Research Directions in Cricket

Tim B. Swartz *

Abstract

This paper concerns quantitative (statistical) research in cricket. It reviews the sport, it reviews the existing literature and it provides suggestions for future work.

Keywords: One-day cricket, Test cricket, Twenty20 cricket.

^{*}Tim Swartz is Professor, Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6. Swartz has been partially supported by grants from the Natural Sciences and Engineering Research Council of Canada. The author thanks Basil de Silva, Charith Karunarathna, Ananda Manage, Saman Muthukumarana, Harsha Perera, Lasantha Premarathna, Pulindu Ratnasekera and Rajitha Silva all of whom provided helpful advice on an earlier version of the paper.

1 INTRODUCTION

The Handbook which you are reading is divided into six main sections covering the five major sports of baseball, basketball, hockey, soccer and football, and a final section on other sports. Since this article on cricket resides in the final section, it is worth reflecting on what constitutes a major sport. In some ways, cricket may be considered a major sport with a long history dating back to the 16th century. In terms of either participation or viewership, cricket is the second most popular sport in the world after soccer (Woods 2011). In terms of online media coverage, cricket ranks fourth worldwide after soccer, basketball and tennis (www.biggestglobalsports.com). In addition, the format of cricket known as Twenty20 is a "big money" sport. Specifically, amongst professional leagues, the IPL (Indian Premier League) had the second highest average player salaries in 2014-2015 trailing only the National Basketball Association (http://www.businessinsider.com/sports-leagues-top-salaries-2015-5). However, in terms of advanced sports analytics, the sport of cricket appears to lag behind the major sports listed above. This article attempts to provide the reader with an up-to-date survey of the quantitative work that has been done in relation to cricket and on the promising research opportunities that are available.

When discussing cricket analytics, it is important to distinguish between the three main formats of cricket. The traditional version of cricket and the "highest" form of the sport is known as *test cricket*. Test cricket matches can take up to five days to complete. Matches are played between two teams representing countries which are Full members of the ICC (International Cricket Council). The 10 Full member cricketing nations of the ICC have connections to England through the expansion of the British Empire and consist of Australia, Bangladesh, England, India, New Zealand, Pakistan, South Africa, Sri Lanka, West Indies and Zimbabwe. In test cricket, each side has two *innings* (opportunities) for batting. Completion of an innings depends on the *dismissal* of the batting team (they are made "out") or the voluntary *declaration* by the batting team that their innings are complete or the time constraints for the match. Test matches often result in draws (roughly 25% of matches in recent years). Teams sometimes bat conservatively to attain a draw when they fear they are on the verge of losing; batting conservatively helps a team avoid dismissal and lengthens the duration of the innings. Hence test cricket tends to be a slow moving game though full of strategy.

To accommodate the constraints of modern living and to infuse more excitement into cricket, the second form of cricket known as *one-day cricket* was introduced in the 1950's or 1960's depending on who is given credit for the origins of the game. However, there is no doubt that the format began in earnest in the 1970's with the first World Cup contested in 1975. One-day cricket is based on *limited overs* where each side has a single innings of batting. Batting is terminated by either dismissal or by completion of 50 overs (300 balls). In one-day international (ODI) cricket, draws are less common (1% compared to 26% in test cricket during the last 10 years). The style of play in one-day cricket is more aggressive with batsmen more willing to swing freely and score runs along with the associated increased risk of *wickets*. One-day cricket is seen by some as a brash upstart version of cricket, departing from tradition with various innovations including colored uniforms, day/night games, etc.

The third form of cricket known as Twenty20 or T20 cricket is the most recent incarnation of cricket. Twenty20 cricket was introduced in 2003 in an English domestic tournament. It is also a form of limited overs cricket and shares most of the characteristics of one-day cricket. The major difference is that Twenty20 cricket is based on 20 overs (120 balls) per side and therefore matches terminate in roughly three hours of playing time. As of 2016, there have been six World Cups in international Twenty20 cricket (2007, 2009, 2010, 2012, 2014 and 2016). Interest in Twenty20 has exploded with the introduction of professional leagues, most notably the IPL which was formed in 2008 and the Australian Big Bash which began in 2011. With fewer overs in which to lose wickets, Twenty20 cricket is even more aggressive than one-day cricket. In particular, batsmen score 4's and 6's at higher rates. It appears that this aspect of the game where the focus is primarily on scoring runs has lead to even greater excitement with perhaps less attention on in-game strategy.

Previously, the only overview of cricket analytics was provided by Clarke (1998). Much has changed since that time. Specifically, advances in computing have provided more tools to address large data sets. The structure of this article consists of sections that are mostly independent and focus on various topics of interest in cricket. The paper concludes with some discussion on where cricket analytics may be heading.

2 DATA

One of the great advantages of doing work in cricket analytics is the availability of reliable data. The primary source for data is the Cricinfo website (www.espncricinfo.com). In the Match series archive, match information reaches back to the 1770's. In recent history, details on all international matches involving Full ICC members have been recorded in the form of results, scorecards and match commentaries. There is also selective information on ICC Associate member's cricket, domestic cricket and women's cricket. The Cricinfo website also hosts a friendly search engine Statsguru which offers a variety of options and provides immediate results.

On the other hand, one of the current limitations of cricket data is that it is typically presented in summary form. For example, although it is easy to find the number of runs that a batsman has scored in a particular match or a particular period of time, the breakdown of run scoring is not so easily accessible. For example, one may be interested in knowing how runs were accumulated by a batsman through the frequency of 1's, 2's, 3's, 4's, and 6's at various stages (i.e. overs and wickets) of a match. We note that some researchers (e.g. Davis, Perera and Swartz 2015a) have developed parsers which scan match commentary logs to extract detailed ball-by-ball data. Detailed data has been helpful in analyzing the batting and bowling characteristics of individual players.

