Cricket Prognostic System: A framework for real-time analysis in ODI cricket

Aman Verma, Masoumeh Izadi
Broadcast Solutions Group
1 Changi North Street 1
Singapore 498789
(+65)65429093
aman.verma.mtl@gmail.com,
masoumeh.izadi@bsgroup.tv

ABSTRACT
In this paper, we present the Cricket Prognostic System (CPS), an analytical research framework for the sport of cricket. Using longitudinal ball-by-ball data and historical data, CPS employs advanced machine learning and statistical methods to analyze and predict events. More than thirty dynamic variables on real-time statistics along with a comprehensive and in-depth set of indicators based on historical data are used in an array of products that aim to engage fans in cricket coverage. This is novel in cricket, as, to our knowledge, no current work has defined the underlying metrics and progression of such models to convey predictions in One Day International (ODI) cricket.

Keywords
Cricket; sports analytics; predictive models; player performance

1. INTRODUCTION
Cricket is a game played between two teams of 11 players each, where the two teams alternate scoring (batting) and defending (bowling and fielding). The 11 players who bat on a team in an innings are selected in an order that is defined based on the team's strategy against the opponent. Cricket is played professionally in three formats: test, One Day International (ODI) and Twenty20 (T20). Test cricket is the most traditional form of cricket, typically lasting several days. Early this century, the International Cricket Council (ICC) has sought to broaden the popularity of the sport by shortening the length of games to a limited number of overs (an over consists six ball deliveries). Shorter matches were intended to make the game more exciting, make it more appealing to broadcasters, and attract a new audience. Internationally, there are two types of limited over formats: Twenty-20 International (T20I) which has a maximum of 20 overs (120 ball deliveries) in each of two innings, and ODI, which has a maximum 50 overs (300 ball delivery) in each of two innings. An innings can also stop if ten of eleven batters are dismissed before the maximum over limit. Often, the second innings will stop before the maximum number of batters because the batting team has scored enough runs to secure victory.

Numerous research papers have considered probabilistic analysis in tennis, baseball, basketball, ice hockey, and American football, yet, hardly any such analysis has previously taken place in cricket, as Clarke [1] describes in a survey of papers written on cricket analytics. Although cricket is not a fast-paced sport like soccer, basketball, or hockey, there are many variables and constraints that complicate probabilistic analysis for cricket as compared to similar sports such as baseball. For instance, after each over has been completed, a different bowler bowls the ball from the opposite side of the field, additionally, an odd number of runs scored on a ball results in a switch of batsmen (unless this is the end of an over), bowlers switch sides at the end of an over and can only bowl a maximum of 10 overs in an ODI match, the first team to bat sets the number of runs to beat in the first inning and this can change the strategy of the second team to bat, and many more rules. Also, because limited overs cricket has been recently introduced, there have been numerous changes to the rules which significantly impact the dynamics of the game. For instance, over the ten-year period of our data set, the rules of ODI Cricket concerned the restrictions on where the bowling captain can place his fielders (power play) changed significantly three times. The complicated and evolving rules, in addition to the fact that cricket is played in an international level only in a few countries in the world has resulted in less cricket research studies, despite being the second most watched sport on the planet, after soccer.

Published research on cricket analytics can be broadly classified into three areas: player performance evaluation [2,3,4], game prediction [1,5,6,7,8,9,10], and batting evaluation [11,12,13,14]. With respect to ODI cricket, Clarke [1] presented a dynamic programming formulation to estimate the expected runs generated in the first inning and the maximum chance of winning in the second innings. Later studies also used dynamic programming, with modifications to take into account the bowling quality and the quality of the cricket ground [7]. The dynamic programming approach was also deployed in a recently used prediction algorithm, WASP [13], by Sky Sport New Zealand for the first time in November 2012 during Auckland's HRV Cup Twenty20 game against Wellington. WASP works the best on picturing how well an average batting team would do against an average bowling team in the conditions under which the game is being played given the current state of the game. Monte Carlo simulations are another approach for simulating cricket matches in test format [14], in T20, and in ODIs [6, 9]. One major drawback of most of these approaches is the strong assumptions about the batters, the bowlers, and their expected performances.

Kaluarachchi and Varde [15] employed association rules and a naive Bayes classifier to analyze the factors contributing to a win,

taking into account the variation in time of day of play (day or day and night). These approaches use a very limited subset of high-level features to analyze the factors contributing to victory. Furthermore, they do not address score prediction, or the progression of the game.

