Assessing the Impact of Fielding in Twenty20 Cricket

Harsha Perera, Jack Davis and Tim B. Swartz *

Abstract

This paper attempts to quantify the importance of fielding in Twenty20 cricket. We introduce the metric of expected runs saved due to fielding which is both interpretable and is directly relevant to winning matches. The metric is assigned to individual players and is based on a textual analysis of match commentaries using random forest methodology. We observe that the best fielders save on average 1.2 runs per match compared to a typical fielder.

Keywords: Random forests, Simulation, Textual analysis, Twenty20 cricket.

*Harsha Perera and Jack Davis are PhD candidates, and 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 authors thank two anonymous reviewers for helpful comments that improved the manuscript.

1 INTRODUCTION

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

In Twenty20 cricket, Davis, Perera and Swartz (2015a) introduced simulation procedures that estimate the expected number of runs that players contribute to their teams in terms of batting and bowling when compared against average players. In cricket, a player's expected run differential (in comparison to an average player) is an unequivocal performance measure of interest as it leads to directly to wins and losses. In Davis, Perera and Swartz (2015a), it is seen that the best Twenty20 players contribute nearly 8-10 additional runs when compared to average players. Yet, there have been no similar investigations on the effects of fielding. An exceptional fielding play in cricket may be described as "fantastic" but there is no sense of the scale of the contribution in terms of runs. This paper is a first attempt to quantify the impact of fielding.

In an ideal world, cricket matches would yield detailed spatial data where fielding contributions could be objectively assessed. For example, one could determine the distance covered by a fielder in making a catch and the time that it takes to cover the distance. In Major League Baseball (MLB), spatial data is facilitated through FIELDf/x technology and papers have been written (e.g. Jensen, Shirley and Wyner 2009) that assess fielding contributions via spatial statistics. Similarly, in the National Basketball Association (NBA), spatial data is provided through the SportVU player tracking system which records the coordinates of the ball and each player on the court 25 times per second. Detailed data of this sort have provided opportunities to investigate lesser-studied basketball characteristics such as defensive proficiency (Franks, Miller, Bornn and Goldsberry 2015). As cricket is the second most popular game in the world (following soccer), it seems only a matter of time until spatial data will likewise be available for analysis.

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

Broadly speaking, this paper introduces "moneyball" concepts which strike at the core of what is important in Twenty20 cricket. The book Moneyball (Lewis 2003) and its ensuing

Hollywood movie starring Brad Pitt chronicled the 2002 season of the Oakland Athletics, a small-market MLB team who through advanced analytics recognized and acquired undervalued baseball players. Moneyball analyses often rely on *relative value statistics*. Relative value statistics have become prominent in the sporting literature as they attempt to quantify what is really important in terms of winning and losing matches. For example, in Major League Baseball (MLB), the VORP (value over replacement player) statistic has been developed to measure the impact of player performance. For a batter, VORP measures how much a player contributes offensively in comparison to a replacement-level player (Woolner 2002). A replacement-level player is a player who can be readily enlisted from the minor leagues. Baseball also has the related WAR (wins above replacement) statistic which is gaining a foothold in advanced analytics (http://bleacherreport.com/articles/1642919). In the National Hockey League (NHL), the plus-minus statistic is prevalent. The statistic is calculated as the goals scored by a player's team minus the goals scored against the player's team while the player is on the ice. More sophisticated versions of the plus-minus statistic have been developed by Schuckers et al. (2011) and Gramacy, Taddy and Jensen (2013). In soccer, McHale and Szczepanski (2014) identify goal scoring ability by removing extraneous features. They use mixed effects models where variables such as team strength, home field advantage and shot opportunity are considered. In this paper, we introduce *expected runs saved due to fielding* in Twenty20 cricket. This may be classified as a relative value statistic as it measures the number of runs that a team saves on average due to the fielding performance of a given player when compared against a baseline player.

To our knowledge, the first and only quantitative investigation of fielding was undertaken by Saikia, Bhattacharjee and Lemmer (2012). They proposed measures that are based on subjective weights. By contrast with the statistic proposed in this paper, their approach requires a video assessment of every fielding play to provide a measure of fielding proficiency.