It now seems only a matter of time until more extensive cricket data will become widely available. Currently, the Hawk-Eye computer tracking system is in use in most major test, ODI and Twenty20 matches. Although its applications could be expanded, it is currently used for reviewing in-game umpiring decisions. For example, *leg before wicket* decisions may be appealed and challenged via Hawk-Eye technology. Hawk-Eye is also regularly used for providing graphics of ball locations and trajectories in live television broadcasts.

3 RESETTING OF TARGETS

For resetting targets in interrupted one-day cricket matches, the Duckworth-Lewis (D/L) method (Duckworth and Lewis, 1998, 2004) supplanted the method of run-rates in the late 1990's and has since been adopted by all major cricketing boards. Given the popularity of cricket, it could be argued that the Duckworth-Lewis method has provided the greatest impact to the sporting world from a mathematical, statistical and operational research perspective. What makes the acceptance of the Duckworth-Lewis method so remarkable is that the method is largely viewed by the public as a black-box procedure. The sporting world tends to like simple rules and simple statistics. From this point of view, the adoption of the Duckworth-Lewis method was a masterpiece in overcoming political hurdles.

The fundamental concept underlying the Duckworth-Lewis method is that of *resources*. In a one-day cricket match, the team batting first begins with 50 overs and 10 wickets at their disposal. They continue batting until either their overs are completed or 10 wickets have fallen. It is the combination of wickets and overs remaining in an innings that provides the capacity for scoring runs. The Duckworth-Lewis quantification of the combination of wickets and overs is known as resources. When innings are reduced due to an interruption, then targets are reset to "fair" values

based on the resources remaining. In one-day cricket, a team has 100% of its resources remaining at the beginning of its innings (50 overs and 0 wickets lost). When a team has used up all of its 50 overs, then 0% of its resources are remaining. Similarly, when 10 wickets are lost, a team has 0% of its resources remaining.

To get a sense of the use of the Duckworth-Lewis method, consider the abbreviated Duckworth-Lewis table presented in Table 1 based on the Standard Edition as taken from the 2014-2015 ICC Playing Handbook found at www.icc-cricket.com. Imagine that Team A is batting and scores 250 runs on completion of its innings. It then rains prior to the resumption of the second innings. When Team B comes to bat, they are only allotted 30 overs. Using the old method of run rates, Team B would need to score 250(30/50) + 1 = 151 runs in order to win the match. The obvious problem with the run rate approach is that Team B can bat more aggressively since their 10 wickets are spread throughout 30 overs rather than 50 overs. Therefore the target of 151 runs is lower than what might be considered fair. However, using the Duckworth-Lewis approach, Table 1 indicates that with 30 overs remaining and zero wickets lost, Team B retains 75.1% of its resources. Therefore the D/L target is rounded up to $250(0.751) \rightarrow 188$ runs. The large difference between 151 runs and 188 runs indicates how unpalatable it was for matches to be determined by run rates. The ICC Playing Handbook describes other scenarios in which D/L can be used to reset targets in interrupted matches.

	Wickets Lost										
Overs Available	0	1	2	3	4	5	6	7	8	9	10
50	100.0	93.4	85.1	74.9	62.7	49.0	34.9	22.0	11.9	4.7	0.0
40	89.3	84.2	77.8	69.6	59.5	47.6	34.6	22.0	11.9	4.7	0.0
30	75.1	71.8	67.3	61.6	54.1	44.7	33.6	21.8	11.9	4.7	0.0
20	56.6	54.8	52.4	49.1	44.6	38.6	30.8	21.2	11.9	4.7	0.0
10	32.1	31.6	30.8	29.8	28.3	26.1	22.8	17.9	11.4	4.7	0.0
5	17.2	17.0	16.8	16.5	16.1	15.4	14.3	12.5	9.4	4.6	0.0
1	3.6	3.6	3.6	3.6	3.6	3.5	3.5	3.4	3.2	2.5	0.0
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

....

Table 1: Abbreviated version of the Duckworth-Lewis resource table (2014-2015 Standard Edition). The table entries indicate the percentage of resources remaining in a match with the specified number of wickets lost and overs available.

The Duckworth-Lewis methodology is a statistical approach based on an exponential decay model in a regression context. We note that in addition to the Standard Edition there is also a Professional Edition of the Duckworth-Lewis method that is used in international matches. Recently, Frank Duckworth and Tony Lewis have ceded the management of D/L to David Stern of Australia. In the tradition of Duckworth-Lewis, Stern (2016) has updated the approach to account for recent changes in scoring. The approach is now sometimes referred to as the Duckworth-Lewis-Stern (D/L/S) method.

It is worth stating that over the years, there have been a long list of competitors to the Duckworth-Lewis method. (e.g. Clarke 1988, Christos 1998, Jayadevan 2002, Carter and Guthrie 2004, McHale and Asif 2013). However, it is the author's opinion that D/L/S will be difficult to supplant as it has passed the test of time and is well rooted in all levels of one-day cricket. Although competitors argue that their approach is better, the exercise of comparison seems futile. First, there is no gold standard to determine what is a "fair" target. Second, most of the alternative approaches provide targets that differ only minimally from the D/L/S targets. Weatherall (2011) compares the adequacy of various methods for resetting targets in interrupted one-day cricket matches.

One area where there may be some legitimate concerns over D/L/S is the application of D/L/S in Twenty20 cricket (Perera and Swartz 2013). In Twenty20, the current solution for resetting targets involves scaling the resource table that was developed for one-day cricket. To be precise, refer again to Table 1 and note that 56.6% of a team's resources remain when 20 overs are available and zero wickets have been lost. Therefore, the modified resource table for Twenty20 is obtained from the original resource table by dividing each cell entry by 0.566. The modified Twenty20 table contains only the rows where the available overs do not exceed 20. To see that the mapping from the one-day format to Twenty20 is logically flawed, one needs to only consider the different stages at which powerplays occur in the two formats. The powerplay in Twenty20 occurs in the first six overs whereas the powerplay in one-day cricket occurs in the first 10 overs. Since run scoring rates are higher during powerplays, it is not reasonable to map the cells corresponding to 15-20 available overs in Twenty20 cricket (a non-powerplay situation) to the cells corresponding to 15-20 available overs in Twenty20 cricket (a powerplay situation). However, at the end of the day, target resetting in Twenty20 is not a vital problem as match interruptions rarely occur. When the weather is bad, matches tend to be abandoned rather than shortened.

