Was Bradman Denied His Prime?

Paul J. Bracewell* Farinaz Farhadieh† Clint A. Jowett‡
Don G. R. Forbes** Denny H. Meyer††

*Offlode, paul@offlode.com
†Offlode and Swinburne University of Technology, farinaz@offlode.com
‡Offlode, clint@offlode.com
**Offlode, don@offlode.com
††Swinburne University of Technology, dmeyer@swin.edu.au

Copyright ©2009 The Berkeley Electronic Press. All rights reserved.
Was Bradman Denied His Prime?

Paul J. Bracewell, Farinaz Farhadieh, Clint A. Jowett, Don G. R. Forbes, and Denny H. Meyer

Abstract

Time series clustering is used to show that, relatively, the career progression of Australian legend Sir Donald Bradman’s test career as a batsman was most similar to West Indian Brian Lara. Consequently, it is likely his peak performance would have occurred while the Second World War disrupted all international cricket.

Data from the 20 international cricketers who played in at least 70 innings over more than 17 years and averaged more than 40 runs per dismissal (as at January 1, 2009) is used to create a number of global measures that indicate the ebb and flow of a career. As is shown in this paper, this clustering methodology, proposed by Wang et al. (2006), generates instinctive clustering results and can be applied on different length time series.

Utilizing the framework created for clustering, Bradman’s batting average is estimated to be 105 if his career had been uninterrupted.

KEYWORDS: time series clustering

*Aspects of this paper are part of a larger piece of work, A Statistical Framework for Quantifying Adaptive Behavioural Risk for the Banking Industry, undertaken by Farinaz Farhadieh for the PhD Degree at Swinburne University of Technology, Melbourne. This research is supervised by Dr. Denny Meyer and Dr. Paul Bracewell, with funding provided by Offlode.*
1. INTRODUCTION

Sir Donald Bradman is an undisputed cricketing legend. Needing four runs in his last test innings for Australia to average at least 100 runs, he was dismissed without scoring. This left Bradman with an unparalleled career batting average of 99.94.

Acknowledged as the greatest batsman of all time, Bradman was 30 when international cricket was suspended due to the outbreak of World War II. After hostilities ceased he returned to the test arena in 1946 and retired, aged 39, after the Australian tour to England in 1948.

Did the Second World War rob the public of a chance to see Bradman in his prime? We use time series clustering to characterize Bradman’s career and compare him with other great players. The similarity of his career progression with other players is used to validate the imputation of performances between 1939 and 1945.

Pragmatically, it could be argued that these results are of little value as potentially there are many unforeseen circumstances which could have affected Bradman’s ability to perform, such as the fibrositis which led to him being invalided out of the Australian Army in 1941. Furthermore, many others lost a great deal more than just the chance to bat in their prime during the war years.

The methodology described in this paper has wider applications, such as comparing players for scouting and selection purposes in a variety of sports. We also use this approach to define changes in customer banking behavior for the purposes of adjusting estimates of credit risk.

Time series clustering is an active area of research, especially with regard to financial and medical applications (Pattarin, Paterlini & Minerva, 2004; Basalto & De Carlo, 2006; Savvides, Promponas & Fokianos, 2008). From a sporting perspective, time series clustering has been used to automatically compile video and audio highlight packages by gauging audience reactions (Radhakrishnan, Divakaran & Xiong, 2004; Radhakrishnan, Otsuka, Xiong & Divakaran, 2005).

In this paper, we review time series clustering before exploring suitable data for analysis. We then perform clustering on the 20 international cricketers who have played in at least 70 innings over more than 17 years (as at January 1st, 2009). A discussion of these clusters follows to provide the historical context and validate the clusters. Finally, we use these results to estimate what Sir Donald Bradman may have scored during 1939-1945 and recalculate his career batting average.
2. OVERVIEW OF TIME SERIES CLUSTERING

Clustering time series has been a successful area of research in a wide range of fields. Clustering is an unsupervised learning task dividing the population into homogenous groups based on their behavior. There have been many studies on different clustering algorithms. The choice of an appropriate clustering algorithm is dependent upon the characteristics of the data and the purpose of clustering (Liao, 2005). Keogh, Lin and Truppel (2003) classified time series clustering broadly into two categories:

1. **Whole clustering:** in this approach similar time series from a set of individual time series are grouped into one cluster.
2. **Subsequence clustering:** sliding windows are extracted from a single time series and these windows are compared to find similarities and difference.

Two of the popular algorithms used for clustering and finding similarities between data points in raw time series data are: Hierarchical clustering and K-mean clustering based on Euclidean Distance.

**K means clustering** is the most commonly used clustering algorithm. The disadvantage of this algorithm is that the number of clusters must be specified by the user. Also this method requires the length of each time series to be equal due to Euclidean distance calculation (Wang, Smith & Hyndman, 2006).

**Hierarchical clustering** produces a nested hierarchy of similar groups of objects according to a pairwise distance matrix of the objects. Hierarchical clustering is widely used since it does not require the user to specify any parameter such as number of clusters and it does not make any assumptions about data distribution. Though this method has great visualization power (Keogh et al, 2003; Keogh & Kasetty, 2002; Mantegna, 1999), its application is limited to small data sets due to complex quadratic calculations (Keogh et al, 2003). This method also requires each time series to be of the same length (Wang et al, 2006).

**Euclidean distance** is the most frequently used metric (Agrawal, Kumar, Agrawal & Yadev, 1993). Empirical comparison by Keogh and Kasetty (2002) has shown that Euclidean distance performs better than other distance measures when applied on the same datasets. However this method requires the time series to be of the same dimension.

