APPLICATION OF KNN ALGORITHM FOR CRICKET TEAM PLAYER’S SUBSTITUTE

Vijay Maurya[1], Anuj Verma[2], Nikita Goel[3]

[1]*Asst. Prof., Electronics Engineering, Institute of Engineering & Technology, Lucknow, U.P., India*
[2]*Asst. Prof., Electrical Engineering, International Maritime Institute, Greater Noida, U.P., India*
[3]*M.Tech. Scholar, Electronics Engineering, AKTU Lucknow, U.P., India*

ABSTRACT

In the world of cricket, various important decisions like choosing substitutes generally based on player’s performance in the last tournament, does not always results best. Now a day’s new scientific methods have been chosen for taking decisions in sports. In this paper we have tried to make the above decision using KNN algorithm. As an illustration we have taken the data set of 377 cricketers of the one day cricket and by giving appropriate weight to the attribute and have tried to implement the above said aim which shows that this is a more practical approach to be followed for the better results.

Keyword: - *clustering analysis; KNN algorithm; data objects; distance; substitute*

1. INTRODUCTION

Cricket is a bat-and-ball team sport first documented as being played in southern England in the 16th century. By the end of 18th century, cricket had developed to the point where it had become the national sport of England. The expansion of the British Empire led to cricket being played overseas and by the mid-19th century the first international matches were being held. Today, the sport is played in more than 100 countries. Standard limited over’s cricket (where an over is a series of six bat-and-ball rounds) was introduced in England in the 1963 season in the form of a knockout cup contested by the first-class county clubs. In 1969, a national league competition was established. The concept was gradually introduced to the other major cricket countries and the first limited over’s international match was played in 1971. The first Cricket World Cup took place in England in 1975.

A "one day match", so called because each match is scheduled for completion in a single day, is the most common form of limited over’s cricket played on an international level. In practice, matches sometimes continue on a second day if they have been interrupted or postponed by bad weather. The main objective of a limited over’s match is to produce a definite result and so a conventional draw is not possible, but matches can be undecided if the scores are tied or if bad weather prevents a result. Each team plays one innings only and faces a limited number of over’s, usually a maximum of 50 over’s (300 deliveries).

[5] Match data since the beginning of the ODI game is available. However our literature search found no previous research and literature on the above mentioned topic. Some work could be found on topic of optimal scoring rates by Clarke [7] and Preston and Thomas [6]. They utilize dynamic programming methods. But in their research they mentioned only about the strategies in the one day cricket. Some studies, such as those conducted by De Silva [8] analyze the magnitude of the victory. It is found that most of these studies describe the factors affecting winning but do not focus on the analysis of the factors with the goal of predicting the probability of victory if some changes are there in the squad.

The K-Nearest Neighbors algorithm (or K-NN for short) classification is the most fundamental and simple classification methods that should be the one of the first choices for a classification study when there is a little or no prior knowledge about the distribution of the data [4]. Knn is a non-parametric method used for classification. When you say a technique is a non-parametric, it means it does not make any assumptions on the underlying data distribution. KNN assumes the data is in features space i.e. the data points are in metric space. The data can be scalars or possibly even multidimensional vectors. Since these points are in future space, they have a notion of distance. This distance can be calculated by Euclidean distance, city block distance, chebychev distance, cosine distance but the distance which we are using for our work is the Euclidean distance.

Method to find KNN:

1.) Calculate distances of all training vectors to test vector

2.) Pick k closest vectors

The training examples are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples. In the classification phase, k is a user-defined constant, and an unlabeled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point. A commonly used distance metric for continuous variables is Euclidean Distance. The Euclidean distance between X=(X_1, X_2, X_3… X_n) and Y = (Y_1, Y_2, Y_3 …Y_n) is defined as:

$$D(X,Y) = \sqrt{\sum_{i=1}^{k}(Xi - Yi)^2} \qquad \text{Where } k = n$$

2. PARAMETERS TO WORK WITH

Number of nodes: 377 cricketers

•**Dimensions:** Matches, Innings, Over's, Maidens, Average, Economy, 4 wickets, 5 wickets, Runs, Highest, Strike Rate, 4s, 6s, 50s, 100s.

•**Value of k:** as requires by user.