4 MATCH SIMULATION

When viewing a cricket match, the question that is continually on the minds of most supporters is how is their team doing. This is a mental exercise that is dynamic and requires the assessment of the current score, the overs remaining and the wickets lost. Formally, the scoring of runs is a stochastic process, and knowing the run scoring distribution would assist in prediction. Somewhat surprisingly, there has been little activity in the fundamental and related problem of developing match simulators. This is likely due to the complexity with which runs are scored in cricket.

Some of the earliest statistical publications concerning sport can be attributed to Elderton (1945) and Wood (1945) who proposed geometric distributions for individual runs scored in test cricket. Modeling test cricket seems particularly challenging as the run distribution depends on a team's objectives; sometimes teams are attempting to score runs at a high rate and sometimes teams are batting defensively.

Although likewise a difficult problem, there is more hope for modeling one-day and Twenty20 cricket. At least in these formats, the batting team is always attempting to score runs. However, the limited number of papers that have considered modeling/simulation in the one-day format tend to suffer from a lack of realism. For example, the papers may not have taken into account differences between batsmen, differences between bowlers, the stage of the match (i.e. overs completed and wickets taken), the home team advantage and modified tactics when chasing in the second innings.

In the author's opinion, the most comprehensive simulator to date which takes into account the above situational effects is due to Davis, Perera and Swartz (2015a). Their simulator was developed for Twenty20 although the ideas are directly transferable to one-day cricket. We now provide a brief overview of their match simulator which has many applications, some of which are described in the following sections. One obvious application that has not been exploited is the calculation of in-match probabilities. In cricket, there are 8 broadly defined outcomes that can occur when a batsman faces a bowled ball. These batting outcomes are listed below:

outcome
$$j = 0 \equiv 0$$
 runs scored
outcome $j = 1 \equiv 1$ runs scored
outcome $j = 2 \equiv 2$ runs scored
outcome $j = 3 \equiv 3$ runs scored
outcome $j = 4 \equiv 4$ runs scored
outcome $j = 5 \equiv 5$ runs scored
outcome $j = 6 \equiv 6$ runs scored
outcome $j = 7 \equiv$ dismissal

In the list (1) of possible batting outcomes, *extras* such as *byes*, *leg byes*, *wide-balls* and *no balls* are excluded. In the simulator, extras are introduced by by generating occurrences at the appropriate rates. Extras occur at the rate of 5.1% in Twenty20 cricket. The outcome j = 3

is not so common and the outcome j = 5 is rare, but both outcomes are retained to facilitate straightforward notation.

According to the enumeration of the batting outcomes in (1), Davis, Perera and Swartz (2015a) suggested the statistical model:

$$(X_{iow0}, \dots, X_{iow7}) \sim \text{multinomial}(m_{iow}; p_{iow0}, \dots, p_{iow7})$$
 (2)

where X_{iowj} is the number of occurrences of outcome j by the *i*th batsman during the *o*th over when w wickets have been taken. In (2), m_{iow} is the number of balls that batsman i has faced in the dataset corresponding to the *o*th over when w wickets have been taken. The dataset that they considered is special in the sense that it consists of detailed ball-by-ball data. The data were obtained using a proprietary parser which was applied to the commentary logs of matches listed on the CricInfo website (www.espncricinfo.com).

The estimation of the multinomial parameters p_{iowj} in (2) is a high-dimensional and complex problem. The complexity is partly due to the sparsity of the data; there are many match situations (i.e. combinations of overs and wickets) where batsmen do not have batting outcomes. For example, bowlers typically bat near the end of the batting order and do not face situations when zero wickets have been taken.

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

Given the estimation of the parameters in (2), first innings runs can be simulated for a specified batting lineup facing an average team. This is done by generating multinomial batting outcomes in (1) according to the laws of cricket. For example, when either 10 wickets are accumulated or the number of overs reaches 20, the first innings is terminated. Davis, Perera and Swartz (2015a) also provide modifications for batsmen facing specific bowlers (instead of average bowlers), they account for the home field advantage and they provide adjustments for second innings batting.

There may be several research avenues for improving the simulator. For example, additional covariates (such as the pitch condition) could be introduced. It might also be possible to augment the dataset for more reliable estimation. For example, can one-day cricket data be combined in some fashion with Twenty20 data? And as all cricket enthusiasts know, wickets are of great

importance in determining the outcome of matches. For the simulator proposed in Davis, Perera and Swartz (2015a), it is essential that the wicket probabilities p_{iow7} in (2) are estimated reliably. Naturally, this is a difficult problem since the occurrence of wickets is rare for even well-established batsmen. Therefore, a future research direction is to consider ways to improve the estimation of batsman-specific wicket probabilities.

5 EVALUATING PLAYER PERFORMANCE

Like all sports, cricket has statistics to measure player performance. However, this is an area where traditional statistics are lacking, and "better" statistics have not yet been embraced or have yet to be developed. Perhaps it is the case that the some of the advanced statistics are too difficult to calculate or are too complicated for the general public to comprehend.

The most commonly reported batting statistic is the *batting average*. The batting average (calculated separately in each of the three formats of cricket) is the total number of runs scored by a player divided by the total number of occasions where he has been dismissed. To see that the sole use of batting average is less than ideal, consider the pathological case of a player who has batted in 100 innings, has scored 100 runs and has been dismissed only once. The player would have a batting average of 100/1 = 100.0 which would be the highest batting average ever attained! Yet such a player would be a detriment to his team since he scores runs at such a low rate. To get a better understanding of this batsman's performance, it is typical to simultaneously assess his *strike rate* which is the average number of runs scored per 100 balls. Similar comments are applicable when discussing *bowling average* and *economy rate* which are the commonly reported bowling statistics.