In a survey study, Liao categorized the previous methods on clustering of time series into three major categories according to the data used for clustering (Liao, 2005):

1. **Raw data based approach:** clustering methods that are reformed on the raw data. (Golay, Kollias, Stoll, Meier, Valavanis & Boesiger, 1998; Liao, Bolt, Forester, Hailman, Hansen, Kaste et al, 2002; Kakizawa, Shumway & Taniguchi, 1998)
2. Model-based approach: in this approach time series are described by a model and clustering is based on model parameters. (AR coefficient model by Maharaj, 2000; Markov chain model by Ramoni, Sebastiani & Cohen, 2000; ARIMA mixture model by Xiong & Yeung, 2002).

3. Feature-based approach: a set of features are extracted from the raw data and these features are used for clustering (perceptually important points by Fu, Chung, Ng & Luk, 2001; cross-correlation function by Goutte, Toft, Rostrup, Nielsen & Hansen, 1999; time-frequency representation of the transient region by Owsley, Atlas & Bernard, 1997)

The approach proposed by Wang and others aims at providing a method of clustering time series which is robust to missing data. It follows the feature extraction principle which measures structure level similarity for the time series data relying upon global features or model parameter extraction. There are methods to extract statistical features such as Discrete Wavelet Transform (DWT), Discrete Fourier Transform (DFT), Auto Regression Moving Average (ARMA) and Compression based Dissimilarity Measures (CDM) but they all have a high computational complexity and require certain conditions to be satisfied for the method to be successful.

Feature extraction can also be considered as a means to obtain dimension reduction in time series. Using global measures such as trend, seasonality, periodicity, serial correlation, skewness and kurtosis along with more advanced measures viz. non-linearity structures, self–similarity and chaos, a useful set of time series characteristic features can be extracted as measures.

Extracting the summarized characteristics of the time series provides a meaningful reduction in dimension. This can be treated statistically so that long length datasets or different length time series can be reduced to a limited number of measures that are less sensitive to noise. These extracted features contain within themselves the distinct characteristic of the parent time series and hence, when applied to a clustering algorithm as a finite set of inputs, they can distinguish the similarity and the differences between the original time series (Wang et al, 2006). Thus the main aim of the feature extraction is to produce a set of input measures containing the original characteristics of the parent time series which can be used in any of the clustering techniques (Wang et al, 2006). The typical global characteristic measures are summarized next.

Trend and Seasonality refer to the general behavior of a series. A trend is said to exist when there is long term change in the mean level and can be estimated by smooth non parametric methods, such as penalized regression spline, whereas seasonality of time series is defined as a pattern that repeats itself over fixed intervals of time. This can be diagnosed with large autocorrelation or large partial autocorrelation coefficient at the seasonal lag.
Periodicity is a fundamental characteristic of the time series data which gives us an idea about how frequently the observations are spaced in time. Sometimes seasonal time series are also called periodic series; although there is a major distinction between the two since periodicity varies in frequency whereas seasonality has periodicity with a fixed frequency.

Serial Correlation describes the correlation between the successive values in the time series. The Box-Pierce statistic can be used to evaluate the degree of serial correlation.

Non-Linear Autoregressive Structure: The non parametric kernel test and the neural network test can be used to check for the presence of non-linearity in time series regression models. Comparative studies have shown that the neural network test yields better results than the non parametric kernel test (Lee, 2001).

Skewness is a measure to check the symmetry of the data. For a normal distribution, skewness is zero. Negative values for skewness tell us that the left tail is heavier than the right tail and the positive values tells us that right tail is heavier than the left tail.

Kurtosis measures the peakedness of a distribution. A high value of kurtosis tells us that the distribution has a distinct peak near the mean, declines sharply and has a heavy tail whereas a low value of kurtosis suggests a flat top near the mean.

Self-Similarity (Long Range Dependence) can be described simply as a phenomenon that looks the same or behaves the same when viewed at different degrees of magnification or different scales on a dimension. The dimension can be space (length, width) or time. The Hurst exponent can be used to get an idea of the importance of long range dependence.

This method of clustering time series based on global characteristic measures is better than other clustering methods in the sense that it does not require many conditions to be true before it can be used. Unlike other alternative methods this approach does not cluster the time series based on a distance measure, rather it clusters global features extracted from individual time series, and therefore it can be applied on different length time series. These features are then fed into any relevant clustering algorithm.

In this study, polynomial functions are used to approximate the time series for each individual using the scaled average contribution per calendar year. The use of polynomial functions enables weighted regression methods to be used allowing more emphasis for seasons where the individual had more innings. The estimated parameters of these models have been used as clustering measures as well as serial correlation at lag one, skewness and kurtosis. The method chosen for our analysis is to initially model the data for each cricketer and to then applying cluster time series analysis based on the global time series measures proposed by Wang and others and the parameters of the fitted model. Clustering based on these measures leads to more instinctive clustering results (Wang et al, 2006).
3. DESCRIBING BATTING PERFORMANCE

Comparisons are often made between current well-performed players and the game’s greats, regardless of the sport. In this paper, we will compare Sir Donald Bradman’s test career with all other test players who have batted in at least 70 innings during a career spanning at least 17 years as of 1st January 2009. Of the players assessed, only the Indian batsman Sachin Tendulkar is currently playing.

