems that only consider 'macro' variables, such as opposition, venue, past [team] performances, home advantage etc. As previously mentioned, no research discussing the development of a team rating measure, utilising individual player ratings [within cricket], was identified during the literature review process, persuading the author to undertake this research.

Research Milestones

Before adopting an optimisation model four key tasks required completion:

1. Identify the batting and bowling metrics that significantly contribute towards a team's ability to win (i.e. winningness).

2. Identify an individual rating system that accurately derives a player's rating, as a function of significant performance metrics.

3. Identify a method to calculate a team's overall rating, as a function of individual player ratings.

4. Identify a method that calculates the probability of team i beating team j utilising the rating of both teams.

3.2 Research Methodology

Given the research objectives the following research methodology was applied:

1. Since the primary research objective was to develop an adaptive rating system that produces the 'optimal' cricket team using individual ratings, and given the definition of 'optimal' - *the set of 11 individual players that produce the greatest probability of winning for team i against any given opponent j*- the author was required to identify individual performance metrics that significantly impact a team's ability to win (i.e. percentage wins, Y) a limited overs cricket match. Additionally, the secondary research objective required the developed system to accurately predict match outcomes (i.e. win or loss) to validate the primary research objective. This meant significant performance metrics in terms of percentage wins, also referred to as 'winningness', had to be identified. The research requirements solidified the use of *winningness* as the dependent variable to identify the significant performance metrics. The fundamental philosophy underpinning this approach is the expectation that (a) better teams are composed of better players and (b) better teams tend to win more often.

2. Evaluate different individual rating methods that utilise performance metrics to derive player ratings. The 'optimal' player rating method will produce the greatest predictive power (i.e. produces the largest proportion of correct match outcomes) when filtering the individual ratings through the adaptive system to generate a team rating measure. Three individual ratings methods were evaluated: (1) Principal Component Analysis (2) Analytical Hierarchy Process and (3) Product Weighted Measure.

    - The product weighted measure ranking (PWM) system required power coefficients, i.e. weights, to be assigned to each significant performance metric when calcu-

lating individual player ratings. Additionally given different metrics have varying effects, for each *player-type* on winningness, a method to establish appropriate metric weights was identified. Identifying an approach to accurately calculate these weightings was critical to the implementation of the PWM ranking system.

- The author introduced a novel method combining the Analytical Hierarchy Process (AHP) and Random Forest technique to calculate these weights. The approach combines prior expert knowledge, gathered from the AHP, with objective inferences drawn from the Random Forest technique (chapter 8, section 8.6.3).

3. Identify and modify an [existing] optimisation system to select the 'optimal' cricket team (i.e. 11 players), defined as: *the set of 11 individual players that produces the greatest probability of winning for team i against any given opponent j.*

4. The optimal team rating was calculated by aggregating individual player ratings. This aggregation approach was justified in [30], the paper stated that cricket is a sport characterised by one-on-one interactions between batsmen and bowlers, and individual player abilities establish the outcome of this interaction. Furthermore match outcomes are defined by the sum of interactions between batsmen and bowler. Therefore summing the individual player ratings, for a given team, provides a fair indication of team strength.

5. The probability of team i beating team j was derived through pairwise comparisons. Since the individual ratings and team ratings were measured on a ratio scale the Bradley and Terry model for comparing winning probabilities from ratings was implemented:

$$\pi_{i,j} = \frac{Rating_i}{Rating_i + Rating_j}$$

6. The predictive accuracy of the adopted *optimisation model + selected individual rating system* (i.e. adaptive system) was benchmarked against the T.A.B[1] and CricHQ's[2] predictive system.

[1]Totalisator Agency Board in New Zealand.
[2]A cricket technology industry pioneer with headquarters in Wellington, New Zealand.

3.3 Previous Research

The research adopted a Binary Integer Programming [optimisation] model, however the author identified previous research in which such a system had been applied for team selection within cricket ([42], [68]). However the research methodology outlined in [42] and [68] suffered many issues. The following research weaknesses were identified:

1. Ad-hoc metric selection

The performance metrics utilised to establish individual player ratings were subjectively chosen with no justification.

2. Unsuited for all-rounders

Equal weights were allocated to an all-rounders batting and bowling ability when deriving player ratings. This leads to inaccurate player ratings because even though all-rounders are well-rehearsed in both batting and bowling, they still possess a dominant skill and therefore should be classified as either batting or bowling all-rounders, and their abilities should be weighted accordingly. Additionally the framework did not consider situations in which an all-rounder only contributed through either batting or bowling, but not both. In this case the method failed to produce an all-rounders rating as the individual rating equation required an all-rounder to bat *and* bowl during a match.

3. Ad-hoc method of developing optimisation model constraints

The model constraints were formulated in an ad-hoc fashion, leading to inaccurate optimal teams generated by the model. For example it is common [cricketing] knowledge that T20 cricket is a batsmen dominated game. Therefore when constructing an optimal T20 team the model constraints should be formulated such that the optimisation method produces a team containing greater batting talent than bowling talent.

4. Lack of team rating measure

Given *optimal* team A and *optimal* team B, as suggested by the model, the research provided no method of comparing the strength of the two teams. For example given optimal team A vs. optimal team B, what is the probability that team A beats team B? who is stronger?

5. Lack of validating the optimal team

The research provided no method of validating whether or not the team produced by the optimisation model was 'optimal'. Furthermore, operationalised concept, 'optimal' was not defined.

6. Performance metrics were subjectively allocated weights

The authors implemented a *product weighted measure* ranking system to derive individual player ratings, however the weights (i.e. power coefficients) allocated to each performance metric, for each player-type (i.e. batsmen, bowlers, all-rounders and keepers), were 'subjectively' chosen. The performance metrics were allocated equal weights, producing inappropriate player ratings because different performance metrics have varying effects on individual player-types, across formats.

7. Lacking of testing different individual rating systems

Individual player ratings were derived using the product weighted measure, however the variability of the 'optimal' team across various individual rating methods was not examined. Moreover, the reasoning for the product weighted measure being identified as the 'optimal' individual player rating method was not described.

3.4 Software and Hardware

Analyses and statistical programming were executed using the SAS language and R (Rgui 64-bit v3.0.2; R Core Team, 2015). R is an S-PLUS statistical programming environment for statistical computing and graphics. The choice of software was determined by the extensibility for modelling packages and the need for flexible object-oriented data manipulation. By using R, which is free, open-source and readily available over the Internet, all procedures carried out can be reviewed and replicated. Formatted tables and figures were generated through R using LaTeX markup language, MiKTeX typesetting system and Pandoc file converter. All research was carried out on a desktop computer equipped with dual Xeon quad core CPU 2.4GHz, 32GB RAM, running 64-bit Windows 10.

Chapter 4

Data Extraction and Processing

The analysis conducted throughout this research required end-of-match scorecard data for limited overs cricket matches. Scorecard data outlines each players batting and bowling performance statistics in the first and second innings of a limited overs cricket match. This data is readily available from the ESPN Cricinfo website (www.espncricinfo.com)[1]. An automated process using the SAS language was developed to extract and parse the scorecard data, and provide a more convenient data structure. The process extracted relevant details on a match-by-match basis and stored the data in a tabular form for easy access; appendix B illustrates data structure after the scorecards were extracted. Since this research focused on limited overs cricket both T20 and one day data was required.

1. T20 scorecards were extracted for each match from the 2015 season of the Indian Premier League (IPL) and Caribbean Premier League (CPL), i.e. *two major domestic T20 competitions.*

2. One day scorecards were extracted for each match from the 2011 and 2015 Cricket World Cup (CWC) competition, i.e. *one day international competitions.*

The IPL and CWC2011 datasets were implemented during the analysis phase (i.e. training sets). The training sets were utilised to identify the performance metrics that significantly

[1]This data was obtained with permission from ESPNCricinfo.com.

effect winningness (Chapter 6). The CPL and CWC2015 scorecards were utilised to validate the reliability and predictive power of the developed *adaptive rating system* (i.e. test set). Table 4.1 illustrates the contents of a cricket scorecard.

Table 4.1: Scorecard elements

Player Info	**Game info**	**Batting metrics**	**Bowling metrics**
Player Name	Cricinfo ID	Dismissal	Overs
Player ID	Innings	Runs Scored	Maidens
Role		Minutes played	Runs Conceded
Order		Balls Faced	Wickets
		Fours Hit	Economy Rate
		Sixes Hit	Boundary 4's
		Strike Rate	Boundary 6's
			Extras
			Dots

4.1 Data Manipulation

The IPL dataset contained scorecards from 60 games (1591 player observations), while the CWC2011 contained scorecards from 49 games (1475 player observations). The following steps were applied to the two scorecard datasets.

After extraction the IPL and CWC2011 scorecards were split into two separate sets:

Dataset 1: Batting metrics

Contained match-by-match player observations with their associated batting metrics and biographic information (i.e. player name, role etc.), for each match. This dataset contained all player observations where *role* = batsman, which is coded in the data as 1.

Dataset 2: Bowling metrics

Contained match-by-match player observations with their associated bowling metrics and biographic information (i.e. player name, role etc.), for each match. This dataset contained all player observations where *role* = bowler, which is coded in data as 2.

Since each player in the two datasets [for both competitions] contained multiple match observations, each performance metric was aggregated and averaged across the entire season by player ID. The output produced season performance statistics, for each player, in the IPL and CWC2011 competitions (Appendix C). Table 4.2 outlines the performance metrics that were calculated[2].

Table 4.2: Performance Metrics

Batting metrics	**Bowling metrics**
Batting Average	Economy Rate
Batting Strike Rate	Strike Rate
Average Contribution	Bowling Average
Percentage Boundaries hit	Percentage Boundaries conceded
Runs Scored	Dot Balls
Balls Faced	Balls Bowled
Total Boundaries	Percentage Dots
Sixes	Runs Conceded
Fours	Wickets
Games Played	Games Played
Number of wins	Fours Conceded
Percentage wins (Y)	Percentage wins (Y)
	Sixes Conceded
	Number of wins
	Total Boundaries
	Total Maidens

Next, a *player-type* (i.e. batsmen, bowler, batting all-rounder, bowling all-rounder or wicket keeper) was assigned to each player. Additionally, each player was tagged to a team. A players *'player-type'* was established by:

1. The position (i.e. *order*) in the batting or bowling line-up a player, on average, occupied. For example 'pure' batsmen, those who specialise in batting, generally bat in the top order of a batting line-up (i.e. *order* = 1-4), while 'pure' bowlers, those who specialise in bowling, generally bowl during the early stages of an innings (i.e. *order* = 1-4).

[2]Definitions of the performance metrics can be found in Appendix A.

2. Manually checking a players biography via ESPNCricinfo. Wicket-keepers and all-rounders were manually obtained through player biographies.

Next, all players classified as batsmen, wicket-keepers and batting all-rounders, across the IPL and CWC2011 datasets were entered into a single dataset, while the bowlers and bowling all-rounders were entered into another dataset. Subsequently the IPL and CWC2011 batting metrics and the corresponding players, across the IPL and CWC2015 datasets were combined into a single dataset, referred to as the *batting dataset*. The same was applied to the bowling metrics, referred to as the *bowling dataset*. The batting dataset contained 321 observations (i.e. players) and 14 columns (i.e. metrics) while the bowling dataset contained 238 observations and 21 columns. The intuition was that the batting and bowling metrics that significantly effect winningness in limited overs cricket, are the same across formats. Although the effect size and significance of each metric, for each player-type, varies across formats.

4.2 Data Limitations

Through data collection and processing, limitations were identified in the extracted scorecards. A major limitation to the data was missingness, for instance a number of IPL scorecards failed to record extras, fours conceded, sixes conceded and/ or minutes played. These scorecard inconsistencies produced misalignments in the data, as the SAS extraction process did not accommodate for occasions where ESPNCricinfo failed to record metrics. The 'missing' metrics were obtained using *ball-by-ball* commentary data from ESPNCricinfo.com.

The difference between scorecard and ball-by-ball data is that the former presents an overall view of match result, while the latter provides information on what happened during each ball of a match. To extract the ball-by-ball data the author developed an additional SAS process which parsed the associated commentary log for each match. The process translated commentary data into numerical data, producing a more convenient data structure; Appendix D illustrates data structure after extraction. The SAS script extracted the relevant details on a ball-by-ball basis, and stored data in a tabular form for easy access. This was then summarised into a scorecard format and stored in a tabular form, as shown in Appendix B. The 'ball-by-ball' based

scorecards were merged with the scorecards containing no missing metrics, and commenced processing the scorecards, i.e. calculating the appropriate performance metrics for each player-type and splitting/ merging the dataset into batting and bowling metric datasets.

Chapter 5

Exploratory Data Analysis and Regression Diagnostics

This chapter evaluates the characteristics of the analysis datasets and establishes validity of the regression assumptions. The data outlined in the previous chapter is used to detect outliers, determine the presence of multicollinearity and interrelationships among the predictor variables, as well as assessing the size, strength and direction of these relationships.

5.1 Summary Statistics

Running summary statistics on the analysis datasets yielded missing values (N/A), however no discrepancies were found within the summaries. Removing missing value observations produced batting and bowling datasets with 321 and 195 observations, respectively. It should be noted that minimum values of zero were observed for the *sixes hit, fours hit, percentage wins, number of wins, total balls bowled, total maidens, total wickets* and *total sixes conceded* within the datasets. A Cook's distance test revealed influential observations (i.e. outliers) that would substantially change the estimate of coefficients, leading to inaccurate conclusions in a regression analysis.

5.2 Multicollinearity and Interrelationships

Using the *car* and *asbio* packages in R, variance inflation factors (VIF) and scatterplot/ correlation matrices were produced, respectively. The presence and strength of multicollinearity and interrelationships among the batting and bowling metrics were determined.

5.2.1 Variance Inflation Factors (VIF)

Running the *VIF* function on model (5.1) produced an 'alias' error, indicating the presence of linearly dependent batting metrics (i.e. perfect multicollinearity)[1]. Conducting an 'alias' analysis revealed that *total balls* and *total boundaries* were linearly dependent across the batting metrics. Removing *total boundaries* (i.e. nullifying alias errors) from the model illustrated strong multicollinearity between the *total runs*, *total balls*, *total dismissals* and *innings played* metrics.

$$\begin{aligned} Winningness = batting\,average \times \beta_1 \; + \; strike\,rate \times \beta_2 \; + \; total\,runs \times \beta_3 \\ + percenatge\,boundaries \times \beta_4 \; + \; total\,balls\,faced \times \beta_5 \\ + total\,boundaries \times \beta_6 \; + \; sixes\,hit \times \beta_7 \; + \; fours\,hit \times \beta_8 \\ + Games\,played \times \beta_9 \; + \; total\,dismissals \times \beta_{10} \end{aligned} \tag{5.1}$$

An alias analysis was also conducted on model (5.2), since running the 'VIF' function on the model revealed the presence of perfect multicollinearity. The alias analysis found that *balls bowled* and *boundaries conceded* were linearity dependent, across the bowling metrics. Removing both *boundaries conceded* and *balls bowled* from the model revealed strong multicollinearity between all metrics except *games played*, *total maidens*, *total overs* and *total wickets*.

[1] VIF values > 10 indicate poor regression coefficient estimates due to multicollinearity.

$$\begin{aligned} Winningness = economy\,rate \times \beta_1 \; + \; strike\,rate \times \beta_2 \; + \; bowling\,average \times \beta_3 \\ + percenatge\,boundaries \times \beta_4 \; + \; dot\,balls \times \beta_5 \\ + balls\,bowled \times \beta_6 \; + \; percentage\,dot \times \beta_7 \; + \; runs\,conceded \times \beta_8 \\ + wickets \times \beta_9 \; + \; games\,played \times \beta_{10} \\ + sixes\,conceded \times \beta_{11} \; + \; fours\,conceded \times \beta_{12} \\ + boundaries\,conceded \times \beta_{13} \; + \; total\,overs \times \beta_{14} \\ + total\,maidens \times \beta_{15} \end{aligned} \tag{5.2}$$

5.2.2 Scatterplots and Correlation matrix

A scatterplot and correlation matrix for the batting metrics illustrated strong positive relationships between the *total dismissals, strike rate, percentage boundaries, total runs scored, total balls* and *games played* metrics. The relationships between these metrics produced correlation values, r, ≥ 0.70. Moreover, all relationships among the batting metrics were statistically significant at the 5% level. These results illustrate strong interrelationships among the batting metrics.

A scatterplot and correlation matrix for bowling metrics illustrated strong positive relationships between the *total balls bowled, total overs* and *total dots* metrics. The relationships between these metrics produced correlation values, r, ≥ 0.70. Out of the 120 bowling metric relationships 105 were statistically significant at the 10% level. These results illustrate a strong interrelationships among the bowling metrics.

5.3 Regression Assumptions

A regression analysis was conducted on the batting and bowling dataset to identify performance metrics that are practically and statistically significant contributors to team "winningness" i.e. percentage wins. The validity of the regression analysis was tested by examining the validity

of the following regression assumptions:

1. Independence of errors
2. Normality of Residuals
3. Constant variance of residuals
4. Residuals Outliers

5.3.1 Independence of Errors

Assumption (1) was tested by conducting a Durbin-Watson test, using the *durbinwatson()* function from the 'car' package in R. The Durbin-Watson test tests whether or not the residuals are serially correlated (i.e. lag one autocorrelation errors). The hypotheses tested are $H_0 : \rho = 0$ (i.e. autocorrelation of zero) $vs.$ $H_1 : \rho \neq 0$. The test statistic can vary between 0 and 4 with a value of 2 meaning that the residuals are uncorrelated. A value greater than 2 indicates a negative correlation between adjacent residuals, whereas a value below 2 indicates a positive correlation" [38, p.874].

The Durbin-Watson test, for the batting metric regression analysis, generated a $p-value$ above the 5% ($p-value$ = 0.8) confidence threshold, suggesting that the null hypothesis, H_0, cannot be rejected, indicating no serial correlation among the residuals (i.e. independence of errors at lag one). The D-W test statistic for the batting metric regression model was 1.97, indicating that the independence of errors assumption holds.

The Durbin-Watson test, for the bowling metric regression analysis, generated a $p-value$ below 5% ($p-value$ = 0.002), suggesting that the null hypothesis should be rejected, indicating lag one serial correlation among the residuals. The D-W test statistic for the bowling regression model was 1.57. The result shows that the independence of error assumption fails.

5.3.2 Normality of Residuals

A *Quantile-Quantile* (Q-Q) plot of residuals tests the assumption of *'normally distributed residuals'*. Residuals that follow a straight line indicate normally distributed errors, while any trends

other than a 'straight line' indicate a violation in the assumption.

The Q-Q plot of batting regression residuals showed the residuals deviating from a straight-line trend (figure 5.1), suggesting that the normality assumption had been violated. Additionally a Shapiro-Wilks test produced $p - value$ less than 5% for the set of residuals reinforcing the claim that the residuals are not normally distributed.

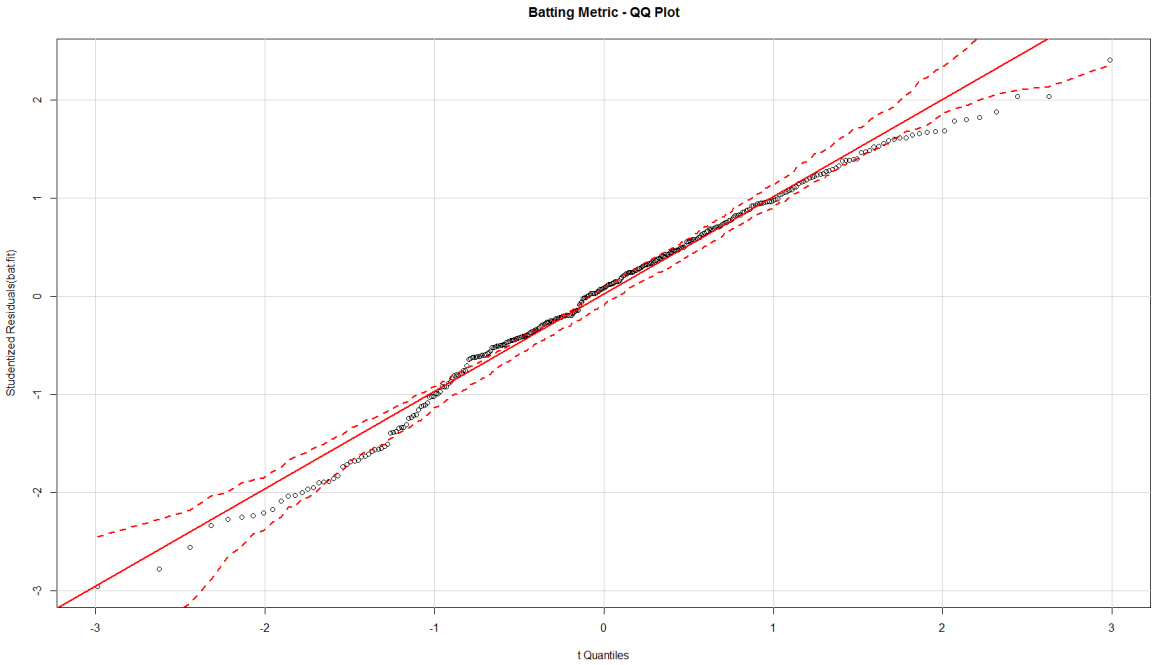

Figure 5.1: Batting Metric Q-Q plot

The Q-Q plot of the bowling regression residuals showed the residuals deviating from a straight line trend (figure 5.2), suggesting that the normality assumption had been violated. Additionally a Shapiro-Wilks test produced a $p - value$ less than 5% for the set of residuals, reinforcing the claim that the residuals are not normally distributed.