In the above discussion, batting average fails to consider the number of balls faced and the strike rate fails to consider the number of dismissals. It is therefore clear that a single batting performance measure should at least take into account the following three important quantities: runs scored, balls faced and the number of dismissals. Alternative statistics have been proposed that utilize all three of the aforementioned quantities and are typically applicable in all formats of cricket. For example, Croucher (2000) proposed the *batting index* which is defined as the product of the batting average and the strike rate. Although the batting index is simple to calculate, it is not readily interpretable. Using a similar rationale, van Staden (2009) proposed graphical displays to assess performance. Manage and Scariano (2013) considered principal component analysis to reduce six/four batting/bowling variables into a single variable using IPL data. Lemmer (2011)

reviewed a number of related statistics and also proposed a statistic for individual matches that takes into match conditions in limited overs cricket.

Rather than assessing runs scored relative to wickets lost or runs scored relative to overs consumed, Beaudoin and Swartz (2003) and Lewis (2005) considered a natural benchmark for batting in one-day cricket. They considered runs scored relative to resources consumed where the Duckworth-Lewis concept of resources (see section 3 of the manuscript) is the standard quantity that measures the dual combination of wickets and overs. With bowlers, both papers explore symmetrical measures based on runs conceded relative to the resources used while bowling. There are at least two avenues where the general approach may be extended: (1) in Twenty20 cricket where the ideas translate directly and (2) the calculation of resources consumed by a player (during a match) which is a tedious exercise that would benefit from automation.

Getting away from formulae based on runs, wickets and overs, Davis, Perera and Swartz (2015b) addressed player evaluation using the simulator described in section 4. The authors asked how does a player i contribute to winning in Twenty20 matches? Via the simulator they estimated the expected increase in the number of runs scored $E_s(i)$ when player i is inserted in the lineup of an average team. They also estimated the expected decrease in the number of runs conceded $E_c(i)$ when player i is inserted in the lineup of an average team. Then $E_s(i) - E_c(i)$ is the proposed measure of player worth as it speaks directly to the number of runs by which the player improves his team compared to an average player. It is seen that the very best players (both batsmen and bowlers) can improve an international Twenty20 side by 8-10 runs. The approach has great potential and flexibility as it is able to investigate combinations of players and hypothetical lineups. Along the same lines, Perera, Davis and Swartz (2016b) considered the insertion of players into lineups to assess their contribution to fielding. In this analysis, it was seen that the best Twenty20 fielders save on average roughly 1.2 runs per Twenty20 match compared to a typical fielder. In addition to the simulator, the fielding analysis used textual analysis and random forests. The first quantitative investigation of fielding was undertaken by Saikia, Bhattacharjee and Lemmer (2012). By contrast, their approach involved the subjective assessment of every fielding play to provide a weighted measure of fielding proficiency.

On the more technical side, Koulis, Muthukumarana and Briercliffe (2014) proposed a Bayesian hidden Markov model for assessing batting in one-day cricket. The basic modeling assumption is that a player's observed performance (runs scored) is based on his underlying form, and that his underlying form changes over time in accordance with a probability transition matrix with an unknown number of states. Through extensive computation, Koulis, Muthukumarana and Briercliffe (2014) estimate parameters associated with underlying form to obtain measures of performance.

There are two further topics related to player evaluation that are worthy of mention. First, there is some discussion in the literature concerning batting *consistency* (Lemmer 2004 and Borooah and Mangan 2010). The general feeling is that not only is a high batting average a measure of quality but it is preferable if the standard deviation of the runs scored across matches is small. Second, there exist commercial rating methods. The most prominent are the Reliance ICC Player Ratings that have gained stature primarily because they are endorsed by the ICC and appear on the ICC website. These ratings have been developed for batting and bowling in all three formats of cricket. On the downside, the ratings do not have straightforward interpretations as they are based on moving averages involving a somewhat arbitrary compilation of measurements. In addition, the precise details of the ICC ratings may be proprietary as they do not appear to be generally available.

An area of performance evaluation that has not been adequately investigated is performance under irregular conditions. For example, in the second innings of limited overs cricket when a team is chasing, batsmen become more aggressive and take more risks. Consequently, their dismissal rate increases. How should performance be assessed for batsmen and bowlers in these and other irregular situations? Some work in this direction has been considered by Akhtar, Scarf and Rasool (2015) in the context of test cricket. They compare the contribution by players with match outcome probabilities using session by session performances.

6 EVALUATING TEAM STRENGTH

There are several complicating factors that come into play when assessing team strength in international cricket:

- teams do not play often, and hence data are limited; for example, Sri Lanka played only 11 test matches in 2015 (cf. Major League Baseball where the regular season schedule consists of 162 games)
- there is a lack of balance in schedules; for example, New Zealand did not play any Twenty20 matches against Australia in 2015
- there is no consistent dependent variable for winning margins which are reported in different scales; for example, a team may win by 25 runs or may win by 3 wickets (with 61 balls

remaining)

The issues above suggest that comparing cricketing sides is a difficult task. Accordingly, there has not been a lot of work done in the area and there may be opportunities for future research. One idea involves taking into account results from the three formats of cricket. It also strikes the author that the rating/ranking problem is similar to what exists in American college basketball where there are relatively few games and unbalanced schedules. In American college basketball, the topic of ranking is of great importance as tournament invitations are on the line.

The ratings/rankings that are widely referenced in cricket are commercial and are based on subjective combinations of performance measures. They do not offer interpretability nor are they readily convertible to probabilities of victory when one team plays another. Specifically, we point again to the ICC team ratings/rankings which are displayed prominently by the ICC, Cricinfo and the BBC. The rankings are available for each of test, one-day and Twenty20 cricket.

