Journal of Sports Analytics 7 (2021) 139–153
DOI 10.3233/JSA-200503
IOS Press

Are today's Test cricket batsmen better than the greats of yesteryears? A comparative analysis

Anil Gulati* and Charles Mutigwe
Western New England University, Department of Business Information Systems, College of Business, Springfield, MA, USA

Abstract. In sports, including Test cricket, athletes from years past serve as performance role models and set benchmarks for subsequent generations of players. Sports fans often wonder: are players of today as good as greats from the past? Alternatively, how do today's athletes compare with greats from yesteryears? This paper attempts to answer that question for Test match cricket. We applied data mining to batting performance of eighty, now retired, Test Cricket Greats (TCG from hereon) from eight major Test cricket countries. Batting performance attributes included batting average, strike rate, numbers of fifties and hundreds scored, among others. Using k-Means cluster analysis, TCG performance records were classified into three clusters which was our Training Model. Two clusters were populated by established batsmen and the third cluster included bowlers, all-rounders with significant bowling, and some batsmen. The Learning Model was applied to predict classifications of thirty two Test Cricket Active (TCA from hereon) players. Statistical tests were performed, cluster wise, to highlight similarities and dis-similarities between TCA and TCG players. Results show that several active players, while still mid-career, have already achieved batting performance records which are at par with the best of TCG.

Keywords: Sports analytics, data mining, k-means clustering, cricket analytics

1. Introduction

Cricket is one of the more popular spectator sports in the world with viewership in billions spread across the globe. The sport is governed by the ICC (International Cricket Council, 2020) which states on its website "*The ICC is the global governing body for cricket. Representing 105 members, the ICC governs and administrates the game and works with our members to grow the sport. The ICC is also responsible for the staging of all ICC Events. The ICC presides over the ICC Code of Conduct, playing conditions, the Decision Review System and other ICC regulations.*"

Today, cricket is played in three formats: Test match (Test), Twenty20 (T20) and One Day International (ODI). Relatively recently introduced, the last two formats provide fans with fast paced sporting entertainment. In the ODI format, team batting first sets a target score for the team batting second, to chase and exceed. Large score targets combined with limited overs result in batsmen taking more scoring risks. This risk taking behavior is magnified in T20 format which is termed by fans as "hit and run" cricket, hence the term "T20 effect." In Tests, each team bats twice and frequently these matches end in a draw. Obviously, the objective is to score more runs than the opposing side. Time only becomes a factor in closely contested matches with large scores or when match duration is truncated by weather.

*Corresponding author: Anil Gulati, Western New England University, College of Business, 1215 Wilbraham Road, Springfield, MA 01119, USA. Tel.: +1 (413) 782 1711; E-mail: agulati@wne.edu.

Like other sports, cricket produces vast amounts of data which has attracted the attention of researchers. Significant mathematical analysis of game statistics and applications of analytics can be found in the literature. Swartz (2016) presents a survey of cricket analytics. According to the author, major research streams related to cricket include: analysis of scoring targets, simulations of match conditions and match progression, evaluation of player and team performance. Other studies can be classified as those explaining on-field strategies responding live to changing conditions in the match, strategies optimized for specific cricket formats, models of past performance to improve player contributions, effective team roster selections and simulations to determine effective batting order, all with the ultimate goal to improve the probability of winning.

Research on batting performance has devoted adequate energy testing the validity and efficiency of batting performance metrics. For batsmen, batting average is still the most widely quoted statistic as it is easily understood and accepted by the average fan. Though a simple formula, there are serious drawbacks in what it actually measures. It is a flawed and insufficient measure of a batsman's true batting potential. Additionally, batting average does not capture scoring consistency expected from batsmen. Several studies have underscored these deficiencies and proposed reformulations. These studies attempt to effectively represent players' true batting strength, which forms the basis of player ratings, widely disseminated to cricket fans as player rankings. These ratings influence how team selectors assemble squads and enable team captains to dynamically change playing strategies in response to proceedings on the field.

The MRF tyres ICC rankings (previous name, Reliance ICC rankings) are the official record of player rankings and are available from the ICC web site. Updated regularly, rankings are produced in three separate categories: batting, bowling, and all-rounder.

Akhtar et al. (2015) proposed a new system for rating players in a Test match. Their rating system accounted for match conditions by calculating player contributions, separately in the 15 sessions in a Test match (morning, lunch and evening sessions; over five playing days). The fifteen measurements capture player contributions in the context of changing match conditions.

Player contributions in batting, bowling and fielding are weighted into a single score which represents the overall contribution of a player towards the final outcome. The proposed metric can be used to rank players for their contributions in the match. In contrast, the traditional approach is to measure performance once, and upon conclusion of the match.

In their thesis titled "Best Players of The Test Cricket in Last Years 2014" Ahmad and Zada (2015) applied Akhtar et al. (2015) rating system to two test series. The first series was played between Australia and South Africa in Feb/Mar, 2013. Based on batting, bowling, and fielding performance over the entire series, MG Johnson, DW Steyn and BJ Haddin captured the top spots with most significant contributions. Similarly, in Dec. 2014 series played between Australia and India the top three contributors were MG Johnson, NM Lyon and SPD Smith.

Borooah and Mangan (2010) adjusted batting averages of the top 50 all-time best Test cricket batsmen (ranked by career batting average) to derive a new measure CAA (consistency adjusted average). The CAA modified the batting average by incorporating batting consistency. Authors then re-ranked the top 50 on CAA and contrasted those CAA rankings with the original. Their results showed significant deviations.

Works cited above rated performances of players relative to teammates, or players on opposing side, or peers at large. In this paper we contrast career batting performance of Test cricket's top ranked batsmen from yesteryears (TCG) with the performance of top ranked among still active batsmen (TCA). As cricket fans, we often wonder: how does the batting performance of today's top ranked Test cricketers such as: Kohli, Smith, Root, Matthews, DM Bravo, Azhar Ali, Taylor and De Kock, measure up to the batting performance of Test cricket's all-time greats like Tendulkar, Ponting, Cook, Sangakara, Lara, Miadad, Vettori and Amla? This study takes a swing at that question.

2. Related work

Motivated by recent popularity of the two short forms of cricket, T20 and ODI are capturing the most attention of current research. Traditional metrics of performance are shown to be flawed in definition or fail to incorporate qualitative aspects of the game. Player performance and team performance related research from all three formats is of interest to the present study. Proposed new metrics and methods from literature are relevant to this work.

Several studies, using cricket data from limited overs formats have proposed playing strategies. Clarke (1988) contrasted batting performance in the

first and second inning to estimate optimal scoring rates and resulting winning percentages. Preston and Thomas (2000) provide an analysis of batting strategies deployed in the two innings in English County Cricket with the significant conclusion that strategies of the opposing sides differed. Clarke and Norman (1999) studied the "when to run and when to forgo" strategy in an attempt to protect weaker batsman by keeping him on the non-striker end.

Based on derived models of player performance, another research stream proposed optimum batting lineups. Swartz et al. (2006) conducted simulation studies to determine optimal batting order. Traditionally, batting orders are static as most players eventually settle into their perceived position of strength. Taking match conditions into account Norman and Clarke (2010) applied dynamic programming and proposed batting orders that changed with state of the match and report higher expected scores resulting from dynamic strategies they proposed.

Time in a Test match is not a limiting factor, at least initially. Batsmen are afforded the luxury of time to settle into their natural batting rhythm. In early minutes on the crease all batsmen start slow until they find their groove. Brewer (2008) quantified this as "early inning effect" and estimated Hazard function parameters for Test players. The study concluded that during early minutes of their inning batsmen are more vulnerable and perform at a fraction of their potential ability.