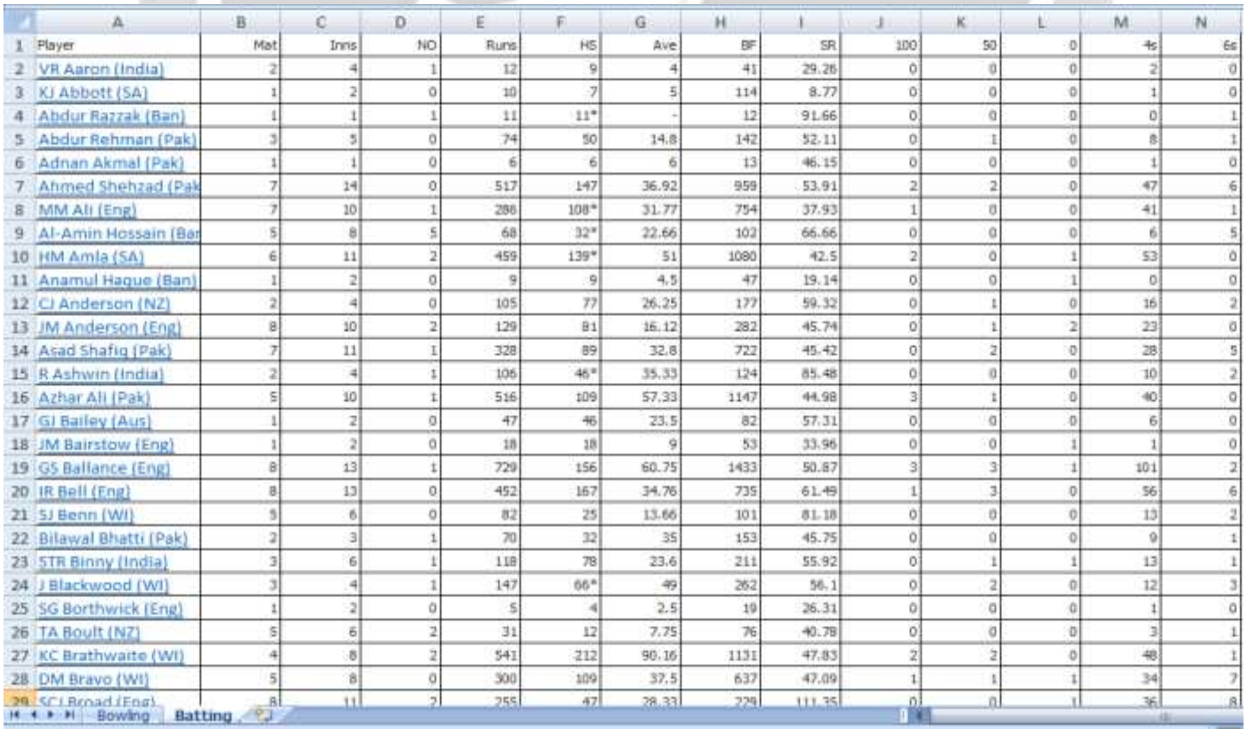
•**Output required:** k appropriate substitute for the player requested by user.

For experimental purpose, we have prepared data set consisting of 377 cricketers along with their statistics