From an academic perspective, there has been little published on the evaluation of team strength. In de Silva, Pond and Swartz (2001), a Bayesian linear model is proposed for one-day cricket where the response variable is run differential. In matches where the team batting second wins, their runs are extrapolated according to the remaining Duckworth-Lewis resources (see section 3). The linear model contains a strength parameter for each team, a home field advantage parameter and an error term where decreasing weight is given to matches occurring further back in time. The interpretation of the team strength parameter is that it is the expected number of runs that the team in question would defeat an average team on a neutral field.

Allsopp and Clarke (2004) proposed models that are very similar to de Silva, Pond and Swartz (2001). Some differences include the use of least squares estimation instead of a Bayesian analysis, decomposing the strength parameter into a batting and bowling parameter, the use of log transforms on the dependent variable and the consideration of first innings results in test matches to provide team ratings for test cricket.

Finally, the match simulator of Davis, Perera and Swartz (2015a) described in section 4 provides a straightforward approach for assessing team strength in Twenty20 cricket. With simulated matches between two teams of interest based on specified lineups, one can calculate various probabilities of interest including the probability that a team defeats its opponent.

7 OPTIMAL LINEUPS

Although fantasy leagues have been around for decades, their popularity has exploded in recent years. A fantasy league is a competition amongst individuals where participants "select" players from a real sports league. A participant's fantasy team then consists of the players that the participant has selected. When a participant's players perform well, they accumulate points, and the fantasy team performs well. Although fantasy leagues have traditionally been an enjoyable past time amongst friends, there is now huge prize money involved through online sites such as DraftKings and FanDuel. Somehow these American enterprises have avoided been classified as gambling sites.

The selection of a fantasy team by a fantasy team owner should obviously depend on how fantasy teams accumulate points. Since fantasy leagues differ in how points are accumulated, there ought to be different objective functions for different fantasy leagues. It is also the author's opinion that a statistical consideration in building fantasy teams that has not been adequately considered is the probability distribution of the points associated with fantasy teams. In terms of winning or placing high in a fantasy league, a fantasy team owner should at least consider the expected number of points that his team accumulates, the variability of points associated with his fantasy team and the number of competitors in the fantasy league. Furthermore, expected fantasy points and point variability should not be straightforward calculations due to the correlations between players. Further complicating the selection problem are constraints imposed by fantasy leagues. For example, sometimes players are assigned a "dollar value" and an owner has a fixed number of dollars to spend.

Clearly, there is considerable strategy involved in selecting fantasy teams. When huge prize money is involved, it is safe to say that the best strategies are kept confidential and are not published in academic journals. In cricket, Brettenny, Friskin, Gonsalves and Sharp (2012) is perhaps the only publication to directly address fantasy team selection.

We now transition to the real sporting problem of team selection; i.e. choosing the 11 cricketers which form a team. In cricket, team selection has traditionally been left to "team selectors" whose methods are largely subjective. There are a number of papers that have been written on team selection algorithms including Lemmer (2013), Ahmed, Deb and Jindal (2013) and Kamble, Venkata Rao, Kale and Savant (2011). The search algorithms used to select teams are varied and include genetic algorithms and binary integer programming. The approaches typically permit constraints on team selection. For example, sometimes a fixed number of pure batsmen, allrounders and bowlers are imposed when forming a team. However, the feature which leads to the greatest differentiation in the approaches lies in the specification of player quality as expressed in the objective function. As we have seen in section 5, there are various measures of performance and these measures are sometimes antagonistic in the context of team selection. For example, selecting a player with a high batting average who seldom makes out reduces the batting opportunities for other players.

In the papers previously mentioned, the problem of interest has been restricted to team selection; i.e. choosing the 11 cricketers to form a team. In the context of Twenty20 cricket, Perera, Davis and Swartz (2016a) extend the problem beyond team selection and also simultaneously consider the determination of the batting order and the bowling order amongst those selected. The objective function that they maximize is expected run differential, the expected number of runs by which one team defeats its opponent. Expected run differential is essentially a proxy for winning, and is therefore a sensible criterion for determining optimal lineups (i.e. team selection, batting order and bowling order). Their approach is again based on the use of the match simulator developed by Davis, Perera and Swartz (2015a) which is briefly described in section 4. The problem of determining optimal lineups is computationally demanding, requiring a search over a vast combinatorial space. For example, given a selected team, there are $11! \approx 40$ million batting orders alone. Perera, Davis and Swartz (2016a) carry out the search through the implementation of a simulated annealing algorithm. The proposed approach has great potential for Twenty20 teams where non-standard yet effective lineups are sometimes revealed.

8 TACTICS

One of the more challenging problems in sports analytics is the discovery and adoption of tactics to improve a team's chance of winning. With cricket being a slow moving game with time for reflection and decision making, it seems well suited for such discoveries.

In test cricket, a common scenario occurs when a team is batting in the third innings and has built a considerable lead. The question that faces this team (Team A) is whether they should voluntarily terminate their innings, and this is known as a declaration. The advantage of declaring early is that there will be more time remaining for Team A to dismiss its opponent and secure a win rather than a draw. The disadvantage is that the target is lower than what it would be if declaration had been delayed, and the opposition may therefore score sufficient runs to defeat Team A. The two relevant papers on the strategic problem of declaration are Scarf and Akhtar (2011) and Perera, Gill and Swartz (2014). Whereas the two approaches are completely different, there is some common ground in terms of the results. Specifically, both papers argue that it is often the case that teams do not declare sufficiently early. This cautiousness causes teams to draw more often and win less often. The misjudgment of declaring late is evident to most fans by the fact that declaring teams rarely lose. Perera, Gill and Swartz (2014) provide specific guidelines (Table 4 of their paper) whether teams should declare at various stages of matches (i.e. overs remaining) and under various circumstances (i.e. weak versus strong opposition, desire to win versus draw).