Building on Brewer's work on slow initial start to batting Stevenson and Brewer (2018) applied survival analysis to predict scoring potential of Test cricketers. Authors concluded that superior "initial batting ability" is not a predictor of a player's career batting average. Factors such as position in the batting order and strike rate may be better predictors of career performance.

Batting average remains to be the most ubiquitous and supreme measure of batting performance. Runs contributed by individual batsmen and accumulated by the team determine winner in a match. Lewis (2005) proposed alternative performance measures for the ODI format. The proposed measures are based on Duckworth/Lewis methodology which adjusts performance for stages in the inning. An alternate batting performance measure was proposed and tested using five years of Test data by Wickramasinghe (2019) who predicted runs scored from a proposed model derived from the physical attributes of batsmen. Usefulness of actionable analytics was discussed by Jain (2015). He argued that analytics generated from historical performance and broadcast live during matches lack granularity and are not actionable for on-field decisions. He proposed using alternative term "Insights" to describe such static analytics. Pai (2020) applied programming tools to generate analytics which could be metrics useful in shaping the on-field decisions during contests for different sports, including cricket.

Using 2019 IPL (Indian Premier League) data, Joshi (2020) performed network analysis to determine more effective partnerships as measured by runs contributed (by the partnerships) to the total team score. This analysis can be used to plan a more productive batting order. Using historical performance data, Mukherjee (2013) concluded that performance of some teams was significantly influenced by the performance of a handful of players while in other cases it was a genuine team effort. Additionally, teams could achieve superior results by placing key players in selected positions, as influencers.

An author writing under alias NSS (NSS, 2016) compared today's four superstars: Virat Kohli, Joe Root, Steve Smith, and Kane Williamson. Using the metrics: Consistency, Dominance, Patience / Hitting Strength, and Winning Contribution (using three or all four, depending on the format) the author provided a framework for contrasting their recent performances in all three formats. In the final analysis, Virat Kohli rules in T20 and ODI while the Test format honors go to Steve Smith. Despite the study lacking in sophisticated analytics tools it proposed a useful framework.

3. ICC membership

ICC member countries may be eligible to play cricket in all three formats. Both teams must be sanctioned to play the match format. Today, most cricket professionals play in more than one format. The focus of this paper is on the application of analytics to the Test format, specifically, batting performance in the Test format. Therefore, all further references made are limited to batting performance in the Test format.

The very first Test match was played between England and Australia in March of 1877. Currently, twelve countries hold Test status granted by the ICC. These countries are designated as Full Status members of the ICC. Member countries are listed in Table 1, which also shows the year they played their first Test along with the to-date win/loss record. As shown in Table 1, Ireland and Afghanistan achieved

Table 1
Test sanctioned national teams and their historical records

		Matches				
Country	Span	Played(P)	Won(W)	Lost(L)	Tied(T)	Drawn(D)
Afghanistan	2018–2019	4	2	2	0	0
Australia	1877–2020	830	393	224	2	211
Bangladesh	2000–2020	118	13	89	0	16
England	1877–2020	1022	371	304	0	347
India	1932–2019	540	157	165	1	217
Ireland	2018–2019	3	0	3	0	0
New Zealand	1930–2020	440	99	175	0	166
Pakistan	1952–2020	428	138	130	0	160
South Africa	1889–2020	439	165	150	0	124
Sri Lanka	1982–2020	289	92	109	0	88
West Indies	1928–2019	545	174	195	1	175
Zimbabwe	1992–2020	109	12	69	0	28

Test status in 2018 and have since played in less than ten Tests each. Based on their limited histories, we eliminated Ireland and Afghanistan from further considerations. Previous to that, two countries admitted to this exclusive club were Bangladesh and Zimbabwe with each having played in about 100 Tests. Among the remaining eight countries, last country granted Test status was Sri Lanka (in 1982) with close to 300 Tests played.

The present study is based on player Test batting data from Australia (AU), Bangladesh (BD), England (EN), India (IN), New Zealand (NZ), Pakistan (PK), South Africa (SA), Sri Lanka (SL), West Indies (WI), and Zimbabwe (ZW). The web site, ESPNcricinfo.com offers one of the more complete statistical records of cricket. The records available go as far back as the first Test match and are current as of the last Test match. Expectedly, older data is of poor quality.

4. k-Means cluster analysis and applications in sports

In data mining, k-Means cluster analysis is a simple and powerful technique used to partition data into clusters. Clustering is an unsupervised classification technique as it is ideally suited for applications with no definable target variable to predict. The algorithm uses available features describing objects and uses some measurement of similarity to generate groupings. The goal is to create groupings in which each group member is more similar to its own cluster siblings than siblings in all other clusters.

Inputs to the algorithm are features describing each object. No additional information is available on the partitions. The analysis begins with k initial clusters selected from input patterns. Each starting cluster has a known centroid. Subsequent observations are assigned to the nearest cluster based on some evaluation metric. Cluster centroids are updated with every change in cluster membership. The algorithm iterates cluster assignments until convergence is reached or some pre-defined stopping criteria are met.

Kaufman and Rousseeuw (1990) provide mathematical formulation of the k-Means algorithm and Praveen et al. (2017) show a simplified presentation supplemented by an illustrative step-by-step example. Cluster analysis has found its way in many domains: banking, music, medicine, customer segmentation, document clustering, recommendation engines, and image segmentation, to name a few.

Kalman and Bosch (2012) provide a robust example of k-Means clustering application in the NBA (National Basketball association). Traditionally, five players on the basketball court each have one of the following five defined roles: point guard, shooting guard, small forward, power forward and center. As the sport is changing with time those "positions" have become inaccurate descriptors of the skills today's players are demonstrating. The authors applied k-Means clustering to 10 seasons of NBA data captured by twenty-three variables. Player records were clustered into nine groupings which authors tagged as the new "positions" they proposed. To test effectiveness of these proposed positions they assembled teams to maximize performance. Their results show that the new positions can be used to assemble game line-ups that deliver superior performance.

We applied k–Means clustering to create a Learning Model, which was then applied to obtain Prediction Model. Both of these Models, and the data

these were generate from are detailed in the Methodology section.

5. Data collection & pre-processing

5.1. Data for the learning model

Data used in the present study were manually scraped from ESPNcricinfo.com. Data were collected from two separate tables and are current as of February, 2020. The "Tests Batting (TB)" table includes rank ordered batting record holders by country and "Batting Innings (BI)" table is the inning by inning list of a player's entire Test match performance history. We detail the data collection process next and define our variables along the way.

For the ten sanctioned Test countries, we downloaded TB listings of top ranked players showing player name and overall career batting stats. From the career span dates, we separated retired and active players. For Bangladesh and Zimbabwe, it was observed that all their retired players had relatively shorter career spans as these two countries started playing Test cricket in years 2000 and 1992, respectively. If included in the Learning Model these two countries will be represented by performance records representing significantly shorter histories. We excluded these two countries from the Learning Model phase which further reduced the list to eight countries.

To get equal representation, we selected the same number of TCG from each country and arbitrarily limited that representation to ten complete and usable records. Starting with the top ranked player, we identified ten retired players. Data from the table TB required minimal processing to populate the following variables:

AVE = Career Batting Average
ZERO = Number of no score outs / Number of innings
FIF = Number of 50 / Number of innings
HUN = Number of 100 / Number of innings

For last three variables, the original variables in table TB were divided by the number of innings. As pointed out earlier, batting average (variable AVE) is the marquee statistic for measurement of batting performance. It has deficiencies. Later in this section we discuss those deficiencies and provide a brief literature review detailing proposed solutions.