The individual performance of players is usually measured in terms of absolute historic metrics, typically the number of runs scored by batsmen and the number of wickets taken by bowlers. A number of correlated metrics with these two measures have been developed and widely used, such as strike rate, career average, balls faced, etc. These measures are undoubtedly valuable and are easy-to-understand for publicly ranking the players and ad hoc benchmarking, however, these summary information are not statistically sufficient statistics to describe how well a player will perform in a given situation.

CPS uses a completely different approach for player comparison based on the individualized, situation-based prognosis. This paper briefly presents the development of CPS for simulating the dynamics of the game and scoring runs in ODI cricket. Although there have been some attempts to build simulators for test cricket and T20, significant differences exist in the rules of the game and in scoring patterns between ODI cricket and the other two formats. In addition, there is not enough evidence on how well these simulators model a real cricket game.

Using CPS one can predict more than the probability of victory for a team in a match. Prognoses can be made about several nested layers of parameters. Predictive models in CPS are developed to assess how well a player (batter or bowler) is likely to perform in a given situation with respect to an opponent team, an opponent player, in a specific partnership, and at a particular time of the game. Future predictions include the number of runs one batter is expected to generate, how the runs are distributed, and the number of balls he is expected to face before being dismissed. These results are then combined in another model to assess how a team is likely to perform in a given match. The predictions can be dynamically updated as the game progresses to give more accurate prognoses based on relevant data.

2. DATA

We extracted ball-by-ball information from ESPN CricInfo (http://espncricinfo.com). These data include 1158 historical international ODI matches during the period May 23, 2006 to February 3, 2016. These ball-by-ball data included the batter, bowler, non-striker, and team, as well as inning information for 333,130 balls bowled. The matches had mean 287 balls bowled and a median 305 balls bowled, reflecting the fact that some games can be greatly shortened (by official decision) due to poor conditions [8]. In addition to the ball-by-ball data, we also collected publicly available batting and bowling performance data for all 1117 players who batted or bowled in any of these games. We collected data for 20 teams, but only 10 of those teams played more than 100 matches.

3. METHODS

To predict probabilities of victory and distributions of interesting player statistics for particular matches, CPS simulates thousands of games, where each simulation uses three classifiers for the outcome of each bowled cricket ball. The first (binary) classifier predicts whether the wicket will be taken (the player is dismissed). The second classifier predicts how many runs will be scored, and the third classifier predicts how many "extra" runs will be scored. Given the two teams, and a batter list for both teams, CPS uses accumulated risk models to simulate a full cricket match. For each bowled ball, if the batter's wicket is not taken, then the runs and extra runs are calculated. CPS always assumes the full game is played; shortened matches due to bad conditions are not simulated.

To parameterize our classifiers, we developed more than three hundred indicators based on descriptive statistics for batters, bowlers, bowling style, cricket ground, teams, and the current state of the game. The "game state" variables include the current over, and the number of wickets taken, and, during the second inning, the run rate required to win the game. Some of these indicators include player consistency, impact, and pressure indices. For example, in the second innings, players may make riskier plays if they need to score a large number of runs in a short number of overs, leading to higher probability of their wicket being taken, but also a higher probability of scoring more runs. Informed by measures of central tendency and variability, we selected indicators as parameters in our models.

The wicket taken model was implemented as a binary logistic regression model. Both the runs and extra runs model were implemented as six independent logistic regression models. For the runs model, each logistic regression model was fit as the probability of scoring 0, 1, 2, 3, 4 or 6 runs (5 runs is exceedingly rare). After the probabilities of each six were calculated, the probabilities were normalized so that they summed to 1. Similarly, the extra runs model predicted whether 0 to 5 extra runs were scored.

In the next section, we describe the experimental results of these analyses.