In Section 2, we provide an overview of the match simulator developed by Davis, Perera and Swartz (2015b). The simulator is the backbone for calculating expected runs saved due to fielding. For the casual reader, this section can be skimmed, as it is only important to know that methodology has been developed for realistically simulating Twenty20 matches. In Section 3, we define the proposed metric of expected runs saved due to fielding and describe its calculation via simulation methodology. The calculation relies on a fielding matrix Λ for a given fielder. The estimation of Λ is carried out by a textual analysis using random forest methodology. An innovation in our estimation procedure involves the amalgamation of matches from International Twenty20 cricket and the Indian Premier League (IPL). In Section 4, we calculate expected runs

saved due to fielding for players with a sufficient Twenty20 history. Some surprising results are revealed and we are able to put into context the importance of fielding in Twenty20 cricket. We conclude with a short discussion in Section 5.

2 OVERVIEW OF SIMULATION METHODOLOGY

We now provide an overview of the match simulator developed by Davis, Perera and Swartz (2015b) which we use for the calculation of expected runs saved due to fielding. There are 8 broadly defined outcomes that can occur when a batsman faces a bowled ball. These batting outcomes are listed below:

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

In the list (1) of possible batting outcomes, *extras* such as *byes*, *leg byes*, *wide-balls* and *no balls* are excluded. In the simulation, extras are introduced by by generating occurrences at the appropriate rates. Extras occur at the rate of 5.1% in Twenty20 cricket. The outcomes $j = 3$ and $j = 5$ are rare but are retained to facilitate straightforward notation.

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

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

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

The estimation of the multinomial parameters in (2) is a high-dimensional and complex prob-

lem. The complexity is partly due to the sparsity of the data; there are many match situations (i.e. combinations of overs and wickets) where batsmen do not have batting outcomes. For example, bowlers typically bat near the end of the batting order and do not face situations when zero wickets have been taken.

To facilitate the estimation of the multinomial parameters p_{iowj}, Davis, Perera and Swartz (2015b) introduced the simplification

$$p_{iowj} = \frac{\tau_{owj}\ p_{i70j}}{\sum_j \tau_{owj}\ p_{i70j}}\ . \tag{3}$$

In (3), the parameter p_{i70j} represents the baseline characteristic for batsman i with respect to batting outcome j. The characteristic p_{i70j} is the probability of outcome j associated with the ith batsman at the juncture of the match immediately following the *powerplay* (i.e. the 7th over) when no wickets have been taken. The multiplicative parameter τ_{owj} scales the baseline performance characteristic p_{i70j} to the stage of the match corresponding to the oth over with w wickets taken. The denominator in (3) ensures that the relevant probabilities sum to unity. There is an implicit assumption in (3) that although batsmen are unique, their batting characteristics change with respect to overs and wickets by the same multiplicative factor which is essentially an indicator of aggression. For example, when aggressiveness increases relative to the baseline state, one would expect $\tau_{ow4} > 1$ and $\tau_{ow6} > 1$ since bolder batting leads to more 4's and 6's.

Given the estimation of the parameters in (3) (see Davis, Perera and Swartz 2015b), 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 (2015b) 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.

3 THE APPROACH

Recall that the match simulator of Section 2 generates batting outcomes according to situational probabilities. Specficially, p_{iowj} is the probability of batting outcome j when batsman i is batting in the oth over having lost w wickets. We first consider an average batting team (i.e. average positional player at each batting position) batting against an average fielding team. Simulating

over many first innings (based on the Twenty20 International dataset in Davis, Perera and Swartz 2015b), we obtained the expected runs scored $\mathrm{E}(R) = 149.9$.

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

$$\mathrm{E}(RSF) \;\; = \;\; \mathrm{E}(R) - \mathrm{E}(R^*) \tag{4}$$

which is the expected runs saved due to fielding. The larger the value of E(RSF) in (4), the better the fielder. We note that E(RSF) is an appealing statistic as it is directly interpretable in terms of runs.