It is the author's opinion that in sport, the ability of teams to invoke strategy depends partially on the ability to retain possession. For example, the author suggests that in the following freeflowing games, strategy/possession is greatest in basketball followed by soccer followed by hockey. Cricket affords considerable strategy as there is an analogy between the ability to retain possession and the ability to vary the batting approach; both affect the pace and scoring in games. When a batsman is cautious, the dismissal probability decreases with the tradeoff of a lower run scoring rate. Since the level of cautiousness/aggressiveness is under the control of the batsman, this surely introduces strategic opportunities. The study of cautiousness/aggressiveness is an important problem in all three formats of cricket. In addition, training batsmen to have the optimal level of aggressiveness does not seem to be an easy task. In Davis, Perera and Swartz (2015a), it is suggested that teams that are trailing in the second innings of Twenty20 matches wait too long to increase their level of batting aggressiveness. In all formats of cricket, Gauriot and Page (2015) conclude that players and captains do not bat optimally from a team perspective, but sometimes focus on individual incentives at various stages of a match.

Scarf, Shi and Akhtar (2011) use negative binomial distributions to model the runs scored in innings and the runs scored in partnerships during test matches. They then take the view that the batting team can control its run rate and its desired target during the third innings. Under these assumptions, probabilities for winning, losing and drawing a match are estimated. Therefore teams can be tactical in their approach to batting by "setting" the run rate and target accordingly.

In a more technical paper, Preston and Thomas (2000) use survival analysis methodology and dynamic programming to investigate batting strategies in limited overs cricket. For both the first and second innings, they obtain optimal run rates which they compare with actual run rates. Although there are a number of simplifying assumptions in their approach, the paper contains some interesting general results. For example, Preston and Thomas (2000) demonstrate that optimal first innings batting should follow the pattern of an increasing run rate.

Silva, Perera, Davis and Swartz (2016) consider a number of radical tactics in Twenty20 cricket. One approach is based on the premise that wickets are less important in Twenty20 cricket than in other formats of cricket. From this premise, modification to batting and bowling orders are suggested. Another approach concerns a weak team facing a stronger team. The basic idea is that the weak team may force increased variability in the match to improve their chance of winning. This may be accomplished by varying the aggressiveness of batting.

In terms of future research ideas concerning tactics, the strategic selection of bowling options has not been carefully explored. For example, it cannot be the case that all batsmen are equally adept at handling the same type of bowled ball. The author is not aware of data collected on individual batsmen consisting of the batting outcome, ball speed, ball action, ball placement, etc. On the flip side, bowlers make choices in their deliveries. Is there a clear understanding about their most effective bowled balls? And does the sequencing of their bowling decisions matter? Another area where there seems to be opportunity is the placement of fielders. The optimal placement of fielders would depend on the batsman, the bowler and possibly the score.

9 MISCELLANEOUS TOPICS

This section covers six miscellaneous topics in cricket that the author has found interesting. As is typical in research, these topics are usually initiated by a question. And in sport, often the question is one that challenges the status quo.

Topic 1: The home team advantage is ubiquitous in sport with various underlying causes (Stefani 2008). The question in this case involves the assessment of the home team advantage in cricket. As was mentioned in the second bullet point of section 6, the problem is complicated by the fact that winning margins are expressed differently depending on whether the team that bats first wins or loses the match. Also, in quantifying home team advantage, some models propose a common home team advantage whereas other models permit a team dependent home team advantage. In de Silva, Pond and Swartz (2001), the home field advantage is estimated to be worth approximately 15.6 runs in ODI cricket (their model D). This is corroborated by Clarke (1998) whose estimate for the common home field advantage in ODI cricket is 14 runs. In Davis, Perera and Swartz (2015a), a simple calculation of the home field advantage in international Twenty20 cricket is given as 9.0 runs. Over a 10 year period in test cricket, home teams have won, lost and drawn 46%, 28% and 26% of their matches, respectively (www.bettingexpert.com/blog/what-is-

home-field-worth-in-test-cricket).

Topic 2: Related to the home field advantage is the effect due to umpiring decisions. Sacheti, Gregory-Smith and Paton (2015) investigate leg before wicket (LBW) decisions based on 1000 test matches. Amongst their interesting findings is that there was previously a systematic bias by officials which favored the home team. However, since the introduction of pairs of neutral umpires (in 2002), the bias has been virtually eliminated. The bias was estimated at approximately one extra LBW decision per innings in favour of the home team when both umpires had the same nationality as the home team. In a novel application, Manage, Mallawaarachchi and Wijekularathna (2010) use receiver operating characteristic (ROC) curves to assess the quality of decisions by umpires in any format of cricket. The approach permits the comparison/ranking of umpires via simple ROC graphics and the corresponding area diagnostics. The ROC curves are based on the percentage of correct and incorrect umpiring decisions. These days, the *decision review system* (DRS) is a technology that is often utilized and enables retrospective judgements of on-field umpiring decisions. With the availability of DRS in test and ODI cricket, the accuracy of umpiring decisions is now at a very high level.

Topic 3: A cricket match begins with a coin toss where the team that wins the toss has the option between batting or fielding. The question in this case is what is the better decision. Pundits will tell you that the decision depends on your opponent, the pitch conditions, whether it is a day/night game, impending weather, etc. One of the interesting observations is that teams have greatly different perspectives on batting first versus batting second. De Silva and Swartz (1997) report that during 1990-1997, Australia chose to bat first 87% of the time in ODI matches upon winning the coin toss whereas Sri Lanka chose to bat first only 36% of the time. Furthermore, de Silva and Swartz (1997) demonstrate no appreciable advantage to winning the coin toss in ODI matches. Allsopp and Clarke (2004) corroborated this finding in both ODI and test cricket. On the other hand, McGinn (2013) argues that it is better to bat first in day/night games where the estimated advantage is roughly 0.2 runs per over (i.e. 10 runs in a 50 over match). An explanation is that the artificial lighting causes difficulties when batting at night. In addition, some blogs have suggested that particular teams are better at target-setting (batting first) than at chasing (fielding first). One wonders whether these team specific results are due to multiple comparisons issues.

