Spatial Comparison of Cricketers

Adhitya Kamakshidasan

▶ **To cite this version:**

Adhitya Kamakshidasan. Spatial Comparison of Cricketers. IEEE VIS 2018 - IEEE Conference on Visualization, Oct 2018, Berlin, Germany. hal-01852138v3

HAL Id: hal-01852138
https://inria.hal.science/hal-01852138v3

Submitted on 2 Jul 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Spatial Comparison of Cricketers

Adhitya Kamakshidasan*

Inria, Saclay

Abstract

There has been an increasing demand from the cricketing community to introduce newer metrics to analyze the game. Data from ball tracking and prediction have urged statisticians to re-look at existing comparative measures. Using methods from topological data analysis, we introduce a new technique to compare cricketers using spatial features. We use data from IPL seasons (2012 -2017) to compare our results with an existing ranking scheme.

Keywords: topological data analysis, spatio-temporal cricket

1 Introduction

When players are hit with injuries in the Indian Premier League (IPL), teams generally scout for like-for-like replacements without changing team composition characteristics. Such replacements are often chosen by using aggregate benchmarks like Strike Rate and Economy, which fail to capture spatial information.

Spatial information associated with a player's performance is usually represented by a pitch-map (see Figure 2). Though such visualizations are keystone in understanding performance over a single match, their interpretation over multiple matches is limited, owing to the high number of deliveries. Cricket analysts are commonly interested in understanding strong and weak scoring regions of a player across the pitch. For example, experts often suggest that Suresh Raina struggles playing short pitched deliveries, despite being the highest aggregate run scorer across all IPL seasons.

Topology-based techniques are ideal for understanding such expert insights due to two main reasons —- they can efficiently identify locations of interesting features over a spatial domain; and their ability to look at the overall distribution of values. From a cricketing perspective, such techniques helps in identifying the abilities of a player across different lines and lengths of deliveries. In a more broader sense, we propose that these techniques can also be used to compare and cluster players of different traits, for injury replacement strategies by team managements.

In this work, using spatio-temporal data and methods from topological data analysis, we present an approach to compare cricketers.

2 Spatial Comparison

Irrespective of the format, cricket is always played on a rectangular pitch that is 22 yards long and 10 feet wide. Every delivery in a match, is bowled on the same pitch and has various attributes associated with it. Some attributes include *delivery type*, *pitch location*, *shot played*, *runs scored/conceded*, *players involved* and *scoring region*. We believe that a player's utility can be directly determined based on their behaviour across the pitch. A key step in identifying such behaviour would be to locate regions corresponding to significant aspects [1].

A scalar function $f : \mathbb{D} \to \mathbb{R}$ maps points in a spatial domain $\mathbb{D}$ to real values. In this work, we are interested in the spatial region corresponding to a pitch, which is represented by a planar domain.

*e-mail: adhitya.kamakshidasan@inria.fr

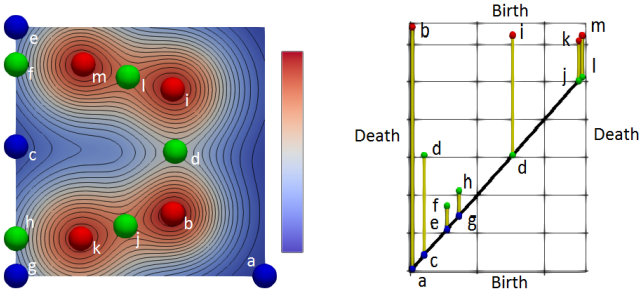

Figure 1: Critical points of a scalar field f defined on a PL 2-manifold (left) and its persistence diagram $\mathscr{D}(f)$ (right). In the persistence diagram, each pair of critical points are represented by a vertical bar (yellow) and their persistence is given by height of the bar.

A function value is defined over each point in this planar domain with the goal to capture the spatial utility of a player. An example of a scalar function, that is used in this work, is the number of *runs scored* off each delivery. We choose this function for illustration, since it intuitively captures the scoring capability of a player over various locations across the pitch. A high function value at a given location implies easier scoring, thereby enhancing/decreasing the utility of batsman/bowler.