In order to carry out the simulations required in (4), it is necessary to obtain the batting characteristics p^*_{iowj} which have been modified from p_{iowj} due to the presence of the fielder of interest. For ease of notation, we temporarily suppress the subscripts *iow* and express

$$p^*_k = \lambda_{0k}p_0 + \lambda_{1k}p_1 + \cdots + \lambda_{7k}p_7 \; . \tag{5}$$

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

$$\Lambda = (\lambda_{jk}) \tag{6}$$

for the fielder of interest. The matrix Λ describes the impact of the fielder in transforming typical batting outcomes to other batting outcomes due to his fielding. From (5), we then have

$$p^* = \Lambda^T p \tag{7}$$

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

be large (i.e. close to unity).

To summarize, we first consider a typical batting lineup against a typical bowling lineup based on the batting characteristics p. Via simulation, this typical team scores $\mathrm{E}(R) = 149.9$ first innings runs on average. With the inclusion of a fielder of interest, the batting characteristics are modified from p to p^* via (7) and the expected runs scored with the presence of the fielder can be simulated to obtain $\mathrm{E}(R^*)$. Therefore, the fielder's contribution in terms of expected runs saved due to fielding is given by $\mathrm{E}(RSF) = 149.9 - \mathrm{E}(R^*)$. In the next section, we use textual analysis and random forests to estimate the fielding matrix Λ.

3.1 Parameter Estimation

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

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

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

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

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

The first example highlights an example corresponding to good fielding whereas the second example corresponds to poor fielding. As can be seen, there is a diversity of language in the match commentaries to describe events. This is suggestive of the use of machine learning tools to reveal patterns in the text. Unfortunately, the commentary logs do not provide information on the names of the announcers; this could lead to a potential bias.

Our approach partitions the 160,247 balls in the dataset (i.e. lines of commentary) into A: the 146,452 cases where no fielders' names are mentioned and B: the remaining 13,796 cases where fielders' names are mentioned. We designated 55 keywords (see Table 1) in dataset A which are features that provide predictive information on the batting outcome (i.e. the observed dependent variable (1)). The keywords were obtained from an initial list containing the most frequently occurring words in the training set A after removing *stop* words (e.g. prepositions, pronouns), words that directly restate the outcome (e.g. "run", "six", "catch", etc.) and player names. We also removed clearly ambiguous words that could be used to inform on both the outcome of the batting performance and the fielding performance (e.g. "brilliant"). The most common keyword is "leg" which appears 17,297 times in dataset A. The least common keyword in Table 1 is "gloves" which appears 184 times in dataset A. We observe the contextual nature of the words in Table 1 which provide insight on the outcome from a bowled ball.

across	chipped	flash	leg	reverse
air	clip	flick	length	short
almost	cut	foot	long	side
angle	deep	forward	low	slog
back	delivery	front	middle	slower
backward	drill	gap	midwicket	square
banged	drive	gloves	pads	strike
bat	easy	hard	pace	stump
boundary	edge	high	pads	sweep
butt	fine	hit	power	swing
charge	firm	knee	pull	swung

Table 1: Contextual words used in random forest to predict batting outcome probabilities.

A random forest (Hastie, Tibshirani and Friedman 2009) was then grown on dataset A using the contextual words as covariates and the batting outcome (1) as the response. The resultant forest provides predictive probability distributions for the outcomes in (1) for any line of commentary. More specifically, we used the `randomForest()` function from the `randomForest` package in R. The random forest procedure has various tuning parameters to optimize predictive performance. For the application discussed here, we trained the random forest on a simple random sample of 100,000 commentary lines from dataset A. We then used the remaining observations as a validation set for choosing the tuning parameters. The optimal predictive performance was found using 2,500 trees, with each leaf including at least five observations. At each node, the best split was found by searching over a random size 15 subset of the total covariates. The covariates are binary variables indicating the presence or absence of the keywords in Table 1. This random subsetting causes the individual regression trees in the random forest to be less correlated and allows the model to identify subtle effects of keywords that might otherwise be missed by individual regression trees.