First, we need to define a suitable measure of performance for comparative purposes. Numerous papers describe batting performance (Bailey & Clarke, 2004; Bracewell & Ruggiero, 2009; Elderton, 1909, 1945; Ganesalingham, Ganeshanandam & Kumar, 1994; Kimber & Hansford, 1993; Pollard, 1977; Reep, Benjamin & Pollard, 1971; Wood 1945). The typical measure used for assessing batting performance is the average, $A_i$. This is defined for the $i$th player as follows:

$$ A_i = \frac{\sum_{j=1}^{n} R_{i,j}}{n-k}, \quad (1) $$

where $R_{i,j}$ is the number of runs scored by the $i$th player in the $j$th innings; $n$ is the total number of innings in which the $i$th player has batted and $k$ is the number of innings in which the $i$th player has remained ‘not out’. However, as we are attempting to compare batsmen across different eras, the batting average is vulnerable to confounding environmental influences, such as ground and pitch conditions and the impact uncovered wickets. That said, imputation is used to estimate the batting average of Bradman in the form outlined above to place the results in context.

We use contribution to reduce the impact of extraneous influences. Contribution is defined as the proportion of the team total that an individual has contributed in an innings, i.e.:

$$ C_{i,j} = \frac{S_{i,j}}{S_{T,j}}, \quad (2) $$

where $S_{i,j}$ is the total number of runs obtained by the $i$th batsman in the $j$th innings ($i = 1, 2, \ldots, 11, j = 1, 2$) and $S_{T,j}$ is team total (i.e. the total number of runs amassed by that team whilst batting in the $j$th innings). This means that our measure for batting performance, $P_i$, becomes:

$$ P_i = \frac{\sum_{j=1}^{n} C_{i,j}}{n}. \quad (3) $$
Contribution is essentially used as a method for smoothing performances. Although contribution enables an individual’s statistics to be influenced by team members, this is a fundamental aspect of cricket. As a team sport the individuals must work together in attempting to win matches. More importantly, we are looking for periods of influence from the various batsmen. Thus it is useful to compare directly with their team mates, which is what contribution does. Bracewell and Ruggiero (2009) outlined several reasons why contribution is a suitable measure.

Average contribution is considered with respect to innings rather than dismissals. Whilst Kimber and Hansford (1993) are critical of this approach with regard to the batting average, as used by Elderton (1945), Wood (1995) and Bracewell and Ruggiero (2009), cricket is a team game and individual performances are influenced by a number of extraneous factors. Bracewell and Ruggiero (2009) outlined several arguments as to why “not out” is essentially the end of an innings (e.g. either run out of batting partners, reached the opposition total, the captain declared the innings closed or bad weather ended play).

To our knowledge, there has been no attempt at imputation of batting scores or clustering of the careers of cricketers. We use data obtained from Cricinfo (www.cricinfo.com).