5.3.3 Constant Variance

The *ncvTest* function [library(car)] was implemented to evaluate homoscedasticity. The function implements the Breush-Pagan test, with hypotheses: H_0 : constant error variance $vs.$ H_1 non-constant error variance.

The Breush-Pagan test generated $p - values$ below 5% for both the batting and bowling re-

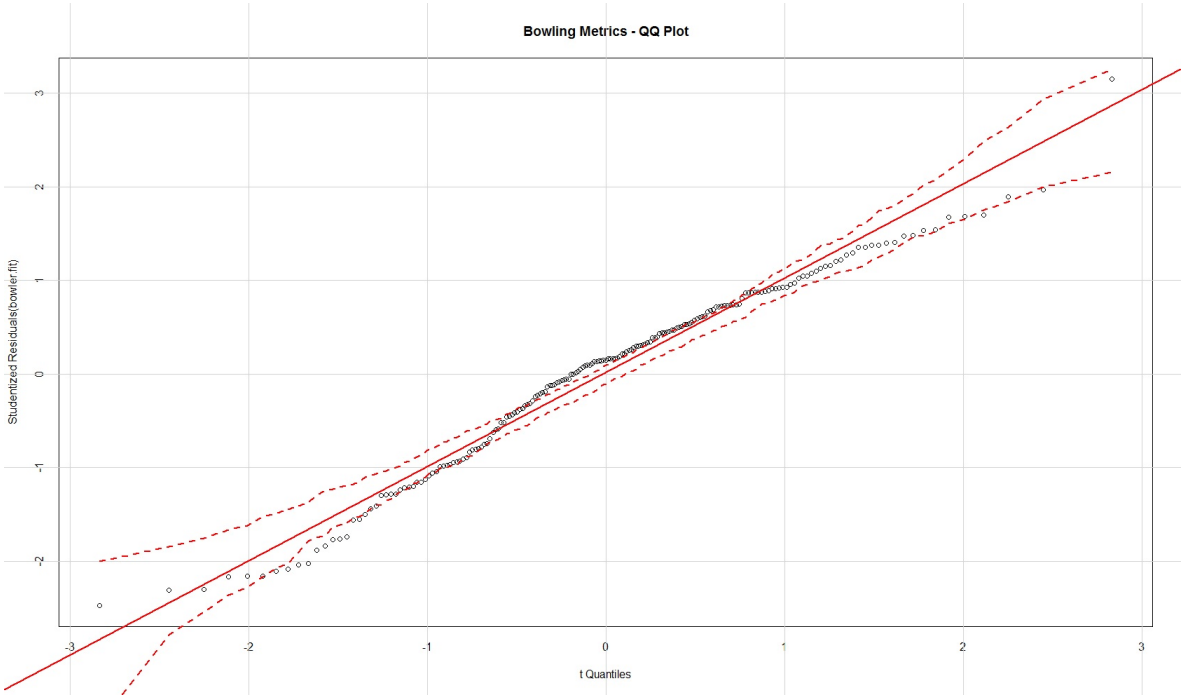

Figure 5.2: Bowling Metric Q-Q Plot

gression residuals, suggesting a rejection of the null hypothesis, H_0, in favour of the alternative hypothesis, H_1, indicating that the variance of the residuals is non-constant. Spread-level plots [absolute studentised residuals residuals vs. fitted values] illustrated decreasing trends (figure 5.3 and figure 5.4), reinforcing the claim that the constant variance assumption had been violated. Additionally the residuals were plotted against the performance metrics which confirmed a violation in the *equal variance of residuals* assumption.

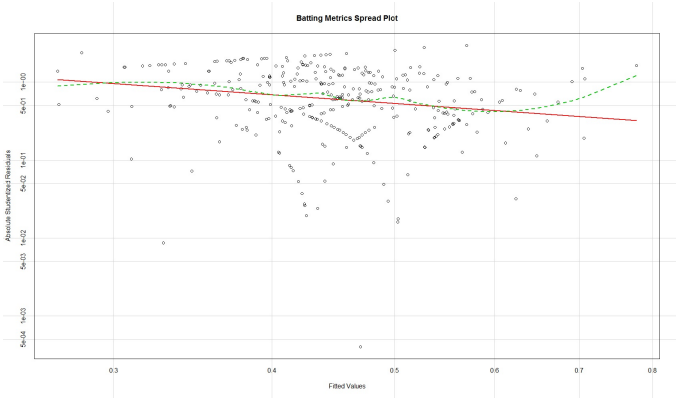

Figure 5.3: Batting Metric Spread Plot

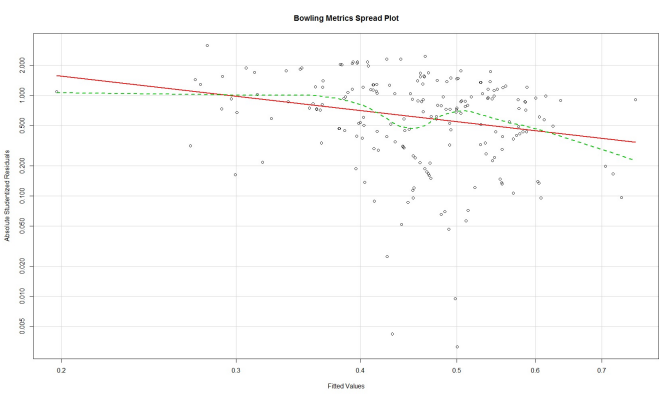

Figure 5.4: Bowling Metric Spread Plot

5.3.4 Residual Outliers

Assumption 4 was tested using the *outlierTest()* function [library(car)] which reports the "Bonferroni $p - value$ for studentised residuals in linear and generalised linear models, based on a t-test for linear models and normal-distribution test for generalized linear models" [3]. The Bonferroni correction measure tests each of the, n, residuals to determine whether or not it is an outlier [45]. The hypotheses tested were: H_0 : No residuals outlier effect $vs.$ H_1 : Residuals outlier effect. The test generated Bonferroni $p - values$ above 5% for both batting and bowling regression residuals, which suggested a failure to reject the null hypothesis indicating no residual outlier effect.

5.4 Chapter Remarks

The regression diagnostics suggest that a regression analysis, to identify significant "winningness" metrics, would produce invalid results due to violations in the constant variance, normality of residuals and independence of errors assumptions. Additionally the strong presence of multicollinearity and interrelationships among the performance metrics suggests that regression results would be invalid and subject to scrutiny and criticism. Since the presence of multicollinearity and interrelationships can produce distorted standard errors of regression coefficients, the succeeding chapter assesses several non-parametric and parametric dimension reduction techniques which account for multicollinearity and interrelationship issues. This ensures the author identifies performance metrics that significantly effect 'winningness' among cricketers, for limited overs cricket.

Chapter 6

Establishing Significant Performance Metrics

This chapter is dedicated to identifying performance metrics that significantly affect a team's ability to win (i.e. winningness), for individual cricketers[1]. Due to the issues of multicollinearity, interrelationships between the performance metrics and the high dimensionality of the data, several dimension reduction techniques are introduced to handle these issues and identify performance metrics that significantly affect a player's contribution to team winningness. To ensure that statistically significant and important metrics are identified two areas of statistical dimension reduction are considered: (1) *Classical Parametric* techniques: Principal Component Analysis, Linear Discriminant Analysis, Stepwise Regression and Hierarchical Variable Clustering, and (2) *Non-parametric* techniques: Regression Trees and Random Forests. Whilst a preliminary regression analysis could have produced an assessment of significance for each metric by evaluating the statistical significance and effect size, such analyses generate unreliable and inaccurate results, when there is multicollinearity and interaction effects. Additionally, given the multitude of performance metrics and the research requirement to produce a highly predictive, practically meaningful, team rating system, an accurate means of assessing variable significance was paramount for research success.

[1] Players are nested within teams and this will also be addressed

Variable selection is a process whereby a heuristic or algorithm identifies the variables that best accomplish a given modelling objective such as explanatory value and prediction accuracy. The three main strategies in variable selection are 'wrappers', 'filters' and 'embedded methods' [43]. With wrappers, the variable importance measures of a supervised learning machine, trained to predict a response variable, are used to determine variable selection in subsequent models. Filters, in contrast, assess importance during a 'pre-processing step' separate from the response variable, while embedded methods are automated and self-contained within the model. Due to the issue of multicollinearity and the high dimensionality of the data, various variable selection techniques are introduced to minimize the presence of such effects and reduce the number of performance metrics that are implemented when evaluating individual player ratings. The succeeding section will provide a brief technical introduction to the proposed dimension reduction techniques, followed by an application section.

The aim of this chapter is to identify the performance metrics that significantly affect winningness for individual cricketers, this could be a problem due to the nesting of players within teams. Achieving this goal addresses issue no. 1 (Chapter 3, Section 3.3).

6.1 Classical Parametric Techniques

6.1.1 Principal Component Analysis

Principal Component Analysis is a dimension reduction technique used in multivariate statistics. It is most useful when there is a high degree of correlation in the predictors (i.e. performance metrics). "The objective of the analysis is to take p variables, $X_1, X_2, ..., X_p$, and find a linear combinations of these to produce indices, $Z_1, Z_2, ..., Z_p$, that are uncorrelated" [58, p.76]. These uncorrelated multi-attribute, Z, components are orthogonal linear combinations of the original, p, variables, measuring different 'dimensions' in the data. The primary goal is to produce components such that the majority of the variance in the indices will be so small as to be negligible, while a small number of Z components explain the largest proportion of variation [58]. According to [58] the steps to produce Principal Components are as follows:

1. Standardise the variables $X_1, X_2, ..., X_p$ such that $\mathbf{X} \sim N(0,1)$[2].

2. Calculate the covariance matrix $\underline{\mathbf{C}}$. This is the correlation matrix if step 1 has been completed. The covariance matrix is used when variables are on similar scales, while the correlation matrix is used when variables are on different scales.

3. Identify the eigenvalues $\lambda_1, \lambda_2, ..., \lambda_p$ and the corresponding eigenvectors $\underline{\mathbf{a}_i}$; $\{i = 1, 2, ..., p\}$. "The coefficient of the ith principal component are then given by $\underline{\mathbf{a}_i}$, while λ_i is its variance" [58, p.80]. Order the eigenvectors by descending eigenvalues to establish the order of significance[3].

4. "Discard any components that only account for a small proportion of the variance in the data" [58, p.80].

5. The significant components (i.e. accounting for a large proportion of variation), Z_i, are expressed as a linear combination of the predictor variables $\Rightarrow Z_i = \lambda_1 x_1 + \lambda_2 x_2 + \therefore + \lambda_i x_i$.

6.1.2 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is " based on the idea of finding a suitable linear combination of the original variables" [58, p.13], with the intention of preserving class discriminating information. Compared to PCA, LDA attempts to provide more class separability by identifying decision regions between classes using available measurements. Additionally, in the presence of 'class' information, supervised approaches, such as LDA, are considered superior (i.e. more effective) than unsupervised techniques [80]. The technique attempts to establish maximum class discrimination by identifying a projection matrix which maximises the between class variance (s_b) and minimises the within class variance (s_w), using a function known as Fisher's criterion. The main objective is to find "an optimal transformation (projection) by minimizing the within-class distance and maximizing the between-class distance simultaneously, thus achieving maximum class discrimination" [80, p.1]. According to [55, p.3] the

[2] It is not necessary to have normality.

[3] A property of eigenvectors of a matrix is perpendicularity meaning that the data can be expressed in terms of the orthogonal eigenvectors, instead of the x and y axes.

steps to conduct a Linear Discriminant Analysis are as follows:

1. Start by calculating, g, the class mean vectors:

$$\bar{x_i} = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{i,j},$$

where $\bar{x_i}$ represents the mean vector for class i (i = index of the class), N_i represents the number of observations in class i and x_{ij}, $\{j = 1, 2, ..., N_i\}$, represents the jth observation in class $i = \{1, 2, ..., g\}$.

2. Calculate a grand mean for the entire dataset:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{g} N_i \bar{x_i} = \frac{1}{N} \sum_{i=1}^{g} \sum_{j=1}^{N_i} x_{i,j}$$

3. Calculate the between class scatter matrix:

$$S_b = \sum_{i=1}^{g} N_i (\bar{x_i} - \bar{x})(\bar{x_i} - \bar{x})^T$$

4. Calculate the within class scatter matrix:

$$S_w = \sum_{i=1}^{g} \sum_{j=1}^{N_i} (x_{i,j} - \bar{x_i})(x_{i,j} - \bar{x_i})^T$$

5. Find the projection matrix,Φ_{lda}. Since the determinant of a covariance matrix establishes the amount of class variance, the projection matrix, Φ_{lda} can be written as:

$$\Phi_{lda} = \frac{|\Phi^T S_b \Phi|}{|\Phi^T S_w \Phi|},$$

where $\Phi = [\phi_1, \phi_2, ..., \phi_m]$ represents the full projection, and the ϕ_i's are vectors defining the direction of one of the new axes. Using this transformation the data-points are converted into a new axis system.

6.1.3 Stepwise Regression

Stepwise regression is a variable 'filtering' technique which implements a combination of the forward selection and backward elimination variable selection techniques. The "stepwise regression algorithm is forward selection followed by backward elimination" [67, p.418]. The technique selects variables by testing various combinations of the variables and evaluates each combinations R^2, and the associated $p - value$ for each variable.

Forward Selection

The forward selection technique starts with the null model and iteratively adds predictors (i.e. performance metrics) to the model and tests variable significance. The predictors with p-values less than α_{crit}, for example 5% significance level, are retained in the model. The forward selection procedure stops when all variable p-values are less than α_{crit} (i.e. keep adding variables into the model until none of the remaining variables are 'significant' to the model). The $p-value$ identifies which performance metrics should be retained or removed from the model.

The predictors with the greatest amount of model contribution are retained. Model contribution is defined as the set of predictors (i.e. performance metrics) that produce the R^2 or the lowest Akaike Information Criteria (AIC).

Backward Elimination

The backward elimination procedure starts with all predictors in the model, $Y = \beta_0 + \beta_1 + ... + \beta_{r-1}X_{r-1} + \epsilon$, and removes the predictors with p-values greater than α_{crit} and refits the model. This process is continued until all p-values are less than α_{crit}. The predictors with the least amount of model contribution are eliminated.

The Backward elimination procedures are as follows:

1. Hypothesis tests ($H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$, $j = 1, 2, ..., r - 1$) are carried out with the corresponding lowest partial F-test (i.e. F_l) or t-test (i.e. t_l) value being compared

with the preselected significance value, F_0 and t_0, respectively [76].

2. Variable X_l is deleted from the original model if $F_l < F_0$ or $t_l < t_0$, and retained otherwise.

Stepwise Regression

Stepwise regression is a modification of the forward selection method in that after each step in which a variable is added, all the candidate variables in the model are checked for statistical significance, if significance is achieved the variable is retained in the model, otherwise it is removed.

1. The algorithm starts with the null model and adds in a single variable using the forward selection method.

2. After each new variable is added into the model the Stepwise regression performs backward elimination,. The smallest partial F-value, F_l, is compared to the preselected significance, F_0; if $F_l < F_0$ then the variable is deleted, otherwise it is retained in the model.

3. The procedure will continue until no variables can be eliminated from the new original model and all next best candidate variables can not be retained in the new original model [67].

The final optimal model minimises the AIC and/ or maximises the adjusted R^2 value. A drawback of the stepwise regression is the stopping criteria only produces a single model whereas there may be a variety of models with a similar goodness-of-fit [67]. An additional drawback is that the forward selection method only selects independent variables that maximise the squared partial correlation coefficient with the dependent variable [10].

6.1.4 Hierarchical Clustering Trees

Clustering methods are an unsupervised learning technique which replaces a group of similar variables by a clustering centroid [56] - an average across all points in the cluster. Clustering

methods can be broken down into 2 sub-groups: (1) Hierarchical methods and (2) Partitioning methods (i.e. $K - means$ clustering). However, only hierarchical clustering will be discussed. Partitioning methods were not implemented during the analysis as partitioning algorithms require the user to specify the number of clusters, defeating the purpose of determining a suitable number of clusters.

Hierarchical clustering methods construct clusters by recursively partitioning in either a top-down or bottom-up fashion, producing a sequence of nested partitions [56]. Usually the value measuring similarity between each pair of documents is stored in a $n \times n$ similarity matrix [56, p.935]. These recursive methods can be divided into 2 sub-groups: (1) *Agglomerative* - assigns each object as its own cluster, these objects are successively merged (i.e. clustered) until a desired result is obtained, and (2) *Divisive* - all objects start off as one cluster, the cluster is then successively divided into sub-clusters until a desired result is obtained. "The merging and division of clusters is performed according to some similarity measure, chosen so as to optimise some criteria, such as sum of squares" [56, p.278]. The clusters are divided and merged based on one of three hierarchical approaches which measure the distance between points and define inter-group similarities:

1. *Single-link:* "The distance between two clusters is considered to be equal to the shortest distance from any member of one cluster to any member of the other cluster" [56, p.279].
2. *Complete-link:* "The distance between two clusters to be equal to the greatest distance from any member of one cluster to any member of the other cluster" [56, p.279].
3. *Average-link:* "The distance between two clusters is considered to be equal to the average distance from any member of one cluster to any member of the other cluster" [56, p.279].

Hierarchical Clustering procedure

Given a dataset D, with n observations to be clustered, and an $N \times N$ distance matrix, the basic process of hierarchical clustering is as follows:

Step 1

***Agglomerative*:** Assign each item as its own cluster, so that if there are N items, there

are N clusters. Let the distances (i.e. similarities) between the clusters be the same as the distances between the items within the clusters.

Divisive: Assign all data-points into one cluster, so that there is one cluster containing all data-points.

Step 2

Agglomerative: Finds the closest pair of clusters and merge into a single cluster, so now there is one less cluster.

Divisive: Find the most dissimilar objects in the cluster and divide into sub-clusters, so now there is an extra cluster.

Step 3

Agglomerative: Merge the single (i.e. dissimilar) objects together and compute the distances (i.e. similarities) between the new cluster and each of the old clusters.

Divisive Compute distances (i.e. dissimilarities) between the new cluster and each of the original clusters.

Step 4

Agglomerative & Divisive Repeat steps 2 and 3 until a desired [4] result is obtained.

6.2 Non-Parametric Techniques

6.2.1 Regression Trees

Regression trees, also known as Decision Trees, are a supervised classification learning technique made up of 'decision nodes' with each decision node containing an individual test function, $f^n(x)$, of discrete outcomes. Given an input the test function, $f^n(x)$, determines the path or branch to follow, depending on the outcome. "Regression Trees organise these nodes in a recursive, unidirectional, hierarchical fashion by repeated application of the test function" [26, p.16]. Tree 'induction' (i.e. training) starts with all data set observations at the 'root' node and corresponding test function. The function splits records into subsets that are input,

[4] In Hierarchical Clustering the desired result is user defined as the number of groups of (similar) objects that best distinguish variable characteristics.

via ‘branches’, to subordinate ‘leaf’ nodes, which in turn split records to lower nodes. The output label of a leaf node constitutes the Regression Trees prediction.