We now describe how the random forest is used to estimate the fielding parameters λ_{jk} for a specified fielder. For a specified fielder, suppose that he has been on the field for n balls and suppose that his name has been mentioned in the commentary lines of dataset B for m balls. Designate his noteworthy plays with the index $i = 1, \ldots, m$ such that the random forest produces predicted outcome probabilities $q_{i0}, \ldots, q_{i7}$. These are the probabilities corresponding to what would have happened had the fielder not made a noteworthy fielding play. However, the ith fielding play did produce a realized batting outcome and we denote this outcome by O_i. This leads to the following estimators

$$\hat{\lambda}_{kk} = \frac{n-m}{n} + \frac{m}{n}\left(\frac{\sum_{i=1}^{m} I(O_i = k) q_{ik}}{\sum_{i=1}^{m} q_{ik}}\right) \tag{8}$$

and

$$\hat{\lambda}_{jk} = \frac{m}{n}\left(\frac{\sum_{i=1}^{m} I(O_i = k) q_{ij}}{\sum_{i=1}^{m} q_{ij}}\right) \tag{9}$$

for $j \neq k$ where I is the indicator function. A detailed construction of (8) and (9) is provided in the Appendix.

We note that equations (8) and (9) make sense theoretically. If a player's name is never mentioned in dataset B, then $m = 0$ and $\hat{\lambda}_{kk} = 1$. This implies that a player never makes mistakes and never makes exceptional fielding plays. Consequently $p^* = p$ for the player of interest, and

through simulation, $\mathrm{E}(RSF) = 0$. Later, in our data analysis, we see that it is easier for players to make mistakes than to make exceptional fielding plays. Therefore $\mathrm{E}(RSF) = 0$ actually corresponds to an elevated standard of play.

Estimators (8) and (9) may be unreliable in the case where a fielder has been on the field for very few balls (i.e. n is small). In this case, one might consider some sort of shrinkage estimator so that estimates are not inflated when a fielder has a limited history. In the analysis of Section 4, our approach to this difficulty is to restrict attention to players who have been on the field for at least 2,000 balls bowled.

A slight difficulty with (8) and (9) is that machine learning algorithms do not take into account the physical aspects of the game. In cricket, there are various scenarios which cannot occur. For example, it would be impossible for a fielder to convert what would have been zero runs to six runs no matter how poor his fielding. In terms of the fielding matrix, this implies the conditional probability $\lambda_{06} = 0$. This also holds true for other scenarios and we therefore introduce the constraints $\lambda_{06} = \lambda_{16} = \lambda_{26} = \lambda_{46} = \lambda_{60} = 0$. When we estimate the Λ matrix using (8) and (9), we make a final adjustment by scaling the non-zero λ's in a given row proportionally so that each row sums to unity.

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

In the next section, we see that MS Dhoni, India's legendary wicketkeeper, has been identified by our methodology as a surprisingly poor fielder according to the $\mathrm{E}(RSF)$ metric. Table 2 provides the estimated fielding matrix for Dhoni. Note that we have combined outcomes $j = 2$ and $j = 3$, and we have combined outcomes $j = 4$ and $j = 5$. This was done since 3's and 5's are rare. One thing we observe from Table 2 is that λ_{70} is the largest of the λ_{jk} values where $j \neq k$. This means that Dhoni occasionally drops potential catches and misses stumping opportunities. One of the other things that we observe from Table 2 is that the relative magnitudes of the λ_{jk}

conform to common sense. For example, it seems that λ_{01} should exceed λ_{02}. The reason is that the consequences of fielding errors from lightly hit balls are easily contained.

		k					
		0	1	2, 3	4, 5	6	7
	0	0.983	0.006	0.001	0.003	0.000	0.006
	1	0.007	0.981	0.004	0.004	0.000	0.004
j	2, 3	0.006	0.007	0.978	0.004	0.000	0.005
	4, 5	0.006	0.004	0.002	0.982	0.000	0.006
	6	0.000	0.006	0.007	0.006	0.976	0.006
	7	0.008	0.006	0.001	0.003	0.000	0.982

Table 2: Estimated fielding matrix $\Lambda = (\lambda_{jk})$ for MS Dhoni.