4. Model Fitting and Summarizing Global Characteristics

To approximate the ebb and flow of a cricketer’s batting performance throughout their career, weighted least squares regression is used to model the scaled average contribution per calendar year, $P$, for all test cricketers whose careers spanned at least 17 years, played more than 70 innings and averaged more than 40 runs per dismissal (20 players). Average contribution is scaled by the range to give each individual a minimum of 0 and a maximum of 1. Four polynomial terms are used: $T^{-2}$, $T^{-1}$, $T$ and $T^2$, where $T$ represents the year of the batsman’s career. The weight used is the number of innings played in any years and no intercept is included in the model. From this model, the predicted scaled average contribution is used to generate a range of other global characteristics. These are skewness, kurtosis and autocorrelation (Auto r). No weight is used as that is already used in estimating the average scaled contribution values. For each batsman, the coefficients for each polynomial term and the other global characteristic values that were used are shown in the table below. Values for these statistics are shown in Table 1, together with a cluster number which is derived as indicated below.
<table>
<thead>
<tr>
<th>Name</th>
<th>$T^{-2}$</th>
<th>$T^{-1}$</th>
<th>$T$</th>
<th>$T^2$</th>
<th>Auto r</th>
<th>Skew</th>
<th>Kurt</th>
<th>P-Value</th>
<th>R-Sq</th>
<th>Clust#</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR Border</td>
<td>-1.71</td>
<td>1.78</td>
<td>0.13</td>
<td>-0.01</td>
<td>0.89</td>
<td>-1.09</td>
<td>0</td>
<td>&lt;.0001</td>
<td>0.94</td>
<td>1</td>
</tr>
<tr>
<td>IVA Richards</td>
<td>-1.29</td>
<td>1.42</td>
<td>0.05</td>
<td>0</td>
<td>0.71</td>
<td>-0.66</td>
<td>-0.63</td>
<td>&lt;.0001</td>
<td>0.68</td>
<td>1</td>
</tr>
<tr>
<td>RB Kanhai</td>
<td>-1.87</td>
<td>1.85</td>
<td>0.07</td>
<td>0</td>
<td>0.75</td>
<td>-0.77</td>
<td>-0.74</td>
<td>&lt;.0001</td>
<td>0.86</td>
<td>1</td>
</tr>
<tr>
<td>G Boycott</td>
<td>-0.52</td>
<td>0.84</td>
<td>0.06</td>
<td>0</td>
<td>0.92</td>
<td>-0.12</td>
<td>-1.6</td>
<td>&lt;.0001</td>
<td>0.85</td>
<td>2</td>
</tr>
<tr>
<td>RB Simpson</td>
<td>0.82</td>
<td>-0.68</td>
<td>0.15</td>
<td>0.01</td>
<td>0.94</td>
<td>-0.76</td>
<td>-0.55</td>
<td>&lt;.0001</td>
<td>0.91</td>
<td>2</td>
</tr>
<tr>
<td>SR Tendulkar</td>
<td>-0.07</td>
<td>0.46</td>
<td>0.13</td>
<td>0.01</td>
<td>0.92</td>
<td>-0.44</td>
<td>-1.09</td>
<td>&lt;.0001</td>
<td>0.93</td>
<td>2</td>
</tr>
<tr>
<td>B Mitchell</td>
<td>0.02</td>
<td>-0.18</td>
<td>0.16</td>
<td>-0.01</td>
<td>0.98</td>
<td>-1.4</td>
<td>1.42</td>
<td>&lt;.0001</td>
<td>0.95</td>
<td>3</td>
</tr>
<tr>
<td>GS Sobers</td>
<td>-0.96</td>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>0.8</td>
<td>-1.32</td>
<td>1.21</td>
<td>&lt;.0001</td>
<td>0.86</td>
<td>3</td>
</tr>
<tr>
<td>JB Hobbs</td>
<td>-0.58</td>
<td>0.66</td>
<td>0.07</td>
<td>0</td>
<td>0.94</td>
<td>-1.21</td>
<td>0.53</td>
<td>&lt;.0001</td>
<td>0.69</td>
<td>3</td>
</tr>
<tr>
<td>J Miandad</td>
<td>-0.05</td>
<td>0.9</td>
<td>0.12</td>
<td>-0.01</td>
<td>0.83</td>
<td>-0.72</td>
<td>1.59</td>
<td>&lt;.0001</td>
<td>0.95</td>
<td>3</td>
</tr>
<tr>
<td>SM Gavaskar</td>
<td>1.19</td>
<td>-0.33</td>
<td>0.12</td>
<td>-0.01</td>
<td>0.75</td>
<td>-0.17</td>
<td>1.85</td>
<td>&lt;.0001</td>
<td>0.91</td>
<td>3</td>
</tr>
<tr>
<td>WR Hammond</td>
<td>0.03</td>
<td>0.66</td>
<td>0.14</td>
<td>-0.01</td>
<td>0.94</td>
<td>-1.38</td>
<td>1.36</td>
<td>&lt;.0001</td>
<td>0.89</td>
<td>3</td>
</tr>
<tr>
<td>BC Lara</td>
<td>-1.31</td>
<td>1.18</td>
<td>0.08</td>
<td>0</td>
<td>0.65</td>
<td>-2.96</td>
<td>10.54</td>
<td>&lt;.0001</td>
<td>0.8</td>
<td>4</td>
</tr>
<tr>
<td>DG Bradman</td>
<td>-0.97</td>
<td>1.13</td>
<td>0.06</td>
<td>0</td>
<td>0.58</td>
<td>-2.58</td>
<td>8.61</td>
<td>&lt;.0001</td>
<td>0.84</td>
<td>4</td>
</tr>
<tr>
<td>DCS Compton</td>
<td>0.77</td>
<td>0.05</td>
<td>0.06</td>
<td>0</td>
<td>0.28</td>
<td>2.53</td>
<td>9.67</td>
<td>&lt;.0001</td>
<td>0.79</td>
<td>5</td>
</tr>
<tr>
<td>M Azharuddin</td>
<td>0.4</td>
<td>0.51</td>
<td>0.06</td>
<td>0</td>
<td>0.74</td>
<td>2.27</td>
<td>8.9</td>
<td>&lt;.0001</td>
<td>0.82</td>
<td>5</td>
</tr>
<tr>
<td>CH Lloyd</td>
<td>0.4</td>
<td>0.54</td>
<td>0.05</td>
<td>0</td>
<td>0.56</td>
<td>1.56</td>
<td>4.11</td>
<td>&lt;.0001</td>
<td>0.88</td>
<td>6</td>
</tr>
<tr>
<td>KD Walters</td>
<td>-1.37</td>
<td>2.24</td>
<td>0.03</td>
<td>0</td>
<td>0.97</td>
<td>2.22</td>
<td>4.06</td>
<td>&lt;.0001</td>
<td>0.86</td>
<td>6</td>
</tr>
<tr>
<td>L Hutton</td>
<td>-2.32</td>
<td>2.22</td>
<td>0.07</td>
<td>0</td>
<td>0.12</td>
<td>-4.31</td>
<td>18.97</td>
<td>&lt;.0001</td>
<td>0.86</td>
<td>-</td>
</tr>
<tr>
<td>SR Waugh</td>
<td>-0.54</td>
<td>0.44</td>
<td>0.08</td>
<td>0</td>
<td>0.92</td>
<td>-1.99</td>
<td>4.97</td>
<td>&lt;.0001</td>
<td>0.91</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1: Summary of Global Characteristics for International Batsmen with Long Careers.

Examples of the results of the polynomial model fitting are shown in Figure 1. Each graph displays a bubble plot overlaid with the fitted polynomial. The size of the bubble indicates the relative number of innings played in that year of an individual’s career.
Figure 1a: Bubble Plot overlaid with fitted polynomial function showing Border’s relative batting contribution by year.

Figure 1b: Bubble Plot overlaid with fitted polynomial function showing Tendulkar’s relative batting contribution by year.

Figure 1c: Bubble Plot overlaid with fitted polynomial function showing Miandad’s relative batting contribution by year.

Figure 1d: Bubble Plot overlaid with fitted polynomial function showing Lloyd’s relative batting contribution by year.

Figure 1e: Bubble Plot overlaid with fitted polynomial function showing Waugh’s relative batting contribution by year.

Figure 1f: Bubble Plot overlaid with fitted polynomial function showing Azharuddin’s relative batting contribution by year.
5. Time Series Clustering

The seven variables that describe the characteristics of career progression, shown in Table 1, were used to cluster the 20 players using Ward’s minimum variance method. Ward’s minimum variance method is a hierarchical approach designed to optimize the variance within clusters. Ultimately, this method was chosen due to quality of the clusters it produced (see figures 3a-3f). Like the clustering methodology, the cut-point for selecting the clusters was determined heuristically to produce clusters that provided results that appeared most similar visually.