For players identified earlier, data on the remaining variables were retrieved from table BI. These variables are discussed in the next section. Records in the BI table included players' complete playing history. A player missing more than 10 percent of data was excluded and replaced with player next in rank. This was true for players whose playing days were farther in the past. We worked our way down in the two tables until we had usable data for ten players from each country. We identify TCG players in Table A1.

5.2. Data for prediction model

For the Prediction Model, batting performance records of active players were obtained and are current as of February, 2020. We included Bangladesh and Zimbabwe as several of their active players appear to have achieved sufficient playing histories which are at par with their contemporaries. Therefore, the Prediction Model included representation from ten Test countries.

We downloaded TB and BI tables for the TCA. To balance the competing forces of sample size sufficiency and sample robustness we limited our selections to players who have played in a minimum of 45 (our arbitrary threshold) Tests. Number of players meeting that threshold varied by country. No active player from Zimbabwe met the threshold. Our final sample included thirty two TCA players from nine countries. We identify TCA players in Table A2.

We obtained following variables from the BI table:

BF = Average balls faced (average of BF in innings played)
SR = Average strike rate (average of SR in innings played)
POS = Average batting position (average of POS in innings played)
DISMISS = Frequency distribution of dismissals. Not outs were recorded as such

First three variables are simple averages, with the first two intimately familiar to cricket fans. Batsmen in the team are assigned a number from 1 to 11 indicating their position in the batting order. With time, players discover and settle into a position. Variable POS is the career average of batting order positions.

DISMISS is a categorical variable and records how a batsman's inning ended. In completed innings, dismissals can happen as: C (Caught), B (Bowled), L (LBW), R (Run out), S (Stumped), or H (Hit wicket). The last three types of dismissals are infrequent and we excluded those from further considerations.

Innings ending without a dismissal are incomplete inning and recorded as "not out (NO)".

Frequency counts of C, B and L were transformed into three ratio scale variables, as follows:

BOW = B / (*B* + C+L)
CAU = C / (*B* + C+L)
LBW = L / (*B* + C+L)

These variables replaced DISMISSAL in the dataset. With these transformations our final list included the following ten variables:

{AVE, BF, SR, ZERO, HUN, FIF, POS, BOW, CAU, LBW}

5.3. Corrections to batting average AVE

Batting average is defined as:

AVE = Total career number of runs scored / career number of times out

Batting average is a biased estimator of a batsman's true batting potential. The numerator is the sum of career runs scored in both, complete innings and incomplete innings while the denominator is a count of complete inning only, thus an upward bias in AVE. Naturally, the extent of overestimation is more pronounced where the number of incomplete innings is relatively large. Normally, that is the case in limited overs formats. This inefficiency has challenged the research community since early days.

A simple fix was proposed by Elderton (1945) to treat the NO as completed innings and include them in the count in the denominator. The modification, though an improvement, still does not accurately estimate the true batting performance. Weighted Batting Average (WBA) computation was proposed by Narayanan (2000) in which incomplete an inning with scores higher than batsman's historical average is assigned a weight of one, the same weight as a complete inning. Otherwise, incomplete innings are assigned a variable weight, ranging from 0 to 1, depending on runs scored.

The obvious complicating factor is the unpredictability of scores in incomplete innings. Scores achieved in previous innings, completed or not, are the only references available. Therefore, attempts at estimation of scores in incomplete innings must start with assumptions about the distribution of such potential scores.

For many of the early years, the prevailing assumption was that batting scores followed a geometric distribution (Elderton, 1945). Using Test cricket data, Kimber and Hansford (1993) challenged that assumption. Brewer (2008), albeit from a different perspective reached a similar conclusion which also challenged the validity of geometric distribution. Kimber and Howard proposed a non-parametric approach which incorporated estimated completed scores for the innings ending in NO and generated measures of batting performance by estimating dismissal probabilities. Other proposed underlying distributions documented in the literature are: log-normal (Bailey and Clarke, 2004), negative binomial (Ganesalingam et al, 1994) and mixed model called 'Ducks and Runs' by Bracewell and Ruggiero (2009). While these distributions have been tested and proven valid in narrow applications, no distribution has proven to be universal.

Additional mathematical re-formulations of batting average have been proposed, which based on certain assumptions, incorporate the estimated score for incomplete innings. Briefly, these include: Product Limit Estimator (PLE) by Danaher (1989), e_2, e_6 and e_{26} by Lemmer (2008, 2011), CALC=(AVE * SR/100) by Basevi and Binoy (2007). Lemmer's e_2 was modified by Van Staden et al. (2010) into a new measure $e_2{}^r$ using different estimates for the NO. Damodran (2006) applied a Bayesian approach and proposed $AV_{Bayesian}$. Maini and Narayan (2007) proposed a method accounting for exposure to the risk and calculated $AV_{exposure}$. Pointing to the validity of two estimation assumptions made by Maini and Narayan, a further improvement, in the form $AV_{survival}$ was proposed by Van Staden et al. (2010).

We elected to use Lemmer's e_{26} calculation for its simplicity and proven efficiency as a better estimator of batting average. Formula for e_{26} is:

$$e_{26} = (\text{sum-runs-out} + (2.1 - 0.005 * \text{avg-runs-no}) * \text{sum-runs-no})/n$$

where, *n* is the total number of innings (including NO), sum-runs-out is the total runs scored in complete innings, avg-runs-no is the average of runs scored in incomplete innings and sum-runs-no is the total runs scored in incomplete innings.

For all players, retired and active, we calculated e_{26} from BI table data and the new variable named $AVEe_{26}$ replaced AVE.

The following is the final list of ten variables for the k–Means analysis.

{$AVEe_{26}$, BF, SR, ZERO, HUN, FIF, POS, BOW, CAU, LBW}

The first six variables capture a player's scoring contributions which have a direct impact on the match outcome. The last four variables, broadly speaking, capture batting style. We grouped these variables as:

$$\text{Scoring} = \{AVEe_{26}, BF, SR, ZERO, HUN, FIF\}$$

$$\text{Style} = \{POS, BOW, CAU, LBW\}$$

6. Methodology

We performed k-Means cluster analysis in two phases. In the first phase, the Training/Learning phase, we classified eighty TCG using their performance records and generated a Learning Model. In second, the Prediction phase, we applied Learning Model results and classified thirty two TCA into the same number of clusters. The ultimate goal was to statistically test for similarities/dis-similarities in performance. The following k-Means Learning model was defined and TCG data were used to generate cluster associations/memberships. In Prediction phase, the Learning Model centroids were used to predict cluster membership for the thirty two TCA.

k-Means Model

$i = 1, 2, 3 \ldots, 80$ representing TCG players in the Learning Model
$j = 1, 2, 3 \ldots, 32$ representing TCA players in the Prediction Model
k = number of clusters.

Learning Model

C = Cluster Centroids, a vector with cluster means of performance on ten features.
$C_A = \{AVEe_{26A}, BF_A, SR_A, ZERO_A, HUN_A, FIF_A, POS_A, BOW_A, CAU_A, LBW_A\}$
$C_B = \{AVEe_{26B}, BF_B, SR_B, ZERO_B, HUN_B, FIF_B, POS_B, BOW_B, CAU_B, LBW_B\}$
$C_C = \{AVEe_{26c}, BF_C, SR_C, ZERO_C, HUN_C, FIF_C, POS_C, BOW_C, CAU_C, LBW_C\}$

Where

DA_i = Euclidean distance of player i from C_A
DB_i = Euclidean distance of player i from C_B
DC_i = Euclidean distance of player i from C_C

Players are assigned to a cluster using:

$$\text{Cluster assignment of player}_i = \begin{cases} 1 & \text{if } DA_i = \min(DA_i, DB_i, DC_i) \\ 2 & \text{if } DB_i = \min(DA_i, DB_i, DC_i) \\ 3 & \text{if } DC_i = \min(DA_i, DB_i, DC_i) \end{cases}$$

Prediction Model

Active players' classification was generated from distances from Learning Model Centroids.