As pointed out by an anonymous Referee, it would be desirable to provide estimates of variability with respect to the estimates in Table 2. We note that this is a most difficult problem. For example, referring to estimators (8) and (9), we observe that there are various sources of variability. We have the variability in m which is the binomial variability due to the number of times a player makes an exceptional play (good or bad) out of the n potential balls. Then there is the variability in O_i which is the variability due to making one of the 8 outcomes listed in (1). Both of these sources of variability are associated with dataset B. Finally, there is the variability in the q_{ij} terms which is due to the fitting of the random forest on dataset A. The amalgamation and quantification of these various sources of variability is something to be considered in future research.

4 PLAYER ANALYSIS

We now consider the analysis of specific players. We restrict our attention to the 157 players who have been on the field for at least 2,000 balls bowled. Therefore our analysis consists of well-established Twenty20 players.

In Table 3, we consider the E(RSF) metric for wicketkeepers. We have included all 13 wicketkeepers in our dataset. Wicket-keepers are considered separately as their positioning on the field allows them greater opportunity to make fielding plays. We observe that Mushfiqur Rahim of Bangladesh is the best fielding wicketkeeper. At the bottom of the list, we were suprised to find Mark Boucher of South Africa who was long considered a top cricketer for a top cricketing nation.

Another observation concerning Table 3 is that there is a negative skewness in the E(RSF) statistic. This can be explained by noting that fielders may make mistakes in various ways. For example, they can drop balls, they can make throwing errors, they can trip, they can fail to stop balls, etc. On the other hand, there are limited opportunites for a fielder to make an exceptional play. For example, if a batted ball is not within reach, nothing can be done.

Name	Team	n	m	E(RSF)
M Rahim	BD	2995	18	0.22
T Taibu	ZIM	2017	24	-0.19
KC Sangakkara	SL	11058	135	-0.46
Q de Kock	SA	3303	43	-0.57
AC Gilchrist	AUS	6788	125	-1.02
D Ramdin	WI	5189	119	-1.11
K Akmal	PAK	6696	138	-1.44
BB McCullum	NZ	6251	120	-2.23
JC Buttler	ENG	3138	60	-2.94
MS Wade	AUS	2473	55	-3.37
BJ Haddin	AUS	4995	137	-3.50
MS Dhoni	IND	15980	467	-3.61
MV Boucher	SA	5490	143	-4.28

Table 3: Expected runs saved due to fielding by wicketkeepers. The variable n refers to the fielder's total number of fielding opportunities and m is the number of notable plays by the fielder such that his name appeared in dataset B.

In Table 4, we list the top 10 fielders who are non-wicketkeepers. It would have been nice to differentiate those players who are positioned in the infield and are more often in positions to make fielding plays. However, the game is fluid and players frequently change positions on the field according to the game conditions. We note that although Nathan Coulter-Nile of Australia is ranked as the best fielder, the fielding effects expressed by E(RSF) are small amongst the top fielders. On the CricInfo website (http://www.espncricinfo.com/australia/content/player/261354.html), Coulter-Nile is described by the national selector John Inverarity as "one of the three or four best fieldsmen in Australia". In the media, there seems to be a belief that AB de Villiers of South Africa is a remarkable fielder. According to our methodology, he ranks 21st of the 144 non-wicketkeepers with E(RSF) $= -0.34$. We note that many of the players in Table 4 have small values of m. This means that while fielding, their names rarely appear. In fact, this is a good thing. Such players are not mentioned because they make few mistakes. As discussed previously, there are more opportunities to make mistakes than to make exceptional fielding plays. The average E(RSF) value amongst the 144 fielders is -1.20 and the lowest value is -4.03 (Brett

Lee of Australia). Therefore, the top fielders who have a positive E(RSF) metric save on average more than 1.2 runs per match compared to a typical fielder.