*Figure 2a: Dendogram showing clusters showing linkages between related careers*
The height of the vertical lines connecting the groups in the dendogram measures the similarity of groups of cricketers. The longer the connecting line the greater the dissimilarity. The resultant groupings can be visualized in the dendogram above. Figure 2b below provides a zoomed view of the dendogram and shows where the cut-point for determining clusters was placed.

Given the uniqueness of the batsmen and the size of the dataset, it is possible to perform this task manually. However, for scouting purposes, a manual approach would not be scalable.

![Dendogram showing clusters with cut point shown using dotted line](image)

**Figure 2b: Close-up view of Dendogram showing clusters with cut point shown using dotted line**

The partial dendogram above shows where the cut-point was made. The 20 players were placed into six clusters with two players, Steve Waugh and Len Hutton remaining unclustered. Table 1 shows the players assigned to each cluster.

It is important to note that these clusters are determining based on the individual’s performance relative to their peers, over their entire career. Each player had an outstanding career, as a test career of 17 or more years illustrates. Thus, these clusters indicate when these batsmen had arguably the greatest impact on the game.

The line plots that follow demonstrate the effectiveness of this clustering approach. For each cluster, the fitted polynomial functions for each player are overlaid. The fitted polynomial is essentially a smoothing function and intended for interpolation only.
Figure 3a: Line plot displaying the fitted polynomial function for relative batting contribution by year for members of cluster 1.

Figure 3b: Line plot displaying the fitted polynomial function for relative batting contribution by year for members of cluster 2.

Figure 3c: Line plot displaying the fitted polynomial function for relative batting contribution by year for members of cluster 3.

Figure 3d: Line plot displaying the fitted polynomial function for relative batting contribution by year for members of cluster 4.

Figure 3e: Line plot displaying the fitted polynomial function for relative batting contribution by year for members of cluster 5.

Figure 3f: Line plot displaying the fitted polynomial function for relative batting contribution by year for members of cluster 6.
6. DISCUSSION OF CLUSTERS

The figures on the previous page demonstrate that the time series clustering worked well at clustering similar career progressions as there is a great deal of congruence in the fitted polynomial functions. The visualizations also reinforce the decision regarding the cut-point. For instance, the Clusters shown in Figure 3a, with Allan Border, and Figure 3b are closely related according to the dendogram. However, each group is clearly quite distinct. For visualization purposes, Mitchell’s fitted polynomial was shifted forward four years.

The cluster analysis generated some interesting groups to review. In this section we will explore the composition of each of the clusters with two or more players with regard to the different patterns and how that relates to the individual careers.

Cluster 1 is illustrated in Figure 3a. The main characteristic of this cluster is the early career dominance, shown by the batsmen belonging to this group, before their relative contribution lessens. The results of Cluster 2 in Figure 3b suggest that the batsmen in this group were at their most dominant through the mid part of their careers. Batsmen belonging to Cluster 3 are shown in Figure 3c. While similar to the trend in cluster 2, these individuals have a greater impact at the start to their careers, while continuing with a relatively high contribution well into the later-half of their career before declining towards the end, suggesting at least a decade of dominance. Figure 3d, displaying the batsmen belonging to the fourth cluster, suggest players who have a somewhat indifferent start to their career, before their contribution improved greatly and stabilized to some extent across the remainder of their careers. The trend apparent in cluster 5, displayed in Figure 3e, suggests that batsmen grouped in this cluster had a stellar start in test cricket, followed by what could be described as a mid career re-emergence. Finally, the batsmen from the sixth cluster in Figure 3f also had a stellar start to their careers, similar to those in cluster 5. However, cluster 6 indicates a strong finish in the batsmen’s career.

Contribution is used as a measure of an individual’s relative influence as a batsman. In the subsections that follow, we explore how the clusters based on relative batting contribution relate to other statistics and expert opinion. We will also highlight Sir Donald Bradman’s career progression, and discuss which batsmen had careers most comparable to him in terms of influence.

Cluster 1 – Early career dominance

The trait associated with Cluster 1, suggests that these batsmen were most likely to have contributed the most in the early part of their playing career, before gradually declining in the later half. This early career dominance is the main
feature of a cluster which features cricket legends Alan Border, Rohan Kanhai and Sir Vivian Richards.

By the time Allan Border called it quits on his distinguished career in 1994, he had managed to take a number of prestigious records with him. At the time his 156 test matches, including 153 in a row, held the record for most tests and most consecutive tests by any player (Meher-Homji, 2003). He finished as the greatest run scorer of all time with a total of 11,174 at an average over 50 and he captained Australia on an unbroken run of 93 tests (Meher-Homji, 2003). Meher-Homji (2003) described Border as ‘an unromantic, uncomplicated but uncompromising workman cricketer. He was an ordinary soul who accomplished extraordinary deeds.’ (Engel, 1995).

Despite a fairly indifferent first few tests in his first year in the Australian test side, Border was a standout performer in his second year in the side, amassing over 1000 runs in 1979 (Cricinfo, 2009). He continued his good form through the first half of his career, where Border had scored close to 7,000 runs at an average close to 53, including 21 centuries (Cricinfo, 2009).

Sir Vivian Richards will be known as one of the greatest cricketers to ever grace the game. In 2000, Wisden listed Richards as one of the top five cricketers of the 20th century (BBC News, 2000). The West Indies legend was one of the most destructive players around. Meher-Homji (2003) went as far as saying only Sir Donald Bradman had a greater crowd appeal than Richards.