Topic 4: Much has been written in the media about the importance of partnerships and the special symbiosis that exists between certain players. However, is there any truth to this sentiment? Batting appears to be an individualistic activity and one would think that any

top-level partner (who knows when to run and when not to run) may neither help not impede a particular batsman. Valero and Swartz (2012) investigate the importance of opening partnerships in test and one-day cricket. They arrive at the conclusion that synergies in opening partnerships may be considered a sporting myth.

Topic 5: In most sports, sporting bodies are reluctant to make frequent rule changes. When a rule change is made, a question of interest is how the change affects the game. This is particularly important for the sanctity of records in sport. One rule that has been tinkered with repeatedly is the powerplay in one-day cricket. Silva, Manage and Swartz (2015) investigated various powerplay rules in one-day cricket. Their approach was based on creating hypothetical parallel matches; i.e. matches where they impute scoring results into actual matches as if there had been no powerplay. They then compare the actual matches with the parallel matches. They conclude that powerplays increase run production but also increase wickets. It is interesting to note that the powerplay rule has changed again (2015), and has been simplified so that teams no longer determine the start of the powerplay.

Topic 6: One of the unique events in cricket is the annual IPL auction that determines the composition of IPL teams. The auction is different from drafts (Tingling 2016) as teams require different strategies to secure their best possible teams. The IPL auction has not been investigated in detail, and it is suggested that this topic has considerable research potential. Karnik (2010) considered the value of player characteristics using hedonic price models whereas Swartz (2011) argued that anomalies have occurred in previous auctions and that drafts may be preferable for dispersing talent.

10 DISCUSSION

Hopefully, this paper (1) has provided a nearly comprehensive picture of the important problems that have been addressed in cricket analytics and (2) may serve as a starting point for those interested in doing work in cricket analytics. It is anticipated that opportunities for research in cricket will flourish in the coming years as more data become available, especially via player and ball tracking technologies.

What research problems are open to an individual who is interested in doing cricket analytics? One place to look is other sports. For example, baseball is a similar game to cricket and shares similar problems. However, unlike cricket, there has been tremendous amount of quantitative research in the sport of baseball where SABR (the Society for American Baseball Research) boasts a membership of more than 6,000 individuals, publishes articles and hosts meetings. One might also look to the sports gambling world as a benchmark for modeling. Whenever money is involved, predictive modeling tends to be as good as possible.

Although women's cricket has been around for a very long time, there are few (if any) quantitative academic articles on the sport. With media attention increasing, it appears that women's cricket is now on the rise. Investigations that distinguish features between the women's game and the men's game may provide various research opportunities.

As was argued in the Introduction, cricket is one of the world's major sports. And what we have seen lately is that other major sports (e.g. baseball, soccer, football, hockey and basketball) are investing heavily in sports analytics. Forward thinking teams have hired quantitative people who are looking at data to inform them on all aspects of their sport. To gain an edge, it seems that many topics are worthy of investigation including nutrition, travel, strategy, injury prevention, player evaluation, salary constraints, drafting, etc. We therefore conclude the paper with a provocative question. When will the gentleman's game of cricket begin to encourage and adopt modern analytics?

11 REFERENCES

- Ahmed, F., Deb, K. and Jindal, A. (2013). Multi-objective optimization and decision making approaches to cricket team selection. Applied Soft Computing, 13, 402-414.
- Akhtar, S., Scarf, P.A. and Rasool, Z. (2015). Rating players in test match cricket. Journal of the Operational Research Society, 66, 684-695.
- Allsopp, P.E. and Clarke, S.R. (2004). Rating teams and analysing outcomes in one-day and test cricket. Journal of the Royal Statistical Society, Series A, 167, 657-667.
- Beaudoin, D. and Swartz, T.B. (2003). The best batsmen and bowlers in one-day cricket. South African Statistical Journal, 37, 203-222.
- Borooah, V.K. and Mangan, J.E. (2010). The "Bradman Class": an exploration of some issues in the evaluation of batsmen for test matches, 1877-2006. *Journal of Quantitative Analysis in Sports*, 6, Article 14.
- Brettenny, W.J., Friskin, D.G., Gonsalves, J.W. and Sharp, G.D. (2012). A multi-stage integer programming approach to fantasy team selection: A Twenty20 cricket study. South African Journal for Research in Sport, Physical Education and Recreation, 34, 13-28.
- Carter, M. and Guthrie, G. (2004). Cricket interruptus: fairness and incentive in limited overs cricket matches. Journal of the Operational Research Society, 55: 822-829.

- Christos, G.A. (1998). It's just not cricket. In *Mathematics and Computers in Sport*, N. de Mestre and K. Kumar (editors), Bond University, Queensland, Australia, 181-188.
- Clarke, S.R. (1988). Dynamic programming in one-day cricket optimal scoring rates. Journal of the Operational Research Society, 39, 331-337.
- Clarke, S.R. (1998). Test statistics. in *Statistics in Sport*, J. Bennett (editor), Arnold Applications of Statistics Series, Arnold: London, 83-103.
- Croucher, J.S. (2000). Player ratings in one-day cricket. In *Mathematics and Computers in Sport*, G. Cohen and T. Langtry (editors). Bond University, Queensland, Australia, 1-13.
- Davis, J., Perera, H. and Swartz, T.B. (2015a). A simulator for Twenty20 cricket. Australian and New Zealand Journal of Statistics, 57, 55-71.
- Davis, J., Perera, H. and Swartz, T.B. (2015b). Player evaluation in Twenty20 cricket. Journal of Sports Analytics, 1, 19-31.
- de Silva, B.M., Pond, G.R. and Swartz, T.B. (2001). Estimating the magnitude of victory in one-day cricket. Australian and New Zealand Journal of Statistics, 43, 259-268.
- de Silva, B.M. and Swartz, T.B. (1997). Winning the coin toss and the home team advantage in one-day international cricket matches. *New Zealand Statistician*, 32, 16-22.
- Duckworth, F.C. and Lewis, A.J. (1998). A fair method for resetting targets in one-day cricket matches. Journal of the Operational Research Society, 49, 220-227.
- Duckworth, F.C. and Lewis, A.J. (2004). A successful operational research intervention in one-day cricket. Journal of the Operational Research Society, 55, 749-759.
- Elderton, W.E. (1945). Cricket scores and some skew correlation distributions. Journal of the Royal Statistical Society, Series A, 108, 1-11.
- Gauriot, R. and Page, L. (2015). I take care of my own: a field study on how leadership handles conflict between individual and collective incentives. American Economic Review: Papers and Proceedings, 105, 414-419.
- Jayadevan, D. (2002). A new method for the computation of target scores in interrupted, limited over cricket matches. *Current Science*, 83, 577-586.
- Kamble, A.G., Venkata Rao, R., Kale, A.V. and Samant, S.P. (2011). Selection of cricket players using analytical hierarchy process. *International Journal of Sports Science and Engineering*, 5, 207-212.
- Karnik, A. (2010). Valuing cricketers using hedonic price models. Journal of Sports Economics, 11, 456-469.
- Koulis, T., Muthukumarana, S. and Briercliffe, C.D. (2014). A Bayesian stochastic model for batting performance evaluation in one-day cricket. *Journal of Quantitative Analysis in Sports*, 10, 1-13.
- Lemmer, H.H. (2004). A measure for the batting performance of cricket players. South African Journal for Research in Sport, Physical Education and Recreation, 26, 55-64.