Name	Team	n	m	E(RSF)
NM Coulter-Nile	AUS	2079	11	0.35
P Utseya	ZIM	2345	6	0.21
Y Khan	PAK	2444	8	0.18
H Masakadza	ZIM	3116	10	0.17
SE Marsh	AUS	5950	42	0.12
RE van der Merwe	SA	3316	24	0.12
AD Mascarenhas	ENG	2571	22	0.11
P Kumar	IND	7031	9	0.08
DP Nannes	AUS	4720	13	0.05
HDRL Thirimanne	SL	2111	7	0.04

Table 4: Expected runs saved due to fielding by the top 10 non-wicketkeepers. The variable n refers to the fielder's total number of fielding opportunities and m is the number of notable plays by the fielder such that his name appeared in dataset B.

In Figure 1, we provide a plot of m/n versus E(RSF). What is interesting is the decreasing trend. The trend implies that players who are infrequently mentioned relative to their fielding opportunities are doing a good job of fielding. In other words, if a player is seldom identified for his fielding, then he is likely not making mistakes and is contributing to his team. Again, in a cricket match, there are more opportunities to make mistakes than to do something exceptional with respect to fielding during a match. We also note that wicketkeepers tend to be mentioned more frequently.

5 DISCUSSION

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

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

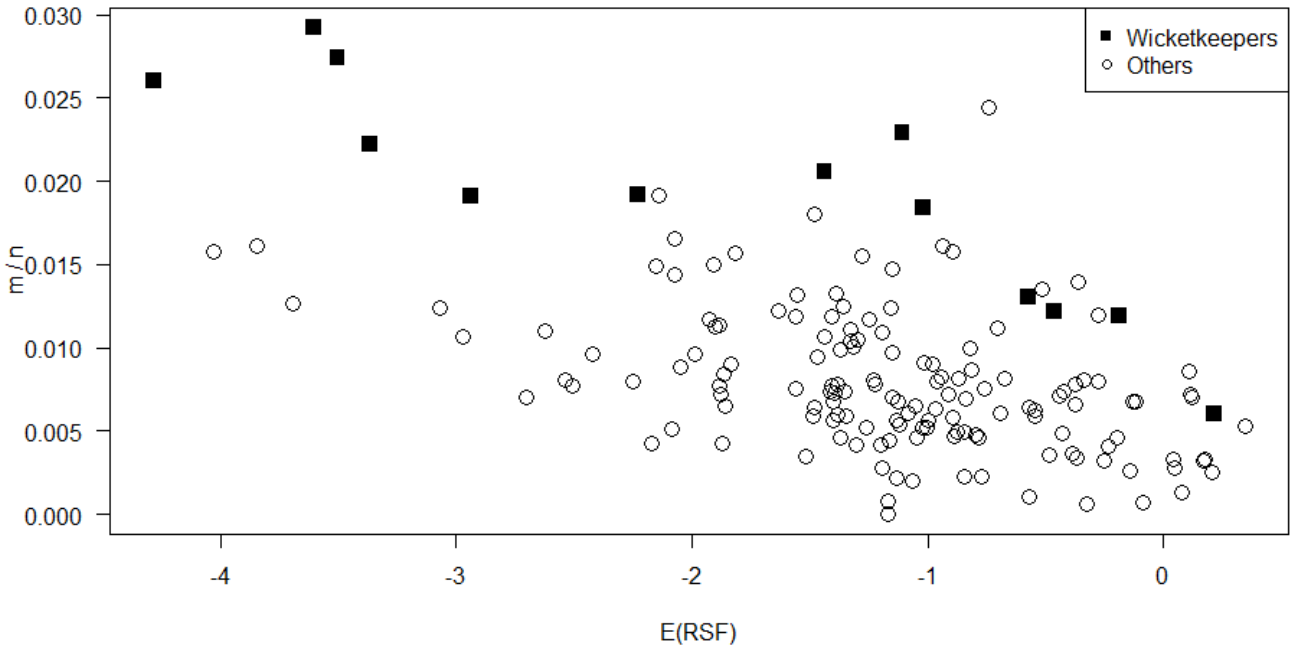

Figure 1: Scatterplot of the ratio of mentions to fielding opportunities (m/n) versus E(RSF).