Richards burst onto the scene in 1974, and it didn’t take long for him to make his mark. He made a terrific 192 not out in just his second test for what was a sign of things to come (Cricinfo, 2009). However, it was Richards third year as a test player that really stood out. In 1976, Richards scored a massive 1,710 runs at an average of 90, including an innings of 291 against England at ‘The Oval’ (Cricinfo, 2009). Richards went on to make over 2,700 runs in his first six years at an average around 58, definitely fitting the idea of early career dominance (Cricinfo, 2009).

Cluster 2 – Mid career dominance

Cluster 2 can be interpreted as a batting career that was most influential throughout the middle part of a player’s career. This trend also suggested that the batsmen from Cluster 2 may have taken a few seasons to become settled at Test level. Mid career dominance is the main feature of a cluster which features legends Englishman Geoffrey Boycott, Australia’s Robert Simpson and Indian Sachin Tendulkar.

In a test career that lasted for the best part of two decades, Geoffrey Boycott was able to create a legacy that places him among the all time greats of the game. Having played over 100 tests for England, Boycott scored more than 8,000 runs
including twenty two centuries (Cricinfo, 2009). Boycott was considered an introvert and somewhat of a lone ranger as a player, but his desire for success is a quality that was often overlooked (Meher-Homji, 2003). Both controversial and at times cantankerous, there is no denying that Boycott was a classical batsman who loved cricket (Meher-Homji, 2003). Although Boycott started his career reasonably well and scored an unbeaten 246 against India in 1967, his best period as a player came in the period between 1971 and 1977, suggesting his peak came during the middle part of his career. In this time Boycott managed to score more than 2,300 runs at an average in the mid sixties, while also scoring eight centuries and 13 fifties in forty two innings (Cricinfo, 2009). Boycott’s stretch from 1971 to 1977 was his 8th through 12th year of playing for England.

Sachin Tendulkar is considered by many as the most complete batsman of his generation, not to mention his status as possibly the biggest icon the sport of cricket has seen (Cricinfo, 2009). His batting is something for all cricket fans to admire including attributes such as balance, great foot movement, precision in shot placement and anticipation (Cricinfo, 2009). Though he has adopted a noticeably conservative approach in the last quarter of his career, which Tendulkar himself put down to his growing role as a senior player, there are no apparent weaknesses in Tendulkar's game (Ezekiel, 2003). He is able to score all around the wicket, off both front foot and back, and has made runs in all parts of the world in all conditions (Ezekiel, 2003). Tendulkar’s career peak is viewed as a golden period between 1994 and 1999, coinciding with his physical peak in his twenties (Ezekiel, 2003). This corresponds with years 6 to 11 of his career. During this six year period Tendulkar amassed more than 4000 runs at an average above 60, while scoring an impressive 16 centuries (Cricinfo, 2009). This career peak fits well with the characteristics associated with Cluster 2.

**Cluster 3 – Decade of dominance then decline**

The traits associated with Cluster 3 suggest that batsmen in this group were most likely to have started contributing at a high level in the early to middle part of their careers, and manage to maintain this level of influence through to the later stages before fading. This decade dominance is the main feature of the biggest cluster and includes Sunil Gavaskar, Walter Hammond, John (Jack) Hobbs, Javed Miandad, Bruce Mitchell and Sir Gary Sobers.

Bombay native Sunil Gavaskar is widely regarded as one of the greatest opening batsmen in test match history, if not the best. This was highlighted when Richie Benaud selected Gavaskar as one of the opening batsmen on his DVD ‘Ritchie Benaud's Greatest XI’ (Gawith, 2004). Throughout his long career Gavaskar broke most batting records by the time he retired in 1987. He held the record for most test runs, having amassed over 10,000 and the most test centuries
with 34, a record that stood for nearly two decades (Cricinfo, 2009). After a great start to his test career, Gavaskar started consistently producing for his side in 1975, with his best period between 1978 and 1980. Gavaskar had a great run of form where he scored close to 3000 runs at an average in the mid sixties (Cricinfo, 2009). In 46 innings over that three year span, Gavaskar scored ten centuries and eleven fifties including two double centuries (Cricinfo, 2009). Gavaskar kept up his good form for the best part of a decade, and posted his highest test score of 236 not out in 1983 (Cricinfo, 2009).

Javed Miandad is considered to be the greatest Pakistani batsman ever. Miandad’s feats place him amongst the best of all time and this talented and prolific batsman is considered one of the true characters of his generation in the sport of cricket (Meher-Homji, 2003). Meher-Homji (2003) went onto describe Miandad as having the ego of Mohammad Ali and the belligerence of John McEnroe, but there was no deny his cricket genius, with only Sir Vivian Richards matching his dazzling stroke play among his contemporaries. Miandad is one of the few batsmen to maintain a batting average over 50 throughout his career, and while Miandad enjoyed a great stretch between 1978 and 1989, his best run of form was between 1983 and 1989 (career years eight through fourteen). During this stretch Miandad scored over 4000 runs at an average of approximately 61.6, including 14 centuries (Cricinfo, 2009). He managed an impressive feat of scoring a double century in five of the seven years.

“The greatest all-round cricketer the world has seen”, those were the words of the great Ritchie Benaud when describing Sobers (Benaud, 2005). Ritchie Benaud’s praise for the West Indian Great didn’t stop there either, writing Sobers was "a brilliant batsman, splendid fielder, particularly close to the wicket, and a bowler of extraordinary skill, whether bowling with the new ball, providing orthodox left-arm spin or over-the-wrist spin" (Benaud, 2005). Sobers career peak came in the 1958 season, his fifth as a test player. In this magical year Sobers scored nearly 1300 runs with an average of over 144 (Cricinfo, 2009). But perhaps the biggest feat was a single innings of 365 not out which became the highest individual score by a test batsman, and a record that would stand for 36 years (Cricinfo, 2009). Sobers influence appears greatest in the early to middle part of his career.

