INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 6.078
(Volume 6, Issue 4)
Available online at: www.ijariit.com

A multivariate statistical approach for ranking the best batsmen in test cricket

Rahul Motipalle
rahulmotipalle@gmail.com
Indian Institute of Information Technology, Sricity, Andhra Pradesh

Sajjanapu Venkat Lokesh Kumar
venkatlokeshkumar.s16@iiits.in
Indian Institute of Information Technology, Sricity, Andhra Pradesh

ABSTRACT

The primary objective of this paper is to devise a ranking system for individual test batsmen (from different eras) considering all possible factors like the difficulty of opposition bowlers, the batsman's consistency, the contribution of the batsman in the success of his team, the era in which he played, etc. Due to the presence of such numerous factors—each one of them having a significance of its own, it is imperative that a multivariate statistical approach is employed in the ranking of Test batsmen. A total of eight factors were zeroed in on which were thought to influence the performance of a batsman. An index was developed for each of the eight factors, and thus each of the batsmen got eight different scores for the eight different factors. Consequently, the eight scores of the batsman act as the coordinates of a point/cluster in an eight-dimensional plane. In this way, each of the batsmen under study is represented by a point in an eight-dimensional plane. Finally, to determine the most exceptional batsmen among the batsmen under study, the concept of multivariate statistical outlier detection using Mahalanobis distances was used. However, the concept of Outlier Detection only gives us an idea of the most exceptional batsmen when compared to the others. In order to determine the best batsmen, the process of outlier detection is followed up with the determination of the efficiencies of each of the batsmen under study. The efficiencies of the batsmen are calculated by adopting the approach of Data Envelopment Analysis (DEA), wherein each batsman is likened to a machine or a Decision-Making Unit (DMU). The higher the efficiency of a batsman, the greater his success in converting low inputs (difficult input parameters) to high outputs.

Keywords— *Ranking Batsmen, Ranking Cricketers, Outlier Detection, Factor Analysis, Data Envelopment Analysis, Test Cricket*

1. INTRODUCTION

With an estimated global following of over 2.5 billion, Cricket is the second-most followed sport in the world[1]. The rapid growth and commercialization of Cricket has necessitated a proper analysis of a player's ability and worth. Ergo, a proper ranking system which can reflect the true picture of the performance of a cricketer on various fronts becomes the need of the hour. In such a popular sport, there are bound to be comparisons between the different players of the game, though they may have played in different conditions and in different situations of the game. To complicate matters further, many of these comparisons are made for batsmen across different generations. Batsmen of two different eras cannot be compared in the same breath without normalizing their statistics in a standardized manner, since there are many factors that differentiate them. Each of these factors will be discussed in considerable depth in this paper and meaningful insights will be provided about how different batsmen score on these factors before ultimately ranking them. We propose a new ranking system based on the scores obtained by batsmen on eight different (and largely independent) factors. The eight factors are enumerated below:

1. Difficulty of Opposition Bowlers
2. Difficulty of Opposition Teams
3. Consistency of the Batsman
4. Statistically Computed Average
5. Batsman's contribution to the success of the team (Team Contribution)
6. Home/Away Conditions
7. Longevity
8. Miscellaneous (Time-varying) Factors

Scarf et al[2] (2014) proposed a new rating system for Test matches based on players' performances (batting, bowling and fielding) during each session of a test match. Borooah and Mangan[3] ranked batsmen based on their adjusted batting average with scores adjusted relative to his team rather than in absolute terms, since cricket is a team sport and the contribution of the batsman to the

success of his team is always to be measured, when trying to develop a ranking system. Sarkar and Banerjee[4] (IIM Calcutta, 2016) considered a total of five criteria to rank players: Statistically estimated batting average, Consistency (or) Dependability, Longevity, Quality of Runs and Opposition Diversity.

In cricket, a batsman should not be measured purely on the basis of the output he generates, whether it is his runs or his Strike Rate. Though highly debatable and subjective, some batsmen are luckier and face easier inputs (could be bowling or teams as a whole) when compared to others. Then, we need to introduce a concept called efficiency in order to determine the magnitude of the output that the batsman generates from the input he gets. Gweshe and Durbach[5] (2013) first used Data Envelopment Analysis to analyze the efficiency of cricketers in the 2011 World Cup. The efficiencies of the batsmen are calculated by adopting the approach of Data Envelopment Analysis (DEA), wherein each batsman is likened to a machine or a Decision-Making Unit (DMU). The higher the efficiency of a batsman, the greater his success in converting low inputs (difficult input parameters) to high outputs.

2. METHODOLOGY

The work done in this paper to rank batsmen can be summarized through the following diagram:

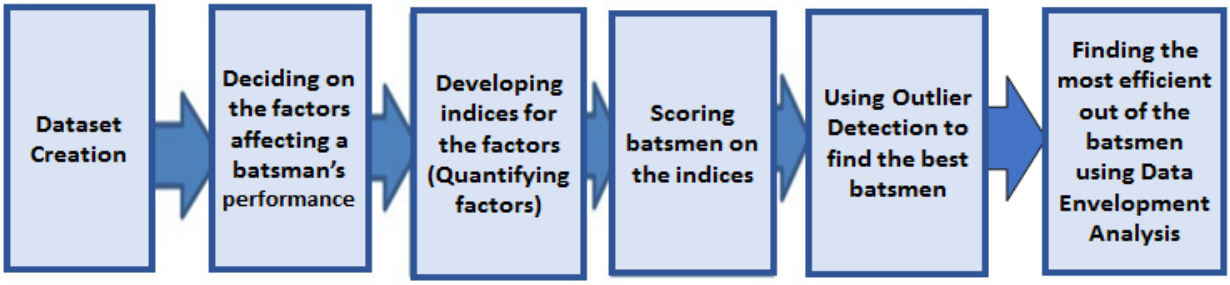

Fig. 1: Workflow Diagram of the Project

Appropriate datasets were created using the Statsguru website provided by ESPNCricInfo. The eight factors that were decided as the ones that will be used to quantify the performance of a batsman have already been discussed in Section 1.

2.1 Developing Indices for the Identified Eight Factors

2.1.1 Difficulty of Opposition Bowlers: For fair comparison between any two batsmen, we need to first quantify the quality or the strength of the bowling attacks faced by the batsmen. If two batsmen have similar statistics, but one batsman faced more difficult bowling relative to the other, he can be judged as the better batsman, as far as batting against quality bowling attacks is concerned.