Table 1: data set classification

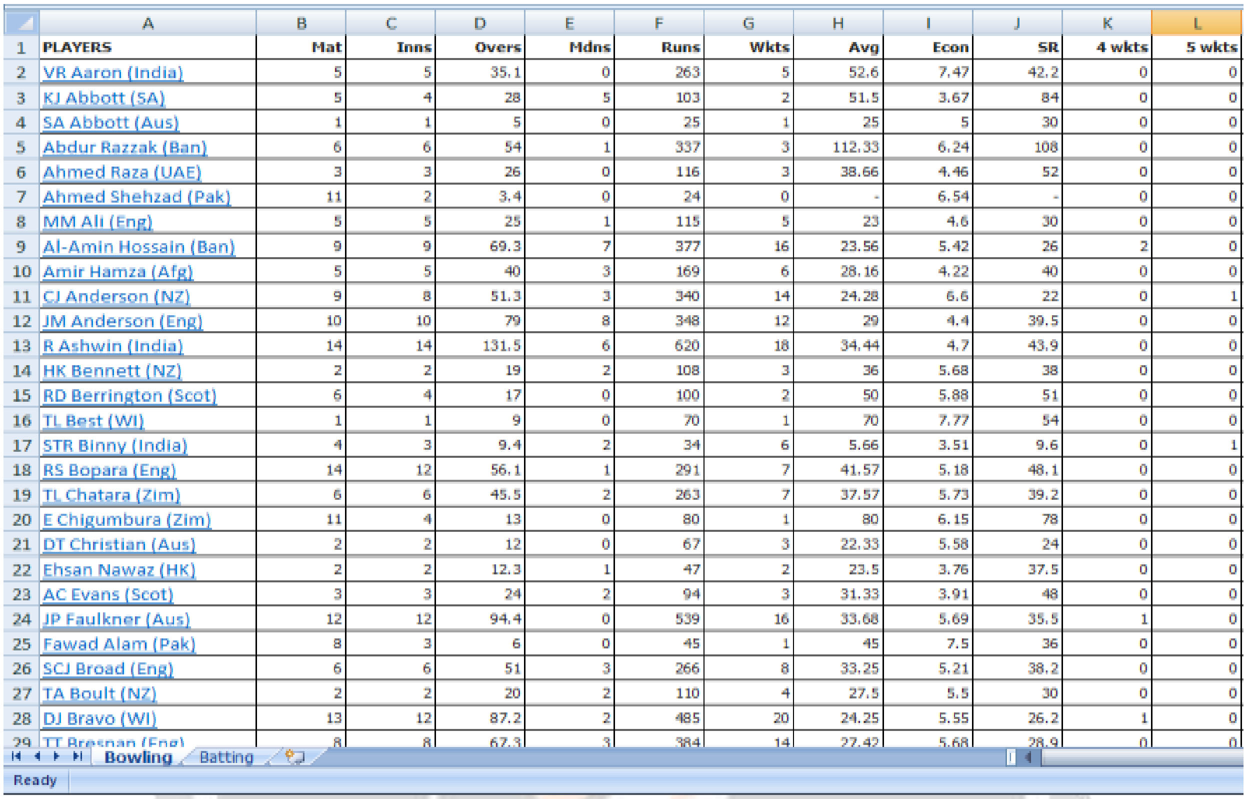
Data sets	No. of Samples	No. of Attributes
Batsmen	69	5
Bowler	308	6

For example we have taken the data set with the following attributes:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Player	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	0	4s	6s
2	VR Aaron (India)	2	4	1	12	9	4	41	29.26	0	0	0	2	0
3	KJ Abbott (SA)	1	2	0	10	7	5	114	8.77	0	0	0	1	0
4	Abdur Razzak (Ban)	1	1	1	11	11*	-	12	91.66	0	0	0	0	1
5	Abdur Rehman (Pak)	3	5	0	74	50	14.8	142	52.11	0	1	0	8	1
6	Adnan Akmal (Pak)	1	1	0	6	6	6	13	46.15	0	0	0	1	0
7	Ahmed Shehzad (Pak	7	14	0	517	147	36.92	959	53.91	2	2	0	47	6
8	MM Ali (Eng)	7	10	1	288	108*	31.77	754	37.93	1	0	0	41	1
9	Al-Amin Hossain (Bar	5	8	5	68	32*	22.66	102	66.66	0	0	0	6	5
10	HM Amla (SA)	6	11	2	459	139*	51	1080	42.5	2	0	1	53	0
11	Anamul Haque (Ban)	1	2	0	9	9	4.5	47	19.14	0	0	1	0	0
12	CJ Anderson (NZ)	2	4	0	105	77	26.25	177	59.32	0	1	0	16	2
13	JM Anderson (Eng)	8	10	2	129	81	16.12	282	45.74	0	1	2	23	0
14	Asad Shafiq (Pak)	7	11	1	328	89	32.8	722	45.42	0	2	0	28	5
15	R Ashwin (India)	2	4	1	106	46*	35.33	124	85.48	0	0	0	10	2
16	Azhar Ali (Pak)	5	10	1	516	109	57.33	1147	44.98	3	1	0	40	0
17	GJ Bailey (Aus)	1	2	0	47	46	23.5	82	57.31	0	0	0	6	0
18	JM Bairstow (Eng)	1	2	0	18	18	9	53	33.96	0	0	1	1	0
19	GS Ballance (Eng)	8	13	1	729	156	60.75	1433	50.87	3	3	1	101	2
20	IR Bell (Eng)	8	13	0	452	167	34.76	735	61.49	1	3	0	56	6
21	SJ Benn (WI)	5	6	0	82	25	13.66	101	81.18	0	0	0	13	2
22	Bilawal Bhatti (Pak)	2	3	1	70	32	35	153	45.75	0	0	0	9	1
23	STR Binny (India)	3	6	1	118	78	23.6	211	55.92	0	1	1	13	1
24	J Blackwood (WI)	3	4	1	147	66*	49	262	56.1	0	2	0	12	3
25	SG Borthwick (Eng)	1	2	0	5	4	2.5	19	26.31	0	0	0	1	0
26	TA Boult (NZ)	5	6	2	31	12	7.75	76	40.78	0	0	0	3	1
27	KC Brathwaite (WI)	4	8	2	541	212	90.16	1131	47.83	2	2	0	48	1
28	DM Bravo (WI)	5	8	0	300	109	37.5	637	47.09	1	1	1	34	7
29	SCJ Broad (Eng)	8	11	2	255	47	28.33	229	111.35	0	0	1	36	8