Cluster 4 – Indifferent start

Batsmen in Cluster 4 began their careers indifferently which belies the enormous impact that these individuals would have on the game. However, it also suggests that their contribution started to climb around the second or third season and stay at a high level throughout their careers, dropping off slightly towards the end. The two batsmen associated with cluster four are none other than Sir Donald Bradman and Brian Lara.
Sir Donald Bradman is considered in most cricketing circle’s as the greatest cricketer of all time, and ever since he left the game after his final test in 1948 a number of cricketers have been compared to Bradman, in what can only be described as a mission to find the ‘next Don Bradman’. Having retired from the sport he loved and played for over two decades with such impressive statistics and a number of records to his name, it can be easy to forget that his career was shortened due to World War Two. To have missed out on six years in the middle of his career has previously made it rather difficult to figure out exactly how Bradman’s career may have developed if his career hadn’t been interrupted. Reviewing his career statistically, Bradman finished his first season in 1928 with a batting average in the low 50s (Cricinfo, 2009), and had arguably his best year in 1932, his fifth year playing for Australia when he averaged over 400 (Cricinfo, 2009). While Bradman did have a few difficult series, including the infamous ‘body line tour’ against England, another reason why he was placed in Cluster 4 was that he didn’t seem to go through any lengthy form slumps, and always managed to bounce back.

In the last 15 years there have been two players who have been compared to each other than anyone else, Indian master Sachin Tendulkar and West Indian great Brian Lara. Despite Lara being a left-handed batsman, both he and Tendulkar have emerged as the greatest batsman of their generation. One difference between the two appears to be the ability to compile big scores. Lara’s first test century was against Australia in 1993 when he scored a massive 277 runs, while it is well noted that it took Tendulkar over ten years to reach his first double century, with some saying that it may have been Tendulkar’s impatience at the crease which saw him fail to build on most of his centuries (Ezekiel, 2003). To add to Lara’s knack for big totals, he scored more than 200 runs in an innings in 8 different calendar years including the first and third highest totals ever. Only Bradman has gone past 200 runs in an innings more times than Lara. Many people believe that Lara peaked in the year of 1994 at age 25 when he broke the record for highest test innings (375) as well as the record for highest first class innings (501 not out) (Andrews & Jackson, 2001). What some may consider as a ‘peak’ in Lara’s career would no doubt be viewed as 1994, but the period from 1994 through 1999 resulted in more than 4700 runs at an average close to 53, including 12 centuries (Cricinfo, 2009). But to Lara’s credit, he managed to keep putting up big scores right through to his retirement, including his world record of 400 not out against England in 2004. Like Bradman, Lara had a knack for producing big scores, and is second only Bradman for most test scores over 200, and sits alongside Bradman as one of just three players to pass 300 twice (the third player being Virender Sehawg of India).
Cluster 5 – Stellar start with mid career re-emergence

Batsmen in Cluster 5 were most likely to have the greatest relative contribution at the very start of their careers, before a reduction in influence leading to a mid-career resurgence. This stellar start, followed by a mid career re-emergence is a characteristic of batsmen Mohammad Azharuddin and Denis Compton.

Azharuddin, the former Indian cricket captain, is considered one of the best middle order batsmen of the 1980s and 1990s. He started his career well, a Cluster 5 characteristic, after he scored a century in each of his first three test matches (Cricinfo, 2009). His relative influence diminished until he had arguably his best year as a Test cricketer in 1990 when he scored three centuries in the one year and averaged over 75 (Cricinfo, 2009). This fine play led to Azharuddin receiving the Wisden Cricketer of the Year award in 1991. This second performance peak around 1990-1991 is also typical of Cluster 5 players.

Cluster 6 – Stellar start with strong finish

Cluster 6 is quite similar to Cluster 5 as both clusters were highly influential early in their careers. However, for Cluster 5 batsmen, their relative influence peaked again towards the end of their career. The two batsmen associated with having a stellar start and strong finish are West Indies great Clive Lloyd and Australian Kevin (Doug) Walters.

Clive Lloyd was an attractive batsman and considered by many to be one of the greatest captains ever (Meher-Homji, 2003). Lloyd made his test debut in 1966, averaging over 80 runs in his first year playing for the West Indies (Cricinfo, 2009). However, most observers of the game would specify the twilight of his career as his best. In Lloyd’s last four years as a test player, between 1981 and 1984, he amassed a total of 2,342 runs in just over 40 innings, giving him a remarkable average of 61.6 (Cricinfo, 2009).

7. IMPUTATION

As Bradman’s career progression is most similar to West Indian legend Brian Lara, it would appear that his peak performances would have occurred in the 12th to 14th years of his career (1939-1941), which coincided with World War II. In this section, we impute what Bradman is likely to have scored during this time frame and test if it would have significantly changed his batting average.

The predicted contribution from the polynomial function used to fit Bradman’s career is used to predict the mean number of runs scored per year using linear regression. The intercept is omitted from these models. The mean number of
runs scored per innings is used to enable fair comparison between years where there were different number of matches played. From this model, we use the standard error of the estimate of the slope to build confidence intervals.

However, the mean number of runs needs to be converted to the traditional batting average, which is the total number of runs scored per dismissal. We achieve this by estimating the number of not outs per calendar year. Assuming that the probability of being not out in an innings remains constant, we use the proportion of observed not outs to create upper and lower estimates of the number of dismissals per year.