In order to quantify the bowling faced by a particular batsman, we need to first determine who were the top bowlers of each of the opposition teams in the matches played by the batsman through the course of his career. In order to do this, the top ten wicket-taking bowlers from each of the eight opposition country teams were added to the opposition bowlers' dataset of each of the 50 batsmen under study. We identified wickets taken by the bowler (Wkts), Strike Rate (SR) and Average (Ave) to be the variables that quantify the strength of a bowler. After identification of the variables, we need to design a composite index that weighs all these 3 variables. This composite index will be known as the 'Strength' of the bowler. The greater the strength of a bowler, the more difficult it is for a batsman to face him.

The method of Factor Analysis was used to construct the composite index called Strength from the three observed variables namely Wickets, Strike Rate and Average. The orthogonal factor model[6] can be written as follows:

$$X_{(3X1)} - \mu_{(3X1)} = \Lambda_{(3X1)} F_{(1X1)} + \varepsilon_{3X1}$$

The observed variables matrix, X, is a 3x1 matrix comprising of the Wickets, Strike Rate, and Average of the bowler in the matches that featured the batsman under study.

$$X_{i,j} = \begin{bmatrix} Wickets_{i,j} \\ SR_{i,j} \\ Ave_{i,j} \end{bmatrix}$$

The mean matrix, μ, is a 3x1 matrix, which is determined by calculating the means of all the three observed variables of the 70 bowlers corresponding to the batsman under study.

The loadings matrix, Λ, is obtained by performing Exploratory Factor Analysis upon the seventy observations of the three variables.

The error term, ε, has an expected value of zero, i.e., $E(\varepsilon) = 0$. Hence, we can neglect the error term.

Neglecting the error term, the equation comes to be:

$$X - \mu = \Lambda F$$

Pre-multiplying both sides of the equation, we get:

$$\Lambda^T (X - \mu) = \Lambda^T \Lambda F$$

Finally, the factor underlying the three observed variables can be written as:

$$F = \frac{\Lambda^T (X - \mu)}{\Lambda^T \Lambda}$$

Hence, the strength of the bowler *i* against batsman *j* can be written as:

$$Strength_{i,j} = \frac{\Lambda_j{}^T (X_{i,j} - \mu_j)}{\Lambda_j{}^T \Lambda_j}$$

A striking point to note is that though many batsmen may have faced the same bowler, the strength of the bowler is the same for no two batsmen. This is because the statistics of the bowler that are taken in the bowler table corresponding to the batsman under study are those that pertain to only the limited matches that featured the same batsman. For example, consider the case of James Anderson, who bowled against both Sachin Tendulkar and Virat Kohli. Now since James Anderson has largely different numbers for the three observed variables namely Wickets, Strike Rate and Average for the matches that featured Tendulkar as opposed to the matches that featured Kohli, he will also possess a different strength for the two batsmen, as shown in Tables 1(a) and 1(b).

Table 1(a) and Table 1(b) compare the strengths of five bowlers, all who played in both the eras of Sachin Tendulkar as well as Virat Kohli. This comparison will help us get an idea of how the strength of even a single bowler changes from time to time, let alone the strength of an overall team. From this comparison, we can observe that James Anderson, who already had a decent strength during Tendulkar's period, went on to become really dangerous during Kohli's period with a much higher strength. On the contrary, Dale Steyn who had a very high strength during Tendulkar's period, registered a very low strength during Kohli's period. Stuart Broad was almost equally effective in both eras, while Nathan Lyon and Morne Morkel were largely ineffective in Tendulkar's era but showed a dramatic improvement in Kohli's era to become one of the strongest bowlers in his era.

Table 1: A comparison of the strengths of five select bowlers against Tendulkar and Kohli

Table 1(a): Strength of 5 selected bowlers against Sachin Tendulkar

Bowler	Wkts	Ave	SR	Strength
J Anderson	57	23.49	57.2	10.32
D Steyn	42	32.75	37.8	23.51
S Broad	27	23.59	50.5	13.04
M Morkel	22	35.72	63.5	-0.55
N Lyon	22	38.68	60.1	-0.54

Table 1(b): Strength of 5 selected bowlers against Virat Kohli

Bowler	Wkts	Ave	SR	Strength
J Anderson	65	29.71	54.3	22.13
D Steyn	12	19.21	67.7	1.12
S Broad	43	30.62	63.2	10.25
M Morkel	28	20.28	51.6	17.93
N Lyon	80	33.25	61.0	16.97

Finally, the bowling difficulty level for a batsman is obtained by averaging the strengths of all the bowlers present in his bowler table.

2.1.2 Difficulty level of opposition teams: To measure the batting performance of a batsman over many years, there is the need to not only take the strength of the opposition bowling into consideration, but also the overall difficulty level of the opposition teams during the career span of the batsman under study. The reason why we need to take this factor into consideration is that there are many cases where the opposition team is not strong enough despite possessing a strong bowling attack.

For determining the strength of a team, we propose a formula based on the number of matches won, drawn/tied and lost (the three possible outcomes of a Test match) for the course of a decade, i.e., a ten-year period. The strength of a team *i* in a decade *j* can be expressed as:

$$Strength_{i,j} = \frac{1 * MW_{i,j} + 0.5 * MD_{i,j} + 0 * ML_{i,j}}{MP_{i,j}}$$

where:

$MW_{i,j}$ = Matches won by team *i* in decade *j*

$MD_{i,j}$ = Matches drawn/tied by team *i* in decade *j*

$ML_{i,j}$ = Matches lost by team *i* in team *j*

$MP_{i,j}$ = Matches played by team *i* in team *j*

Theoretically, the maximum strength of a team is 1, which is in case they won all games they played during that decade, whereas the minimum strength of a team is 0, which is in case they lost all games they played during that decade. Hence, the *teamStrength* of a team during a particular decade is a numerical value between 0 and 1.

Refer to Table-2 to get an idea of what the strength of each of the eight teams looked like in the last five decades. Chart-2 gives us an idea of where each team stood with respect to other teams in a decade. It also indicates how the strength of the teams improved or deteriorated from one decade to the next.

The *teamDiff* score for a batsman is calculated as the sum of the strengths of all the opposition teams he played against in a decade. If the span of a batsman's career extended for more than a decade, then the average of the difficulty scores across the decades that he played is taken.