4. RESULTS

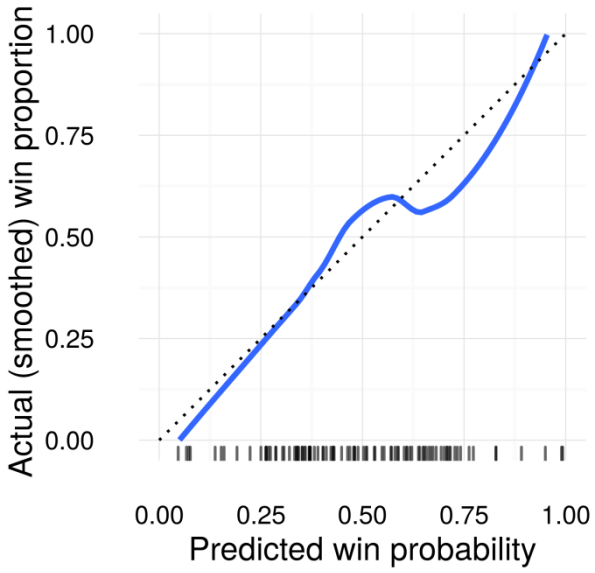

Figure 1. Calibration plot for win probability in CPS. Along the horizontal axis, the "rug plot" shows the predicted probability for each of the 93 matches, with the position jittered slightly to reveal the point density. The dotted line represents perfect calibration. The actual win proportion (solid line) is calculated with a LOESS smoother, with a neighborhood (α) of 0.8, using tricubic weighting (proportional to $(1 - \text{distance}^3)^3$).

We evaluated CPS by predicting the winner of 93 randomly selected ODI matches played between the May 23, 2006 and February 3, 2016 that did not end in a "no result". We predicted

the probability of each game by simulating 100 games for each randomly selected match, which allowed us to estimate the probability of each team winning as the proportion of simulated games won. The predictive results for each match in this set are compared against the actual result of the match played. We found that CPS could accurately pick the winner about 70% of the time. In Figure 1, we show the calibration of the win probability. This figure shows a great calibration result for CPS.

Using current batter rankings, we selected some top-rated ODI players, and measured some performance characteristics using our models. We found that some performance characteristics were indeed very different even among similar top-rated players. Figure 2 shows a comparison of the cumulative risk of being dismissed for three batters, holding all other things constant. AB de Villiers is the top rated batter, JE Root is the 10th best, and AD Hales in an average batter.

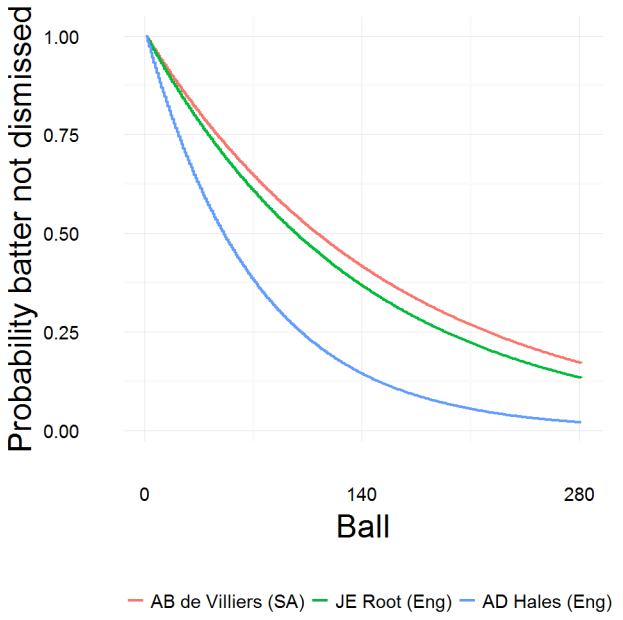

Figure 2. Not dismissed probability by number of balls bowled. For three sample batters, the estimated probability that a player is not dismissed by the number of bowled balls is shown.

We also experimented with CPS for the 2016 Tri-Nation series in West Indies. In this experiment, we used CPS to predict balls faced and runs generated for all the players involved in the tournament. Three countries are taking part in TriNation series in West Indies. The publically announced batting list from Australia was used for selection of Australian players, but because the batter list was not announced for West Indies and South Africa at the time of the experiment, we used previous ODI games to form a plausible batting list. Table 2 presents the list of South Africa batters in the order we selected for the matches against West Indies and against Australia. Based on 5000 simulations, South Africa has an 82% chance of winning against West Indies and Australia has a 63% chance of winning against South Africa. As Table 2 shows, a player can have a completely different performance in terms of runs and balls faced against different teams, which confirms that it is not reasonable to assume even a player-specific average in modeling ODI cricket.