Fig-1: Bowlers' Statistics



	A	B	C	D	E	F	G	H	I	J	K	L
1	PLAYERS	Mat	Inns	Overs	Mdns	Runs	Wkts	Avg	Econ	SR	4 wkts	5 wkts
2	VR Aaron (India)	5	5	35.1	0	263	5	52.6	7.47	42.2	0	0
3	KJ Abbott (SA)	5	4	28	5	103	2	51.5	3.67	84	0	0
4	SA Abbott (Aus)	1	1	5	0	25	1	25	5	30	0	0
5	Abdur Razzak (Ban)	6	6	54	1	337	3	112.33	6.24	108	0	0
6	Ahmed Raza (UAE)	3	3	26	0	116	3	38.66	4.46	52	0	0
7	Ahmed Shehzad (Pak)	11	2	3.4	0	24	0	-	6.54	-	0	0
8	MM Ali (Eng)	5	5	25	1	115	5	23	4.6	30	0	0
9	Al-Amin Hossain (Ban)	9	9	69.3	7	377	16	23.56	5.42	26	2	0
10	Amir Hamza (Afg)	5	5	40	3	169	6	28.16	4.22	40	0	0
11	CJ Anderson (NZ)	9	8	51.3	3	340	14	24.28	6.6	22	0	1
12	JM Anderson (Eng)	10	10	79	8	348	12	29	4.4	39.5	0	0
13	R Ashwin (India)	14	14	131.5	6	620	18	34.44	4.7	43.9	0	0
14	HK Bennett (NZ)	2	2	19	2	108	3	36	5.68	38	0	0
15	RD Berrington (Scot)	6	4	17	0	100	2	50	5.88	51	0	0
16	TL Best (WI)	1	1	9	0	70	1	70	7.77	54	0	0
17	STR Binny (India)	4	3	9.4	2	34	6	5.66	3.51	9.6	0	1
18	RS Bopara (Eng)	14	12	56.1	1	291	7	41.57	5.18	48.1	0	0
19	TL Chatara (Zim)	6	6	45.5	2	263	7	37.57	5.73	39.2	0	0
20	E Chigumbura (Zim)	11	4	13	0	80	1	80	6.15	78	0	0
21	DT Christian (Aus)	2	2	12	0	67	3	22.33	5.58	24	0	0
22	Ehsan Nawaz (HK)	2	2	12.3	1	47	2	23.5	3.76	37.5	0	0
23	AC Evans (Scot)	3	3	24	2	94	3	31.33	3.91	48	0	0
24	JP Faulkner (Aus)	12	12	94.4	0	539	16	33.68	5.69	35.5	1	0
25	Fawad Alam (Pak)	8	3	6	0	45	1	45	7.5	36	0	0
26	SCJ Broad (Eng)	6	6	51	3	266	8	33.25	5.21	38.2	0	0
27	TA Boult (NZ)	2	2	20	2	110	4	27.5	5.5	30	0	0
28	DJ Bravo (WI)	13	12	87.2	2	485	20	24.25	5.55	26.2	1	0
29	TT Bresnan (Eng)	8	8	67.3	3	384	14	27.42	5.68	28.9	0	0

Fig-2: Batsmen’s Bowlers’

3. NEED OF SUBSTITUTE

A substitute in the game of cricket is basically a replacement that umpires allows when a player has been injured or become ill after the nomination of the players at the start of the game or at the start of the tournament. But now the question arises, how we will choose the substitute for a particular player. So, our work could be one of the options.