Table-2: Computing the strengths of each of the 8 main test-playing nations for the 5 decades

COUNTRY	1970s	1980s	1990s	2000s	2010s
Australia	0.51	0.48	**0.63**	**0.77**	0.57
England	**0.56**	0.41	0.42	0.57	0.55
India	0.48	0.44	0.49	0.56	0.60
South Africa	--	--	0.62	0.60	**0.64**
New Zealand	0.30	0.52	0.41	0.44	0.50
Pakistan	0.48	0.56	0.57	0.49	0.48
West Indies	0.52	**0.71**	0.51	0.31	0.38
Sri Lanka	--	0.26	0.44	0.57	*0.44

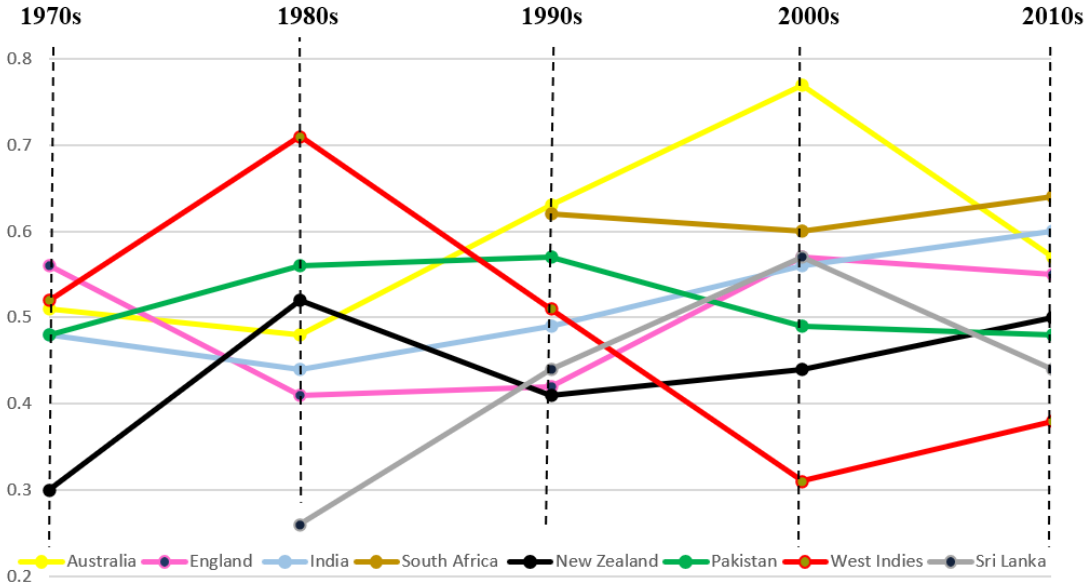

Fig. 2: Plotting the strengths of the main test-playing nations for the last 5 decades

2.1.3 Consistency: The biggest drawback of using the conventional batting average as a measure of ranking batsmen is that the measure of a batsman's consistency is totally missed. According to the concept of conventional batting averages, two batsmen having similar averages should be considered equally consistent or dependable. However, we know that is not the case. Borooah and Mangan (2007) illuminate some primary issues of Batting Averages. Consistency cannot be measured by aggregate measures like mean or median. One needs to have a look at all the observations of the sample to determine the consistency.

The next step is to come up with an index that can help us measure the consistency level of every single batsman under study. As mentioned earlier, for developing an index for a parameter like consistency, one needs to consider all the observations of the sample, which in our case is all the scores of the batsman under study. We took all the scores that a batsman scored in his career and fitted it to a distribution. When plotted on a graph, the scores of a batsman appear like they arise out of an exponential distribution. One can see that the lowest scores, i.e., 0-10 have the highest number of occurrences in the sample set of the batsman's scores and the number of occurrences go on decreasing as the score range goes on increasing.

To check whether the scores of a batsman really arise out of an exponential distribution, we performed hypothesis testing using the famous Pearson's chi-square test. The null hypothesis is the statement that the scores of a batsman arise out of an exponential distribution whereas the alternative hypothesis is the statement that the scores of a batsman do not arise out of an exponential distribution. The chi-square test yielded a negligible p-value of 0.04 in the case of Sachin Tendulkar, which means that only 4% of Tendulkar's scores could be explained by an exponential distribution. Similarly, when the same hypothesis testing was conducted on the scores of the other batsmen under study, low p-values were obtained once again indicating that the scores of all the batsmen could not be explained by an exponential distribution. Due to this result of ours, the null hypothesis that the scores of a batsman arise out of an exponential distribution was rejected.

After the failure of the exponential distribution to explain the scores of a batsman, other distributions including the likes of the lognormal distribution were explored. However, these distributions also yielded poor p-values. Finally, we stopped at the Weibull

distribution. While the Weibull distribution yielded the highest p-values when compared to all other distributions, it was also chosen because of its ability to fit all types of data – left-skewed data, right-skewed data, symmetric data etc.

Klammer et Al[7] (2009) explain in their paper that if the null hypothesis that the observations are arising out of a particular distribution is true, then all the obtained p-values form a uniform distribution when fitted. The justification of this point lies in the fact that the whole point of using the correct distribution (for example, normal, t, f, chisq, Weibull etc.) is to transform the test statistic to a uniform p-value. In case the null hypothesis is false, the distribution of the obtained p-values will be more weighted or skewed towards 0.

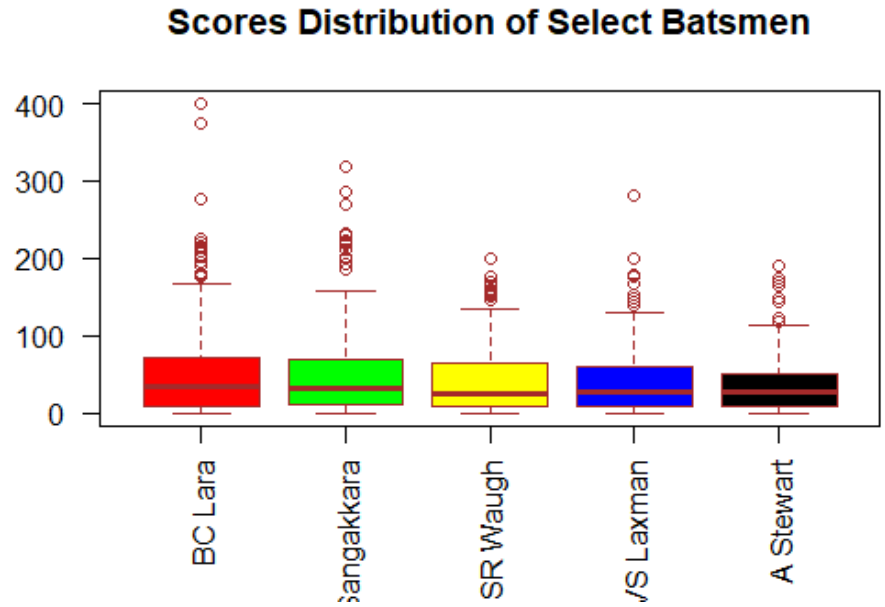

Fig. 3: Box and Whisker Plots depicting the distribution of scores of select batsmen—the greater deviation of scores around the mean indicates lesser consistency