A drawback of our approach is that we only consider the actions of fielders. For example, it is possible that the mere presence of a good fielder has an impact. Batsmen may try to avoid hitting balls to such a fielder. In the data analysis, we have distinguished the fielding contributions made by wicketkeepers and non-wicketkeepers. It is also possible that some fielding positions (e.g. cover) are more difficult from the perspective of fielding than other positions. It may therefore be advisable to only compare fielders who play similar positions.

Despite the above caveats, we have quantified the impact of fielding. Whereas the best individual batters and bowlers contribute roughly 10 runs per match on average in Twenty20 cricket (Davis, Perera and Swartz 2015a), we have found that a good fielder may only save 1.2 expected runs for his team. Whereas this may seem small, it is possible to have multiple good fielders on a team, and therefore the impact of fielding becomes more meaningful.

Now that the impact of fielding can be described in terms of an easily understood quantity (runs), it is possible that better decision making can be made with respect to team selection and salaries. Furthermore, there seems to be a divergence of popular opinion concerning the very best fielders; our Tables 3 and 4 can shed some quantitative light on this topic.

6 REFERENCES

Davis, J., Perera, H. and Swartz, T.B. (2015a). Player evaluation in Twenty20 cricket. *Journal of Sports Analytics*, 1(1), 19-31.

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

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

Gramacy, R.B., Taddy, M.A. and Jensen, S.T. (2013). Estimating player contribution in hockey with regularized logistic regression. *Journal of Quantitative Analysis in Sports*, 9, 97-112.

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

Jensen, S.T., Shirley, K.E. and Wyner, A.J. (2009). Bayesball: A Bayesian hierarchical model for evaluating fielding in Major League Baseball. *The Annals of Applied Statistics*, 3(2), 491-520.

Lewis, M. (2003). *Moneyball: The Art of Winning an Unfair Game*, WW Norton, New York.

McHale, I.G. and Szczepanski, L. (2014). A mixed effects model for identifying goal scoring ability of footballers. *Journal of the Royal Statistical Society, Series A*, 177, Part2, 397-417.

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.

Schuckers, M.E., Lock, D.F., Wells, C., Knickerbocker, C.J. and Lock, R.H. (2011). National Hockey League skater ratings based upon all on-ice events: An adjusted minus/plus probability (AMPP) approach. Unpublished manuscript.

Woolner, K. (2002). Understanding and measuring replacement level. In *Baseball Prospectus 2002*, J. Sheehan (editor), Brassey's Inc, Dulles, VA, pp. 55-66.

7 APPENDIX

Here we provide detailed construction of the estimators (8) and (9). For clarity, we introduce some additional notation. Let N denote a notable fielding play for a given fielder, one that would appear in dataset B of the match commentary. Let A_k denote that the actual batting outcome is k and let S_j denote that the standard batting outcome is j. A standard batting outcome is the outcome that would have occurred without the impact of a notable fielding play. Then using the

rules of conditional probability, the probability of outcome k due to the presence of the fielder is given by

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

where we recall that p_j is the probability of batting outcome j when a typical fielder is present. Referring to (5) and matching coefficients, the fielding characteristics are therefore given by

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

and

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

for $j \neq k$. Finally, the estimators (8) and (9) are obtained by using the sample proportions $\hat{\Pr}(N) = m/n$ and $\hat{\Pr}(A_k \mid S_j \cap N) = \sum_{i=1}^{m} I(O_i = k)q_{ij} / \sum_{i=1}^{m} q_{ij}$. We recall that n is the observed number of fielding opportunities by the fielder of interest. From these n opportunities, he made m notable fielding plays, O_i is the batting outcome associated with the ith notable fielding play and q_{ij} is the probability obtained from the random forest that the ith notable fielding play would have resulted in batting outcome j had the fielding play not been notable.