The technique is a robust non-parametric alternative to classical parametric models and it creates models that are robust to the distorting influences of complex variable interactions and interrelationships that would render a parameter model unreliable. Moreover, classical parametric models are replete with assumptions and distribution restrictions. Regression Trees, however, are “immune to the potential model-defeating characteristics of these effects and are a useful tool in identifying terms for the regression model to help the models perform better” [32, p.27].

The technique applies binary recursive partitioning to the sample space which minimises the training error to improve the fit. The recursive technique is a partitioning method “whereby the data are successively splits along coordinates axes of the explanatory variables so that, at any node, the split which maximally distinguish the response variables in the left and right branches is selected” [28, p.686], these sequences of splits define a binary tree. The optimal split (i.e. minimises the residual sum of squares) is found over all variables and all possible split points that bring about the largest drop in the residual sum of squares. To produce better statistical performance the full tree may be pruned using a ‘pruning’ technique, which “recursively ‘snips’ off the least important splits based upon the cost-complexity measure [28], such as the Gini index, Shannon’s Information and reduced error, which reflects the trade-off between fit and explanatory power. These cost-complexity measures prune the tree based on a given cut-off threshold, such as misclassification rate, information gained etc., for each decision node. For each decision node if the criteria is not met the node and subsequent tree is pruned. Overall pruning the regression tree reduces the complexity and over-fitting, increasing predictive accuracy.

Bootstrap Aggregation

Bootstrap Aggregation, also known as bagging, takes an arbitrary classifier and aggregates copies of that classifier to improve its performance. “Bagging predictors is a method of generating multiple versions of a predictor and using these to get an aggregated predictor” [18, p.123].

The algorithm works as follows:

Given a dataset $D = [(x_1, y_1), ..., (x_n, y_n)] \sim (p)$ a new non-random x_i is generated and a prediction on Y, associated with a given x, is produced. Supposing there is a true underlying function $f(x) = y$, a prediction on y is feasible for any given x_i. The intuition behind bagging is that for a collection of datasets a prediction on Y is generated on a particular x value. If each dataset is independently drawn from p then the average of the predicted $y_i's$, for $i = 1, ..., n$ datasets, will be close to the true value of Y. Given that only one dataset is accessible, a bootstrap aggregation technique is implemented. The technique utilises the original dataset and approximates p by randomly re-sampling, with replacement, n data points from D and for each re-sampled dataset, m, a value for $f^m(x)$ associated with a particular value of x can be generated. Supposing that $f(x)$ is an unbiased estimator of the true prediction of y the error can be measured according to a loss function, $\Phi(f(x) - y)^2$, and the corresponding 'risk' can be measured by the expectation of the squared distance from the true value, $E[(f(x) - y)^2] = E[(f(x) - E(f(x))^2)] = \sigma_y^2$. Next a new, unbiased estimator, variable Z is generated, defined as:

$$\begin{aligned} Z &= \frac{1}{m}\sum_{i=1}^{m} f(x)^i \\ E[Z] &= \frac{1}{m}\sum_{i=1}^{m}(y) = y \end{aligned} \tag{6.1}$$

The risk for each Z is computed via:

$$\begin{aligned} &E[(Z-y)^2] = E[Z - E[(Z)^2]] \\ &= \sigma_y^2(Z) = \sigma_y^2\left(\frac{1}{m}\sum_{i=1}^{m} f(x)^i\right) \\ &= \left(\frac{1}{m}\right)^2 \sigma_y^2\left(\sum_{i=1}^{m} f(x)^i\right) \\ &= \left(\frac{1}{m^2}\right)\sigma_y^2\left(\sum_{i=1}^{m} f(x)^i\right) \\ &= \frac{1}{m}\sigma_y^2(f(x)). \qquad \square \end{aligned} \tag{6.2}$$

Notice that the expected loss/ risk, $Z = \frac{\sigma_y^2}{m}$, is dependent on m, therefore as m tends towards infinity the expected loss tends towards zero. The intuition is that p is approximated by the empirical distribution, $\hat{p}$, and the uniform bootstrapped samples are drawn from $\hat{p}$.

6.2.2 Random Forest

A 'wrapper' feature selection technique, Random Forest, is a [meta-learning] ensemble technique consisting of a collection of uncorrelated and unpruned regression trees. Random Forests "are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest" [19, p.5]. Random Forest splits base-classifier Regression Trees on a random sub-sample of variables. "The generalization error for forests converge asymptotically to a limit as the number of trees in the forest become larger" [19, p.5].

The Random Forest algorithm:

1. Given a dataset $D = [(x_1, y_1), ..., (x_n, y_n)]$ the Random Forest technique constructs T_i regression trees using a bootstrapped sample D_i, for $i = 1, ...B$, where B is the number of bootstrapped samples.

2. Using D_i, regression tree T_i is constructed such that at each node a random subset, m, of features is selected, and only splits on the m features from that particular subset are considered.

3. After the B trees have been constructed a 'majority vote' is taken over $T_1, ..., T_B$ to generate an aggregated predictor. Due to the law of large numbers random forests do not over-fit. The disadvantage of a single regression tree is that it has high variance and is highly sensitive to the particular arrangement of the data points. Averaging over an ensemble of trees reduces the variance, leading to increased performance and reduced error.

6.3 Dimension Reduction Application

6.3.1 Principal Component Analysis

Principal Component Analysis was applied to identify a small number of Z components that adequately explained a large proportion of the variation in the analysis datasets[5]. If such results are obtained the components would then be used to produce 'new', Z_p', variables which are linear combinations of the eigenvalues (obtained from the eigenvectors) and the original performance metrics.

Applying the method to the batting dataset it was found that two components explain approximately 82% of data variation, with the first component explaining 66% of variance. These findings were reinforced by examining a scree plot which indicated that approximately two components sufficiently explained the variation. However, two major issues were encountered with these results.

1. The metric coefficients within the components varied in directions producing contradictory components. For example, the batting strike rate coefficient would be positive, while the batting average would be negative, generating counter intuitive components.

2. The new, Z_p', variables lacked interpretability, this was a major drawback as the research required results that were understandable and easily communicated to coaching, management staff and other non-technically inclined interested parties.

Applying the method to the bowling dataset it was found that three components explained 82% of variation, with the first two components explaining 71% of variance (48% and 23%, respectively). These findings were reinforced by a scree-plot which indicated that approximately three components adequately explained data variation. However the result of this analysis also suffered from interpretability issues and counter-intuitive results.

Given these problems it was concluded that PCA was an inappropriate dimension reduction

[5]Principal Component Analysis was executed using the *principal()* function in *library(psych)*. The components were based on the correlation matrix.

technique to ascertain the significant performance metrics. However given the findings it was assumed that approximately two-five performance metrics, across batsmen and bowlers, would be adequate, as the PCA results suggest that a range of 3-5 components adequately explain winningness among cricketers.

6.3.2 Linear Discriminant Analysis

Linear Discriminant Analysis requires a "class" variable to discriminant against[6]. Accordingly a 'class' attribute was added to each observation across the two datasets based on the players' world ranking. The rankings were extracted from the official International Cricketing Council (ICC) website [4].

Five "classes" were established with the following classification criteria:

1. A player ranked in the top 20 was classified as *class = 1*
2. A player ranked between 21-50 was classified as *class = 2*
3. A player ranked between 51-75 was classified as *class = 3*
4. A player ranked between 76-100 was classified as *class = 4*
5. A player ranked above $>$ 100 was classified as *class = 5*

The LDA equation applied to the batting datasets was:

$$\begin{aligned} Discriminant\ Function\ (i.e\ Class) = batting\ average \times \beta_1 \ +\ strike\ rate \times \beta_2 \\ + total\ runs \times \beta_3 \ +\ percenatge\ boundaries \times \beta_4 \ +\ total\ runs \times \beta_5 \\ + total\ balls\ faced \times \beta_6 \ +\ total\ boundaries \times \beta_7 \ +\ sixes\ hit \times \beta_8 \\ + fours\ hit \times \beta_9 + \ inning\ played \times \beta_{10} \ +\ total\ dismissals \times \beta_{11} \end{aligned} \tag{6.3}$$

[6]Linear Discriminant Analysis was applied using the *lda()* and *partimat()* function in the *library(MASS)* and *library(klaR)*, respectively.

The LDA equation applied to the bowling datasets was:

$$\begin{aligned} Discriminant\ Function\ (i.e\ Class) = economy\ rate \times \beta_1\ +\ strike\ rate \times \beta_2 \\ +\ bowling\ average \times \beta_3\ +\ percenatge\ boundaries \times \beta_4\ +\ dot\ balls \times \beta_5 \\ +\ balls\ bowled \times \beta_6\ + percentage\ dot \times \beta_7 + runs\ conceded \times \beta_8 \\ +\ wickets \times \beta_9\ +\ games\ played \times \beta_{10}\ +\ sixes\ hit \times \beta_{11} \\ +\ fours\ hit \times \beta_{12}\ + boundaries\ conceded \times \beta_{13}\ +\ total\ overs \times \beta_{14} \\ +\ total\ maidens \times \beta_{15} \end{aligned} \tag{6.4}$$

To test the predictive accuracy of the LDA technique both the batting and bowling datasets were randomly split into training and test sets. Once a batting and bowling model was trained it was applied to the test set to determine the prediction accuracy.

Applying LDA to the batting dataset it was found that players in class 1 and 2 tended to have greater *batting averages*, *percentage boundaries* and *total runs* metrics, compared to those in class 5. These findings were reinforced by assessing the accuracy of the predictions, the following classification accuracy was found established:

Table 6.1: Batting Metrics LDA classification accuracy

Class	**1**	**2**	**3**	**4**	**5**
Prediction Accuracy	0.61	0.41	0.16	0.067	0.65

The results illustrate that the batting metrics are 'adequate' for discriminating players in class 1,2 and 5. Applying an LDA to the bowling dataset and assessing the accuracy of predictions the following classification accuracy was established:

Table 6.2: Bowling Metrics LDA classification accuracy

Class	**1**	**2**	**3**	**4**	**5**
Prediction Accuracy	0.56	0.29	0.31	0.23	0.50

The results illustrated that the bowling metrics are adequate for discriminating players in class 1 and 5. *partimat()* plots could not be generated due to the number of performance metrics and

margin size issues. However observing the data, bowlers in class 5 tended to have lower strike rates, economy rates and percentage boundaries compared to those in classes 1-4.

The LDA results did not establish performance metrics that significantly contribute towards winningness. Rather the technique revealed the metrics that significantly contribute towards a players world ranking. The LDA results were not meaningful in the context of this research and do not provide a means of distinguishing metric significance. As previously mentioned the significance of each metric must be evaluated in terms of winningness, as the research objectives were geared around producing an adaptive rating system which accurately predicts match outcome.

6.3.3 Stepwise Regression

The stepwise regression model[7] applied to the batting dataset was:

$$\begin{aligned} Winningness\,(i.e.\,percentage\,of\,wins) = batting\,average \times \beta_1 + strike\,rate \times \beta_2 \\ + total\,runs \times \beta_3 + percenatge\,boundaries \times \beta_4 + total\,runs \times \beta_5 \\ + total\,balls\,faced \times \beta_6 + total\,boundaries \times \beta_7 + sixes\,hit \times \beta_8 \\ + fours\,hit \times \beta_9 + inning\,played \times \beta_{10} + total\,dismissals \times \beta_{11} \end{aligned} \tag{6.5}$$

The final model results showed that *total runs scored*, *total dismissals*, *sixes*, *batting average*, *balls faced* and *strike rate* were significant metrics and produced the greatest AIC value. A regression analysis, using the results suggested by the stepwise regression, indicated that all performance metrics were statistically significant at the 5% level. However only *strike rate*, *total runs* and *total dismissals* were practically significant. Additionally the significant metrics only explained 22% of the variation in the model.

[7] Stepwise regression was executed using the *StepAIC* function in *library(MASS)*.

The stepwise regression model applied to the bowling dataset was:

$$\begin{aligned} Winningness = economy\ Rate \times \beta_1 \ + \ strike\ rate \times \beta_2 \ + \ bowling\ average \times \beta_3 \\ + percenatge\ boundaries \times \beta_4 \ + \ dot\ balls \times \beta_5 \ + \ balls\ bowled \times \beta_6 \\ + percentage\ dot \times \beta_7 \ + \ runs\ conceded \times \beta_8 \ + \ wickets \times \beta_9 \\ + games\ played \times \beta_{10} \ + \ sixes\ hit \times \beta_{11} \ + \ fours\ hit \times \beta_{12} \\ + boundaries\ conceded \times \beta_{13} \ + \ total\ overs \times \beta_{14} \ + \ total\ maidens \times \beta_{15} \end{aligned} \tag{6.6}$$

The final model results showed that *games played*, *total dots*, *total maidens*, *total runs conceded*, *total wickets* and *total sixes* were significant metrics and produced the greatest AIC value. A regression analysis, using the final model results, indicated that all performance metrics were statistically significant at the 5% level. However only *games played*, *total dots*, *total runs* conceded and *total wickets* were practically significant. Additionally the significant metrics explained an inadequate amount of variance ($r - squared$ = 18%).

The stepwise regression results were unreliable as such parameter classical techniques are ill-equipped to handle multi-collinearity and interaction effects. Additionally the analyses were unable to produce a practical and parsimonious model. However the stepwise regression did provide insightful results, stating that *scoring efficiency* (i.e. strike rate), *scoring consistency* (i.e. total runs scored), and *run restriction* (i.e. total runs conceded and total dots) are key 'winningness' metrics.

6.3.4 Hierarchical Cluster Analysis

Applying a Hierarchical Clustering technique to the batting dataset produced a dendogram with four distinct clusters[8,9]. The dendogram illustrated that the four clusters focused on *scoring efficiency*, *scoring consistency*, *scoring volume* and *games played*. A *stability of partitions plot* was generated to establish the appropriate number of clusters. The stability plot was produced by taking, $B = 50$, bootstrap samples of 321 observations and creating 50 dendograms. "The

[8]Hierarchical clustering was executed using the *hclustvar()* function in *library(ClustOfVar)*.
[9]Clusters are formed by optimising the squared Pearson correlation.

partition of these B dendograms are compared with the partitions of initial hierarchy using the corrected Rand index" [21, p.7], which measures the similarity between cluster[10]. The stability plot showed that four clusters produced the smallest mean adjusted rand criterion reinforcing the claim that four key features characterise batting metrics.

Applying the method to the bowling metrics produced a dendogram with five distinct clusters. A stability plot ($B = 50$) showed that five clusters produced a small mean adjusted rand criterion, indicating that five key features characterise bowling metrics: (1) *run restriction* (2) *wicket-taking efficiency* (3) *balls bowled* (4) *total wickets* and (5) *boundary prevention*.

The clustering results provided insight into the relationship between performance metrics and identified the key features of the batting and bowling metrics. However the analysis produced very little in terms of establishing significant *winningness* metrics.

Parametric Reduction remarks

The following inferences were drawn from the parametric analysis:

1. Evidence suggests that there are three to four performance metrics that adequately explain variance in winningness, among limited overs cricketers.

2. The batting and bowling metrics adequately discriminate between high and low *quality* players.

3. Four key features characterise the batting metrics: (1) *scoring efficiency*, (2) *scoring volume* (3) *scoring consistency* and (4) *games played*.

4. Five key features characterise the bowling metrics: (1) *wicket-taking efficiency* (2) *run restriction* (3) *volume of balls bowled* (4) *boundaries conceded* and (5) *total wickets.*

5. The stepwise regression indicated that strike rate, batting average and total runs were significant contributors to winningness. Interestingly these three metrics are geared around *scoring efficiency*, *scoring consistency* and *scoring volume*. Moreover, these results indicate that winningness is highly influenced by the *efficiency*, *consistency* and *magnitude*

[10]Stability plots were produced using the *stability()* function in *library(ClustOfVar)*.

at which runs are accumulated. Moreover, the results indicated that among the bowling metrics *wickets*, *boundary prevention* and *run restriction* were significant contributors to 'winningness'.

Given the large presence of multicollinearity and interaction effects among the metrics, and the inability of classical parametric technique to handle high degree of multicollinearity, the lack of statistically robust and valid results were expected. As a consequence of the conflicting results, and the lack of variance explained, the capability of non-parametric reduction techniques to handle the issue of multicollinearity and interactions effects were evaluated.

6.3.5 Regression Trees

Applying a regression tree analysis to the batting metrics found that *total runs scored*, *total dismissals*, *balls faced*, *total boundaries*, *batting average* and *strike rate* significantly contribute towards winningness[11,12]. Additionally the $rsq.rpart()$ plot of the regression tree illustrated that 13 splits produced the greatest $r - squared \approx 0.35$. However, the regression tree produced counter intuitive results. The results suggested that *lower strike rates* lead to greater winningness. Therefore the regression tree was pruned, using a cp (complexity parameter measure) of 0.035 (i.e. 3 nodes), as suggested by the relative error plot. The pruned tree illustrated that three splits produced an $r - squared \approx 0.18$. However, the results produced were counter-intuitive stating that *lower strike rates* and *greater dismissals* lead to greater winningness.

The application of a regression tree analysis to the bowling metrics found that *total balls bowled*, *total dots*, *total runs conceded*, *total wickets*, *strike rate*, *economy rate*, *total boundaries*, *percentage boundaries* and *percentage dots* significantly contribute towards winningness. The $rsq.rpart()$ plot illustrated that 15 splits produced the greatest $r - squared \approx 0.60$. However, again, the regression tree results were counter-intuitive. Pruning the tree ($cp = 0.05$, nodes = 3) revealed sensible results, illustrating that *lower economy rates* lead to greater winningness. The pruned tree illustrated that five splits produced an $r - squared \approx 0.40$.

[11] Regression trees were created using the *rpart* function in *library(rpart)*.

[12] The *method* parameter was set to *anova* as the response was a continuous variables and specifies a splitting criteria based on within-node residual sum of squares.

However due to a regression trees susceptibility to high variance and sensitivity to the particular data-point arrangement, a random forest technique was applied.

6.3.6 Random Forest

Applying the Random Forest technique to the batting datasets, using the $importance()$ [library(randomforest)] function[13], the five most important metrics were:

1. Strike Rate
2. Balls Faced
3. Batting Average
4. Total Runs Scored
5. Percentage Boundaries

Interestingly these important metrics are associated with *scoring efficiency* (i.e. strike rate and percentage boundaries), *scoring consistency* (i.e. batting average) and *scoring volume* (i.e. total runs scored). Moreover four out of the five metrics (strike rate, percentage boundaries, batting average and runs scored) were identified as statistically and practically significant throughout the application of classical techniques. Applying the random forest technique to the bowling dataset, the five most important metrics were:

1. Economy Rate
2. Bowling Average
3. Strike Rate
4. Percentage Boundaries
5. Percentage Dots

[13]Random forests were applied using the *randomforest* function in *library(randomforest)*. The 'ntree' parameter was set at 5000, indicating 5000 trees were grown ensuring that every input row was predicted a sufficient number of times. The *importance* function produced an influence score for each performance metric indicating its importance to the model.