EA_j = Euclidean distance of player j from C_A
EB_j = Euclidean distance of player j from C_B
EC_j = Euclidean distance of player j from C_C

Players are assigned to a cluster using:

$$\text{Cluster assignment of player}_j = \begin{cases} 1 & \text{if } EA_j = \min(EA_j, EB_j, EC_j) \\ 2 & \text{if } EB_j = \min(EA_j, EB_j, EC_j) \\ 3 & \text{if } EC_j = \min(EA_j, EB_j, EC_j) \end{cases}$$

7. Analysis and results

We used SPSS for all statistical analyses: DA, ANOVA, independent sample *t*-tests, and k-Means cluster analysis and related test. We used Excel for data preparation and calculations for the Prediction phase.

Table 2 shows descriptive stats for TCG and TCA. Note the scale differences among variables. Before applying the k-Means clustering algorithm in SPSS, we normalized all variables to the standard (0, 1) distribution.

Table 2
Descriptive stats, TCG) and TCA players

	TCG Mean(St dev)	TCA Mean(St dev)
Scoring		
$AVEe_{26}$	44 (6.86)	39.5 (10.14)
BF	79.7 (16.46)	68.75 (20.16)
SR	47.56 (8.14)	49.7 (11.05)
ZERO	0.0701 (0.0212)	0.0807 (0.04)
HUN	0.1027 (0.0394)	0.0842 (0.0517)
FIF	0.1932 (0.0452)	0.1795 (0.0628)
Style		
POS	3.92 (1.8)	4.78 (2.27)
BOW	0.1635 (0.04)	0.1676 (0.05)
CAU	0.6743 (0.05)	0.6757 (0.06)
LBW	0.1622 (0.05)	0.1567 (0.04)

Table 3
Learning Model (TCG) Cluster Centroids[%]

	Cluster		
	Elite A	Elite B	Elite C
Scoring			
$AVEe_{26}$	33.43	44.79	47.92
BF	53.38	83.26	87.1
SR	53.55	45.72	47.4
ZERO	0.0900	0.0700	0.0600
HUN	0.0400	0.1100	0.1200
FIF	0.1600	0.1900	0.2200
Style			
POS	6.3	3.41	3.54
BOW	0.1700	0.1600	0.1500
CAU	0.7000	0.6800	0.6300
LBW	0.1200	0.1500	0.2000

[%]Centroids are shown using the original scales.

7.1. Learning model results

In clustering, discovering the most efficient number of clusters (k) is important to draw meaningful conclusions. We ran k-Means analysis for $k = 3, 4$, and 5. Using silhouettes statistics generated by SPSS we determined $k = 3$ as the more efficient classification. Given that these players (for both, TCG and TCA) are the best of the best and occupy highest ranks in the batting history of Test cricket, we labeled the three resulting clusters as Elite A, Elite B, and Elite C.

With "Cluster membership" option in SPSS, each player was associated with a cluster. For eighty TCG final cluster member counts are: Elite A (13), Elite B (40), and Elite C (27.) We present TCG cluster memberships in Appendix A, Table A1. Table 3 shows final cluster centroids using the original scales.

The SPSS reported inter-cluster distances are: between Elite A & Elite $B = 3.65$, A & $C = 4.83$, and B & $C = 1.97$. These distances imply that clusters Elite B and Elite C are "more similar" with each other and "more dissimilar" to Elite A. For the thirteen members of Elite A-TCG we checked the "playing role" listed by ESPNCRICinfo.com, which includes: three bowlers, five all-rounders, and five batsmen. In the conclusions section we discuss bowling contributions of Elite A-TCG. Members in the other two clusters are all batsmen. As shown in Table 3, AVEe26 for Elite A is the lowest and Elite C is slightly higher than Elite B.

We note that a handful of players received classifications that Test cricket fans may find to be anomalies. This was true for both, TCG and TCA. This is an appropriate place for us to take a pause from presentation of results and explain these cases. The main thrust of our analysis was to perform inter-group comparisons. We made no attempt to compare individual players. KC Sangakara (TCG), the highest scorer in the history of Sri Lankan Test cricket and KS Williamson (TCA), the second highest scoring active player for New Zealand were both associated with their respective Elite B. As mentioned previously, five TCG batsmen ended up as Elite A. We explain all these classifications through the algorithm.

The k-Means clustering algorithm classified each player into a cluster based on "similarity" between player's attributes pattern and cluster centroids, considering all competing clusters. Similarity is quantified as a high order distance measure which aggregates equally weighted deviations from cluster centroids, on all attributes. Therefore, larger deviations are a negative while smaller deviations are favorable. Relatively larger deviations in a handful of attributes, as few as one, can influence cluster assignment. Additionally, smaller (favorable) deviations in one attribute may be nullified or overcome by larger (unfavorable) deviations in another. Presence of larger deviations in one cluster comparison forces algorithm towards other competing clusters for potential membership. The algorithm does not discriminate between attributes contributing smaller or larger deviations.

For each specific player, we can identify attributes contributing larger deviations to the aggregated distance, thus justifying their classification. For Sangakara and Williamson, larger deviations in CAU and LBW explain why these two players were classified away from Elite C. Similar observations can explain why batsmen: WJ Cronje, A Ranatunga, MS Dhoni, BB McCullum, and TM Dilshan and all-rounder ST Jayasuriya were classified as Elite A-TCG. The underlying attributes contributing larger deviations are specific to, and different for each player. The algorithm determined that attribute patterns of these players bear greater resemblance to Elite A.

Continuing with the results, with cluster membership as the grouping variable, we performed one-way ANOVA to compare the three attributes patterns. Table 4 shows ANOVA results. For space reasons we do not report the customary Totals row which can be derived from information presented. With the exception of one variable (BOW) the three patterns differ which indicates dis-similarities in at least one of the three pairs of Elite A, Elite B, and Elite C.

Next, we performed three independent sample *t*-tests, pairwise comparing means of Elite A, Elite

Table 4
ANOVA results comparing Elite A, B and C - Learning Model (TCG)

	Between Groups		Within Groups			
	SS	MS	SS	MS	F	p
Scoring						
$AVEe_{26}$	1891.95	945.98	1824.43	23.69	39.92	0.0000#
BF	10995.11	5497.55	10414.79	135.26	40.65	0.0000#
SR	603.49	301.74	4634.31	60.19	5.01	0.0090#
ZERO	0.0089	0.0044	0.0266	0.0003	12.8424	0.0000#
HUN	0.0572	0.0286	0.0657	0.0009	33.5669	0.0000#
FIF	0.0315	0.0157	0.1299	0.0017	9.3307	0.0002#
Style						
POS	88.15	44.08	167.57	2.18	20.25	0.0000#
BOW	0.00	0.00	0.13	0.00	1.28	0.2830
CAU	0.0426	0.0213	0.1197	0.0016	13.7138	0.0000#
LBW	0.0695	0.0347	0.1121	0.0015	23.8744	0.0000#

df (Between Groups) = 2, df (Within Groups) = 77. #Significant at 1%.