5. Conclusion

In this study, we presented our Cricket Prognosis System, CPS, and some evaluation results in the forms of estimated win probabilities and statistics for given batters that can be used for player evaluation and benchmarking. CPS has a reasonably good accuracy to be used in various applications. To our knowledge, there is no systematic evaluation of a simulator or a prediction model in previous studies in the literature. Most published studies show the results over a hand full of matches or compare their predictions in a few matches against the betting odds and still not reaching an acceptable accuracy. The in-game prognoses for players define a novel set of metrics for player comparison. These metrics also help on-line planning for individual teams to strategize the game better for more favorable results. Our experimental results showed that these prognosis for balls to be faced and runs to be generated capture the reality of batting in ODIs to a good extent.

Table 1. Predicted and actual mean balls faced for South African batters in the West Indies TriNation tournament. The table shows the mean number of balls faced and runs scored in simulations of matches between South Africa and both West Indies and Australia. Additionally, the actual "Ave" mean the player average balls faced and runs scored in an inning over previous matches is shown. We excluded the instances where the batter was not dismissed at the end of batting.

South Africa Batters	Runs WI	Runs Aus	Runs Ave	BF WI	BF Aus	BF Ave
Q de Kock	62	18	51	69	18	56
HM Amla	58	47	58	69	54	61
F du Plessis	44	49	36	54	56	44
RR Rossouw	37	12	42	36	14	42
AB de Villiers	32	33	52	32	32	48
F Behardien	24	27	27	23	25	24
D Wiese	17	19	13	18	19	16
CH Morris	14	31	5	16	37	7
K Rabada	10	11	6	14	12	16
KJ Abbot	9	14	7	12	16	11

Imran Tahir	9	9	6	11	10	8

6. REFERENCES

[1] S. R. Clarke (1998). *Test statistics*, Statistics in Sport, J. Bennett (editor), Arnold, London, 83-103.

[2] Danielle Catherine MacDonald (2015). *Performance analysis of fielding and wicket-keeping in cricket to inform strength and conditioning practice*. PhD Thesis, Auckland University of Technology.

[3] Satyam Mukherjee (2014). *Quantifying individual performance in Cricket-A network analysis of batsmen and bowlers*. Physica A 393, 624–637.

[4] Paul J. Bracewell, Farinaz Farhadieh, Clint A. Jowett, Don G. R. Forbes, Denny H. Meyer (2009) *Was Bradman Denied His Prime?* Journal of Quantitative Analysis in Sports, 5(4).

[5] Koulis, T., Muthukumarana, S. and Briercliffe, C. (2014). *A Bayesian Stochastic Model for Batting Performance Evaluation in One-Day Cricket*. Journal of Quantitative Analysis in Sports 10 (1), 1-13.

[6] Swartz, T.B., Gill, P.S. and Muthukumarana, S. (2009). *Modeling and simulation for one-day cricket.* The Canadian Journal of Statistics, 37(2), 143-160.

[7] Scott Brooker and Seamus Hogan (2011*). A Method for Inferring Batting Conditions in ODI Cricket from Historical Data*. Technical report No. 44/2011. Department of Economics and Finance College of Business and Economics. University of Canterbury.

[8] F. C. Duckworth and A. J. Lewis.(1998). *A fair method for resetting the target in interrupted one-day cricket matches.* The Journal of the Operational Research Society, 49(3):pp. 220–227.

[9] Perera, H., Davis, J. and Swartz, T.B. (2015). *Optimal lineups in Twenty20 cricket*. To appear in Journal of Statistical Computation and Simulation.

[10] Vignesh Veppur Sankaranarayanan, Junaed Sattar and Laks V. S. Lakshmanan (2014*) Auto-play: A Data Mining Approach to ODI Cricket Simulation and Prediction.* SDM.

[11] Matthews Ovens and Bruce Bukiet (2006) *A mathematical modelling approach to One Day International cricket batting orders.* Journal of Sports Science and Medicine, 5, 495-502.

[12] Cohen, G. L. (2002) *Cricket chances*. Mathematics and Computers in Sports, 1-13.

[13] Banare, Abhijit (2014). WASP: *Winning and Score Predictor makes for an interesting watch on television*. Cricket Country.

[14] Uday Damodaran (2006). *Stochastic dominance and analysis of ODI batting: 1989-2005*, Journal of Sports Science and Medicine 5, 503-508.

[15] A. Kaluarachchi and A. Varde (2010). *CricAI: A classification based tool to predict the outcome in ODI cricket*. In 5th International Conference on Information and Automation for Sustainability, pages 250–255.