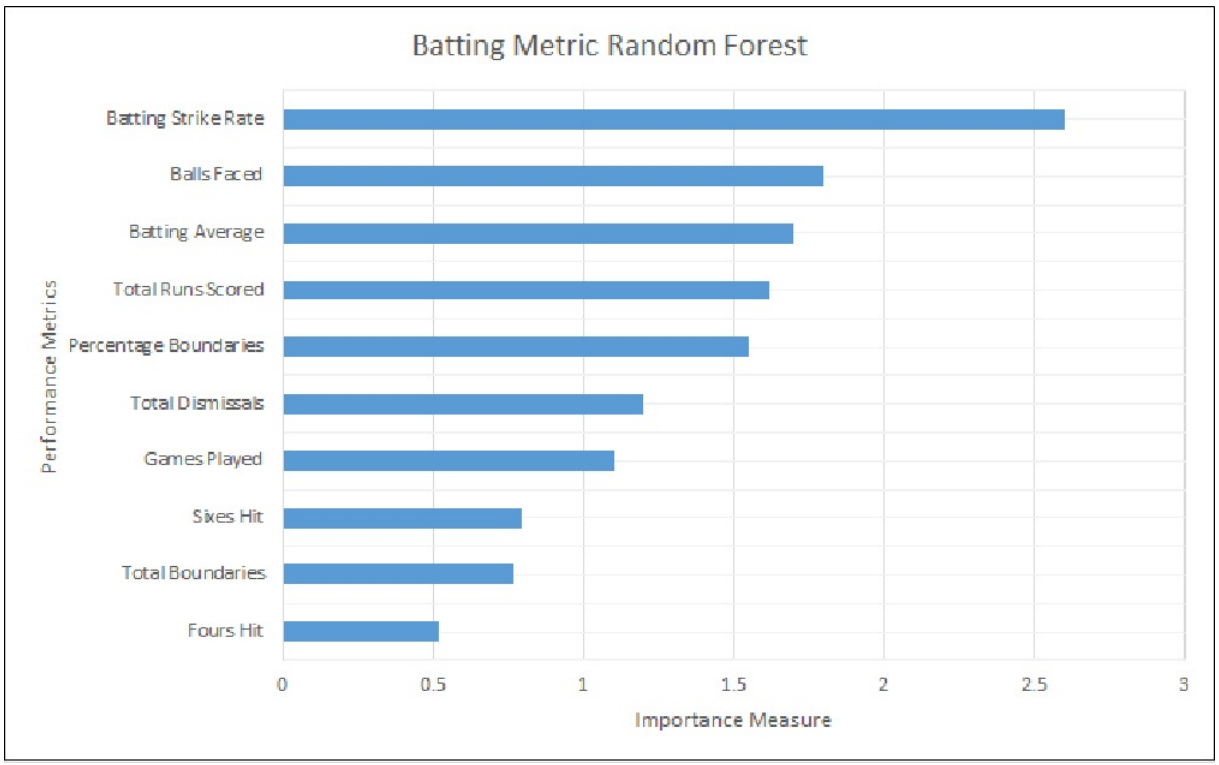

Figure 6.1: Batting Metrics Random Forest

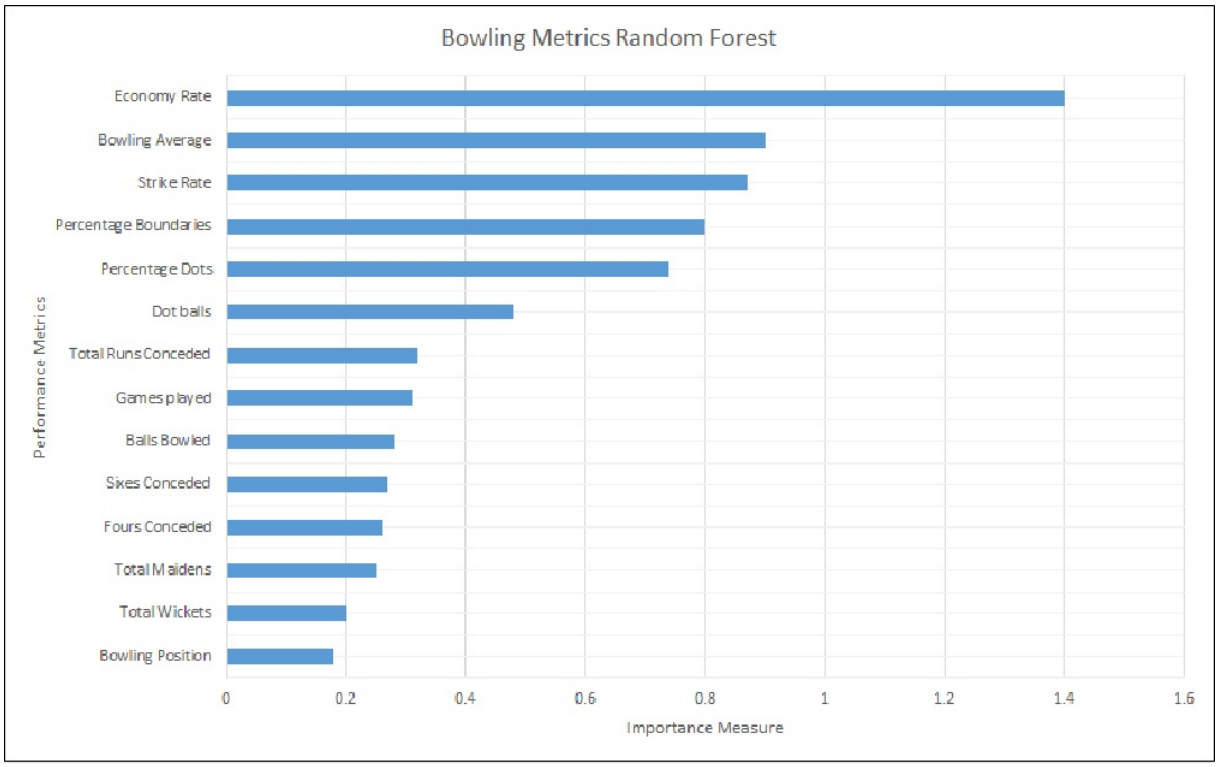

Figure 6.2: Bowling Metrics Random Forest

Interestingly out of the five features that characterise bowling metrics, only three were regarded as important for "winningness": *wicket-taking efficiency* (strike rate and bowling average), *boundary prevention* (i.e. percentage boundaries) and *run restriction* (i.e. economy rate and percentage dots). The results show that reducing the number of runs conceded and increasing the rate at which wickets are taken are significant to winningness.

6.4 Summary of Dimension Reduction results

Table 6.3: Summary Table of Dimension Reduction Results

Summary of Results		
Method	**Batting Metrics**	**Bowling Metrics**
PCA	N/A	N/A
LDA	batting average percentage boundaries total runs	economy rate strike rate percentage boundaries
Stepwise Regression	strike rate total runs total dismissals	total dots total runs total wickets
Hierarchical Clustering	strike rate percentage boundaries total runs	economy rate percentage boundaries strike rate
Regression Tree	strike rate total dismissals	economy rate
Random Forest	strike rate balls faced batting average total runs percentage boundaries	economy rate bowling average strike rate percentage boundaries percentage dots

Table 6.1 shows the significant/ important batting and bowling performance metrics established by each technique. Based on a combination of the results of the six dimension reduction tech-

niques the following performance metrics were selected to evaluate a individual player ratings (i.e. chapter 8):

Table 6.4: Significant performance metrics

Batting metrics	**Bowling Metrics**
Strike rate	Economy rate
Percentage Boundaries	Percentage boundaries
Batting average	Strike rate
Total Runs	Bowling Average
Total Balls Faced	Percentage Dots

6.5 Performance Metric Validation: Lorenz Curve and Linear Discriminant Analysis

A Lorenz curve was implemented to validate metric performance and examine discriminatory power[14] of the selected performance metrics. A Lorenz curve graphically "relates the cumulative proportion of income units to the cumulative proportion of income received when units are arranged in ascending order of income" [48, p.719] . Applying this method to the batting and bowling dataset, the players represent the cumulative percent of people in the population and percentage wins represent the cumulative percentage of events. The gap between the curve and the line of equality represents (i.e. AUC) the disparity between larger income groups and smaller income groups. In this case the gap represents disparity between high and low percentage wins[15] (i.e. winningness), and measures the classifiers (i.e. important performance metrics) discriminatory performance.

Applying a Lorenz curve to the 5 most important batting metrics produced an area under the curve (AUC) of 0.64 (figure 6.3), illustrating 'good' discrimination between players with high and low percentage wins. A Lorenz curve for the five most important bowling metrics produced an AUC of 0.63 (figure 6.4), illustrating 'good' discriminatory power.

[14]Lorenz curves were generated using the *rocr* function in *library(ROCR)*.
[15]High percentage wins ≥ 65; low percentage wins < 65.

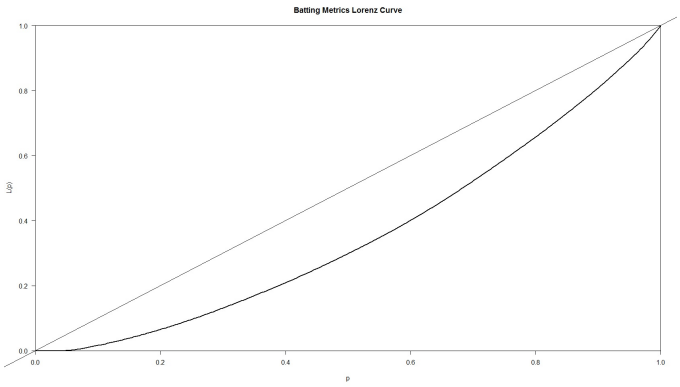

Figure 6.3: Batting Metrics Lorenz Curve

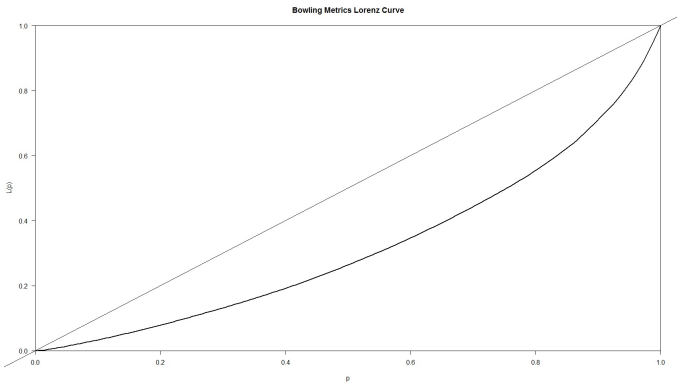

Figure 6.4: Bowling Metrics Lorenz Curve

Applying a Linear Discriminant Analysis there was a slight improvement in predictive accuracy across the 5 classes, from those results reported in tables 61. and 6.2.

Table 6.5: Linear Discriminant Analysis Accuracy

Class	**1**	**2**	**3**	**4**	**5**
Batting Predictions	0.62	0.27	0.09	0.10	0.68
Bowling Predictions	0.56	0.29	0.31	0.23	0.50

These results reinforce the selected metrics as *good* winningness metrics and influential to a players rating.

6.6 Chapter remarks

This chapter identified the performance metrics that have a significant effect on "winningness". Consequently, this chapter remedied research flaw no. 1 (Chapter 3, section 3.3). Interestingly, the significant batting metrics are geared around *scoring efficiency*, *scoring consistency* and *scoring volume*, while the significant bowling metrics are geared around *wicket-taking efficiency* and *run restriction*. The validity of the 5 most *important* batting and bowling metrics was established by an AUC of 0.64 and 0.63, respectively. Given that practically and statistically significant metrics have been established, optimisation methods, individual rating systems and forecasting techniques are now considered.

Chapter 7

Optimisation System

This chapter summarises the framework and the mathematical formulation of the adopted optimisation technique. The aim of this chapter is to determine an optimisation system that identifies the best players for a team utilising individual ratings, across each player-type, based on a set of player and team constraints.

Formally "an optimisation algorithm is an iterative numerical procedure for finding the values of the vector **x** that maximises or minimises the objective function $f(\mathbf{x})$ subject to constraints **c**" [65, p.100]. The goal of any optimisation problem is to identify values of the unknown variables which optimise the objective function, based on a set of constraints.

7.1 Mathematical Formulation

The mathematical formulation of an optimisation problem has three key components:

1. **Decision Variables; x** - A vector of unknown parameters (i.e one or more variables) on which to make a decision, and defines the ultimate decision of the optimisation.

   In the context of this project the decision variable is binary:

$$x_{ij} = \begin{cases} 1, & \text{if } player\ j\ is\ selected\ for\ role\ i \\ 0, & \text{otherwise} \end{cases} \tag{7.1}$$

2. **Constraints; c** - "A vector of constraints which the unknown parameters must satisfy" [65, p.60] and limits the values to a feasible region.

   In the context of this project model constraints are based on team construction for limited overs cricket. The constraints specify team balance, in terms of the number of bowlers, batsmen, all-rounders and keepers selected in the optimal team.

3. **Objective Function** - "A quantitative measure of performance of the system, more commonly known as the *objective function* which is typically minimised or maximised" [65, p.48]. The objective function mathematically represents a measure for the 'goodness' of values for the decision variables.

Given these components an optimisation problem can be written as:

$$\min_{x \in \mathbb{R}} f(\mathbf{x}) \; or \; \max_{x \in \mathbb{R}} f(\mathbf{x}) \;\; s.t \;\; \mathbf{c}^{1} \tag{7.2}$$

In the context of this project the overall team rating is required to be optimised (i.e. maximised), therefore the best players (i.e. players with highest rating) in each category (i.e. player-type) are required to be selected, subject to model constraints (i.e. team and player-type constraints).

7.2 Determining the optimisation system

There are various factors to consider when defining an optimisation problem. The following section discusses four general issues that may arise:

1. Is the optimisation problem discrete or continuous or a combination of the two?

- In *discrete (or combinatorial)* optimisation the model variables are either binary or integer, drawn from a finite set of feasible solutions. "The optimal solution to such problems is derived from a finite set of feasible solutions, that is, a vector of integers" [65, p.61].

[1] The domain of **x** is not necessarily **R**. In fact, in this problem, each **x** can only take 0 or 1.

- In *continuous* optimisation the model variables adopt a continuous range of values, usually real numbers. "The optimal solution to such problems is derived from an infinite set of feasible solutions, that is, a vector of real numbers" [65, p.61]. Models with continuous and discrete variables are *Mixed Integer* programs.

2. Is the optimisation problem stochastic or deterministic?

- *Stochastic* optimisation problems arise when the model is subject to randomness, that is, the model is not fully specified, at the time of formulation.
- "*Deterministic* optimisation problems are models that are fully specified, that is, there is no unknown quantity at time of formulation" [65, p.61].

3. Is the optimisation constrained or unconstrained?

Constrained optimisations refer to problems in which the objective function is optimised with respect to [unknown] parameters in the presence of constraints. These constraints, on the unknown parameters, must be satisfied for the objective function to be feasible. According to [65] a constraint could simply be:

1. A boundary placed on a variable declaring that a variable must take integer value
2. A general linear constraint
3. A non-linear inequality

4. Is the local solution also the global solution?

According to [65] many computer algorithms only identify local solutions, with no in-built functionality to check for global solutions. "However, many non-linear functions have several local minimums in which case one would be interested in which one of these local minimums is also the global minimum, which is, the best solution of all such minima" [65, p.61].

7.3 Binary Integer Programming

As mentioned previously the optimisation method for this research required the implementation of a binary decision variable, assigning a value = 1 to selected players and value = 0 otherwise

(i.e. not selected). Since the adaptive rating system requires selecting players associated with the largest individual ratings, given a set of team and player-type constraints, a maximisation objective function is implemented.

Given the binary nature of the decision variable:

$$x_{ij} = \begin{cases} 1, & \text{if } player\ j\ is\ selected\ for\ role\ i \\ 0, & \text{otherwise} \end{cases}, \tag{7.3}$$

a Binary Integer Programming Model (BIPM) was adopted. "In such a model each decision is modelled with binary decision; setting the variable equal to 1 corresponds to making the 'yes' or 'selected' decision, while setting it to 0 corresponds to going with 'no' or 'not-selected' decision" [12, p.76]. The BIPM technique utilises a *branch and bound* algorithm to solve the optimisation problem.

7.4 Branch and Bound Algorithm

The Branch and Bound algorithm "finds the optimal solution to an integer program by efficiently enumerating the points in a sub-problem's feasible region" [79, p.476]. Essentially, searching the complete space of solutions for a given problem, for the best solution. Branch and Bound algorithms are most commonly used tool for solving discrete (or combinatorial) optimisation problems, specifically large scale NP-hard (non-deterministic polynomial-time) [discrete] optimisation problems.

The method begins by solving the Linear Program (LP) relaxation of the integer program [79]. The solution to the integer problem (IP) is considered 'optimal' if the decision variables assume integer values to the LP relaxation. However if the optimal solution to the LP relaxation are not all integers values then the algorithm partitions the feasible region of the LP, in an attempt to establish more information regarding the location of the optimal solution. An arbitrary variable, x_1, is selected, that is fractional in the optimal solution to the LP relaxation [79]. The algorithm then branches on the arbitrary value and creates two additional sub-problems, sub-problem 2

and 3, known as nodes. "The constraints associated with any node of the tree are the constraints of the *LP relaxation + constraints* associated with the arcs leading from sub-problem 1 to the node" [79, p.480]. If the optimal solution to the sub-problems does not yield an all integer solution, then the sub-problems are used to create a new set of sub-problems. This process continues until: 1. further sub-problems can not yield any useful information, in this case the sub-problem is **fathomed**, or 2. when an optimal all integer solution has been obtained.

7.4.1 Binary Integer Programming Framework

The BIPM objective function:

$$Z = \max \sum_{i=1}^{n} \sum_{j=1}^{n_i} c_{ij} x_{ij} \tag{7.4}$$

where c_{ij}[2] represents the player rating for player j in role i, $\{i = 1, 2, 3, 4\}$,

$$where\ role\ i = \begin{cases} 1, & \text{if batting ability} \\ 2, & \text{if bowling ability} \\ 3, & \text{if all-rounder ability} \\ 4, & \text{if wicket keeping ability} \end{cases} \tag{7.5}$$

Decision Variable:

$$x_{ij} = \begin{cases} 1, & \text{if } player\ j\ is\ selected\ for\ role\ i \\ 0, & \text{otherwise} \end{cases} \tag{7.6}$$

The decision variables are binary identifiers for player-type i and player j, where $(i = 1, 2, 3, 4)$ and $(j = 1, 2, ..., n_i)$.

[2]The player rating coefficient, c_{ij}, calculations are outlined in chapter 8

7.4.2 Model Constraints

Given the flaws associated with the optimisation constraints outlined in [42] and [68], constraints that were specific to T20 and one day cricket were formulated. That is, constraints that accurately reflect a team's composition and the type of talent required to win limited overs cricket matches were assessed.

Constraints that were team and player orientated were formulated. Taking into account the number of batsmen, bowlers, all-rounders, wicket-keepers and number of players required to build a limited overs cricket team. Given the requirements of a cricket team the following model constraints were formulated:

1. Team Constraint

$\sum_{i=1}^{4} \sum_{j=1}^{n_i} x_{ij} = 11$ - 11 players must be selected in the optimal team.

2. Selection Constraint

$\sum_{i=1}^{4} x_{ij} \leq 1$ - Restricts players from being selected twice in the optimal team.

3. Batsmen Constraint

$\sum_{j}(x_{1j} + x_{3j}) \geq X$ - Ensures atleast X batsmen (i.e. specialist batsmen or batting all-rounders) are selected.

4. Bowler Constraints

$\sum_{j}(x_{2j} + x_{3j}) \geq X$ - Ensures atleast X bowlers (i.e. specialist bowlers or bowling all-rounders) are selected.

5. Keeper Constraint

$\sum_{j}(X_{4j}) = 1$ - Ensures a keeper is selected in the optimal team.

6. All-rounder Constraint

$\sum_{j}(x_{3j}) \geq 0$ - An all-rounders is automatically selected if the number of batsmen, bowlers and wicket-keepers exceed the allowable limits.

It was assumed the performance metrics that significantly influence team winningness are the same across formats (Chapter 3). However, the effect of each performance metric on winnin-

genss varies across formats. For example if batting metrics have a greater effect on winningness for T20 cricket relative to one day format, then T20 teams would prefer to have a 'batting' heavy team than a 'bowling' heavy or balanced team. Given such [team] variation across formats, the model constraints (i.e. batsmen, bowlers and all-rounders) required objective identification[3]. The constraints were formulated such that the 'optimal' team produces the greatest probability of winning.

7.4.3 Establishing Model Constraints

Given that model constraints are team orientated rather than individual player constraints, performance metrics that contribute significantly towards winningness at the team level, as opposed to the individual level, were established. The analysis datasets as described in chapter 4 were utilised, however the recorded performance metrics were aggregated and averaged on the 'team' variable. This manipulation step created IPL and CWC2011 datasets, containing both batting and bowling metrics with team level observations (Appendix E).

Applying a random forest technique to the CWC2011 team dataset the top 10 important performance metrics, for one day cricket, were: 1. *percentage boundaries (bowl)*, 2. *economy rate*, 3. *bowling average*, 4. *total wickets*, 5. *strike rate (bowl)*, 6. *batting average* 7. *total boundaries*, 8. *total dots*, 9. *total runs scored* 10. *games played*. The results indicate that bowling metrics are of greater importance than batting metrics, for winningness, among one day teams. The results show that seven out of the top ten metrics are bowling orientated, and are predominately geared around *run restrictions* and *wicket-taking efficiency* (interestingly the top metrics were geared to run restriction and wicket-taking efficiency) illustrating that strong run restricting and efficient wicket-taking abilities are necessary to increase a teams chance of winning a one day cricket match. The results indicate that model constraints, for one day cricket, should be formulated such that the optimal team has a greater bowling focus than batting focus. Additionally it can be inferred that bowling all-rounders are preferred over batting all-rounders, for one day cricket.