Table 5
t- tests of means, comparing Learning Model Elite A, B and C (TCG)

	A	B	C	A vs B	A vs C	B vs C
		Mean / St Dev			t / p	
Scoring						
$AVEe_{26}$	33.43	44.79	47.92	–7.16	–8.84	–2.63
	5.24	4.88	4.67	0.0000#	0.0000#	0.0107&
BF	53.38	83.26	87.10	–8.57	–8.22	–1.30
	10.35	11.10	12.90	0.0000#	0.0000#	0.1981
SR	53.56	45.73	47.40	3.02	2.35	–0.90
	9.21	7.74	7.01	0.0039#	0.0242&	0.3701
ZERO	0.0853	0.0744	0.0563	1.84	4.87	3.79
	0.0150	0.0195	0.0187	0.0709	0.0000#	0.0003#
HUN	0.0430	0.1099	0.1208	–7.26	–7.79	–1.50
	0.0290	0.0288	0.0298	0.0000#	0.0000#	0.1397
FIF	0.1643	0.1851	0.2191	–1.59	–3.67	–3.47
	0.0500	0.0378	0.0412	0.1184	0.0007#	0.0009#
Style						
POS	6.30	3.41	3.54	5.85	5.58	–0.36
	1.74	1.48	1.32	0.0000#	0.0000#	0.7168
BOW	0.1779	0.1640	0.1560	1.06	1.63	0.79
	0.0394	0.0413	0.0399	0.2931	0.1112	0.4320
CAU	0.6989	0.6878	0.6424	0.80	4.35	5.00
	0.0527	0.0403	0.0298	0.4273	0.0001#	0.0000#
LBW	0.1233	0.1482	0.2016	–2.21	–5.25	–5.90
	0.0466	0.0311	0.0431	0.0316&	0.0000#	0.0000#

df (A vs B) = 51, df (A vs C) = 38, df (B vs C) = 65. #Significant at 1%. &Significant at 5%.

B, and Elite C. Table 5 presents those results. The *t*-tests show that Elite A-TCG vs Elite B-TCG performances are different on six attributes, Elite A-TCG vs Elite C-TCG differ on eight, and Elite B-TCG vs Elite C-TCG differ on five.

Elite A vs B & C (TCG)

On the scoring indicators, Elite A scored the fewest runs while Elite C scored more than Elite B. Elite A faced the fewest balls with no difference between Elite B and Elite C. Elite A batted with significantly higher strike rate than Elite B and Elite C. Test cricket fans will conclude that these results are in line with the way batsmen lower in the batting order bat. As indicated by POS, Elite A batted lower in the order, had more ZERO outs and scored fewer HUN as compared to Elite B and Elite C. For FIF, Elite A scored the same number as Elite B, but scored fewer than Elite C.

Elite B vs C (TCG)

Elite C scored more runs, recorded fewer ZERO outs, scored more FIF, were CAU less frequently and recorded more LBW, as compared to Elite B. There

Table 6
ANOVA results comparing Elite A, B and C - Prediction Model (TCA)

	Between Groups		Within Groups			
	SS	MS	SS	MS	F	*p*
Scoring						
$AVEe_{26}$	1926.82	963.41	1261.20	43.49	22.15	0.0000#
BF	8281.82	4140.91	4323.31	149.08	27.78	0.0000#
SR	320.17	160.08	3465.61	119.50	1.34	0.2777
ZERO	0.0082	0.0041	0.0415	0.0014	2.87	0.0730
HUN	0.0480	0.0240	0.0348	0.0012	20.01	0.0000#
FIF	0.0588	0.0294	0.0634	0.0022	13.44	0.0001#
Style						
POS	82.38	41.19	77.69	2.68	15.37	0.0000#
BOW	0.00	0.00	0.09	0.00	0.70	0.5032
CAU	0.0192	0.0096	0.1102	0.0038	2.53	0.0973
LBW	0.0101	0.0051	0.0348	0.0012	4.21	0.0248&

df (Between Groups) = 2, df (Within Groups) = 29. #Significant at 1%. &Significant at 5%.

is no difference in position in the batting order, strike rates and number of hundreds.

In summary, these results support the conclusion that performance patterns varied across the three clusters of TCG.

7.2. Prediction model results

In Prediction Model phase, TCA were assigned to a cluster based on the smallest distance to the three Learning Model centroids. We present TCA cluster memberships in Appendix A, Table A.2. Thirty two TCA players received classifications as: Elite A (6), Elite B (17), and Elite C (9) with "playing role" of six players in Elite A listed as: two bowlers, two all-rounder, one bowler all-rounder, and one batting all-rounder. All six Elite A have made significant contributions with the ball. In the conclusions section, we present their bowling records.

Using cluster as the grouping variable in SPSS, we performed one-way ANOVA to compare the three attributes patterns and Table 6 shows summarized results. Again, we omit reporting the Totals Row. The three patterns are dis-similar on four of the six scoring attributes, $AVEe_{26}$, BF, HUN and FIF. The patterns are similar on SR, ZERO, BOW and CAU. These results indicate that there are significant performance differences among players in Elite A-TCA, Elite B-TCA, and Elite C-TCA. These results are consistent with results for the TCG.

We performed independent sample *t*-tests of means to pairwise compare performance attributes of Elite A-TCA, Elite B-TCA, and Elite C-TCA. As shown in Table 7, the three performance patterns are different.

Elite A vs B & C (TCA)

On the scoring indicators, Elite A scored the fewest runs, while Elite C scored the most. Elite A faced the fewest balls, batted lower in the order, scored fewer fifties, and scored fewer hundreds as compared to Elite B and Elite C.

Elite A had indistinguishable difference in strike rate and scores of ZERO as compared to Elite B and Elite C. And, Elite A dismissals were similar to Elite B but different from Elite C in that Elite A were dismissed CAU less frequently, and out on LBW more often.

Elite B VS C (TCA)

Elite C scored more runs, scored more fifties, scored more hundreds, were out CAU less frequently and recorded more LBW, as compared to Elite B. There is no difference in outs with a ZERO, balls faced, strike rate, position in the batting order, and bowled out between Elite B and Elite C.

The research question posed in this paper was: How do Test cricketers of today compare with Test cricket's greats from the past? We parse this question into four analyses whose results collectively answer it. We looked deeper inside the clusters. We pose a sub-question and support the answer with results from our analysis.

Questions one: At the group level, how does performance of TCG compare with that of TCA?

We performed independent sample *t*-tests of means comparing performance attributes of TCG with TCA. Those results are presented in Table 8, which includes descriptive statistics and independent sample *t*-tests results.

This analysis shows that, as a group, TCG performance was significantly different from that of TCA

Table 7
t-tests of means, Prediction Model Elite A, B and C (TCA)

	A	B	C	A vs B	A vs C	B vs C
		Mean / St Dev			*t/p*	
Scoring						
$AVEe_{26}$	24.18	41.05	46.79	–5.83	–5.27	–2.30
	8.72	4.99	7.77	0.0000#	0.0002#	0.0306&
BF	36.21	73.20	82.04	–7.37	–5.71	–1.83
	14.43	9.03	15.73	0.0000#	0.0001#	0.0790
SR	56.26	47.98	48.56	1.55	1.10	–0.16
	17.30	8.56	9.98	0.1369	0.2914	0.8776
ZERO	0.1139	0.0744	0.0705	1.99	1.76	0.35
	0.0687	0.0286	0.0245	0.0601	0.1018	0.7316
HUN	0.0172	0.0822	0.1325	–5.01	–5.38	–3.32
	0.0219	0.0288	0.0488	0.0001#	0.0001#	0.0029#
FIF	0.0961	0.1865	0.2218	–4.04	–4.28	–2.11
	0.0689	0.0378	0.0458	0.0006#	0.0009#	0.0456&
Style						
POS	8.06	3.78	4.48	5.11	5.64	–1.00
	1.14	1.92	1.24	0.0000#	0.0001#	0.3294
BOW	0.1889	0.1586	0.1703	1.20	0.73	–0.48
	0.0264	0.0591	0.0578	0.2427	0.4769	0.6324
CAU	0.6568	0.6984	0.6454	–1.28	0.43	2.13
	0.0672	0.0691	0.0373	0.2159	0.6766	0.0436&
LBW	0.1543	0.1430	0.1843	0.66	–1.47	–3.23
	0.0482	0.0309	0.0315	0.5134	0.1652	0.0036#

df (A vs B) = 21, df (A vs C) = 13, df (B vs C) = 24. #Significant at 1%. &Significant at 5%.