In order to efficiently compute the topological features of the scalar function f, it is represented as a piecewise linear (PL) function f: $K \to \mathbb{R}$. The planar domain $\mathbb{D}$ of the function is represented by a 2D regular grid K. The function is defined on the vertices of the grid and linearly interpolated across each box. To account for the high density of a point cloud representing a player, the set of points are first interpolated to a regular grid using scattered data interpolation techniques and then a scalar function is computed for each player.

The critical points of a smooth real-valued function are exactly where the gradient of a scalar field becomes zero. The critical points of PL function are classified based on the behaviour of the function within a local neighbourhood. An important measure that is often associated with critical points is its persistence. Persistence diagrams encode the persistence of features as points in a plane.

Understanding player performance across regions of the pitch, is akin to understanding the persistent features of super-level components of the scalar field. Pitch regions having high persistence indicate easier scoring ability. Here, the specific high and low function values of a scalar field are not too important, but rather their distribution is key for comparing the behaviour of two players. We capture this, by finding similarity between persistence diagrams.

For two scalar fields, the similarity of persistence diagrams is generally measured by using a distance function like the Wasserstein distance or the Bottleneck distance. Low persistent features can be considered to be topological noise in the scalar field [2]. In our case, two players can be compared to one another by their respective persistence diagrams after removal of low persistent features.

We also compare the players using Earthmovers distance. This distance reflects the minimal amount of work that must be performed to transform one distribution into the other by moving "distribution mass" around. For this, grayscaled pixel values of the scalar field are used to construct histograms for similarity detection between players.

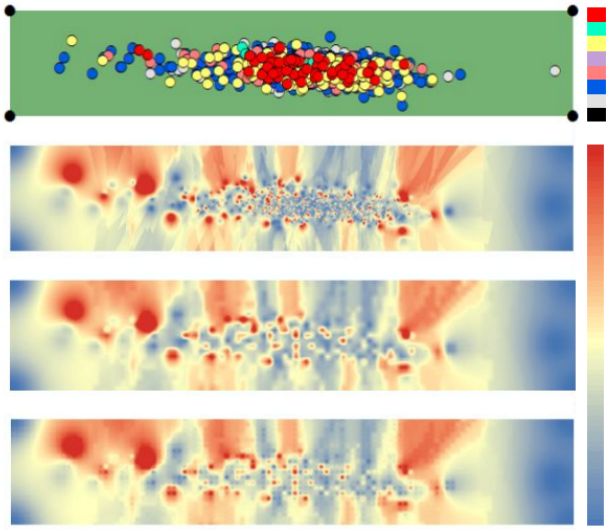

Figure 2: Sunil Narine's bowling across six IPL seasons. Pitch-map visualization for *runs conceded* function (first row). Delaunay triangulation of pitch locations with z-coordinate as scalar function value (second row). Scalar field after Shepard's interpolation and removal of topological noise respectively (third, fourth rows).

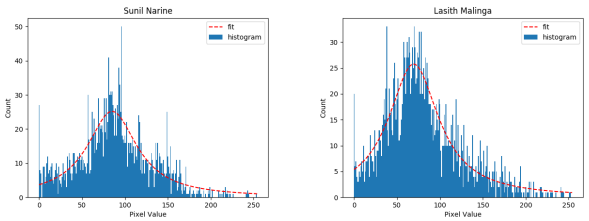

Figure 3: Grayscale pixel distribution of Sunil Narine (left) and Lasith Malinga (right) used by Earthmovers distance for player comparison.

3 Method

Hawk-Eye is a commercial vendor that captures real-time spatio-temporal data for assisting officials with decision making processes. We demonstrate our method using Hawk-Eye data scrapped from various sources on the internet. The pair *(x, y)* specifies the pitching location of a delivery on a cricket pitch, where y is the horizontal coordinate and x is the vertical coordinate relative to an origin at the centre of off-stump for a right-handed batsman. The positive y-axis points to the right from the bowler's perspective and the positive x-axis points towards the non-strikers stumps.

Our analysis includes ball-by-ball data from all IPL matches across 2012 and 2017. In this work, we look only at *pitching locations* and *runs scored*.