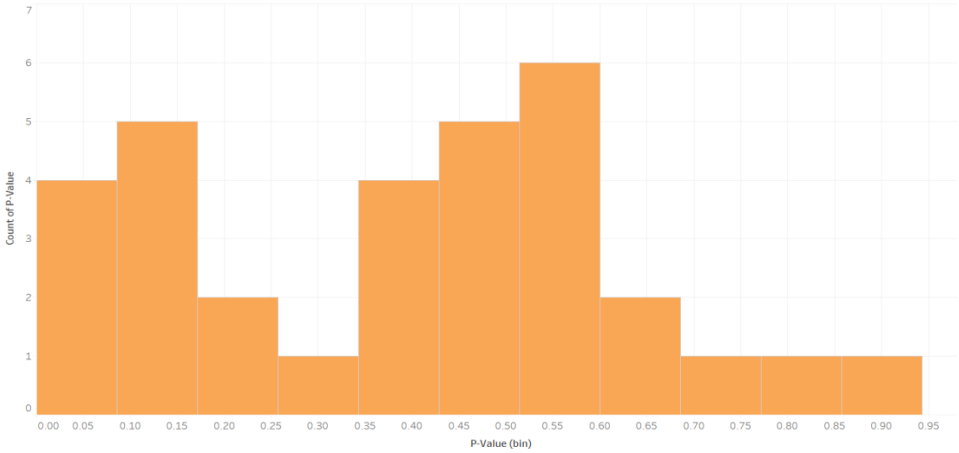

Fig. 4: The distribution of the p-values of the scores of all batsmen under study

When a distribution of the p-values of all the batsmen's scores was taken, they were more or less found to follow a uniform distribution. Chart-4 gives us an idea of how the distribution of the p-values looks like. Hence, we fail to reject our null hypothesis which states that all the scores of a particular batsman under study arise out of a Weibull distribution.

Table 3: A comparison of the p-values of the exponential and Weibull distributions of the plotted scores of a few selected batsmen

	Player	p-value (Exponential)	p-value (Weibull)
1	S Tendulkar	0.00	0.27
2	RT Ponting	0.14	0.89
3	JH Kallis	0.04	0.43
4	BC Lara	0.17	0.53
5	M Jayawardene	0.01	0.51
6	Younis Khan	0.23	0.75
7	MJ Clarke	0.02	0.52

The probability distribution function (pdf) of a Weibull distribution can be expressed in terms of its shape and scale parameters (α and θ) as:

$$f(z) = (\alpha z^{(\alpha-\frac{1}{\theta})} * e^{-z^{\alpha/\theta}});$$

The log likelihoods of the pdf were computed and then maximized upon to obtain the Maximum Likelihood Estimate (MLE) of the mean μ and the standard deviation σ. The MLE of the mean of a Weibull distribution came out to be:

$$E(Z) = (\theta^{\frac{1}{\alpha}})\Gamma(\frac{1}{\alpha}+1)$$

While the MLE of the standard deviation can be mathematically expressed as:

$$V(Z) = \theta^{(2/\alpha)}*\Gamma((2/\alpha)+1)-((\theta\text{^}(1/\alpha)*\Gamma(1/\alpha+1))$$

We thus define the consistency of a batsman in terms of the standard deviation of the Weibull distribution of all his individual scores in Test cricket. The more the standard deviation of the batsman's scores, it means that the batsman's scores fluctuate to extreme values on either sides of the mean. Consequently, a higher standard deviation means a lesser consistency or dependability of the batsman. This means that the standard deviation of a batsman's scores is a negative influencer on his performance, i.e., the performance of a batsman is inversely proportional to the standard deviation of his scores.

Since we want all our indices to be of positive influence, the reciprocal of the standard deviation is taken for the consistency. The final index for consistency was obtained by computing the reciprocal of the aforementioned formula for standard deviation of a Weibull distribution.

2.1.4 Statistically Computed Average: The usage of solely the conventional batting average for ranking batsmen has severe shortcomings. However, there is also the need to factor in how much each batsman is expected to score in any given match. For this purpose, we will not use the conventional batting average, but instead turn to the expected value or the mean of the Weibull distribution of the batsman's scores, the mathematical expression for which we had derived earlier. We call this mean of the Weibull distribution the statistically computed average and when used in conjunction with the other indices, it turns out to be a very powerful tool.

Lemmer (2008a) statistically proved that if a batsman had a not out score and assuming that the batsman would be allowed to continue until he gets dismissed, then, on average, he could have been expected to double his score. This is more so in the case of a Test match, where the batsman always tries to minimize risk when batting. Thus, all the not-out scores of a batsman are doubled before all the scores are fitted to a Weibull distribution. Eventually, the statistically computed average of each batsman in the study is obtained by the formula:

$$E(Z) = (\theta^{\frac{1}{\alpha}})\Gamma(\frac{1}{\alpha}+1)$$

where E(Z) is the expected value or the mean of the Weibull distribution of all the scores of a batsman.

2.1.5 Team Contribution: When one is developing a methodology for ranking batsmen, it is imperative to look at how much of the batsman's runs have helped the team win, or helped the team wriggle out of difficult situations. In a test, there are four possible scenarios: a win, a loss, a drawn test or a tied test. We consider a tied test to be equal to a drawn test (though technically it is not the case).

The interpretation of the *teamCont* index goes this way: the batsman who has scored a majority of his runs whenever his team has won or at least managed a draw gets a higher *teamCont* score when compared to a batsman who did a bulk of his scoring in the matches in which his team lost. Through this index, we can get an idea of who were the batsmen who influenced the success of their team the most.

$$teamCont_i = (win\%)_i + 0.5 * (draw\%)_i$$

2.1.6 Home/Away Percent: In a recent article by BBC titled "Ashes: Is cricket the hardest sport to win away from home?", the author Marc Higginson statistically proves that test matches are always heavily biased in favor of the home side. In the past 100 years, the percentage of tests won by the away team has never exceeded 28%; it has always remained in the range of 20% to 30% and in the year 2013, this particular statistic registered an abysmal low of 6.8% of wins for the visiting teams.

The aforementioned facts and statistics illuminate the great difficulty for an away team to win a test match. Against such a backdrop, a batsman who scores heavily in away conditions is definitely a superior batsman compared to the others. Hence, we consider home/away percentage to be our sixth factor in our ranking methodology.

2.1.7 Miscellaneous (Time-varying) factors: The game of cricket is a constantly changing one. The only constant in cricket apart from bat and ball is change. Right from the technological advancements in umpiring equipment to the increase in safety and protection for all cricketers on the field, the game of Cricket presents us with starkly different pictures from era to era. Some of the rules also keep on changing from time to time. It is important for us to determine the batsmen who played without the luxury of the three subfactors that we considered in the miscellaneous factors were: (i) Helmets; (ii) Decision Review System (commonly abbreviated as DRS); and (iii) Bouncers.