Finally, to give adequate weighting to the likely runs scored in the missing seasons, the observed average number of innings per year, rounded down to the nearest integer, is used to obtain the expected total runs per calendar year. Based on this the lower limit for number of not outs is the number of observed not outs (this means that the batsman would not have been not out in the time frame being imputed).

Thus our estimate for the \(i\)th batsman in the \(s\)th career year becomes:

\[
A_{i,s} = \frac{T_{i,s}}{\hat{n}_{i,s}} - \hat{k}_{i,s},
\]  

(4)

where \(T_{i,s}\) is either the observed total runs scored, \(T_{i,s}\), or,

\[
T_{i,s} = \hat{M}_{i,s} \hat{n}_{i,s}
\]  

(5)

The estimate for the number of innings played in a calendar year, \(\hat{n}_{i,s}\), is either the number of innings played in that year or the average for the \(i\)th player per calendar year. The estimated mean runs for the \(i\)th player in the \(s\)th season is:

\[
\hat{M}_{i,s} = k_i \hat{C}_{i,s}
\]  

(6)

where \(k_i\) is the slope coefficient and \(\hat{C}_{i,s}\) is the estimated average contribution per calendar year obtained from the fitted polynomial function.

The upper estimate for total runs uses the estimate for the upper confidence limit for runs scored and the upper confidence limit for not outs (giving the lowest likely number of dismissals). The converse applies to estimating the lower limits.

This process is performed for both Bradman and Lara. In both instances, the regression used to estimate the observed mean runs per year from the estimated contribution per year was highly significant (\(p<0.0001\)).

The graphs that follow plot the observed cumulative mean runs against the career year, overlaid with the upper and lower 95% confidence limits for the cumulative predicted mean runs. In both instances, after the first third of their careers the observed mean runs falls inside the confidence intervals obtained from the estimation process. This indicates that the estimated model fits the data adequately. Thus, the estimates obtained for Bradman are reasonable.
Using this model to impute Bradman’s likely performances for 1939-1945 we estimate his batting average to be 105.41 [95% CI (90.48,123.44)]. The comparable result for Lara is much lower, with his observed batting average of 52.88 contained within the estimated 95% confidence interval (44.81, 67.16).

![Graph of Bradman's cumulative mean runs](image1)

**Figure 4:** Plot of the observed cumulative mean runs against the career year of Bradman with 95% confidence interval limits from the estimation model overlaid.

![Graph of Lara's cumulative mean runs](image2)

**Figure 5:** Plot of the observed cumulative mean runs against the career year of Lara with 95% confidence interval limits from the estimation model overlaid.
Using the upper 95% confidence limit estimate and mean to obtain the standard deviation for the average runs scored per season enables us to test if Bradman’s batting average would have been higher had the Second World War not interrupted his career. Comparing the estimated average (105.41) with the observed average (99.94) gives a p-value of 0.2763. Whilst it would appear that Bradman’s estimated average is higher than his observed average there is insufficient evidence at the 5% significance level to conclude that his average would have been significantly higher had the war not interrupted his career.

It is acknowledged that there could be alternative methods used to model player performance with respect to career development, such as modeling performance as a function of age (or stage of career). These have not been explored in this paper, mainly due to two underlying reasons.

Firstly, very few individuals have played 17 or more years of cricket at the highest level. This implies that to sustain selection, an individual must have continued to perform sufficiently to justify inclusion, which in itself is an interesting attribute of their performances. As a consequence, it is important to view these careers in their entirety and not in a piecemeal manner as a function of age.

Secondly, the ebb and flow of a career (and ongoing selection), is conditional on previous performance and as such, characterizing performance requires the conditional behavior of the individual to be included in the underlying model. The polynomial functions used to model annualized performance per year of career that are presented in this paper encapsulate the interaction between years and without this interaction it could lose some of its applicability. For characterizing player performance over time, the polynomial functions performed adequately for our needs (enabling meaningful clusters to be constructed) and the exploration of other methods was not seen as necessary.
8. Conclusion

Time series clustering was used to show that, relatively, the impact of Australian legend Sir Donald Bradman’s Test career as a batsman was most similar to West Indian Brian Lara in terms of career progression and not in terms of overall caliber.

Data from the 20 international cricketers who played in at least 70 innings over more than 17 years (as at January 1st, 2009) was used to create a number of global measures to indicate the ebb and flow of a career.

The methodology chosen to cluster time series was based on both a model and feature based approach. These global time series measures as proposed by Wang et al (2006) generated instinctive clustering results based on different length time series (Wang et al, 2006).

To standardize the data, the average contribution per year was used to reflect the relative influence of each batsman. The use of contribution rather than runs scored enabled extraneous factors, such as pitch conditions, to be negated by using team mates as a frame of reference. The average contribution per year for each batsman was then standardized using the range. This smoothed, standardized data was used to fit a polynomial function for each batsman. The parameters of this model, along with the skewness, kurtosis and autocorrelation were used to generate six meaningful clusters.

Importantly, the player groupings were meaningful, in terms of career progression, with Bradman and Lara identified as having careers that developed in a similar fashion. This highlights the usefulness of time series clustering for player comparison.

The model fitting and clustering suggested that Bradman’s prime may have been most likely to occur during the Second World War when international sport was disrupted. The study then imputed Don Bradman’s likely performances during this period to estimate what his batting average. Although the estimated average was calculated to be 105.41 this was not significantly different from his actual average of 99.94.

Without doubt, Sir Don Bradman was an exceptional talent and viewed as the gold standard of international batsmanship. However, it is often difficult to compare players across eras. This paper has introduced a methodology for comparing sporting careers that is used in the financial sector and has demonstrated that time series clustering is an effective technique for comparing player careers.
REFERENCES