[3] Keeper constraints remain the same, regardless of format, as a wicket-keeper is always required.

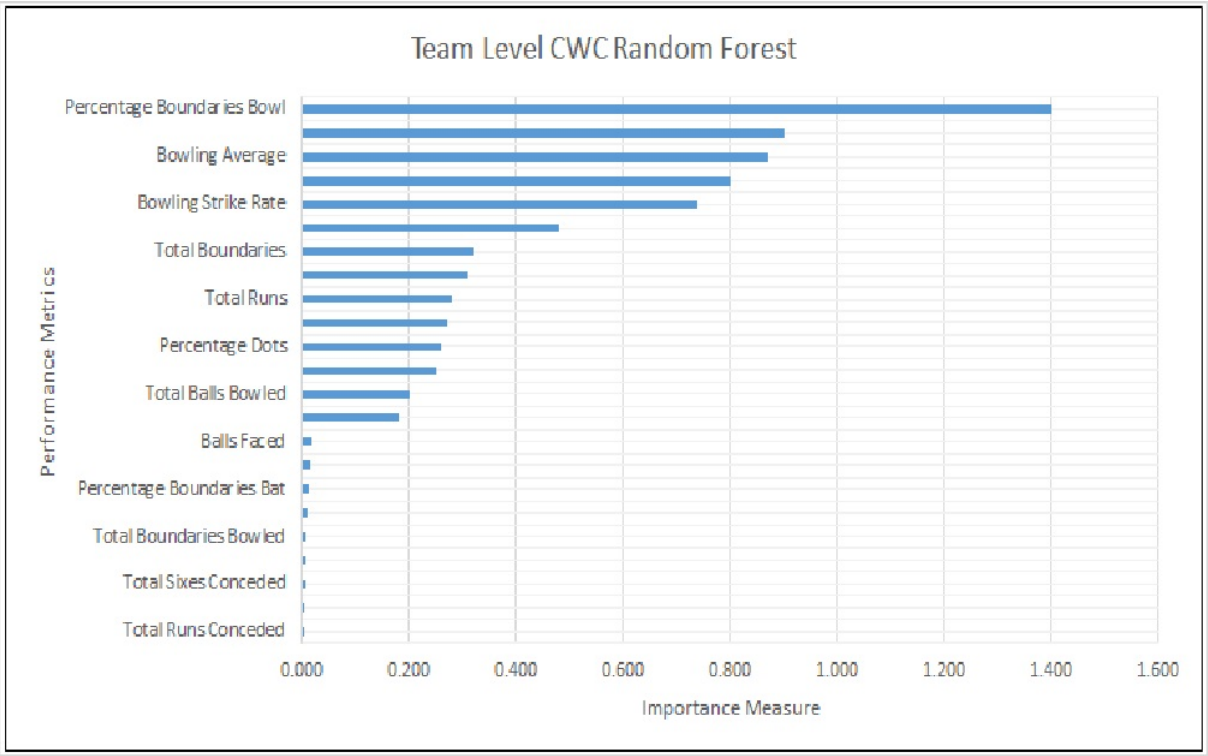

Figure 7.1: CWC Team Random Forest

Applying a random forest technique to the IPL team dataset the top 10 important performance metrics, for T20 cricket, were: 1. *strike rate (batting)* 2. *total runs scored* 3. *Total fours (batting)*, 4. *percentage boundaries (bat)*, 5. *total boundaries* 6. *batting average*, 7. *percentage boundaries (bowl)*, 8. *economy rate*, 9. *total dismissals (bat)*, 10. *total maidens*. These results indicate that batting metrics are of greater importance than bowling metrics for winningess among T20 teams. The results show that seven of the top ten metrics are batting orientated, and are predominately geared around scoring efficiency and consistency. It is revealed that batsmen with high scoring efficiency and scoring consistency are necessary to increase a teams chance of winning a T20 cricket match. Moreover, the results indicate that the model constraints, for T20 cricket, should be formulated such that the optimal team generated by the optimisation system has a greater batting focus than bowling focus. Additionally it can be inferred that batting all-rounders are preferred over bowling all-rounders, for T20 cricket.

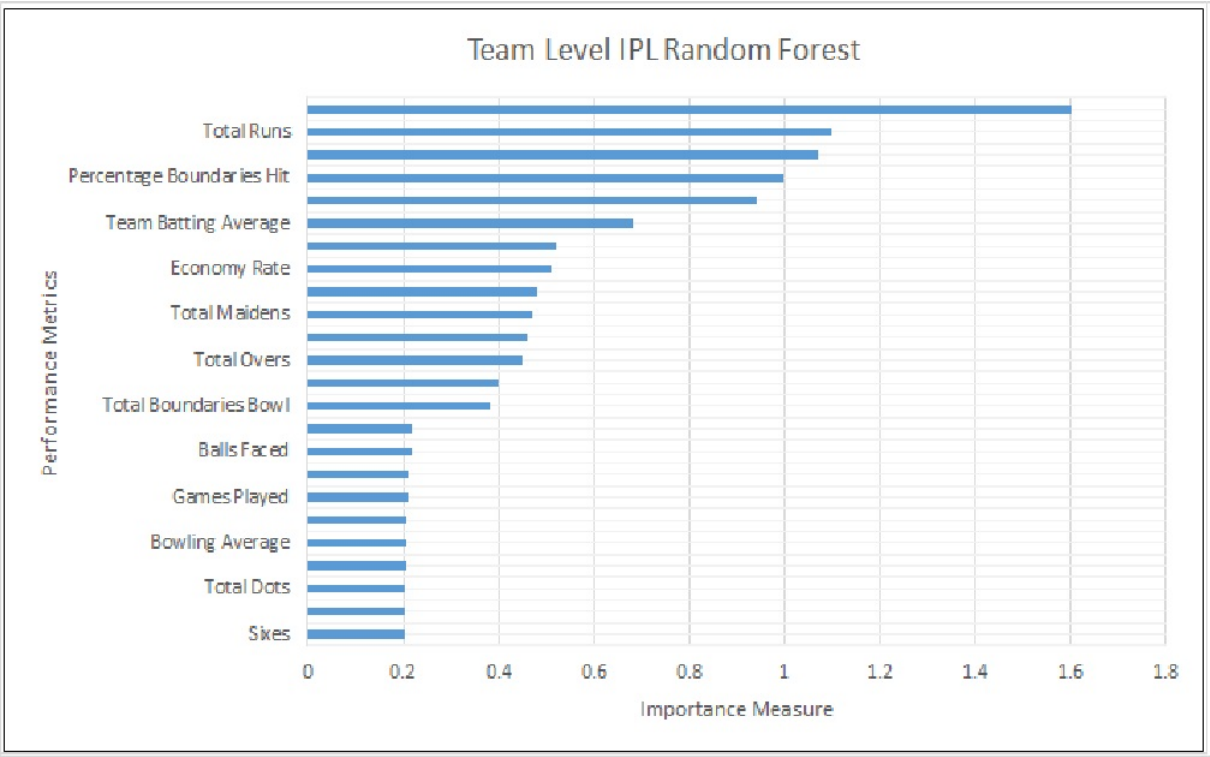

Figure 7.2: IPL Team Random Forest

Based on these findings the following model constraints were developed for T20 and one day cricket:

Table 7.1: Player-type Constraints

Player-type	**T20**	**One Day**
Batsmen	$\sum_j (x_{1j} + x_{3j}) \geq 7$	$\sum_j (x_{1j} + x_{3j}) \geq 5$
Bowler	$\sum_j (x_{2j} + x_{3j}) \geq 4$	$\sum_j (x_{1j} + x_{3j}) \geq 6$
All-Rounders	$\sum_j x_{3j} \geq 0$	$\sum_j x_{3j} \geq 0$
Wicket-Keeper	$\sum_j x_{4j} = 1$	$\sum_j x_{4j} = 1$

These model constraints obey the findings established above. The T20 constraints persuade the model to produce an optimal team with a heavy focus on batting ability, while the one day constraints persuade the model to produce an optimal team with a heavy focus on bowling ability.

7.4.4 Optimisation Formulation

Given these findings a *Discrete Deterministic Constrained* optimisation model with the following framework was adopted:

1. Objective Function

$$Z = \max \sum_{i=1}^{4} \sum_{j=1}^{n_i} c_{ij} x_{ij} \tag{7.7}$$

where c_{ij} is player rating for player j in role i,

$$where\, role\, i = \begin{cases} 1, & \text{if batting ability} \\ 2, & \text{if bowling ability} \\ 3, & \text{if all-rounder ability} \\ 4, & \text{if wicket keeping ability} \end{cases} \tag{7.8}$$

2. Decision Variable

$$x_{ij} = \begin{cases} 1, & \text{if } player\ j\ is\ selected\ for\ role\ i \\ 0, & \text{otherwise} \end{cases} \tag{7.9}$$

3. Model Constraints

Table 7.2: Overall model constraints

Constraints	**T20**	**One Day**
Team	$\sum_{i=1}^{4}\sum_{j=1}^{n_i} x_{ij} = 11$	$\sum_{i=1}^{4}\sum_{j=1}^{n_i} x_{ij} = 11$
Player	$\sum_{i=1}^{4} x_{ij} \leq 1$	$\sum_{i=1}^{4} x_{ij} \leq 1$
Batsmen	$\sum_{j}(x_{1j} + x_{3j}) \geq 7$	$\sum_{j}(x_{1j} + x_{3j}) \geq 5$
Bowler	$\sum_{j}(x_{2j} + x_{3j}) \geq 4$	$\sum_{j}(x_{2j} + x_{3j}) \geq 6$
All-Rounders	$\sum_{j} x_{3j} \geq 0$	$\sum_{j} x_{3j} \geq 0$
Wicket-Keepers	$\sum_{j} x_{4j} = 1$	$\sum_{j} x_{4j} = 1$

7.5 Chapter Remarks

This chapter identified an appropriate optimisation model for the adaptive rating system and objectively established the model constraints. Consequently, the chapter identified a solution to research flaw no. 3 (Chapter 3 , section 3.3). The following chapter assesses three individual rating systems, and identifies the system producing the most accurate team ratings and the greatest proportion of correct match outcomes. This is achieved by filtering the individual player ratings through the optimisation model and calculating the optimal team rating.

Chapter 8

Evaluating Individual Rating Systems

This chapter discusses and evaluates three individual player rating systems that were identified during the literature review process. The chapter outlines the *adaptive rating system* developed for predicting match outcome in the CPL2015 and CWC2015 competitions. The objective of this chapter is to establish the optimal individual rating system to implement into the *adaptive rating system* (i.e. optimisation system + individual rating system). The optimal [individual] rating method is defined as *the system that produces the greatest predictive accuracy, observed as the largest proportion of correct match outcomes* when integrated into the adaptive system. The 'optimal' individual rating system was identified by adopting the procedure outlined on the following page (section 8.2).

8.1 Optimal team rating using individual player ratings

The optimal team rating was calculated by aggregating individual player ratings. This aggregation approach was justified in [30]. It was stated that cricket is a sport characterised by one-on-one interactions between batsmen and bowlers, and that a players ability establishes the outcome of this interaction. Moreover the match outcome is defined by the interactions between batsmen and bowlers, therefore summing the individual player ratings provides a fair indication of team strength.

8.2 Establish the optimal team and optimal team rating

The development of the optimisation framework, the appropriate model constraints and the identification of important performance metrics, enabled implementation of the adaptive system to establish the optimal team and the associated team rating. However the 'optimal' individual rating system needed to be established first. The following process was applied to the Caribbean Premier League 2015 (CPL) and Cricket World Cup 2015 (CWC2015) competitions:

1. Extract scorecard data from the CPL and CWC2015 competition, once each team has played at least three games. This three game 'buffer' generates ratings that are indicative of a players 'true' underlying ability, since during the early stages of a competition players may be 'rusty [that is, not performing near their potential due to reduced recent game time].

2. For each competition split the scorecard data into four datasets (i.e. batsmen, bowlers, all-rounders and keepers) and calculate each players season performance metrics[1].

3. Calculate each players individual ratings, based on their player-type, by applying one of the following three individual rating systems (section 8.6 discusses method of application).

   (a) Analytical Hierarchy Process with Technical Order Preference by Similarity to Ideal Solution (AHP-TOPSIS) and Analytical Hierarchy Process with Complex Proportion Assessments (AHP-COPRAS)

   (b) Product Weighted Measure

   (c) Principal Component Analysis

4. Split the players by their respective team and merge by team. This creates a dataset for each team containing each players individual rating by player-type (Appendix F).

5. Input each teams player rating dataset through the Binary Integer Programming model and generate the optimal team. The output dataset contains 0's (i.e. not selected) and 1's (i.e. selected).

[1]The performance metrics calculated for each player-type are outlined in chapter 4, section 4.1.

6. Calculate the optimal team rating by summing the ratings of the selected players.

7. Apply the Bradley-Terry model to calculate the probability of team i beating team j.

$$\pi_{i,j} = \frac{R_i}{R_i + R_j}$$

This seven step iterative process was applied after every match played in the CPL and CWC2015 competition. At the end of every match individual player ratings were updated and the optimal team rating, for each team, was reproduced, using the *adaptive system* (i.e. adaptive team rating system = optimisation model + individual ratings). "This rating process represents an adaptive system as it updates player and team ratings based on historic performances upon the availability of data about current performances" [51, p.3]. Moreover, it was stated that adaptive systems provide the most suitable rating measures [73].

8.2.1 Bradley-Terry Model

The Bradley-Terry model predicts the outcome of a comparison. Given a pair of individuals i and j drawn "The Bradley-Terry model assumes that in a contest between two players, say player i and player j, $(i, j \in \{1, ..., k\})$, the odds that i beats j is $\frac{\alpha_i}{\alpha_j}$, where α_j and α_i are positive valued parameters which might be thought of as representing 'ability'" [40, p.1]. The model estimates the probability team i beats team j (i.e. pairwise comparison) using:

$$p(i > j) = \frac{p_i}{p_i + p_j}, \tag{8.1}$$

where p_i represents the relative ability (i.e. ratings) of object, or team in this instance, i. The outcome of a match depends on the current ability of the two competitors. It was stated in [51] that the outcome of many sporting disciplines can be determined by pairwise comparisons, and that the outcome of a match or game is dependent on the current ability of the two teams. Applying the Bradley-Terry methodology to the team ratings produces the probability of team i beating team j, addressing issue no. 4 (chapter 3, section 3.3).

The following sections of this chapter discusses the mathematical aspects of the three individual rating systems and the steps implemented to derive respective player ratings. Then each rating system is applied to the CPL and CWC2015 competitions. The ratings are then filtered through the optimisation model to generate the optimal team and the associated team rating. The accuracy of each *adaptive system* (i.e. optimisation system + individual rating method) will be benchmarked against the predictive accuracy of the CricHQ algorithm[2] and TAB outcomes. The T.A.B was utilised as a benchmarking tool as it incorporates collective opinion and subjectively evaluates risk, while the CricHQ algorithm incorporates objective measures to evaluate risk. Moreover, since the CricHQ algorithm incorporates 'macro' variables such as past match results, venue and opposition, it was possible to test the research hypothesis: *team based ratings system that consider 'micro' variables (i.e. individual player ability) should outperform rating systems that only consider 'macro' variables*. The predictive accuracy of the adaptive system will validate:

1. The player and team rating system
2. The model constraints
3. The use of the important performance metric
4. The aggregation method applied to generate team ratings

8.3 Analytical Hierarchy Process

The Analytical Hierarchy Process (AHP) is a multi-criteria decision making tool developed by Thomas Saaty [64]. Given a user defined pairwise comparison matrix, the AHP translates the matrix into a vector of relative weights for each criterion element (i.e. performance metrics) using a mathematical model. The pairwise comparison matrix provides a numerical comparison of each attribute with respect to the other attributes being evaluated. These matrix entries are determined using the fundamental AHP scale (table 8.1) and are based on prior experience or expert knowledge. Applying the AHP to the pairwise comparison matrix translate the subjective weights into objective weights, representing the importance of the attribute relative to

[2] [15] outlines the algorithm.

the other attributes. Moreover the method implements a consistency measure for each attribute to ensure that the 'user' defined weights are consistent and reduces *bias* in the decision making process. " The aim is to provide the decision maker a precise reference in order to make adequate decisions and reduce the risk of making biased decisions by decomposing the problem into a hierarchy of more easily comprehended sub-problems" [70, p.74]. According to [6, p.4] the follows steps are computed to conduct an AHP:

1. Compute the value of criteria weights

The user defines an $n \times n$ pairwise comparison matrix, A, where n represents the number of evaluation criteria (i.e. performance metrics) [3]. Each a_{ij} entry evaluates the importance of performance metrics i with respect to j. The entries a_{ij} and a_{ji} must satisfy: $a_{ij} \times a_{ji} = 1$, while criteria with the same level of importance must satisfy: $a_{ij} = a_{ji} = 1$. The importance of criteria i relative to j can be established via the fundamental scale of the AHP:

Table 8.1: Fundamental Scale of AHP

Value of a_{ij}	**Interpretation**
1	i and j are equally important
3	i is slightly more important than j
5	i is more important than j
7	i is strongly more important than j
9	i is absolutely more important than j

2. Synthesis Judgement

Derive the normalized pairwise comparison matrix, A_{norm}, by equating the sum of column entries to 1. The entries in matrix A_{norm} are computed as:

$$\bar{a}_{ij} = \frac{a_{ij}}{\sum_{j=1}^{n} a_{ij}} \quad \forall\, j = 1, 2, \ldots, n$$

[3] The evaluation criteria were the significant performance metrics outlined in table 6.2 (chapter 6).

3. Create a criteria weight vector

$$w_i = \frac{\sum_{j=1}^{n} \bar{a}_{ij}}{n} \quad \forall\, i = 1, 2, \ldots, n$$

A relationship exists between the pairwise comparison matrix A and the weights vector, w, such that $Aw = \lambda_{\max} w$. The maximum eigenvector $\lambda_{\max}$ can be found by computing a consistency check:

$$CV_i = \frac{\sum_{j=1}^{n} a_{ij} \times w_j}{w_j}, \forall\, i = 1, 2, \ldots, n$$

and dividing the summation of consistency check values by, n, the number of criteria:

$$\lambda_{\max} = \frac{\sum_{i=1}^{n} CV_i}{n}$$

4. Consistency check of pairwise comparison matrix

The $\lambda_{\max}$ parameter enables the derivation of a consistency ratio (CR) which validates the consistency of the estimated vector:

$$CI = \frac{\lambda_{\max} - n}{n - 1}$$

$$CR = \frac{CI}{RI}$$

n	2	3	4	5	6	7	8	9	10
RI	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.51

The random index[4] (RI) is dependent on n. If $CR \leq 0.1$ then the values of subjective judgement (i.e. pairwise comparison matrix) and the weights generated in step 3 are regarded as acceptable.

[4]The random index (RI) is a predetermined table produced by Thomas Saaty.

8.3.1 TOPSIS

Developed by Hwang and Yoon [22], TOPSIS is a multi criteria decision making tool which evalutes various options (i.e. players) based on their similarity (i.e. distance) to the optimal solution by generating weights using the AHP and loading the weights into the TOPSIS process. Formally, TOPSIS " is a multiple criteria method to identify solutions from a finite set of alternatives. The basic principle is that the chosen alternative should have the shortest distance from the ideal solution and the farthest distance from the negative-ideal solution" [46, p.1138]. The 'optimal' alternative (i.e. player) has the shortest geometric distance from the positive ideal solution and the longest geometric distance from the negative ideal solution. According to [6, p.5] the following steps are computed to conduct an AHP-TOPSIS:

1. Form a decision matrix:

$$D = \begin{array}{c} \\ L_1 \\ L_2 \\ \vdots \\ L_m \end{array} \begin{array}{c} \begin{array}{ccccc} C_1 & C_2 & C_3 & \dots & C_n \end{array} \\ \begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{i2} & x_{i3} & \dots & \dots & x_{mn} \end{pmatrix} \end{array}$$

where $L_1, L_2, ..., L_m$ represents each player and $C_1, C_2, ..., C_n$ represents the criteria. Moreover, i represents the criteria (i.e. performance metric) index $\{i = 1, ..., m\}$, m is the number of alternatives (i.e. players) index. In the context of this research, D, represents the season performance metric dataset (Appendix C). x_{ij} represents performance metric j for player i.