2.1.7.1. Helmets: Though Graham Yallop of Australia, the first batsman to wear a helmet to a cricket field, wore a helmet to an Australia-West Indies test match as early as in the year 1978, it was not until the year 1980 that helmets were introduced. Helmets prevented injuries to the head, thus allowing the batsmen donning helmets to play fast bowlers more freely without the

fear of getting injured. Once helmets became an integral part of a batsman's cricket kit, they allowed batsmen to focus only on hitting the ball without having the unwanted thought of having to protect oneself against the ball. In fact, legendary Australian opening batsman Matthew Hayden once said, "The helmet is one of the greatest assets that a cricketer can have". Helmets give a reassurance to the batsman that even in the case of missing the ball when a bouncer is bowled, the helmets are there to provide him the much-required protection.

Having understood the importance of helmets for batsmen in the game of cricket, we need to devise a subfactor for helmets wherein the batsmen who scored big runs despite not playing with a helmet on get a higher helmet score when compared to similar batsmen who played with helmets.

For computing the helmet index, we take the proportion of the batsman's career for which the batsman wore a helmet. In the end, the helmet scores thus obtained for all the batsmen are subtracted from the number 1.

$$HelmetIndex = 1 - \frac{No.of\ years\ played\ with\ \ helmet}{Total\ no.of\ years\ of\ career\ span}$$

2.1.7.2. Decision Review System (DRS): The Decision Review System (DRS) which was first used in a Test match between India and Sri Lanka in the year 2008, can be considered a gamechanger for Cricket. Both the batting team and the bowling team can make use of the DRS facility to refer the decision of the on-field umpires to the third umpire, in case they feel that the decision taken by the on-field umpires is incorrect.

Before the arrival of the DRS system on the stage of international cricket, there were several instances wherein despite the batsman being not-out, the on-field umpires made an error in judgement and consequently gave the batsman to be out. In several cases, these wrong decisions on the part of the umpire had a great bearing on the game and in some cases, they proved to be the difference between victory and defeat for the losing side. For a batsman, a wrong decision from the umpire could mean a potential big innings being cut short.

However, since the introduction of the Decision Review System in the international arena of Test Cricket, batsmen have effectively used it to successfully overrule the wrong decisions taken by the on-field umpires.

Very similar to how we computed the helmet index, the DRS index is also calculated by first calculating the proportion of the batsman's career for which DRS was in place, and then subtracting the thus obtained result from 1.

2.1.7.3. Bouncers: In cricket, a bouncer is a special type of delivery bowled by fast bowlers in which the ball is made to bounce at a height greater than the chest of the batsman. Bouncers are usually directed at the bodyline of the batsman. In test matches, bouncers are bowled to intimidate the batsman whereas they may be used in the limited formats to put a lid on the scoring rate, or to push the batsman on to the backfoot in case the batsman is playing very freely on the front foot. In any case, a bouncer is probably one of the toughest balls to face in cricket.

In any case, even if a batsman has a good defensive technique to protect both his wicket and his body from the dangerous bouncers of the bowler, it is really difficult to score good runs of bouncers.

Against such a backdrop, if the pitch provides good assistance to the fast bowlers, there is the possibility that the bowler keeps on repeatedly bowling bouncers in order to handicap him from scoring any runs, and to intimidate the batsman to such an extent that the batsman keeps on anticipating the bouncer and on a delivery when the batsman is anticipating the bouncer yet again, the bowler bowls a wicket-taking delivery and gets the batsman out. In fact, this style of intimidatory tactics was very much in vogue in the 1970s and the 1980s (remember that the batsmen did not wear helmets in this era) where teams like the West Indies intimidated the opposition batsmen with fast-paced bouncers. Keeping these intimidatory tactics in mind, the International Cricket Council in 1991 introduced a "one bouncer per batsman per over" rule in an attempt to discourage the use of intimidation. However, in the wake of the negative reception it got by both players and umpires, the ICC increased it to two bouncers per over in 1994.

There is the need to address the issue of the number of bouncers that were allowed per over during the career spans of different batsmen to assess how good each one of them were, given a certain number of bouncers allowed in an over during their time.
Period 1: 1970-1990— 6 bouncers allowed per over
Period 2: 1991-1994— 1 bouncer allowed per over
Period 3: 1995-2019— 2 bouncers allowed per over

The bouncer index for an index is calculated using the following formula:

$$bouncerIndex_i = \frac{6 * (years\ in\ P1)_i + 1 * (years\ in\ P2)_i + 2 * (years\ in\ P3)_i}{Total\ years\ of\ career\ span}$$

2.1.8 Longevity: Sustaining both fitness levels and performance levels over prolonged periods of time is a hallmark of any sportsman, and a batsman in Cricket is no exception. Hence, we introduce the last factor called longevity. Longevity is a weighted average of the number of years and number of innings over which the batsman's career spanned, with both weights being equal to 0.5.

The only problem with considering longevity as a factor is that contemporary batsmen like Virat Kohli, Steven Smith, Kane Williamson, Joe Root, and David Warner are disadvantaged—they have not yet retired from the game and they may go on to play many more matches and score many more runs. To solve this problem, the number of matches that they may go on to play till the end of their careers was estimated by extrapolating the number of matches that they have already played to the number of years they are predicted to play.

The longevity L_i of a batsman i who has played x_i matches and over y_i years is mathematically defined as:

$$L_i = 0.5 * \frac{x_i}{\frac{1}{n}(\sum_{i=1}^{n} x_i)} + 0.5 * \frac{y_i}{\frac{1}{n}(\sum_{i=1}^{n} y_i)}$$

2.2 Computing Mahalanobis distances

Let $a_i, b_i, c_i, d_i, e_i, f_i, g_i, and\ h_i$ denote the batsman's difficulty of opposition bowlers faced, difficulty of opposition teams, consistency, statistically computed average, team contribution, home-away index, the miscellaneous index, and longevity respectively. It is to be noted that all of the aforementioned eight quantities are non-negative and all of them are positive influencers on the ranking of a batsman, meaning a higher value of any of the indices for a batsman would mean better performance.