3.1 The high rate of injuries in first class games

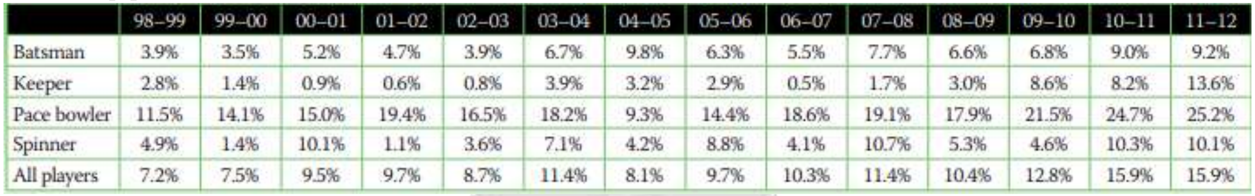
In 34 percent of first class games, a team will have at least one player suffer an injury that either prevents continued participation in the game or causes him to miss the following game. With one third of teams affected, it cannot be argued that injuries in professional cricket are rare events that do not need to be catered for. Cricket has injury prevalence rates similar to football and therefore has the same need to consider substitute players.

3.2 Increasing fast bowler injury prevalence

Fast bowlers are clearly not coping with the new make-up of the cricket calendar, which is here to stay given the eight-year forward planning of the Future Tours Program and the popularity of the T20 tournaments. The key factor is that bodies are generally not designed for participating in the cricket equivalent of a sprint event (four over’s of extreme pace are called for in a T20 match) rapidly followed by the cricket equivalent of a marathon (30–40+ over’s are required from each bowler in a first class match). This transition would be made far easier if a first class or Test match workload was shared between a greater numbers of bowlers.

3.3 Risk of injuries worsening if players push through pain

Cricket has prided itself on the ‘tough’ environment of the Test match arena where players are required to push through pain and minor injuries for the benefit of the team. A player is expected to continue to bat or bowl even if suffering from cramp, for example. However, serious injuries do occasionally occur in cricket and the expectation that a player should always push through pain for the benefit of the team could, in rare cases, be catastrophic. Cricketers can suffer concussion, cardiac conditions, severe dehydration (especially if playing with gastroenteritis in hot, humid environments) and medically, this could lead to dire consequences if there is major pressure from the rules not to pull out and leave the team short. More commonly, fast bowlers can suffer stress fractures of the lumbar spine. Although hard data is difficult to come by, the clinical impression is that if a bowler stops bowling early in the cycle of a lumbar stress fracture, the bone can heal nicely, but permanent non-unions which can affect the entire playing career can occur if the bowler pushes through the pain barrier for too long. Rules which encourage pushing through pain probably lead to worse outcomes for back injuries in young fast bowlers, which is a blight on the game of cricket.[2]



	98–99	99–00	00–01	01–02	02–03	03–04	04–05	05–06	06–07	07–08	08–09	09–10	10–11	11–12
Batsman	3.9%	3.5%	5.2%	4.7%	3.9%	6.7%	9.8%	6.3%	5.5%	7.7%	6.6%	6.8%	9.0%	9.2%
Keeper	2.8%	1.4%	0.9%	0.6%	0.8%	3.9%	3.2%	2.9%	0.5%	1.7%	3.0%	8.6%	8.2%	13.6%
Pace bowler	11.5%	14.1%	15.0%	19.4%	16.5%	18.2%	9.3%	14.4%	18.6%	19.1%	17.9%	21.5%	24.7%	25.2%
Spinner	4.9%	1.4%	10.1%	1.1%	3.6%	7.1%	4.2%	8.8%	4.1%	10.7%	5.3%	4.6%	10.3%	10.1%
All players	7.2%	7.5%	9.5%	9.7%	8.7%	11.4%	8.1%	9.7%	10.3%	11.4%	10.4%	12.8%	15.9%	15.9%

Fig-3: Injury prevalence (players missing through injury) by player position by season

4. EXPERIMENTAL RESULTS AND ANALYSIS

We have found k-appropriate substitutes for a player as required by user. For finding the substitute we have normalized the data and give the appropriate weight to each and every attribute. The user can find substitute for a Batsman or a Bowler by the following steps:

Step1: Enter '1' for Batsman and '2' for Bowler.