- Lemmer, H.H. (2011). The single match approach to strike rate adjustments in batting performance measures in cricket. *Journal of Sports Science and Medicine*, 10, 630-634.
- Lemmer, H.H. (2013). Team selection after a short cricket series. *European Journal of Sport Science*, 13, 200-206.
- Lewis, A.J. (2005). Towards fairer measures of player performance in one-day cricket. *Journal of the Operational Research Society*, 56, 804-815.
- Manage, A.B.W., Mallawaarachchi, K. and Wijekularathna, K. (2010). Receiver operating characteristics (ROC) curves for meauring the quality of decisions in cricket. *Journal of Quantitative Analysis* in Sports, 6, Article 8.
- Manage, A.B.W. and Scariano, S.M. (2013). An introductory application of principal components to cricket data. *Journal of Statistics Education* [electronic journal], 21, Article available online at http://www.amstat.org/publications/jse/v21n3/scariano.pdf.
- McGinn, E. (2013). The effect of batting during the evening in cricket. Journal of Quantitative Analysis in Sports, 9, 141-150.
- McHale, I.G. and Asif, M. (2013). A modified Duckworth-Lewis method for adjusting targets in interrupted limited overs cricket. *European Journal of Operational Research*, 225, 353-362.
- Perera, H., Davis, J. and Swartz, T.B. (2016a). Optimal lineups in Twenty20 cricket. *Journal of Statistical Computation and Simulation*, To appear.
- Perera, H., Davis, J. and Swartz, T.B. (2016b). Assessing the impact of fielding in Twenty20 cricket. Journal of the Operational Research Society, Under revision.
- Perera, H., Gill, P.S. and Swartz, T.B. (2014). Declaration guidelines in test cricket. Journal of Quantitative Analysis in Sports, 10, 15-26.
- Perera, H. and Swartz, T.B. (2013). Resource estimation in Twenty20 cricket. IMA Journal of Management Mathematics, 24, 337-347.
- Preston, I. and Thomas, J. (2000). Batting strategy in limited overs cricket. *The Statistician*, 49, 95-106.
- Saikia, H., Bhattacharjee, D. and Lemmer, H.H. (2012). A double weighted tool to measure the fielding performance in cricket. *International Journal of Sports Science and Coaching*, 7(4), Article 6.
- Sacheti, A. and Gregory-Smith, I. (2015). Home bias in officiating: evidence from international cricket. Journal of the Royal Statistical Society, Series A, 178, 741-755.
- Scarf, P.A. and Akhtar, S. (2011). An analysis of strategy in the first three innings in test cricket: declaration and the follow-on. *Journal of the Operational Research Society*, 62, 1931-1940.
- Scarf, P.A., Shi, X. and Akhtar, S. (2011). On the distribution of runs scored and batting strategy in test cricket. Journal of the Royal Statistical Society, Series A, 174, 471-497.

- Silva, R., Manage, A.B.W. and Swartz, T.B. (2015). A study of the powerplay in one-day cricket. European Journal of Operational Research, 244, 931-938.
- Silva, R., Perera, H., Davis, J. and Swartz, T.B. (2016). Tactics for Twenty20 cricket. South African Statistical Journal, To appear.
- Stefani, R. (2008). Measurement and interpretation of home advantage. In Statistical Thinking in Sports, R. Koning and J.H. Albert (editors), Chapman and Hall/CRC: Boca Raton, 203-216.
- Stern, S.E. (2016). The Duckworth-Lewis-Stern method: Extending the Duckworth-Lewis methodology to deal with modern scoring rates. *Journal of the Operational Research Society*, doi:10.1057/jors.2016.30.
- Swartz, T.B. (2011). Drafts versus auctions in the Indian Premier League. South African Statistical Journal, 45, 249-272.
- Tingling, P. (2016). Educated guesswork drafting in the National Hockey League. In Handbook of Statistical Methods and Analyses in Sports, J.H. Albert, M.E. Glickman, R. Koning and T.B. Swartz (editors), Chapman and Hall/CRC: Boca Raton, xxx-xxx.
- Valero, J. and Swartz, T.B. (2012). An investigation of synergy between batsmen in opening partnerships. Sri Lankan Journal of Applied Statistics, 13, 87-98.
- van Staden, P.J. (2009). Comparison of cricketers' bowling and batting performance using graphical displays. *Current Science*, 96, 764-766.
- Weatherall, D. (2011). Cross-validation of rain rules for interrupted one-day cricket matches. MSc Thesis, Department of Mathematical Statistics and Actuarial Science, University of the Free State, South Africa.
- Wood, G.H. (1945). Cricket scores and geometrical progression. Journal of the Royal Statistical Society, Series A, 108, 12-22.