The Mahalanobis distance[8] can be defined as the distance between any two points in multivariate space. The Mahalanobis distance always measures distance relative to the 'centroid'. A centroid can be thought of as a base or central point—a point that servers as an overall mean for multivariate data. Mahalanobis distances are commonly used in the detection of multivariate outliers, which is exactly the method we have employed to determine the greatest batsmen in the span of the last 50 years. For a player i, the Mahalanobis distance can be defined as:

$$MD(X_I, X_J) = \left\{(X_i - X_j)^T S^{-1}(X_i - X_j)\right\}^{1/2}$$

The above equation gives us the mathematical representation of the Mahalanobis distance or the squared statistical distance of the eight-dimensional vector $(a_i\ b_i\ c_i\ d_i\ e_i\ f_i\ g_i h_i)$ from the origin vector (0 0 0 0 0 0 0 0) in the setting of the axes rotated in the direction of correlation between the eight values comprising the vector, viz., $a_i, b_i, c_i, d_i, e_i, f_i, g_i, and\ h_i$. However, a major drawback with the concept of Mahalanobis distance is that the inverse of the correlation matrix is required for computation of the Mahalanobis distance. If the variables are highly correlated[9] (Varmuza & Filzmoser, 2016), the correlation matrix becomes a singular matrix which makes it impossible for us to compute its inverse. Hence, we first computed the correlation matrix to confirm that the variables are not highly correlated. In our case, the correlation between the eight variables was found to be low.

2.3 Performing Data Envelopment Analysis (DEA) to compute the efficiencies of all batsmen

When ranking batsmen by Data Envelopment Analysis, the eight factors are divided into input and output factors--four input and five output factors. (One additional input factor namely Home-Away Played ratio is added to the eight factors taken earlier—this factor is the ratio of the number of away matches played by the batsman to the number of home matches played by the batsman). The segregation of factors into input factors and output factors is done in this way: the factors that influence the difficulty or the relative ease for a batsman to score are the input factors, while the output factors are those that the batsman himself generates through his performances with the bat.

Table 4: Input and Output factors of a batsman

INPUT FACTORS	OUTPUT FACTORS
Difficulty of Opposition Bowlers	Consistency
Difficulty of Opposition Teams	Statistically Computed Average
Miscellaneous Factors (Helmets, DRS, Bouncers)	Team Contribution
Home-Away Matches Played Ratio	Home-Away Averages Ratio
	Longevity

Each batsman is considered to be a machine or a Decision-Making Unit (DMU). The DEA technique calculates the efficiencies of all DMUs by taking a set of input and output variables and then set a benchmark. The most important advantage of this technique is that it can handle the multiple input and output variables which are generally not comparable to each other. The efficiency is found by solving a Linear Programming (LP) problem for each batsman and maximizing outputs to get the weights for the final LP formulation. When performing DEA, the inputs are reversed (i.e., they are subtracted from 1 after normalization—since we want the tougher inputs to be lower in magnitude and the concept of efficiency is to get high outputs from low inputs)

The efficiency of a batsman *i* is defined in mathematical terms as:

$$\boldsymbol{efficiency_i} = \frac{(\boldsymbol{u1} * \boldsymbol{consistency_i}) + (\boldsymbol{u2} * \boldsymbol{StatAverage_i}) + (\boldsymbol{u3} * \boldsymbol{teamCont_i}) + (\boldsymbol{u4} * \boldsymbol{homeAway_i}) + (\boldsymbol{u5} * \boldsymbol{longevity_i})}{(\boldsymbol{v1} * \boldsymbol{bowlDiff_i}) + (\boldsymbol{v2} * \boldsymbol{teamDiff_i}) + (\boldsymbol{v3} * \boldsymbol{Miscellaneous_i}) + (\boldsymbol{v4} * \boldsymbol{homeAwayPlayed_i})}$$

where u1, u2, u3, u4, and u5 are the weights of the output factors, while v1, v2, v3, and v4 refer to the weights of the input factors

3. DISCUSSION OF RESULTS

The Mahalanobis distances of each of the batsmen from the mean cluster was calculated to obtain a measure of how exceptional he was when compared to the others. Since we have taken all our indices to be non-negative positive influencers, the greater the Mahalanobis distance of a batsman's eight-dimensional position vector from the mean cluster vector, the more exceptional the batsman is relative to others, i.e., batsman performance is directly proportional to his Mahalanobis distance from the origin vector. Table-5 gives us the recalculated relative scores of the batsmen on a scale of 1 (this is done so that no one factor is more influential than another factor), along with their Mahalanobis distances from the mean vector. The efficiencies of the batsmen thus obtained are listed in Table-6. Batsmen with an efficiency of 1 are the most efficient, as they have effectively converted low inputs to high outputs. The final rankings in Table-7 are obtained by ranking the outlier batsmen who have the highest efficiencies at the top.

4. CONCLUSION

Table-7 enumerates the ten all-time great batsmen in Test Cricket. Unsurprisingly, Sir Don Bradman occupies the No.1 spot on the rankings. His superior statistically computed average, high miscellaneous score (he did not have helmets or DRS in his time, nor was there any limit on bouncers), high team contribution (a very high proportion of his runs came in wins or drawn matches of Australia), and a reasonably high bowling difficulty, all contributed to his undisputed No.1 spot. In fact, the distance between Bradman and the second-best Haynes is so large that Bradman is at a distance of 47.94 from the mean cluster while Haynes is at a distance of 16.6361, which is not very large compared to the distance of the greats following him. Chanderpaul, Richards, Gavaskar, Tendulkar, Lara, Gower, Greenidge and Gooch complete the list of the 10 all-time great batsmen in Test Cricket.

Table 5: Batsmen ranked in descending order of their Mahalanobis distances from the mean batsman vectorRank