Table 8
t-tests of means, all retired vs all active

	TCG Mean(St dev)	TCA Mean(St dev)	TCG vs TCA t (p)
Scoring			
$AVEe_{26}$	44.00 (6.86)	39.5 (10.14)	2.72 (0.0077#)
BF	79.70 (16.46)	68.75 (20.16)	2.98 (0.0036#)
SR	47.56 (8.14)	49.7 (11.05)	–1.12 (0.2631)
ZERO	0.0701 (0.0212)	0.0807 (0.04)	–1.83 (0.07)
HUN	0.1027 (0.0394)	0.0842 (0.0517)	2.05 (0.0429)
FIF	0.1932 (0.0452)	0.1795 (0.0628)	1.29 (0.1993)
Style			
POS	3.92 (1.8)	4.78 (2.27)	–2.1 (0.0383&)
BOW	0.1635 (0.04)	0.1676 (0.05)	–0.43 (0.669)
CAU	0.6743 (0.05)	0.6757 (0.06)	–0.13 (0.8948)
LBW	0.1622 (0.05)	0.1567 (0.04)	0.58 (0.5658)

df = 110. #Significant at 1%. &Significant at 5%.

Table 9
t-tests of means, Elite A-TCG vs Elite A-TCA

	Elite A-TCG Mean(St dev)	Elite A-TCA Mean(St dev)	Elite A-TCG vs Elite A-TCA t (p)
Scoring			
$AVEe_{26}$	33.43 (5.24)	24.18 (8.72)	2.9 (0.0099#)
BF	53.38 (10.35)	36.21 (14.43)	2.97 (0.0085#)
SR	53.56 (9.21)	56.26 (17.3)	–0.45 (0.6588)
ZERO	0.0853 (0.015)	0.1139 (0.0687)	–1.47 (0.1596)
HUN	0.043 (0.029)	0.0172 (0.0219)	1.93 (0.0704)
FIF	0.1643 (0.05)	0.0961 (0.0689)	2.46 (0.0248&)
Style			
POS	6.3 (1.74)	8.06 (1.14)	–2.24 (0.0391&)
BOW	0.1779 (0.04)	0.1889 (0.03)	–0.62 (0.5429)
CAU	0.6989 (0.05)	0.6568 (0.07)	1.49 (0.1553)
LBW	0.1233 (0.05)	0.1543 (0.05)	–1.33 (0.1998)

df = 17. #Significant at 1%. &Significant at 5%.

on some attributes, and similar on others. TCG scored significantly more runs, faced more balls and batted higher in the batting order. On all other metrics, the performances were similar.

Question two: Do Elite A-TCA perform as well as Elite A-TCG? Table 9 presents descriptive statistics and results of the independent sample *t*-tests of means.

These results show that Elite A-TCA players performed poorly as compared to Elite A-TCG, scored fewer runs (24.18 compared to 33.43), faced fewer balls (36.21 to 53.36), scored fewer fifties (0.0961 to 0.1653) and generally batted lower in the batting order (8.06 to 6.3). These results must be viewed from the algorithm perspective. As stated earlier, five of the TCG batsman are classified as Elite A. The algorithm does not consider playing roles and classifies cases solely on the basis of attributes patterns.

On other metrics, though differences were marginal and statistically not significant, TCA member have higher strike rate, are more likely to get out with no score and scored fewer hundreds. The unknown in

Table 10
t-tests of means, Elite B-TCG vs Elite B-TCA

	Elite B-TCG Mean(St dev)	Elite B-TCA Mean(St dev)	Elite B-TCG vs Elite B-TCA t (p)
Scoring			
$AVEe_{26}$	44.79 (4.88)	41.05 (4.99)	2.63 (0.0111[&])
BF	83.26 (11.1)	73.2 (9.03)	3.3 (0.0017[#])
SR	45.73 (7.74)	47.98 (8.56)	–0.97 (0.3341)
ZERO	0.0744 (0.0195)	0.0744 (0.0286)	0 (0.9975)
HUN	0.1099 (0.0288)	0.0822 (0.0288)	3.32 (0.0016[#])
FIF	0.1851 (0.0378)	0.1865 (0.0378)	–0.13 (0.9005)
Style			
POS	3.41 (1.48)	3.78 (1.92)	–0.78 (0.441)
BOW	0.164 (0.04)	0.1586 (0.06)	0.4 (0.693)
CAU	0.6878 (0.04)	0.6984 (0.07)	–0.73 (0.4691)
LBW	0.1482 (0.03)	0.143 (0.03)	0.58 (0.5618)

df = 55. [#]Significant at 1%. [&]Significant at 5%

Table 11
t-tests of means, Elite C-TCG vs Elite C-TCA

	Elite C-TCG Mean(St dev)	Elite C-TCA Mean(St dev)	Elite C-TCG vs Elite C-TCA t (p)
Scoring			
$AVEe_{26}$	47.92 (4.67)	46.79 (7.77)	0.53 (0.5984)
BF	87.1 (12.9)	82.04 (15.73)	0.96 (0.3414)
SR	47.4 (7.01)	48.56 (9.98)	–0.39 (0.7022)
ZERO	0.0563 (0.0187)	0.0705 (0.0245)	–1.82 (0.077)
HUN	0.1208 (0.0299)	0.1325 (0.0488)	–0.86 (0.3938)
FIF	0.2191 (0.0412)	0.2218 (0.0458)	–0.17 (0.8651)
Style			
POS	3.54 (1.32)	4.48 (1.24)	–1.88 (0.0685)
BOW	0.156 (0.04)	0.1703 (0.06)	–0.83 (0.4117)
CAU	0.6424 (0.03)	0.6454 (0.04)	–0.24 (0.8093)
LBW	0.2016 (0.04)	0.1843 (0.03)	1.1 (0.277)

df = 34.

these differences may be the impact of five batsmen in TCG.

This analysis naturally leads to the next two questions, how do the other two groups compare? We answer those questions next.

Question three: Do Elite B-TCA perform as well as Elite B-TCG? The following Table 10 presents descriptive statistics and independent sample *t*-tests of means results.

The results show that Elite B-TCA did not perform as well on three scoring metrics: scored fewer runs (41.05 compared to 44.79), faced fewer balls (73.2 to 83.26) and scored fewer hundreds (0.0822 to 0.1099). Though not significant, Elite B-TCA have slightly higher strike rate and batted slightly lower in the batting order.

Question four: Do Elite C- TCA perform as well as Elite C-TCG? We present descriptive stats and results of *t*-tests comparing means of the two groups, in Table 11.

Results show that performance of Elite C-TCA players was not different from that of Elite C-TCG. On the margins, though insignificant, Elite C-TCA scored slightly fewer runs, faced fewer balls, more likely to get out without scoring and batted slightly lower in the batting order.

8. Conclusions

There are significant similarities and significant differences in the on-field performance of the TCA and the TCG. Our analysis showed that, with the exception of two anomalies, performance patterns evaluated by the clustering algorithm resulted in groupings which are consistent with impressions of Test cricket fans.