2. Normalise the decision matrix:

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}}$$

where r_{ij} represents the relative performance of player i, $(i = 1, ..., m)$ for performance metric j, $(j = 1, .., n)$.

3. Construct the weighted normalized decision matrix:

Multiply, **r**, the normalized decision matrix by the associated weights, w_j, representing the AHP weight for performance metric j:

$$v_{ij} = r_{ij} \times w_j$$

4. Determine the positive and negative ideal solution:

Determine the worst alternatives and the best alternatives

$$Positve\ Ideal\ Solution;\ A^+ = \{v_1^+, v_2^+, ..., v_n^+\}$$

$$v_j^+ = [(\max v_{ij} | j \in J), (\min v_{ij} | j \in J')]$$

$$Negative\ Ideal\ Solution;\ A^- = \{v_1^-, v_2^-, ..., v_n^-\}$$

$$v_j^- = [(\min v_{ij} | j \in J), (\max v_{ij} | j \in J')]$$

$J = \{j = 1, 2, ..., n | j\}$ is associated with the benefit criteria and $J' = \{j = 1, 2, ..., n | j\}$ is associated with cost criteria.

5. Calculate the separation measure:

Calculate the distance between the largest alternative i and the worst alternative:

$$S_i^+ = \sqrt{\sum_{j=1}^{n} (v_j^+ - v_{ij})^2}$$

Calculate the distance between the alternative i and the best condition:

$$S_i^- = \sqrt{\sum_{j=1}^{n} (v_j^- - v_{ij})^2},$$

where $i = 1, 2, 3, ..., m$

6. Calculate the relative closeness to the ideal solution:

Measure the relative closeness of each player to the ideal solution:

$$C_i = \frac{S_i^-}{S_i^+ + S_i^-}$$

The larger the C_i the better the performance of the alternatives (i.e. player). This step calculates the similarity to the worst solution.

The disadvantage of the AHP-TOPSIS is that it either maximises or minimises performance metrics. The technique lacks the ability to evaluate both maximising and minimising performance metrics.

8.3.2 COPRAS

COPRAS is a multi-criteria decision making tool utilised to evaluate both maximizing and minimising criteria values. Introduced in [81] to solve various problems in the construction industry, "The COPRAS method uses a stepwise ranking and evaluating procedure of the alternatives in terms of significance and utility degree." [63, p.24]. The technique generates weights using the AHP and inputs them into the COPRAS method. According to [63, p.25-27] the following steps are computed to conduct an AHP-COPRAS:

1. Form a decision matrix:

$$D = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & x_{d3} & \dots & x_{dn} \end{bmatrix}$$

where d represents the number of players and n represents the number of performance metrics.

2. Normalise the decision matrix:

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2}}$$

3. Construct the weighted normalized decision matrix:

Multiply, **r**, the normalized decision matrix by the associated weights, w_{ij}, which were calculated through the AHP:

$$v_{ij} = r_{ij} \times w_j$$

4. Calculate sums, R_i, of attributes values for which larger values are preferred:

$$P_i = \sum_{j=1}^{K} v_{ij},$$

where K is the number of attributes which must be maximized (i.e. beneficial criteria).

5. Calculate sums, R_j, of attribute values for which smaller values are preferred:

$$R_j = \sum_{j=k+1}^{n} v_{ij},$$

where $n - k$ is the number of attributes which must be minimized (i.e. non-beneficial criteria).

6. Calculate the relative weights of each alternative, Q_i:

$$Q_i = P_i + \frac{\sum_{j=1}^{n} R_j}{R_j \sum_{i=1}^{n} \frac{1}{R_j}}$$

Note: Q_i refers to player i and performance metric j.

7. Calculate the utility degree of each alternative, N_i:

$$N_i = \frac{Q_i}{Q_{\max}} \times 100\%$$

8.4 Product Weighted Measure

The Product Weighted Measure (PWM) was proposed in [29]. The author developed four individual rating methods for the four different player-types (i.e. batsmen, bowlers, all-rounders and keepers). The method produces *raw ratings* for each player and then calculates the *actual ratings* relative to other players within their player-type. The rating calculations for each player-type are calculated as follows:

8.4.1 Batsmen Ratings

Batting ratings were calculated using eqn. 8.2 and eqn. 8.3:

$$U_{1j} = (Y_{1j}^{\alpha_1})(Y_{2j}^{\alpha_2})(Y_{3j}^{1-\alpha_1-\alpha_2}), \tag{8.2}$$

where U_{1j} represents the raw ratings for batsmen j using batting performance metrics Y_{1j} Y_{2j} and Y_{3j}, while $\alpha_i's$ represents the importance weightings allocated to each performance metric (see subsection 8.6.3). The batting performance metrics ($Y_{1j}, Y_{2j}\ and\ Y_{3j}$) vary across formats, please refer to table 8.2 and 8.3 to see which metrics were utilised. The raw ratings, U_{1j}, were then scaled to produce *actual* batsmen ratings relative to other batters in the league:

$$C_{1j} = \frac{U_{1j}}{\sum_{j=1}^{n} U_{1j}} \times n, \tag{8.3}$$

where n represents the number of batsmen in the competition.

8.4.2 Bowler Ratings

A bowlers rating was calculated via eqn. 8.4:

$$U_{2j} = (Y_{4j}^{\alpha_1})(Y_{5j}^{\alpha_2})(Y_{6j}^{1-\alpha_1-\alpha_2}), \tag{8.4}$$

where U_{2j} represents the raw ratings for bowler j using bowling performance metrics, Y_{4j}, Y_{5j} and Y_{6j}. The bowling performance metrics ($Y_{4j}, Y_{5j}\ and\ Y_{6j}$) vary across, please refer to table 8.2 and 8.3 to see which metrics were utilised. The optimisation model outlined in the previous chapter incorporates a maximisation objective function and 'low' values of Y_{3j}, Y_{4j} and Y_{5j}, indicate 'good' bowlers. As such the U_{2j} values were scaled, such that higher values represent 'good' bowlers, using a technique outlined in [42]:

1. $V_{2j} = K - (\frac{U_{2j}}{\sum_{j=1}^{n} U_{2j}})$, where K is a positive value such that $K - (\frac{U_{2j}}{\sum_{j=1}^{n} U_{2j}}) > 0$

2. The bowler ratings were defined as: $C_{2j}^{1} = \frac{V_{2j}}{\sum_{j=1}^{n} V_{2j}} \times n_2$, where n_2 represents the number of bowlers in a competition. This transformation ensured that higher ratings represented a better bowler.

Moreover, to ensure that the bowler ratings, C_{2j}^{1}, had an equivalent variance compared to the batting ratings, the bowler ratings were scaled using a technique outlined in [52]:

$$C_{2j}^{p+1} = C_{2j}^{p\ \frac{\sigma_{C_1}}{\sigma_{C_2}^{p}}} \tag{8.5}$$

where σ_{C_1} and $\sigma_{C_2}^{p}$ represents the standard deviation of the batting ratings and standard deviation of the bowler ratings for the pth iteration, respectively. To ensure equivalent spread of the batting and bowling ratings, equation 8.5 is an iterative process which stops when it has converged to an accepted lower limit, therefore $C_{2j}^{p+1} = C_{2j}$.

8.4.3 All-rounder Ratings

All-rounder ratings were calculated by multiplicatively combining their batting and bowling ratings:

$$C_{3j}^{1} = (C_{1j}^{\beta})(C_{2j}^{1-\beta}), \tag{8.6}$$

where C_{1j} and C_{2j} represents the batting and bowling ratings, respectively, and β represents the weightings associated with the batting and bowling ratings. The scale adjusted measure (eqn. 8.5) was also applied to the all-rounder ratings, C_{3j} to ensure equivalent spread.

8.4.4 Wicket-Keepers Ratings

Wicket-keepers were treated as batsmen and therefore their ratings were calculated using the method specified in section 8.4.1. Due to data limitations wicket keeper metrics such as *byes* and *catches* could not be utilised to derive ratings.

8.5 Principal Component Analysis

The coefficients of the first component, for each player-type, are used to weight each of the performance metrics, $\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_i x_{ij}$, where λ_i represents the component coefficient for performance metric i and x_{ij} represents the value for metric, i, for player j. The steps to conduct a PCA were outlined in Chapter 6, section 6.1 (page 49).

8.6 Application of Individual Rating Systems

The following performance metrics were utilised to rate each player-type across formats[5]:

Table 8.2: Performance metrics for one-day cricket by player-type

One Day			
Batsmen	**Bowlers**	**All-rounders**	**Wicket-Keepers**
Batting Average	Economy Rate	Batting Average	Batting Average
Total Runs Scored	Percentage Boundaries	Total Runs Scored	Total Runs Scored
Batting Strike Rate	Bowling Strike Rate	Batting Strike Rate	Batting Strike Rate
		Economy Rate	
		Percentage Boundaries	
		Bowling Strike Rate	

Table 8.3: Performance metrics for T20 cricket by player-type

T20			
Batsmen	**Bowlers**	**All-rounders**	**Wicket-Keepers**
Total Runs Scored	Economy Rate	Batting Average	Total Runs Scored
Percentage Boundaries	Percentage Boundaries	Percentage Boundaries	Percentage Boundaries
Batting Strike Rate	Bowling Strike Rate	Batting Strike Rate	Batting Strike Rate
		Economy Rate	
		Percentage Boundaries	
		Bowling Strike Rate	

8.6.1 Analytical Hierarchy Process

A relevant application of the AHP in a sporting context was applied to 16 soccer teams in Israel's National League to predict team rankings [70]. Using facility quality, coach level, player levels, fans, previous season performance and current performance, an expert defined pairwise comparison matrix was created and AHP weights, for each criteria, were generated. AHP-TOPSIS and AHP-COPRAS were applied to rank IPL (2012) players [33].

The AHP pairwise comparison matrices for each player-type for each competition were developed by ex first-class cricketer and Wellington Firebirds selector, Jason Wells[6].

[5]The individual rating methods were applied to each player by player-type after every game in the CPL and CWC2015 competition.

[6]73 First class matches and 81 List A games between 1989 and 2001.

AHP-TOPSIS to rank Batsmen, Bowlers and Wicket-Keepers

As mentioned, the TOPSIS method finds solutions from a finite set of alternatives that simultaneously minimise the distance from an ideal solution and maximises the distance from a negative ideal solution [69]. To determine the *ideal solution* for batsmen and wicket-keepers the positive ideal solution, A^+, was implemented since their performance metrics were benefit criteria (i.e. higher values represent better batsmen). The negative ideal solution, A^-, was applied to rate bowlers since their performance metrics were cost criteria (i.e. lower values represent better bowlers) $\Rightarrow$ the idea is to reduce cost. The relative closeness, C_i, represents player $i's$ rating at the end of match k, for each competition.

AHP-COPRAS to rank all-rounders

The AHP-COPRAS technique was utilised to evaluate projects (i.e. players) with criteria (i.e. metrics) that must be maximised and minimised to produce sensible ratings. Given these aspects the technique was applied to all-rounders, as both batting (i.e. benefit criteria) and bowling (i.e. cost criteria) performance metrics identify an all-rounders ability. The degree of utility, N_i, represents each players rating at the end of match k, for each competition. Higher Values of N_i indicate better all-rounders.

8.6.2 Principal Component Analysis

The PCA ranking method was utilised in [57] to rate batsmen and bowlers in the IPL (2012). The author claimed that if the first principal component explained at least 70% of variation, the component coefficients could be used to weigh the associated player performance metrics and produce a player rating, representing a type of weighted average, $R_i = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \therefore +$. However, the methodology outlined in [57] ignored all-rounders and wicket-keepers, and the performance metrics implemented were selected in an ad-hoc manner. A new principal component analysis was conducted on each dataset (i.e. batsmen, bowlers, all-rounders and wicket-keepers), across the two competitions after every match.

8.6.3 Product Weighted Measure

The Product Weighted Measure (PWM) was developed and applied in [29] to rank batsmen, bowlers, wicket-keepers and all-rounders in international one day cricket. However the performance metrics used to rank the players were selected in an adhoc manner, and the weightings, α, were subjectively chosen. As mentioned previously the importance of each performance metrics on winningness varies across T20 and one day cricket. It was established that T20 cricket is a batsmen orientated game, with greater preference for highly scoring efficient batsmen. Given the difference in importance of each performance metrics across formats, the author introduced a novel method for determining the appropriate weightings, α, for each important performance metric, for each player-type, across formats.

Random Forest + AHP Weightings

The system for determining the appropriate weightings, α, is outlined as follows:

1. Identify the order of importance for each performance metric, for each player-type, across the two formats. The order of importance for each performance metric is established by the random forest(RF) importance plot, for each player-type, across formats.
2. Use the RF order of importance plot to create an $n \times n$ pairwise comparison matrix, for each player-type, where each entry, a_{ij} represents the importance of criteria i with respect to j. The relative importance of each performance metric, a_{ij}, follows the logic (i.e. importance order) established by the random forest importance plot. For example, if *percentage boundaries* are of greater importance to winningness than *batting average*, among batsmen, the relative importance of percentage boundaries vs. batting average $>$ 1. A pairwise comparison matrix was produced for each player-type and their associated performance metrics, across T20 and One Day cricket. The T20 and one day pairwise comparison matrices, for each player-type, can be found in appendix F[7]. The order of importance for each performance metric, across each format, was established through the Random Forest importance plot in chapter 7 (figure 7.1 and figure 7.2).

[7] As mentioned in chapter 7 wicket-keepers are treated as batsman, across both formats, therefore the batsman and keepers comparison matrices are identical.

3. Run the AHP on the pairwise comparison matrices and generate the weights associated with each performance metric for each player-type[8]. The following α weightings were generated:

Table 8.4: AHP performance metric weightings by player-type for one-day cricket

One Day Cricket, $alpha$, weightings				
Performance Metrics	**Batsmen**	**Bowlers**	**All-rounders**	**Wicket-Keepers**
Batting Average	0.35	-	0.35	0.35
Total Runs Scored	0.38	-	0.38	0.37
Batting Strike Rate	0.27	-	0.27	0.28
Percentage Boundaries	-	0.32	0.35	-
Bowling Strike Rate	-	0.27	0.27	-
Economy Rate	-	0.41	0.38	-

Table 8.5: AHP performance metric weightings by player-type for T20 cricket

T20 Cricket, $alpha$, weightings				
Performance Metrics	**Batsmen**	**Bowlers**	**All-rounders**	**Wicket-Keepers**
Total runs scored	0.33	-	0.34	0.33
Percentage Boundary (batsmen)	0.30	-	0.30	0.30
Batting Strike Rate	0.37	-	0.36	0.37
Percentage Boundary (bowler)	-	0.30	0.35	-
Bowling Strike Rate	-	0.33	0.27	-
Economy Rate	-	0.37	0.38	-

The weightings represented in table 8.3 and 8.4 align with findings established in the previous chapter (section 7.4, page 77-79), stating that a winning T20 team requires players with high scoring efficiency, high scoring consistency and high run restricting ability. A winning one day team requires players with high run restriction ability, high wicket-taking efficiency and high scoring consistency.

The weightings, β, allocated to an all-rounders ability were dependent on the type of all-rounder being evaluated, for example, if the player being ranked was a batting all-rounder, C_1 (i.e. batting rating), received a weighting of 0.60, while bowling all-rounders had a bowling rating, C_2, weight of $\beta = 0.60$. The logic behind this modification was that even though all-rounders

[8] AHP was executed using the *ahp()* function in *library(pmr)*.

are capable of batting and bowling, majority of all-rounders are stronger in one aspect relative to the other, and therefore should be recognised accordingly.

8.6.4 Application Results

Applying the individual rating methods and optimisation system (i.e. adaptive system) across the CPL and CWC2015 competitions, and applying the adaptive rating system, for each match, the following results were produced[9].

Table 8.6: Adaptive System Accuracy of match predictions

Adaptive System Results			
Competition	**PWM**	**AHP**	**PCA**
Cricket World Cup 2015	76%	65%	35%
Caribbean Premier League 2015	70%	61%	30%

Table 8.4 represents the accuracy of each adaptive rating system. The accuracy was calculated via:

$$System\ Accuracy = \frac{number\ of\ correct\ outcomes}{total\ matches\ played}$$

8.7 Optimal team vs. Playing team

Applying the adaptive rating system across the two competitions, highlighted that on occasion the optimal team generated by the optimisation model would differ from that selected by coaches and managers; meaning the 'optimal' team rating would not relate to the playing team. To counter this issue, rather than using the optimal team rating, the player ratings of those selected by coaches, were aggregated to generate a team rating. Even though this did not represent the *optimal* team rating it did provide an indication of strength of the playing team.

[9]The adaptive system was 'run' for 23 CPL matches and 34 CWC2015 matches.

8.8 Adaptive Rating System flaws

8.8.1 PCA ranking flaws

The results of the adaptive system utilising the PCA method produced the worst prediction accuracy. The reason for poor results was that in majority of the matches, across CPL and CWC2015, the PCA method was not applicable, as the first component failed to explain at least 70% of the variation, predominately for bowlers and all-rounders. Additionally, on occasion, the component coefficients had a counter-intuitive direction effect, for example in some instances coefficients associated with the economy rate, λ_{ER}, would have a positive effect (+), while strike rate and percentage boundaries would have negative coefficients (-). This produced counter-intuitive rankings as all component coefficients were required to have the same direction effect, across each player-type.

8.8.2 AHP-TOPSIS and AHP-COPRAS flaws

Although the results produced by the adaptive system utilising AHP-TOPSIS and AHP-COPRAS were considerably better the methods frequently produced rankings that either over or under rated a players ability.

8.8.3 Product Weighted Measure flaws

Although the adaptive system utilising the PWM produced the best predictive accuracy the author did identify method flaws. Similar to the AHP-TOPSIS and AHP-COPRAS, the system had the tendency to 'over-rate' players with strong performance metrics, especially during the early stages of a season. If a player had an abnormally *good* start to the season relative to the others within their player-type the ratings produced were too high. To counter this issue the performance metrics were scaled and bound between 0 and 1.

Another flaw to the PWM was that it failed to produce ratings for all-rounders who *only* participated in either a batting or bowling capacity. It was found that during the early stages of a

season each player only has a few opportunities to fulfil their role; for example an all-rounder may only need to bat or bowl, but not both. This creates situations where an all-rounder can significantly contribute towards a match outcome, but ratings are not produced as only one ability was utilised. This produced under-rated players and teams. To counter this issue the PWM was modified as follows:

1. If a batting all-rounder has not taken a wicket, during the season, the players batting rating, C_1, is regarded as their all-rounder rating.
2. If a bowling all-rounder has not batted, during the season, but did bowl, the players bowling rating, C_2, is regarded as their all-rounder rating.

As a validation method the modified PWM was applied to CPL and CWC2015 matches, the method offered slight improvements, predicting correctly 74% and 82% of matches, respectively, outperforming the TAB and CricHQ algorithm.

Predictive Systems			
Competition	**TAB**	**CricHQ**	**Adaptive System**
Cricket World Cup 2015	71%	76%	82%
Caribbean Premier League 2015	49%	62%	74%

8.9 Forecasting Methods

Since the PWM ratings are generated relative to the sum of the other ratings, for a given player-type, this enables the ability to track player performance on a match-by-match basis, and assesses a players progression as the season matures. This increases the *adaptive* nature of the developed rating system. The time-stamped ratings enabled the application of forecasting methods to player ratings. An area for future research is exploring optimal forecasting methods.

In [31] Exponentially Weighted Moving Average (EWMA) control charts were applied to individual batting performances. The study results appeared to produce sensible performance predictions. Moreover, exponential smoothing was applied in [24] to predict tennis player ratings. It was found that exponential smoothing produced predictive player ratings. [17] utilised

control charts to monitor batting performances of New Zealand domestic cricketers, and found that control charts such as EWMA accurately forecasted a batsmen's *form*.

8.9.1 Exponentially Weighted Moving Average

According to [75] the formal definition for EWMA test statistic is given by:

$$z_t = \alpha \bar{x}_t + (1 - \alpha) z_{t-1},$$

where α is a constant weight, representing the level of importance placed on current observations, $\bar{x}_t$ is the sample mean at time t, and z_{t-1} is the test statistic from time $t - 1$. "Exponentially Weighted Moving Averages (EWMA) are known for exhibiting optimal properties for some forecasting and quality control applications" [75, p.1]. The technique averages the data and allocates less and less importance to older observations. In the context of this research EWMA is adopted to forecast player and team ratings and measure their quality (i.e. form).