Rank	Player	*bowlDiff*	*teamDiff*	consistency	*statMean*	*teamCont*	homeAway	miscellaneous	Longevity	Distance
1	Sir DG Bradman (AUS)	0.6000	0.5282	0.0000	1.0000	0.9556	0.0939	1.0000	0.2606	47.9388
2	DL Haynes (WI)	0.9889	0.7143	0.8391	0.0427	0.7603	0.3540	0.6061	0.4245	16.6361
3	S Chanderpaul (WI)	0.2000	0.8658	0.6313	0.2671	0.0000	0.3394	0.2600	0.7808	15.4541
4	IVA Richards (WI)	0.3200	0.1991	0.2609	0.1958	0.7171	1.0000	0.7557	0.4091	14.6334
5	SM Gavaskar (INDIA)	0.2500	0.2013	0.2036	0.2219	0.1771	0.4047	0.8495	0.4537	14.1701
6	SR Tendulkar (INDIA)	0.7000	0.5974	0.2440	0.2551	0.4051	0.5761	0.3137	1.0000	14.1580
7	RT Ponting (AUS)	0.1000	0.5325	0.3142	0.2107	0.9308	0.2141	0.3011	0.6648	12.9787
8	BC Lara (WI)	0.6000	0.8701	0.0000	0.2430	0.0048	0.3433	0.3421	0.4976	12.6424
9	DI Gower (ENG)	0.7296	0.1991	0.8224	0.0980	0.2146	0.2821	0.6532	0.3628	11.3524
10	CG Greenidge (WI)	0.6000	0.1991	0.4742	0.1008	0.9636	0.6980	0.7557	0.4164	11.2279
11	CH Lloyd (WI)	0.3200	0.0000	0.6746	0.1411	0.6511	0.6313	0.8778	0.2119	11.1015
12	GA Gooch (ENG)	0.8333	0.3506	0.5551	0.0611	0.3323	0.0000	0.6534	0.5893	10.2986
13	ME Waugh (AUS)	0.0000	0.4935	0.8154	0.0913	0.7843	0.4047	0.3158	0.2752	10.2081
14	DB Vengsarkar (INDIA)	0.2500	0.2013	0.6563	0.0948	0.1714	0.2784	0.6970	0.3831	9.3578
15	SPD Smith (AUS)	0.2866	0.6104	0.1621	0.3669	0.7133	0.4166	0.0000	0.3628	9.1052
16	MEK Hussey (AUS)	0.4000	0.5779	0.5006	0.2437	0.8951	0.6387	0.2138	0.0000	9.0629
17	Mohammad Yousuf (PAK)	0.9000	0.7143	0.2944	0.2339	0.3240	0.7801	0.3421	0.1794	8.8969
18	KC Sangakkara (SL)	0.2000	0.7922	0.0550	0.2964	0.5136	0.2554	0.2292	0.4667	8.7908

19	V Kohli (INDIA)	0.7167	0.5714	0.1265	0.2317	0.5604	0.5133	0.0000	0.3628	8.2354
20	AJ Stewart (ENG)	0.6000	0.7597	1.0000	0.0273	0.1863	0.2589	0.3421	0.4050	7.7568
21	AR Border (AUS)	0.4833	0.2511	0.6589	0.2015	0.3044	0.3620	0.6058	0.5779	7.3757
22	SP Fleming (NZ)	0.7000	0.8506	0.6809	0.0682	0.1952	0.7624	0.3368	0.3263	7.0953
23	AJ Strauss (ENG)	0.3000	0.7208	0.6952	0.0428	0.6905	0.1932	0.2566	0.0998	7.0532
24	CH Gayle (WI)	0.6000	1.0000	0.5731	0.0734	0.0558	0.4880	0.2429	0.3093	6.9110
25	JL Langer (AUS)	0.9000	0.4935	0.4749	0.1209	0.9695	0.1917	0.3316	0.3093	6.8429
26	DPMD Jayawardene (SL)	0.7000	0.7576	0.1817	0.1720	0.5216	0.1182	0.2600	0.5795	6.6026
27	BB McCullum (NZ)	0.3000	0.8506	0.6575	0.0000	0.1837	0.2401	0.1711	0.2281	6.5761
28	HH Gibbs (SA)	1.0000	0.6104	0.4795	0.0580	0.5003	0.6569	0.3421	0.1745	6.4913
29	GP Thorpe (ENG)	0.3000	0.7597	0.8396	0.1167	0.5116	0.4209	0.3300	0.2354	6.1095
30	DA Warner (AUS)	0.7014	0.6104	0.4682	0.1463	0.6210	0.1775	0.0000	0.3628	6.0490
31	JH Kallis (SA)	0.8000	0.5801	0.3491	0.3059	0.6204	0.3156	0.2839	0.6810	5.9989
32	ML Hayden (AUS)	0.7045	0.4935	0.3835	0.1796	1.0000	0.1240	0.3372	0.3474	5.9750
33	MA Atherton (ENG)	0.8029	0.7597	0.7219	0.0215	0.2199	0.1104	0.3664	0.3157	5.7637
34	KS Williamson (NZ)	0.8029	0.7013	0.1984	0.2700	0.6292	0.4642	0.0000	0.3628	5.7281
35	Javed Miandad (PAK)	0.8333	0.2554	0.3686	0.2338	0.4843	0.3828	0.6742	0.4261	5.5069
36	DC Boon (AUS)	0.3667	0.3831	0.7890	0.1234	0.5997	0.1069	0.4879	0.2622	5.4900
37	VVS Laxman (INDIA)	0.6000	0.6710	0.7001	0.1550	0.4619	0.6782	0.3011	0.4805	5.4397
38	JE Root (ENG)	0.7000	0.6364	0.3257	0.1715	0.5551	0.2221	0.0000	0.3628	5.1845
39	PA de Silva (SL)	0.3910	0.7013	0.4549	0.0596	0.2031	0.4438	0.4418	0.3864	5.0704
40	R Dravid (INDIA)	0.7000	0.6710	0.3469	0.1875	0.4814	0.7131	0.3011	0.6291	5.0674
41	ST Jayasuriya (SL)	0.8000	0.7468	0.6048	0.0016	0.4405	0.1564	0.3235	0.3904	4.9031
42	Younis Khan (PAK)	0.6000	0.8182	0.1331	0.2735	0.5098	0.6708	0.2018	0.4846	4.8943
43	SR Waugh (AUS)	0.5000	0.4372	0.4393	0.2589	0.7298	0.3357	0.4211	0.6656	4.6957
44	KP Pietersen (ENG)	0.8000	0.7208	0.3234	0.1369	0.5605	0.2452	0.1916	0.1404	4.4947
45	SC Ganguly (INDIA)	0.5000	0.7208	0.7932	0.0613	0.3681	0.6266	0.3421	0.2573	4.3972
46	V Sehwag (INDIA)	0.7000	0.6948	0.1741	0.1878	0.4893	0.2787	0.2566	0.2378	4.2318
47	LRPL Taylor (NZ)	0.5000	0.8506	0.5381	0.1796	0.3979	0.3805	0.0855	0.1964	3.9379
48	Inzamam-ul-Haq (PAK)	0.7000	0.7143	0.4503	0.2113	0.6205	0.7192	0.3273	0.3864	3.8601

49	HM Amla (SA)	0.3000	0.6429	0.3499	0.1762	0.7104	0.2489	0.1368	0.4229	3.7043
50	GC Smith (SA)	0.7000	0.6429	0.3861	0.1391	0.7737	0.6729	0.2292	0.2987	3.5771
51	G Kirsten (SA)	0.6000	0.6104	0.3670	0.1066	0.7449	0.5328	0.3289	0.1948	3.5073
52	MJ Clarke (AUS)	0.4000	0.5779	0.2746	0.1742	0.7087	0.2894	0.1882	0.2484	3.4136
53	M Azharuddin (INDIA)	0.7000	0.5000	0.5209	0.1022	0.2582	0.4282	0.4536	0.2906	2.8687
54	MA Taylor (AUS)	0.6000	0.4416	0.5527	0.0839	0.7434	0.3121	0.3708	0.1859	2.7283
55	AB de Villiers (SA)	0.7000	0.6429	0.4209	0.2095	0.6255	0.2676	0.1471	0.3312	2.2516
56	AN Cook (ENG)	0.7000	0.7208	0.4392	0.0990	0.5078	0.3238	0.1129	0.5081	2.2512