For a given interpolation method and a distance measure, the scalar field of one player can be compared to all other players in the consideration set, to obtain a unique *rank descriptor* for the player. The *rank descriptor* of a player showcases similarity/dis-similarity between all other players. For a given player, the descriptor would rank the player most similar (himself) as *numero uno*, and rank the most dissimilar player as equivalent to the number of players in the consideration set. This can be extended to obtain a *rank matrix*

Table 1: Median correlation values - Spatial Ranking vs Cricmetric

Distance Measure	Interpolation Method			
	Sibson's	Spline	Kriging	Shepard's
Earthmovers	0.279	0.301	0.153	0.229
Bottleneck	0.16	0.094	0.094	0.028
Wasserstein - 1	-0.185	0.112	0.078	-0.168

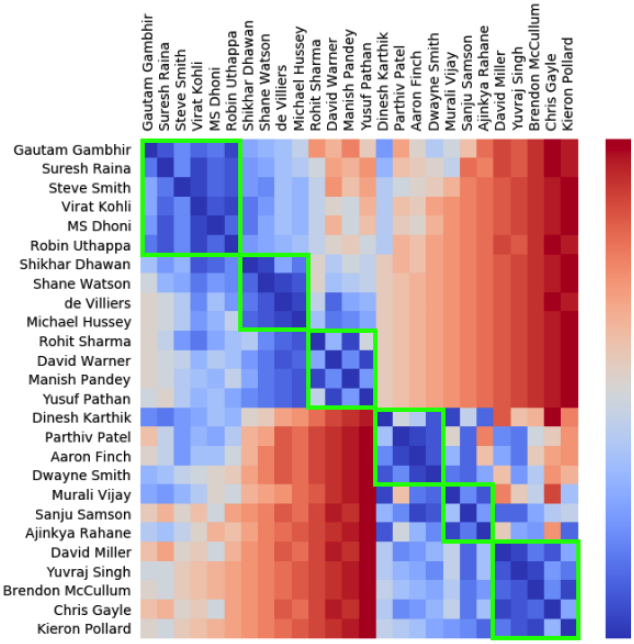

Figure 4: Rank matrix for batsmen obtained using Shepard's interpolation and Earthmovers distance. Each batsman on the left is ranked with a batsman on the top. Similar batsmen are highlighted in green.

for all players (see Figure 4) wherein each row represents the *rank descriptor* for a player. For the construction of the *rank descriptor*, either Bottleneck/Wasserstein distance measures are computed between the persistence diagrams or an Earthmovers distance measure is computed between the respective scalar fields.

We compare our results with Cricmetric's ranking system. Cricmetric provides a non-parametric way of ranking players by computing an eigenfactor score (EFscore). A similar *rank matrix* is constructed for Cricmetric by averaging EFscores across seasons followed by the L_1 norm. One common way to compare ranking schemes is by using the Pearson Correlation Coefficient (PCC). This measures the linear relationship between two data-sets. Like most other correlation coefficients, this one also varies between -1 and +1 with 0 implying no correlation. We compare each player's *rank descriptor* with that of Cricmetric's by using PCC. Table 1 reports the median correlation values for each ranking scheme across players in the comparison set.

4 Results

Our method identifies six different clusters of players from Figure 4. We asked a cricket expert to help us identify and group several characteristics of these players to interpret our results. The player types found along the diagonal of the *rank matrix*, are listed from top-left to bottom-right as follows — "game closers", "slow-starting quick-scorers", "hit-or-miss players", "multiple role players", "stroke makers", and "extremely aggressive starters".

From Table 1 we can clearly see that the correlation values with Cricmetric are not high. Since there is no "ground truth" to ranking systems, we cannot argue that our method performs better/worse than another one. However, the table indicates a shortcoming in that, our method is highly susceptible to the interpolation technique and the distance measure used for comparison.

References

[1] J. Gudmundsson and M. Horton. Spatio-temporal analysis of team sports. *ACM Comput. Surv.*, 2017.

[2] J. Tierny, G. Favelier, J. A. Levine, C. Gueunet, and M. Michaux. The Topology ToolKit. *IEEE TVCG*, 2017.