EWMA Application

The EWMA methodology was embedded into the developed adaptive system allocating an α of 0.72. This method predicted a players rating for the following match, and filtered the predicted ratings through the optimisation system to generate a forecasted team rating. Applying this method to the CPL and CWC2015 matches the following predictive accuracy was established:

Predictive Systems			
Competition	**TAB**	**CricHQ**	**Adaptive System**
Cricket World Cup 2015	71%	76%	86%
Caribbean Premier League 2015	49%	62%	80%

8.10 Chapter Remarks

Given the predictive accuracy of the modified PWM + EWMA individual rating method, it was established as the optimal player rating method. Moreover, the results generated by the

adaptive rating system validate the optimisation framework, model constraints, the utilisation of important performance metrics and the aggregation method applied to generate the optimal team rating.

This chapter identified an appropriate method to derive the optimal team rating as a function of individual player ratings. Moreover a technique was established to calculate the probability of team i beating team j, as a function of overall team ratings. Finally a technique to produce appropriate weights, to allocate to each [important] performance metric for the PWM rating system, was established. Consequently, this chapter remedied research flaw no. 2, 4, 5, 6 and 7 (Chapter 3, section 3.3, page 31-32).

The work presented throughout this chapter has been accepted for the 13th ANZIAM MATH-Sport Conference under the authorship of Patel, Bracewell and Rooney to be published in 2016.

Chapter 9

Future Research, Discussion and Conclusion

9.1 Further Research

Although the developed adaptive rating system produces high predictive accuracy there are situations in which the system falls short. Future research may address these model flaws:

1. Undefined Ratings for bowler and all-rounders with no wickets

During the early stages of a season a bowler is less likely to take wickets due to the bowling opportunities received or "rusty" playing ability. However because a bowlers and an all-rounders ratings are a function of a 'wicket-taking' metric (i.e. bowling strike rate, bowling average etc.), if these player-types fail to produce wickets their ratings would be undefined (i.e. N/A). However just because a bowler does not take a wicket it does not mean the player has failed to make a significant impact or contribution to the team rating. As seen in Chapter 7 a bowlers ability to restrict runs is considered more important than wicket-taking efficiency in terms of winningness for one day cricket.

Utilising reject inference (a technique primarily found in Banking and Finance journals for building credit scorecards) [14] developed a method of inferring a bowlers strike rate, given that the player has not taken a wicket. Therefore its is suggested that future itera-

tions of the adaptive rating system implement the reject inference technique to generate strike rate for non-wicket taking bowlers, enabling the derivation of a players rating.

2. Comparing teams across divisions

Applying the adaptive system to the Cricket World Cup 2015 the author identified two stages in which the model produced the least number of correct match outcomes:

1. The *early stages of the competition* where players have not played a sufficient number of games in order for the individual rating system to produce appropriate ratings, or ratings that are indicative of a players true ability.
2. During the *post round robin matches* (i.e. quarters, semi-finals and finals) in which teams across pools (i.e. divisions) compete. It was found that the modified PWM + EWMA method produced ratings relative to other players within their *player-type* and *division*, and therefore comparing players and teams across divisions is inappropriate. To rectify this issue the author suggests adopting the recalibration method developed in [47]. This method allows the comparison of players and teams across division, by accounting for the strength of each division and appropriately recalibrates ratings. Adopting such methods would allow the user to apply it to competitions with a divisional structure such as the NATWEST T20 competitions (i.e. UK T20 competition) potentially increasing the predictive power of the adaptive rating system.

9.2 Discussion

The lack of academic literature surrounding team rating systems utilising individual ability within cricket, the absence of the application of predictive techniques to forecast match outcome and the growing popularity of sports betting, established an entry point in the market for this research. This research successfully developed a roster-based optimisation model (i.e. adaptive rating system), for limited overs cricket. The developed system incorporates an optimisation framework and individual player rating system.

The research hypothesised that a team based [adaptive] rating system, accounting for individual player performances would outperform systems that only consider ‘macro’

variables such as opposition, venue, past performances home advantages etc. An adaptive system should possess greater ability to account for a larger proportion of variation in match outcomes. The results presented through this research validated the hypothesis.

Although the application of a Binary Integer Programming technique for optimal team selection within cricket had been previously researched ([42], [68]), the research framework lacked depth and statistical rigour:

1. Model constraints and performance metrics were selected in an ad-hoc fashion.
2. Inability to validate the 'optimal' team generated by the optimisation technique.
3. Inability to generate the probability of winning.
4. Inability to generate an overall team rating measure.

Through this research each issue was addressed in an objective manner and validated through application. Moreover addressing these issues the author developed an adaptive rating system which successfully incorporates an individual rating system with an in-built forecasting technique, and an optimisation system that generates the optimal team. The 'optimal' individual rating system was established by applying 3 different systems and evaluating each systems predictive accuracy.

Applying the adaptive system to the Caribbean Premier League (2015) and Cricket World Cup (2015) revealed its ability to outperform well-known predictive algorithms. The results validate the choice of performance metrics that were used to evaluate a players rating, the weights allocated to each performance metrics, for each player-type, and the optimisation constraints that were formulated, reflecting the team make-up required to win a limited overs match. Moreover the adaptive system was regarded as successful, as "a successful predictive system can select the correct outcome 17% better than for random chance in professional sports" [73, p.38]. Given that the developed system generated predictive accuracy well above this threshold, the research was considered successful.

Given Cricket's exponential growth into a multi-billion dollar industry, it has become more critical than ever to introduce analytical methods for team selection. The adaptive system is useful for decision making among coaching and managerial staff, in terms of player selection, and can be implemented to identify the optimal team for limited overs

cricket. The adopted forecasting method is the differentiating factor of the developed system as it accurately depicts a players projected rating/ performance for the upcoming match and enables the coaching staff to make selections accordingly. Moreover, the system allows coaching staff to evaluate an opponents projected performance, at both the team and individual level.

9.3 Conclusion

Due to the nature of human contest, sport lends itself to fluctuations and discrepancies in game outcomes. This in turn generates interest. However, given the monetary growth of the sporting and sports betting industry over the past decade, there are strong incentives for managers, coaches and players to accurately measure and monitor performance, and understand the root cause of match outcome fluctuations. Key stakeholders can not solely rely on subjective views and personal beliefs to make team and player selection decisions.

One sport which has recently seen an exponential rise in the use of objective rating systems to make informed and strategic decisions regarding player and team performances is cricket. Cricket is an ideal sport to isolate individual team member contribution with respect to winning. This is due to the volume of digital data available, combined with the relatively isolated nature of the batsman versus bowler contest observed per ball.

The objective of this research was to develop a roster-based optimisation system for limited overs cricket by deriving a meaningful, overall team rating using a combination of individual ratings from a playing eleven. The research hypothesis was that an adaptive rating system accounting for individual player abilities, outperforms systems that only consider macro variables such as home advantage, opposition strength and past team performances. The assessment of system performance is observed through the prediction accuracy of future match outcomes. This is based on the expectation in elite sport that better teams are expected to win more often. To test the hypothesis, an adaptive rating system was developed. This framework was a combination of an optimisation system and an individual ratings system. The adaptive rating system was selected due to its ability to update player and team ratings based on past performances.

A Binary Integer Programming model was the optimisation method of choice, while a

modified product weighted measure (PWM) with an embedded exponentially weighted moving average (EWMA) functionality was the adopted individual rating system. The weights for this system were created using a combination of a Random Forest and Analytical Hierarchical Process. The model constraints were objectively obtained by identifying the player's role and performance outcomes a limited over cricket team must obtain in order to increase their chances of winning.

Utilising a random forest technique, it was found that players with strong scoring consistency, scoring efficiency, runs restricting abilities and wicket-taking efficiency are preferred for limited over cricket due to the positive impact those performance metrics have on a team's chance of winning. These practically significant variables reinforce the relevance of the findings as they intuitively make sense. In order to define pertinent individual player ratings, performance metrics that significantly affect match outcomes were identified. Random Forests proved to be an effective means of optimal variable selection. The important performance metrics were derived in terms of contribution to winning, and were input into the modified PWM and EWMA methods to generate a player rating.

The underlying framework of this system was validated by demonstrating an increase in the accuracy of predicted match outcomes compared to other established rating methods for cricket teams. Applying the Bradley-Terry method to the team ratings, generated through the adaptive system, calculated the probability of $team_i$ beating $team_j$.

The adaptive rating system was applied to the Caribbean Premier League 2015 and the Cricket World Cup 2015, and the systems predictive accuracy was benchmarked against the New Zealand Totalisator Board Agency (TAB) and the CricHQ algorithm. The results revealed that the developed rating system outperformed the TAB by 9% and the commercial algorithm by 6% for the Cricket World Cup (2015), respectively and outperformed the TAB and CricHQ algorithm by 25% and 12%, for the Caribbean Premier League (2015), respectively. These results demonstrate that cricket team ratings based on the aggregation of individual player ratings are superior to ratings based on summaries of team performances and match outcomes; validating the research hypothesis. The insights derived from this research also inform interested parties of the key attributes to win limited over cricket matches and can be used for team selection. This demonstrated that rating

systems that consider micro variables generate greater predictive accuracy than systems that only consider macro variables.

The results show that cricket team ratings based on the aggregation of individual playing ratings with attributes weighted towards winning limited over matches are superior to ratings based on summaries of team performances and match outcomes. Given the adaptive systems predictive ability to reasonably predict the result of a limited overs cricket match, based on the combination of individual player ratings, this research was considered successful. Moreover, this research achieved all supplementary aims; specifically the development of a roster-based optimisation model, for limited overs cricket, using individual player ratings. In addition, this thesis exploited a unique data set and a proprietary algorithm to make an original contribution, and provided directions for future research.

Appendices

Appendix A

Performance Metric Definitions

Performance Metrics	Definitions	Dataset	Recoded
Batting Average	Total runs divided by total dismissals	Batting	L
Total Dismissals	Number of times a batsmen been dismissed	Batting	F
Batting Strike Rate	Total Runs divided by total balls	Batting	M
Percentage Boundaries (Bat)	Total Boundaries divided by total balls faced	Batting	K
Batting Position	Position of the batting line-up a player occupies on average	Batting	C
Total Runs Scored	Total number of runs a batsmen has contributed to the batting side total	Batting	E
Balls Faced	Number of deliveries faced	Batting	G
Total Boundaries	Total fours hit + total sixes hit	Batting	J
Sixes Hit	Total number of balls hit over the field's boundary in the air	Batting	H
Fours Hit	Total number of balls hit over the field's boundary along the ground	Batting	I
Games Played	Total number of participated matches	Batting	D
Percentage Wins	Total number matches won divided by total number of matches played	Batting	N
Economy Rate	Total of runs conceded by overs bowled	Bowling	Q
Bowling Strike Rate	Total balls bowled divided wickets	Bowling	O
Bowling Position	Position of the bowlers in the bowling line-up	Bowling	C
Bowling Average	Total Runs conceded divided by wickets	Bowling	P
Percentage Boundaries (Bowl)	Total boundaries conceded divided by total balls bowled	Bowling	S
Dot Balls	Total balls bowled in which no runs were conceded	Bowling	G
Balls Bowled	Total number of delieveies by a bowler	Bowling	F
Percentage Dots	Total dots divided by total balls bowled	Bowling	T
Total Runs Conceded	Total number of runs contributed to the batting side	Bowling	I
Total Wickets	Total number of batsmen a bowler has dismissed	Bowling	J
Total Maidens	Total number of overs bowled in which no runs were conceded	Bowling	H
Fours Conceded	Total number of balls bowled in which the ball was hit over the field's boundary along the ground	Bowling	M
Sixes Conceded	Total number of balls bowled in which the ball was hit over the field's boundary in the air	Bowling	N
Total Boundaries Conceded	Total fours conceded + total sixes conceded	Bowling	R
Number of Wins	Total number of games won	Both	
Role	Identifies whether a player was batting or bowling	Both	
Player ID	Unique player identification number	Both	
Game ID	Unique match identification number	Both	
Team	The side in which a player resides	Both	

Appendix B

Scorecard Data Structure

cricinfo_id	player_id	player	dismissal	uns_score	minutes	balls	fours	sixes	strike_rate	overs	maidens	runs_conceded	wickets	economy_rate	extra	dots	undary_fo	undary_s	innings	role	order	team
829705	34102	RG Sharma	not out	98	85	65	12	4	150										1	1	1	MI
829705	5334	AJ Finch	Yadav b M	5	5	5	1	0	100										1	1	2	MI
829705	333904	AP Tare	av b Shakib	7	7	7	1	0	100										1	1	3	MI
829705	33141	AT Rayudu	than b Mo	0	1	2	0	0	0										1	1	4	MI
829705	277662	CJ Anderson	not out	55	61	41	4	3	134										1	1	5	MI
829705	376116	UT Yadav								3	0	36	0	12.		6	6	1	1	2	1	KK
829705	46538	M Morkel								4	1	18	2	4.5		15	3	0	1	2	2	KK
829705	56143	Shakib Al Hasan								4	0	48	1	12.	(2w	7	4	3	1	2	3	KK
829705	230558	SP Narine								4	0	28	0	7.0		10	3	1	1	2	4	KK
829705	276298	AD Russell								3	0	21	0	7.0		7	0	2	1	2	5	KK
829705	32966	PP Chawla								2	0	16	0	8.0		3	2	0	1	2	6	KK
829705	35582	RV Uthappa	an Singh b	9	14	12	0	1	75										2	1	1	KK
829705	28763	G Gambhir	yudu b Bur	57	66	43	7	1	132										2	1	2	KK
829705	290630	MK Pandey	b Harbhaj	40	39	24	2	3	166										2	1	3	KK
829705	446507	SA Yadav	not out	46	36	20	1	5	230										2	1	4	KK
829705	32498	YK Pathan	not out	14	24	12	1	1	116										2	1	5	KK

Appendix C

Season Performance Metrics

player	team	player type	batting_pos	games	runs_scored	total_dismissals	balls_faced	fours	sixes	total_boundaries	pctg_boundaries	batting_ave	ave_strike_rate	pctg_wins
A Bagai	CAN	Batsmen	4	6	225	5	322	27	0	27	8.385093168	37.5	69.8757764	0.17
AA Obanda	KEN	Batsmen	1	3	43	3	65	4	2	6	9.230769231	14.3333333	66.15384615	0
AB de Villiers	SA	Batsmen	4	5	353	4	326	31	7	38	11.65644172	70.6	108.2822086	0.72
AB de Villiers	RC	Batsmen	4	14	513	11	293	60	22	82	27.98634812	36.6428571	175.0853242	0.5
AD Mathews	SL	Batting-allrounder	6	5	94	3	90	10	1	11	12.22222222	18.8	104.4444444	0.67
AD Mathews	DD	Batting-allrounder	5	10	144	7	104	12	6	18	17.30769231	14.4	138.4615385	0.36
AD Russell	KK	Batting-allrounder	5	11	326	9	169	35	19	54	31.95266272	29.6363636	192.8994083	0.5
Ahmed Shehzad	PAK	Batsmen	2	5	44	5	91	7	0	7	7.692307692	8.8	48.35164835	0.75
AJ Finch	MI	Batsmen	2	3	23	3	33	4	0	4	12.12121212	7.66666667	69.6969697	0.625
AJ Strauss	ENG	Batsmen	1	7	334	7	357	34	3	37	10.36414566	47.7142857	93.55742297	0.43
AM Rahane	RR	Batsmen	1	13	540	11	413	53	13	66	15.98062954	41.5384615	130.7506053	0.47
AN Kervezee	NET	Batsmen	3	6	81	6	126	10	0	10	7.936507937	13.5	64.28571429	0
AP Tare	MI	Batsmen	3	2	14	2	17	1	0	1	5.882352941	7	82.35294118	0.625
AR White	IRL	Batsmen	5	2	15	2	37	1	0	1	2.702702703	7.5	40.54054054	0.33
AS Hansra	CAN	Batsmen	5	6	215	5	373	16	5	21	5.63002681	35.8333333	57.64075067	0.17
Asad Shafiq	PAK	Batsmen	3	3	154	2	217	14	0	14	6.451612903	51.3333333	70.96774194	0.75

player	team	player type	bowling_pos	games	total_over	balls_bo	total_dots	tal_maide	runs_con	otal_wicke	games_o	wickets_o	fours_con	sixes_con	e_strike_r	owling_av	onomy_ra	l_bounda	g_bounda	pctg_dots	pctg_wins
LL Tsotsobe	SA	Bowler	2	1	5	30	26	2	14	3	5	3	3	0	10	4.666667	2.8	3	0.1	0.866666667	0.72
TT Samaraweera	SL	Batsmen	6	1	2	12	8	0	4	1	2	1	0	0	12	4	2	0	0	0.666666667	0.67
SK Raina	IND	Batsmen	5	1	2	12	4	0	12	1	2	1	1	0	12	12	6	1	0.083333	0.333333333	0.78
A Ashish Reddy	SH	Batting All-rounder	6	4	6	36	10	0	50	3	6	3	1	3	12	16.66667	8.333333	4	0.111111	0.277777778	0.5
DR Smith	CS	Bowling All-rounder	7	1	2	12	4	0	17	1	2	1	0	1	12	17	8.5	1	0.083333	0.333333333	0.59
KC Cariappa	KK	Bowler	3	1	2	12	3	0	28	1	2	1	3	2	12	28	14	5	0.416667	0.25	0.5
DJ Bravo	CS	Batting All-rounder	5	16	54	324	97	0	426	26	54	26	30	17	12.46154	16.38462	7.888889	47	0.145062	0.299382716	0.59
YS Chahal	RC	Bowler	4	14	48	288	91	0	415	23	48	23	21	28	12.52174	18.04348	8.645833	49	0.170139	0.315972222	0.5
JP Duminy	DD	Batting All-rounder	4	9	17	102	34	0	126	8	17	8	8	4	12.75	15.75	7.411765	12	0.117647	0.333333333	0.36
MA Starc	RC	Bowler	1	12	43	258	113	1	291	20	43	20	32	2	12.9	14.55	6.767442	34	0.131783	0.437984496	0.5
S Aravind	RC	Bowler	2	5	18	108	47	0	132	8	18	8	15	3	13.5	16.5	7.333333	18	0.166667	0.435185185	0.5
MC Henriques	SH	Bowler	4	9	25	150	60	0	158	11	25	11	13	2	13.63636	14.36364	6.32	15	0.1	0.4	0.5
GB Hogg	KK	Bowling All-rounder	4	6	21	126	54	0	144	9	21	9	8	9	14	16	6.857143	17	0.134921	0.428571429	0.5
Bipul Sharma	SH	Bowling All-rounder	5	3	7	42	17	0	50	3	7	3	1	4	14	16.66667	7.142857	5	0.119048	0.404761905	0.5
Imran Tahir	DD	Bowler	5	10	36	216	65	0	314	15	36	15	13	18	14,4	20.93333	8.722222	31	0.143519	0.300925926	0.36
SL Malinga	MI	Bowler	1	15	60	360	149	3	444	24	60	24	47	9	15	18.5	7.4	56	0.155556	0.413888889	0.625

Appendix D

Ball-by-Ball Data Structure

game	cricinfo_id	home	away	venue	dates	year	innings	over	ball	Description	Bowling	Facing	Out	Runs_scored	Sundry_Type	Batting_Pos	Bowling_Pos
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	0	1	ngth delivery outside o	Kulasekara	Guptill	0	0		1	1
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	0	2	outside off on the same	Kulasekara	Guptill	0	0		1	1
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	0	3	utside off as Guptill get	Kulasekara	Guptill	0	0		1	1
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	0	3	de off as he looks for th	Kulasekara	Guptill	0	1	WD	1	1
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	0	4	iddle and off as this is d	Kulasekara	Guptill	0	0		1	1
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	0	5	gets this away on the le	Kulasekara	Guptill	0	1	LB	1	1
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	0	6	length outside off. Was	Kulasekara	McCullum	0	4		2	1
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	1	1	s glanced behind squar	Malinga	Guptill	0	2		1	2
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	1	2	h delivery outside off a	Malinga	Guptill	0	0		1	2
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	1	3	y as this is punched awa	Malinga	Guptill	0	0		1	2
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	1	4	o do is clip this through	Malinga	Guptill	0	3		1	2
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	1	5	ide as he is able to scyt	Malinga	McCullum	0	4		2	2
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	1	6	bat down to dig this ou	Malinga	McCullum	0	0		2	2
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	2	1	of this length delivery	Kulasekara	Guptill	0	0		1	1
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	2	2	side off through the reg	Kulasekara	Guptill	0	0		1	1
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	2	3	off into the batsman as	Kulasekara	Guptill	0	0		1	1
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	2	4	t to the midwicket field	Kulasekara	Guptill	0	1		1	1
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	2	5	ler delivery just outside	Kulasekara	McCullum	0	4		2	1
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	2	6	ullish length outside off	Kulasekara	McCullum	0	0		2	1
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	3	1	pads, as this is eased av	Malinga	Guptill	0	0		1	2
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	3	2	e off which is dead-batt	Malinga	Guptill	0	0		1	2
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	3	3	h as Guptill has no othe	Malinga	Guptill	0	0		1	2
1st	656399	New Zealand	Sri Lanka	Christchurch	14-Feb	2015	1	3	4	e this through mid-on t	Malinga	Guptill	0	1		1	2