Table-6: Efficiencies of all batsmen under study

Player	Efficiency	Player	Efficiency	Player	Efficiency
Sir DG Bradman (AUS)	1.0000	**KC Sangakkara (SL)**	0.9350	**HM Amla (SA)**	0.7371
DL Haynes (WI)	1.0000	**KS Williamson (NZ)**	0.9294	**PA de Silva (SL)**	0.7349
S Chanderpaul (WI)	1.0000	**DI Gower (ENG)**	0.9207	**JL Langer (AUS)**	0.7324
IVA Richards (WI)	1.0000	**GP Thorpe (ENG)**	0.9207	**ST Jayasuriya (SL)**	0.7160
SR Tendulkar (INDIA)	1.0000	**SM Gavaskar (INDIA)**	1.0000	**SPD Smith (AUS)**	0.7106
BC Lara (WI)	1.0000	**LRPL Taylor (NZ)**	0.9084	**BB McCullum (NZ)**	0.7083
AJ Stewart (ENG)	1.0000	**AJ Strauss (ENG)**	0.8974	**IR Bell (ENG)**	0.6774
CG Greenidge (WI)	1.0000	**GC Smith (SA)**	0.8907	**AB de Villiers (SA)**	0.6712
CH Gayle (WI)	1.0000	**VVS Laxman (INDIA)**	0.8719	**DC Boon (AUS)**	0.6682
CH Lloyd (WI)	1.0000	**SC Ganguly (INDIA)**	0.8663	**MJ Clarke (AUS)**	0.6381
GA Gooch (ENG)	1.0000	**MEK Hussey (AUS)**	0.8546	**V Sehwag (INDIA)**	0.6162
HH Gibbs (SA)	1.0000	**RT Ponting (AUS)**	0.8430	**ME Waugh (AUS)**	0.6157
Inzamam-ul-Haq (PAK)	1.0000	**AR Border (AUS)**	0.8337	**V Kohli (INDIA)**	0.6156
Mohammad Yousuf (PAK)	1.0000	**DB Vengsarkar (INDIA)**	0.8069	**DA Warner (AUS)**	0.5915
SP Fleming (NZ)	1.0000	**G Kirsten (SA)**	0.7900	**JE Root (ENG)**	0.5799
Younis Khan (PAK)	1.0000	**MA Atherton (ENG)**	0.7676	**M Azharuddin (INDIA)**	0.5784
R Dravid (INDIA)	0.9820	**AN Cook (ENG)**	0.7657	**MA Taylor (AUS)**	0.5269
DPMD Jayawardene (SL)	0.9784	**ML Hayden (AUS)**	0.7648	**Javed Miandad (PAK)**	0.9711

Table-7: Final Ranking of Top 10 batsmen after ranking the most exceptional batsmen in order of their efficiencies

Rank	Player
1	Sir DG Bradman (AUS)
2	DL Haynes (WI)
3	S Chanderpaul (WI)
4	IVA Richards (WI)
5	SM Gavaskar (INDIA)
6	SR Tendulkar (INDIA)
7	BC Lara (WI)
8	DI Gower (ENG)
9	CG Greenidge (WI)
10	GA Gooch (ENG)

5. REFERENCES

[1] Narayan, Paresh & Rath, Badri Narayan & kp, Prabheesh. (2016). What is the Value of Corporate Sponsorship in Sports?. Emerging Markets Review. 26. 10.1016/j.ememar.2016.02.003.

[2] Sohail Akhtar, Philip Scarf and Zahid Rasool(2017). *Rating players in test match cricket, Journal of the Operational Research Society, Volume 16, 2015-Issue-4.*

[3] Borooah V. K., Mangan J. E. (2010). The "Bradman Class": An exploration of some issues in the evaluation of batsmen for test matches. *Journal of Quantitative Analysis in Sports*, 6(3), 1877−2006.

[4] Sarkar S., Banerjee A., *Measuring Batting Consistency and Comparing Batting Greats in Test Cricket: Innovative Applications of Statistical Tools,* IIM Calcutta WPS No.784 (2016)

[5] Gweshe, T & Durbach, Ian. (2013). An analysis of the efficiency of player performance at the 2011 Cricket World Cup. ORiON. 29. 137-153. 10.5784/29-2-123.

[6] Johnson R. A., Wichern D. W. (2015). *Applied Multivariate Statistical Analysis*. 6th Ed., Pearson; *680-692*

[7] Klammer, A. A., Park, C. Y., and Stafford Noble, W. (2009) Statistical Calibration of the SEQUEST XCorr Function. *Journal of Proteome Research*. 8(4): 2106–2113

[8] Mahalanobis P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2, 49–55.

[9] Varmuza, K. & Filzmoser, P. *Introduction to Multivariate Statistical Analysis in Chemometrics*. CRC Press.

[10] Lemmer, H.H. (2008a). Measures of batting performance in a short series of cricket matches. South African Statistical Journal, 42(1): 83-105.

[11] Rohde N. (2011). *An "Economic" ranking of batters in test cricket. Economic Papers: A Journal of Applied Economics and Policy, 30(4), 455–465.*

[12] Alexander M.Petersen, Orion Penner,and H.Eugene Stanley(2011). *Methods for detrending success metrics to account for inflationary and deflationary factors.*

[13] Richard J. Boy,Peter M.Philipson(2018). *On the ranking of test match batsmen, A Journal of the Royal Statistical Society in Applied Statistics.*

[14] Scott M.Berry, C.Shane Reese and Patrick D.Larkey (2012). *Bridging Different Eras in Sports, Journal of the American Statistical Association, Volume 94, 1999-Issue-447*

[15] Brown, H.Shelton (2001). *Comparing Batsmen Across Different Eras: The Ends of the Distribution Justifying the Means.Discusion Paper No.289.School of Economics,The University of Queensland.*

[16] Daniyal Muhammad et al (2012), Analysis of Batting Performance in Cricket using Individual and Moving Range (MR) Control Charts. International Journal of Sports Science and Engineering Vol. 06 No. 04, pp. 195-202

[17] Shah, Sricharan & Hazarika, Partha & Hazarika, Jiten. (2019). A Study on Performance of Cricket Players using Factor Analysis Approach. International Journal of Advanced Research. 8. 656 - 660.