Previously, we have listed "playing role" of members of Elite-A, for both, the TCG and the TCA. To underscore that presentation, we retrieved career bowling records of these players. The six players in Elite A–TCA (3 bowlers, 3 all-rounders), on average bowled 13,726 balls in their careers. For reference, the Elite B-TCA and Elite C-TCA averaged 1,863 and 611 balls respectively. Similar figures for the TCG are: Elite A (11,999), Elite B (1,929) and Elite C (1,770). The bowlers obviously, and the all-rounders are making significant contributions with the ball. These bowlers have achieved sufficient success with the bat also, which earns them a spot in the table TB. The all-rounders, while contributing from both sides of the ball, bat lower in the batting order and contribute fewer runs, as compared to higher order batsmen.

That leaves the five batsmen in Elite A-TCG. These batsmen have attribute patterns which deviate significantly from those of Elite B-TCG and Elite C-TCG. As discussed earlier, the cluster classifications are a net effect of the push and pull between favorable and unfavorable deviations from the centroids. In other words, attribute patterns of these batsmen deviate sufficiently from those of Elite B-TCG and Elite C-TCG. As cricket fans, we are comfortable with the classification assigned to them by the k-Means algorithm.

Among the batting specialists, there were two distinct groupings. These two groupings were similar in performance patterns for the TCA, as well as the TCG. Before concluding that batsmen in one "batting specialists" cluster are superior performers than batsmen in the other, we obtained independent confirmation by performing analysis on official player rankings published by the ICC.

We retrieved current "MRF Tyres ICC Player Ratings" for the Test format (ICC test Match Player Rankings). Previously, and from 2008 to 2016, these rankings were known to cricket fans as "Reliance Player Rankings" named for its sponsor (reliance ICC Player Rankings). For all active players, these ratings are updated regularly and show their current rating along and the career highest rating achieved to-date. For TCG, the rating shown is the highest rating achieved over the entire playing career. For similarity in comparison, we compared the to-date highest career ratings of the TCA with highest career ratings of the TCG.

For each of the three clusters, we calculated average of Reliance/MRF highest ratings. For Elite A-TCG, the average rating was 645.77, and 458.17 for Elite A-TCA. The Elite B-TCG averaged 794.20 as compared to 755.05 for Elite B-TCA while Elite C-TCG came in at 854.30 vs 844.78 for Elite C-TCA. Rather than using these figures for their absolute magnitude, we used them for their relative orders. Conclusions drawn from these averages support our analysis. Elite A, with bowling specialists as members, achieved the lowest average of highest career ratings. Among the established batsmen, Elite C achieved higher averaged career ratings peaks as compared to Elite B.

Comparing Elite C-TCG with Elite B-TCG, players in the Elite C faced more balls which demonstrates patience on the crease. Since time is a less important factor in the Test format, players do not have the urgency to score on every ball. Batsmen can be patient in their shot selection. This patience may have some role in achieving lower number of ZERO outs. Elite C also scored more fifties but did not score significantly more centuries. Elite C were caught out less frequently and were out LBW more often. We will not venture a guess on the impact of DRS (umpire Decision Review System) on this metric. A similar contrast is observed between Elite C-TCA and Elite B-TCA, with one exception that Elite C scored significantly more hundreds.

Arguably (and reluctantly) we conclude that Elite C are better batsmen than Elite B. After all, these are the best of the best batsmen in Test cricket. This conclusions holds for both, the TCG and the TCA.

One potential drawback of the present study is that we compared completed career records of TCG with partially completed career records of the TCA. On average, the TCG played in 179 innings over their entire careers as compared to 117 for the TCA. Even though TCA players have opportunities to improve their performance, it is always an uphill task. It is a mathematical reality that with every additional superior performance, improvements diminish in magnitude. To improve career averages, a player has to achieve superior results (than the past) and the resulting improvements have to carry the weight of entire history. We provide anecdotal evidence, in support.

Selecting one highest ranked TCG player from each of the eight countries, and based on career number of innings played, we calculated their averages at three specific points in their respective careers: at the one third point, the two-thirds point and during the last one-third of the innings played. In presenting these results, we understand this analysis lacks rigor and leave conclusions for the reader to draw. Eight player averages at the one third and two-thirds point are: $AVEe_{26}$ (47.46 vs 51.32), BF (85.47, 90.77), SR (48.27, 48.29), FIF (0.2207, 0.2156), HUN (0.1106, 0.1315), POS (3.62, 3.53), BOW (0.1254, 0.1491), CAU (0.6832, 0.6715), and LBW (0.1914, 0.1794). Eight player averages during the last one third of career innings are: $AVEe_{26}$ (51.68), BF (89.90), SR (47.77), FIF (0.1836), HUN (0.1499), POS (3.44), BOW (0.1668), CAU (0.6645), and LBW (0.1687). Comparing performance at the two-thirds point with performance during the last one third of career innings, deviations are small and most likely statistically insignificant. Some players improved performance in some metrics while a similar number saw declines. Singling out one player, Babar Azam, based on the fact that he is early in his Test career and has already earned a spot in Elite B-TCA, we project that if he continues on this performance trajectory, he will go on to earn a spot in Elite C-TCA. All other Elite B-TCA players are further along in their careers and coupled with the fact that any delta from future superior performance will have to lift the entire history, makes the task that much difficult.

The Elite C-TCA players have already achieved performance levels at par with the sports' greatest. Performance gaps between Elite C and Elite B are relatively small for TCG. With superior future performance and despite it being an uphill task, players in upper echelons of Elite B-TCA can move up and earn a spot in Elite C-TCG. Conversely, poor future performance of players in the lower echelons of Elite B–TCA poses smaller risk of slipping into Elite A-TCG as those performance gaps are much wider.

References

Akhtar, S., Scarf, P. and Rasool, Z. 2015, 'Rating players in test match cricket', *Journal of Operations Research Society*, *66*, p684–695.

Ahmad, F. and Zada, M. 2015. 'Best player of the Test cricket in last years 2014, Thesis, 2015, International Islamic University, Islamabad.

Bailey, M. J. and Clarke, S. R. 2004. 'Market inefficiencies in player head to head batting in 2003 cricket world cup', in P. Pardalos and S. Butenko (edited), Economics, Management and Optimization in sports, New York: Springer.

Basevi, T. and Binoy, G. 2007. 'The world's best Twenty20 players, available at www.espncricinfo.com/story/_/id/22912282/the-world-best-twenty20-players, Accessed Feb 29, 2020.

Boorah, V. K. and Mangan, J. E. 2010. 'The "Bradman Class": An exploration of some issues in the evaluation of batsmen for test matches, 1877–2006', *Journal of Quantitative Analysis in sports*, The Berkley Electronic Press, *6*(3).

Bracewell, P. J. and Ruggiero, K. 2009. 'A parametric control chart for monitoring individual batting performance in cricket', *Journal of Quantitative Analysis in Sports*, *5*, 2016.

Brewer, B. J. 2008. 'A Bayesian analysis of early dismissals in cricket', arXiv preprint: 0801.4408v2.

Clarke S. R. 1988. 'Dynamic programming in one-day cricket-optimal scoring rates', *Journal of the Operational Research Society*, *39*, p331–337.

Clarke S. R., Norman J. M. 1999. 'To run or not?: Some dynamic programming models in cricket', *Journal of the Operational Research Society*, *50*, p536–545.

Danaher, P. J. 1989. 'Estimating a cricketer's batting average using the product limit estimator', *New Zealand Statistician*, *24*(1), p2–5.

Damodran, U. 2006. 'Stochastic dominance and analysis of ODI batting performance: the Indian cricket team, 1989–2005', *Journal of Sports Science and medicine*, *5*, p503–508.