Step2: Then after this user chooses the basis of selection as:

For Batsmen:

1: average 2: strike rate
3: runs 4: overall performance

For Bowlers:

1: average 2: economy
3: wickets 4: overall performance

We used MATLAB 13 for implementing this idea and getting the output values. Then we will get the k- substitute for a particular cricketer based on their performances in 2014. The result which we get after performing our analysis: suppose we have Suresh Raina in our team and he gets injured or for comparison purpose I want to know, which players in the other side are in good form as Raina based on overall performance, so the result of comparison is:

Table 2: result to get substitute for Suresh Raina

Player	Substitute Player 1	Substitute Player 2
Suresh Raina	T M Dilshan	Mushfiqur Rahim

Similarly, after comparing the bowlers we get the result as:

Table 3: result to get substitute for Mohammed Shami

Player	Substitute Player 1	Substitute Player 2
Mohammed Shami	S L Malinga	B.A.W Mendis

The above table has three columns. Column number one is for the player, column number two is for the first nearest substitute and column number three for the second nearest substitute. Our experimental result completely agrees with practical data because comparing the statistics of all the three players we find great similarity among them.

So, the user can choose among the k-substitutes based on availability in their team or can simply compare the performance of team players with the other side.

Applications:

1.) Required in case of player injury.
2.) Used in bidding.
3.) Used to find appropriate substitutes in big tournaments.

5. RELATED WORK

From our work, we found that very limited work has been done on game of cricket. Though cricket shares some attributes with other sports such as baseball, it still remains unique in certain respects and deserves to be analyzed independently. Most of analyzing studies on cricket so far have been conducted using statistical methods. Furthermore, many of them have addressed the five day long test matches but not the One Day Internationals. We present some relevant studies below.

The statistical research on Cricket has been started very early stage of the cricket. In 1945 Wood used the geometric distribution to model the total score in cricket [9]. This was not a study on the ODI form of the game but has been recognized among the pioneering research in the game of cricket. Chedzoy studied the issue of umpiring errors in cricket matches [12]. An umpire is the term used for a referee in cricket. This article focuses on umpiring decisions and how they affect the outcome of the match. This study was also based on a statistical approach. Moreover, it focused only on one aspect of the game, namely, the effect of umpires.

Bailey and Clarke conducted a study to predict the outcome in one day international cricket while the game is in progress [10]. This study was performed using statistical models. The interesting fact about this article is that the authors have statistically proved how the match resources (number of overs and batsmen left) affect the final result. However, they deal with analysis during the current game. They do not predict in advance the chances of winning a new game based on previous matches.

A. Bandulasiri [1] has written an interesting article on predicting the winner in an ODI cricket match. In this paper, the author has used statistical methods to find wining factors for an ODI match.

6. CONCLUSION

In this paper we have improved the decision making in the world of cricket by using KNN algorithm and customizing it as required by the user. The result that we have got is completely agreed with the practical result and we get k appropriate substitute for each player and also the comparison between different players.

7. REFERENCES

[1]. Bandulasiri, “Predicting the Winner in One Day International Cricket”, Journal of Mathematical Sciences & Mathematics Education, Vol. 3, No. 1.

[2]. https://www.johnorchard.com/resources/article-Injury-data-SportsPhysioarticle.pdf

[3]. Cric-info [2014,November], website for cricket data, [online] http://www.cricinfo.com

[4]. Scholarpedia, website for knn article, [Online] http://www.scholarpedia.org/article/Knearest_neighbor

[5]. Wikipedia on the Game of Cricket, website [Online] http://en.wikipedia.org/wiki/Cricket

[6]. Preston and J. Thomas “Batting Strategy in Limited Overs Cricket”, The Statistician (Journal of the Royal Statistical Society: Series D), 2000, 49, 95-106

[7]. S.R. Clarke, “Dynamic programming in one day cricket— optimal scoring rates”, Journal of the Operational Research Society, 1988, Vol. 39, No. pp. 331–337.

[8]. B.M De Silva, and T.B. Swartz, Estimation of the magnitude of the victory in one-day cricket. Australia and New Zealand Journal of Statistics, 2001, Vol. 43, pp. 1369-1373

[9]. G.H. Wood, “Cricket scores and geometrical progression”, Journal of the Royal Statistical Society, 1945, Series A, 108: pp. 12–22

[10]. M. Bailey and S.R. Clarke, “Predicting the match outcome in one day international cricket matches, while the game is in progress”, Journal of Sports Science and Medicine, 2006, Vol. 5, pp. 480-487

[11]. Cricket Australia, [online] www.**cricket**.com.au/**stats**

[12]. O. B. Chedzoy, “Issue of the effect of umpiring errors in cricket Statistician”, 1997, Vol. 46, No. 4, pp. 459-527.