Appendix E

Team Level Dataset

team	balls_faced	total_runs	fours	sixes	al_boundar	g_boundar	m_batting_	team_SR	pctg_wins	tal_dismiss	l_balls_bo	total_dots	fours_con	tal_maide	total_over	runs_conc	sixes_con	tal_wicke	boundarie	e_strike_r	owling_av	onomy_ra	oundaries	pctg_dots	Games_played
AUS	1594	1342	127	17	144	9.03388	39.4706	84.1907	0.58	34	1866	1098	134	18	311	1396	14	45	148	41.4667	31.0222	4.48875	0.07931	0.58842	7
BAN	1423	930	88	5	93	6.53549	18.2353	65.3549	0.5	51	1530	793	102	12	255	1248	12	36	114	42.5	34.6667	4.89412	0.07451	0.5183	6
CAN	1582	963	96	14	110	6.95322	17.8333	60.8723	0.17	54	1686	846	149	9	281	1560	25	43	174	39.2093	36.2791	5.5516	0.1032	0.50178	6
ENG	2058	1720	141	13	154	7.48299	30.7143	83.5763	0.43	56	1992	1100	167	19	332	1742	32	43	199	46.3256	40.5116	5.24699	0.0999	0.55221	7
IND	2564	2338	235	29	264	10.2964	37.7097	91.1856	0.78	62	2628	1323	183	12	438	2218	29	69	212	38.087	32.1449	5.06393	0.08067	0.50342	9
IRL	1634	1306	138	21	159	9.73072	25.6078	79.9266	0.33	51	1782	894	144	8	297	1563	21	40	165	44.55	39.075	5.26263	0.09259	0.50168	6
KEN	1426	848	79	14	93	6.52174	15.1429	59.467	0	56	1338	694	140	9	223	1347	17	23	157	58.1739	58.5652	6.04036	0.11734	0.51868	6
NET	1563	1123	107	15	122	7.8055	20.0536	71.849	0	56	1656	815	159	14	276	1610	28	29	187	57.1034	55.5172	5.83333	0.11292	0.49215	6
NZ	1935	1581	140	36	176	9.09561	31	81.7054	0.63	51	2046	1199	146	23	341	1513	15	60	161	34.1	25.2167	4.43695	0.07869	0.58602	8
PAK	1990	1534	152	11	163	8.19095	28.9434	77.0854	0.75	53	2148	1327	118	24	358	1468	21	67	139	32.0597	21.9104	4.10056	0.06471	0.61778	8
SA	2003	1699	140	21	161	8.03794	33.98	84.8228	0.72	50	1740	1043	102	17	290	1222	17	65	119	26.7692	18.8	4.21379	0.06839	0.59943	7
SL	2343	2073	226	18	244	10.414	45.0652	88.4763	0.67	46	2118	1182	127	14	353	1563	9	65	136	32.5846	24.0462	4.42776	0.06421	0.55807	9
WI	1748	1320	118	33	151	8.63844	22.3729	75.5149	0.43	59	1584	860	121	14	264	1219	6	51	127	31.0588	23.902	4.61742	0.08018	0.54293	7
ZIM	1631	1209	115	11	126	7.72532	23.25	74.1263	0.33	52	1488	804	110	13	248	1172	12	32	122	46.5	36.625	4.72581	0.08199	0.54032	6

Appendix F

AHP Pairwise Comparison Matrix

One Day Pairwise Compirson Matrices

	Batting Performance Metrics		
	Strike Rate	Total Runs Scored	Batting Average
Strike Rate	1	0.833333333	0.666666667
Total Runs Scored	1.2	1	1.25
Batting Average	1.5	0.8	1

	Bowling Performance Metrics		
	Economy Rate	Percentage Boundaries	Strike Rate
Economy Rate	1	1.3	1.5
Percentage Boundaries	0.769230769	1	1.25
Strike Rate	0.666666667	0.8	1

	All-rounder Batting Performance Metrics		
	Strike Rate	Total Runs Scored	Batting Average
Strike Rate	1	0.833333333	0.666666667
Total Runs Scored	1.2	1	1.25
Batting Average	1.5	0.8	1

	All-rounder Bowling Performance Metrics		
	Strike Rate	Economy Rate	Percentage Boundaries
Strike Rate	1	0.833333333	0.666666667
Economy Rate	1.2	1	1.25
Percentage Boundaries	1.5	0.8	1

Twenty-twenty Pairwise Compirson Matrices

	Batting Performance Metrics		
	Strike Rate	Percentage Boundaries	Total Runs Scored
Strike Rate	1	1.25	1.15
Total Runs Scored	0.8	1	1.2
Batting Average	0.869565217	0.833333333	1

	Bowling Performance Metrics		
	Economy Rate	Strike Rate	Percentage Boundaries
Economy Rate	1	1.25	1.15
Strike Rate	0.8	1	1.2
Percentage Boundaries	0.869565217	0.833333333	1

	All-rounder Batting Performance Metrics		
	Strike Rate	Percentage Boundaries	Total Runs Scored
Strike Rate	1	1.1	1.2
Percentage Boundaries	0.909090909	1	1.2
Total Runs Scored	0.833333333	0.833333333	1

	All-rounder Bowling Performance Metrics		
	Strike Rate	Economy Rate	Percentage Boundaries
Strike Rate	1	0.833333333	0.666666667
Economy Rate	1.2	1	1.25
Percentage Boundaries	1.5	0.8	1

Appendix G

Player-Type Ratings by Team

player	Team	Batsman_Ratings	Bowler_Ratings	Overall_Rounder_Ratings	Keepers_Ratings
KS Williamson	NZ	1.10527472	0	0	0
MJ Guptill	NZ	2.311881717	0	0	0
LRPL Taylor	NZ	1.007052278	0	0	0
BB McCullum	NZ	1.727597642	0	0	0
GD Elliott	NZ	1.52409682	0	0	0
AF Milne	NZ	0	0.762708727	0	0
DL Vettori	NZ	0	1.735992566	0	0
MJ Henry	NZ	0	0.778135193	0	0
RML Taylor	NZ	0	0.671894696	0	0
TA Boult	NZ	0	1.587566648	0	0
TG Southee	NZ	0	1.143293519	0	0
L Ronchi	NZ	0	0	0	2.510527638
CJ Anderson	NZ	0	0	1.924377597	0

Bibliography

[1] Big bash's big boom: Will success see its sponsorship and broadcast dollars double? http://www.theaustralian.com.au/business/media/big-bash-final-delivers-a-ratings-winner-for-ten/news-story/991b01dc61040596c415d77fce5978cc. Accessed: 2016-02-16.

[2] Billions of dollars at stake: Why is the international cricket council changing its revenue sharing model? http://www.ibtimes.com/billions-dollars-stake-why-international-cricket-council-changing-its-revenue-sharing-1554078. Accessed: 2016-02-05.

[3] Bonferroni outlier test. http://www.inside-r.org/packages/cran/car/docs/outlierTest. Accessed: 2015-10-11.

[4] International cricket council. http://www.icc-cricket.com/world-t20. Accessed: 2015-07-11.

[5] AKHTAR, S., SCARF, P., AND RASOOL, Z. Rating players in test match cricket. *Journal of the Operational Research Society 66*, 4 (2014), 684–695.

[6] AL MALIKI, A., OWEN, G., AND BRUCE, D. *Combining AHP and TOPSIS Approaches to Support Site Selection for a Lead Pollution Study*. PhD thesis, IACSIT Press, 2012.

[7] ALAMAR, B., AND MEHROTRA, V. Beyond'moneyball': The rapidly evolving world of sports analytics. *Analytics Magazine*.

[8] ALLSOPP, P., AND CLARKE, S. R. Rating teams and analysing outcomes in one-day and test cricket. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 167*, 4 (2004), 657–667.

[9] ANNIS, D. H., AND CRAIG, B. A. Hybrid paired comparison analysis, with applications to the ranking of college football teams. *Journal of Quantitative Analysis in Sports 1*, 1 (2005).

[10] BENDEL, R. B., AND AFIFI, A. A. Comparison of stopping rules in forward "stepwise" regression. *Journal of the American Statistical Association 72*, 357 (1977), 46–53.

[11] BIRNBAUM, P. A guide to sabermetric research, 2013.

[12] BOSCH, R., AND TRICK, M. Integer programming. In *Search methodologies*. Springer, 2014, pp. 67–92.

[13] BRACEWELL, P. Monitoring meaningful rugby ratings. *Journal of Sports Sciences 21*, 8 (2003), 611–620.

[14] BRACEWELL, P., COOMES, M., NASH, J., MEYER, D, H., AND ROONEY, S. Rating the performance of a non-wicket taking bowler in limited overs cricket. 2016.

[15] BRACEWELL, P., DOWNS, M., AND SEWELL, J. The development of a performance based rating system for limited overs cricket. In *MATHSPORT 2014* (2014).

[16] BRACEWELL, P. J., FORBES, D. G., JOWETT, C. A., AND KITSON, H. I. Determining the evenness of domestic sporting competition using a generic rating engine. *Journal of Quantitative Analysis in Sports 5*, 1 (2009).

[17] BRACEWELL, P. J., AND RUGGIERO, K. A parametric control chart for monitoring individual batting performances in cricket. *Journal of Quantitative Analysis in Sports 5*, 3 (2009).

[18] BREIMAN, L. Bagging predictors. *Machine learning 24*, 2 (1996), 123–140.

[19] BREIMAN, L. Random forests. *Machine learning 45*, 1 (2001), 5–32.

[20] BUKIET, B., AND OVENS, M. A mathematical modelling approach to one-day cricket batting orders. *Journal of Sports Science & Medicine 5*, 4 (2006), 495.

[21] CHAVENT, M., KUENTZ, V., LIQUET, B., AND SARACCO, L. Clustofvar: an r package for the clustering of variables. *arXiv preprint arXiv:1112.0295* (2011).

[22] CHING-LAI, H., AND YOON, K. *Multiple attribute decision making: methods and applications*. Springer-Verlag, 1981.

[23] CLARKE, S. R. Dynamic programming in one-day cricket-optimal scoring rates. *Journal of the Operational Research Society 39*, 4 (1988), 331–337.

[24] CLARKE, S. R. Rating non-elite tennis players using team doubles competition results. *Journal of the Operational Research Society 62*, 7 (2011), 1385–1390.

[25] COOK, W. D., GOLAN, I., AND KRESS, M. Heuristics for ranking players in a round robin tournament. *Computers & Operations Research 15*, 2 (1988), 135–144.

[26] COOMES, M. Comparison of reject inference methods on complete data with gradient boosting machine variable selection. 2014.

[27] COOPERS, P. W. Pwc outlook for the global sports market to 2015. Report P-25, Price Waterhouse Coopers, 2015.

[28] CRAWLEY, M. J. *The R book*. John Wiley & Sons, 2012.

[29] CROUCHER, J. Player ratings in one-day cricket. In *Proceedings of the fifth Australian conference on mathematics and computers in sport* (2000), Sydney University of Technology Sydney, NSW, pp. 95–106.

[30] DAMODARAN, U. Stochastic dominance and analysis of odi batting performance: The indian cricket team. *Journal of Sports Science and Medicine 5* (2006), 503–508.

[31] DANIYAL, M., NAWAZ, T., MUBEEN, I., AND ALEEM, M. Analysis of batting performance in cricket using individual and moving range (mr) control charts. *International Journal of Sports Science and Engineering 6*, 4 (2012), 195–202.

[32] DE VILLE, B., AND NEVILLE, P. *Decision Trees for Analytics Using SAS Enterprise Miner*. SAS Institute, 2013.

[33] DEY, P. K., GHOSH, D. N., AND MONDAL, A. C. A mcdm approach for evaluating bowlers performance in ipl. *Journal of Emerging Trends in Computing and Information Sciences 2*, 11 (2011), 563–73.

[34] DI SALVO, V., BARON, R., GONZÁLEZ-HARO, C., GORMASZ, C., PIGOZZI, F., AND BACHL, N. Sprinting analysis of elite soccer players during european champions league and uefa cup matches. *Journal of Sports Sciences 28*, 14 (2010), 1489–1494.

[35] DUCKWORTH, F. C., AND LEWIS, A. J. A fair method for resetting the target in interrupted one-day cricket matches. *Journal of the Operational Research Society 49*, 3 (1998), 220–227.

[36] DYTE, D., AND CLARKE, S. R. A ratings based poisson model for world cup soccer simulation. *Journal of the Operational Research Society 51*, 8 (2000), 993–998.

[37] FARINAZ, F., ET AL. Was bradman denied his prime? *Journal of Quantitative Analysis in Sports 5*, 4 (2009), 1–26.

[38] FIELD, A. *Discovering statistics using IBM SPSS statistics*. Sage, 2013.

[39] FINANCE, B. A study by brandfinance on ipl v to value ipl brand and its nine franchisee brands. Report P-21, Brand Finance India, 2013.

[40] FIRTH, D. Bradley-terry models in r. *Journal of Statistical Software 12*, 1 (2005), 1–12.

[41] GAMING, E., AND ASSOCIATION, B. Sporst betting: Commercial and integrity issues. Report P-5, European Gaming and Betting Association, 2015.

[42] GERBER, H., AND SHARP, G. D. Selecting a limited overs cricket squad using an integer programming model. *South African Journal for Research in Sport, Physical Education & Recreation (SAJR SPER) 28*, 2 (2006).

[43] GUYON, I., AND ELISSEEFF, A. An introduction to variable and feature selection. *The Journal of Machine Learning Research 3* (2003), 1157–1182.

[44] HARVILLE, D. A. The selection or seeding of college basketball or football teams for postseason competition. *Journal of the American Statistical Association 98*, 461 (2003), 17–27.

[45] JACOBY, W. G. Regression iii: Advanced methods, September 2014.

[46] JAHANSHAHLOO, G. R., LOTFI, F. H., AND DAVOODI, A. Extension of topsis for decision-making problems with interval data: Interval efficiency. *Mathematical and Computer Modelling 49*, 5 (2009), 1137–1142.

[47] JOWETT, C. A. Tracking the historical performance of provincial unions in new zealand domestic rugby between 1976 and 2008. 2008.

[48] KAKWANI, N. C. Applications of lorenz curves in economic analysis. *Econometrica: Journal of the Econometric Society* (1977), 719–727.

[49] KIMBER, A. C., AND HANSFORD, A. R. A statistical analysis of batting in cricket. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* (1993), 443–455.

[50] KOULIS, T., MUTHUKUMARANA, S., AND BRIERCLIFFE, C. D. A bayesian stochastic model for batting performance evaluation in one-day cricket. *Journal of Quantitative Analysis in Sports 10*, 1 (2014), 1–13.

[51] LEITNER, C., ZEILEIS, A., AND HORNIK, K. Forecasting sports tournaments by ratings of (prob) abilities: A comparison for the euro 2008. *International Journal of Forecasting 26*, 3 (2010), 471–481.

[52] LEMMER, H. H. A measure for the batting performance of cricket players. *South African Journal for Research in Sport, Physical Education and Recreation 26*, 1 (2004), 55–64.

[53] LEMMER, H. H. The single match approach to strike rate adjustments in batting performance measures in cricket. *Journal of Sports Science and Medicine 10* (2011), 630–634.

[54] LEWIS, A. Towards fairer measures of player performance in one-day cricket. *Journal of the Operational Research Society 56*, 7 (2005), 804–815.

[55] LONDON, I. C. Lecture 16: Linear discriminant analysis. University Lecture, 2015.

[56] MAIMON, O., AND ROKACH, L. *Data mining and knowledge discovery handbook*, vol. 2. Springer, 2010.

[57] MANAGE, A. B., AND SCARIANO, S. M. An introductory application of principal components to cricket data. *Journal of Statistics Education 21*, 3 (2013).

[58] MANLY, B. Multivariate statistical methods. a primer, 2nd edition chapman & hall, london. *Pavoine, Doledec 136* (1994).

[59] MASSEY, K. Statistical models applied to the rating of sports teams. *Bluefield College* (1997).

[60] MEASE, D. A penalized maximum likelihood approach for the ranking of college football teams independent of victory margins. *The American Statistician 57*, 4 (2003), 241–248.

[61] PATEL, A., BRACEWELL, P., GAZLEY, A., AND BRACEWEL, B. Identifying fast bowlers likely to play test cricket based on age-group performances. 2015.

[62] PRICEWATERHOUSECOOPERS. Changing the game, 2011.

[63] RAO, R. V. Improved multiple attribute decision making methods. In *Decision Making in Manufacturing Environment Using Graph Theory and Fuzzy Multiple Attribute Decision Making Methods*. Springer, 2013, pp. 7–39.

[64] SAATY, T. The analytic hierarchy process. *Math Modelling 9* (1987), 161–176.

[65] SARGENT, J. *Player ratings in continuous and discrete team sports*. PhD thesis, RMIT University, 2013.

[66] SCHUMAKER, R. P., SOLIEMAN, O. K., AND CHEN, H. *Predictive modeling for sports and gaming*. Springer, 2010.

[67] SEBER, G. A., AND LEE, A. J. *Linear regression analysis*, vol. 936. John Wiley & Sons, 2012.

[68] SHARP, G., BRETTENNY, W., GONSALVES, J., LOURENS, M., AND STRETCH, R. Integer optimisation for the selection of a twenty20 cricket team. *Journal of the Operational Research Society 62*, 9 (2011), 1688–1694.

[69] SIBALIJA, T. V., AND MAJSTOROVIĆ, V. D. *Advanced Multiresponse Process Optimisation: An Intelligent and Integrated Approach*. Springer, 2015.

[70] SINUANY-STERN, Z. Ranking of sports teams via the ahp. *Journal of the Operational Research Society* (1988), 661–667.

[71] SORENSEN, S. P. An overview of some methods for ranking sports teams. *University of Tennessee. Knoxville* (2000).

[72] STEFANI, R. The methodology of officially recognized international sports rating systems. *Journal of Quantitative Analysis in Sports 7*, 4 (2011).

[73] STEFANI, R. Predictive success of official sports rating in international competition. In *Proceedings of 10th Australasian Conference on Mathematics & Computers in Sport* (2012), pp. 35–40.

[74] STEFANI, R. T. Survey of the major world sports rating systems. *Journal of Applied Statistics 24*, 6 (1997), 635–646.

[75] STEINER, S. H. Ewma control charts with time-varying control limits and fast initial response. *Journal of Quality Technology 31*, 1 (1999), 75.

[76] STUDYLIB. Backward elimination and stepwwise regression. Online Lecture Notes, 2014.

[77] VROOMAN, J. The economic structure of the nfl. In *The Economics of the National Football League*. Springer, 2012, pp. 7–31.

[78] WEST, B. T., AND LAMSAL, M. A new application of linear modeling in the prediction of college football bowl outcomes and the development of team ratings. *Journal of Quantitative Analysis in Sports 4*, 3 (2008).

[79] WINSTON, W. L., AND GOLDBERG, J. B. *Operations research: applications and algorithms*, vol. 3. Duxbury press Boston, 2004.

[80] YE, J. Lecture notes in numerical linear algebra for data exploration - linear discriminant analysis, September 2007.

[81] ZAVADSKAS, E., AND KAKLAUSKAS, A. Determination of an efficient contractor by using the new method of multicriteria assessment. In *International Symposium for "The Organization and Management of Construction". Shaping Theory and Practice* (1996), vol. 2, pp. 94–104.