Elderton, W. 1945. 'Cricket scores and some skew Correlation distributions', *J. R. Statistics Soc*., *108*, p1–11.

Ganesalingam, S., Ganeshananda, S. and Kumar, K. 1994. 'A statistical look as cricket', *Mathematics computing in Sport,* Queensland Australia.

Jain, K., 2015. 'Importance of actionable insights in analytics – with case from ICC Cricket world Cup', available at www.analyticsvidhya.com/blog/2015/03/actionable-insights-analytics-cricket-world-cup/, Accessed Feb 29, 2020.

Joshi, P. 2020. 'Who is the best IPL batsman to bat with? Finding the answer with Network analysis?', available at www.analyticsvidhya.com/blog/2020/02/network-analysis-ipl-data/, Accessed Feb 29, 2020.

Kalman S. and Bosch, J. 2012. 'NBA lineup analysis on clustered player tendencies: A new approach to positions of basketball & modeling lineup efficiency of soft lineup aggregates', ID: 1548738. MIT Sloan Sports Analytics Conference, Boston, MA, USA.

Kaufman, L. and Rousseeuw, P. J. 1990. 'Finding groups in data: An introduction to cluster analysis' New York: John Wiley & sons.

Kimber A.C., Hansford A.R. 1993. 'A statistical analysis of batting in cricket', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *156*, p443–455.

Lemmer, H. H. 2011. 'The single match approach to strike rate adjustments in batting performance measures in cricket', *Journal of sports science & medicine*, *10*(4), p630–634.

Lemmer, H. H. 2008. 'An analysis of players' performance in the first Twenty20 World Cup Series', *South African Journal for Research in Sport, Physical Education and Recreation*, *30*, p73–79.

Lewis A. J. 2005. 'Towards fairer measures of player performance in one-day cricket', *Journal of the Operational Research Society*, *56*, p804–815.

Maini, S. and Narayanan, S. 2007. 'The flaw in batting averages', *The Actuary*, May-07, p30–31.

Mukherjee, S. 2013. 'Complex network Analysis in cricket: Community Structure, Player's profile and performance Index', *Advances in Complex Systems*, 1350031, https://doi.org/10.1142/S0219525913500318.

Narayanan, A. 2000. 'Weighted batting average', available at www.espncricinfo.com/story/_/id/28974421/the-weighted-batting-average-(wba)-odis, Accessed Feb 29, 2020.

Norman, J. M. and Clarke, S. R. 2010. 'Optimal batting order in cricket', *Journal of the Operational Research Society*, *61*, 980–986.

NSS 2016. 'Who is the superhero in the cricket battlefield? An in-depth analysis', Available at www.analyticsvidhya.com/blog/2016/12/who-is-the-superhero-of-cricket-battle field-an-in-depth-analysis/, Accessed Feb 29, 2020.

Pai, A. 2020. 'Sports Analytics – Generating Actionable Insights using Cricket Commentary', available at www.analyticsvidhya.com/blog/2020/02/sports-analytics-generating-actionable-insights-using-cricket-commentary/?utm_source=feedburner&utm_medium=email&utm_campaign=Feed%3A+AnalyticsVidhya+%28Analytics+Vidhya%29, Accessed Feb 29, 2020.

Praveen, P. and Rama, B. 2017. 'A k-Means Clustering Algorithm on Numeric Data', *International Journal of Pure and Applied Mathematics*, *117*(7), p157–164.

Preston I., Thomas J. 2000. 'Batting strategy in limited overs cricket', *Journal of the Royal Statistical Society: Series D (The Statistician)*, *49*, p95–106.

Reliance ICC Player Rankings, available at www.relianceicc rankings.com/about.php, Accessed Feb 29, 2020.

Stevenson, O. G. and Brewer, B. J. 2017. 'Bayesian survival analysis if batsmen in test cricket', *journal of quantitative analysis in sports* arXiv preprint:1609.04078.

Swartz, T.B., Gill, P.S., Beaudoin D. and de Silva, B.M. 2006. 'Optimal batting orders in one-day cricket', *Computers and Operations Research, 33*, p1939–1950.

Swartz, T.B. 2016. 'Research directions in cricket', *Handbook of Statistical Methods and Analysis in Sports,* editors J.H. Albert, M.E. Glickman, T.B. Swartz and R.H. Koning, Chapman & Hall/CRC Handbooks of Modern Statistical Methods: Boca Raton, FL.

The International Cricket Council, available at www.icc-cricket.com/, Accessed Feb 29, 2020.

van Staden, P.J., Meiring, A.T., Steyn, J.A. and Fabris-Rotelli, I.N. 2010. 'Meaningful batting averages in cricket', Proceedings of the 52nd *Annual Conference of the South African Statistical Association* (SASA 2010), North-West University, Potchefstroom, South Africa, p75–82.

Wickramasinghe, I. P. 2019. 'Predicting the performance of batsmen in test cricket', *Journal of Human Sport and Exercise*, *9*(4), p744–751.

Wood, G. H. 1945. 'Cricket Scores and geometric progression', *J. R. Statistics Soc*., *108*, p12–22.

Appendix A

Table A1 serves two purposes. First, it identifies TCG (retired players) used for the Learning Model. Second, it presents classification results from the Learning Model.

Table A2 serves two purposes. First, it identifies TCA (active players) used for Prediction Model. Second, it presents classification results from the Prediction phase.

Table-A1
Cluster memberships of retired players - TCG

Elite A			
MV Boucher	SM Pollock	WJ Cronje	BB McCullum
DL Vettori	CL Cairns	RJ Hadlee	ST Jayasuriya
TM Dilshan	A Ranatunga	WPUJC Vaas	N Kapil Dev
MS Dhoni			
Elite B			
G Kirsten	HH Gibbs	DJ Cullinan	SP Fleming
MD Crowe	JG Wright	NJ Astle	DL Haynes
CH Gayle	CL Hooper	Mohammad Yosuf	Saleem Malik
Zaheer Abbas	Mudassar Nazar	Saeed Anwar	Mohammad Hafeez
MA Atherton	IR Bell	AJ Strauss	GP Thorpe
AR Border	SR Waugh	MJ Clarke	ML Hayden
ME Waugh	JL Langer	DC Boon	GS chappell
KC Sangakara	DPMD Jayawardane	PA De Silva	MS Atapattu
TT Samaraweera	HP Tilakaratne	SM Gavaskar	VVS Laxman
V Sehwag	SC Ganguly	DB Vengsarkar	M Azharuddin
Elite C			
JH Kallis	HM Amla	GC Smith	AB deVilliers
CD McMillan	AH Jones	BC Lara	S Chandrapaul
IVA Richards	CG Greenidge	CH Lloyd	RB Richardson
RR Sarwan	Younis Khan	Javed Miadad	Inzamam ul-Haq
Misbah-ul-haq	AN Cook	GA Gooch	AJ Stewart
DI Gower	KP Pietersen	G Boycott	RT Ponting
MA Taylor	SR Tendulkar	R Dravid	

Table-A2
Cluster memberships of active players - TCA

Elite A			
VD Philander	TG Southee	MM Ali	RAS Lakmal
R Ashwin	RA Jadega		
Elite B			
F Du Plessis	D Elgar	KS Williamson	TWM Latham
DM Bravo	KC Brathwaite	Babar Azam	BA Stokes
JM Bairstow	DA Warner	AD Mathews	FDM Karunaratne
LD Chandimal	AM Rahane	T Iqbal	Mushfiqur Rahim
Shakib al Hasan			
Elite C			
Q dE Kock	LRPL Taylor	BJ Watling	Azhar Ali
Asad Shafiq	JE Root	SPD Smith	V Kohli
CA Pujara			