Quantifying a Cricketer’s ability to adapt to different formats of the game and to remodel the current player rating system based on it

Sreedhar Radhakrishnan, V.R Chittaranjan under the guidance of Dr. Kavi Mahesh.
Centre For Knowledge Analytics And Ontological Engineering, PES University, Bangalore, India.

Abstract: This project introduces a metric to evaluate the adaptation of a player to all the three formats of Cricket which are Test, ODI as well as the Twenty20 format. The metric introduced is based on a ‘Performance Matrix’ developed after Performance Analysis of the players across the multiple cricketing formats.

This quantified ‘Adaptation Factor’ for a player is further used to develop a formula to remodel the current system of player rating in a particular format.

An applied set theory approach to find single format specialists as well as multi-format specialists is introduced in this project as well.

The analysis is supplemented by investigating the impact of team rating taken as an average of individual player ratings against team performance which provides a stronger correlation than the existing Reliance ICC ratings.

1. Introduction

The game of cricket has seen a prodigious evolution since the inception of the Test format of the game in 1877. With the ODI or 'One Day International' format introduced in 1971 and the Twenty20 (T20) format introduced in 2003, there has been a strong change in the way cricket players approach the game. T20 leagues such as the Indian Premier League introduced in 2008 have further acted as strong catalysts in this evolution.

For readers new to the sport of cricket we suggest viewing the Wikipedia Page For Cricket to understand more about the sport and its various formats which will make the understanding of this project report stronger.

With the evolution of the game, there has been a strong need for players to adapt to various situations of the game. A run chase of 300 plus runs in the ODI format is not as intimidating as it was a decade ago because players in the chasing team quickly adapt to the T20 format of the game in the last ten overs or so. It is this 'adaptation factor' of a player that makes the difference between winning and losing.

While the sport of cricket is evolving, the rating system is not. There is no Quantitative Metric for player evaluation which takes into account his adaptation factor. It is this very problem that motivated this research work where we apply data analytics and machine learning to quantify a cricketer's ability to adapt to different formats of the game and remodel the current player rating system based on it.

This project has three logical sections which are :-

1. Performance analysis of the players across multiple cricketing formats.
2. Quantifying the adaptation of a player across multiple cricketing formats and developing a new formula for player evaluation/rating.
3. Evaluating the impact of the formula by finding the correlation between team rating and team performance.

2. Performance analysis of the players across multiple cricketing formats

Before performance analysis of any sort, a reliable and substantial data set was required. This was obtained by means of a web scraping script, which retrieved player statistics for both Batting as well as Bowling across all formats from www.cricketarchive.com for all players who have been a part of the IPL in one of its many seasons.

Once the data set was scraped and cleaned, the concept of unsupervised learning clustering was used and the simple k-means algorithm was applied on the set of players, for 4 formats namely ODI, Test, Domestic T20 and International T20.

Name	Matches	Innings	Not Outs	Runs	High Score	Average	# of 100s	# of 50s	SR	Catches
A B de Villiers	197	189	34	8524	162	54.99	24	47	99.96	161
A C Voges	31	28	9	870	112	45.78	1	4	87.17	7
A C Gilchrist	286	278	11	9595	172	35.93	16	55	96.89	416
A F Milne	33	12	6	130	36	21.66	0	0	98.48	16
A B Barath	14	14	1	394	113	30.3	1	1	64.37	3
A D Mascarenhas	20	13	2	245	52	22.27	0	1	95.33	4
A M Rahane	67	65	2	2093	111	33.22	2	15	79.61	37
A R Patel	22	15	6	91	17	10.11	0	0	95.78	12
A T Rayudu	31	28	7	952	124	45.33	2	5	78.09	10

Table T1 : ODI Batting Stats - Showing 9 of 268 records

The value of k is the number of clusters desired, and since there was no foolproof way to determine the number of clusters the project needed (to be explained in further sections) a trial-error approach was followed.

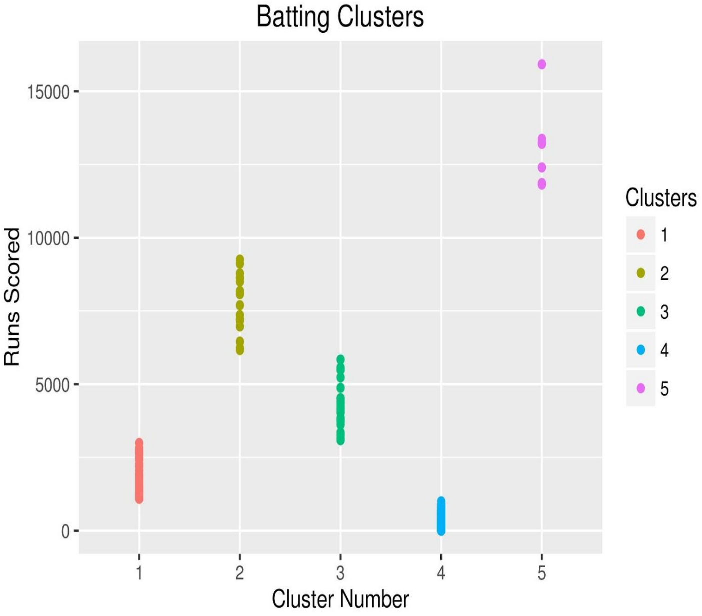
When it came to clustering based on batting statistics, k was set to 5, and in case of bowling statistics k was set to 4. The sole reason of trying different values of k was that the project required some amount of diversity among the clusters. Initially, one particular cluster of batsmen contained bowlers (since their batting figures are generally low), but at the same time, this same cluster contained upcoming batsmen who were just beginning their career. Their statistics too are comparable to that of the bowlers, but there still is a logical distinction between the two sets of players - Bowlers and upcoming Batsmen. Increasing the number of clusters solved this problem and provided more intuitive groupings of players.

Name	Balls	Maidens	Runs	Wickets	Best Bowling	Average	5W	10W	SR	Economy
A Mishra	4229	95	2208	65	5-71	33.96	1	0	65.06	3.13
A D Russell	138	2	104	1	1-73	104	0	0	138	4.52
A Nel	7630	280	3919	123	6-32	31.86	3	1	62.03	3.08
A B McDonald	732	40	300	9	3-25	33.33	0	0	81.33	2.45
A Flintoff	14747	502	7303	219	5-58	33.34	3	0	67.33	2.97
A Symonds	2094	81	896	24	3-50	37.33	0	0	87.25	2.56
A D Mathews	3312	129	1534	30	4-44	51.13	0	0	110.4	2.77
A Kumble	40850	1576	18355	619	10-74	29.65	35	8	65.99	2.69
A Nehra	3447	122	1866	44	4-72	42.4	0	0	78.34	3.24

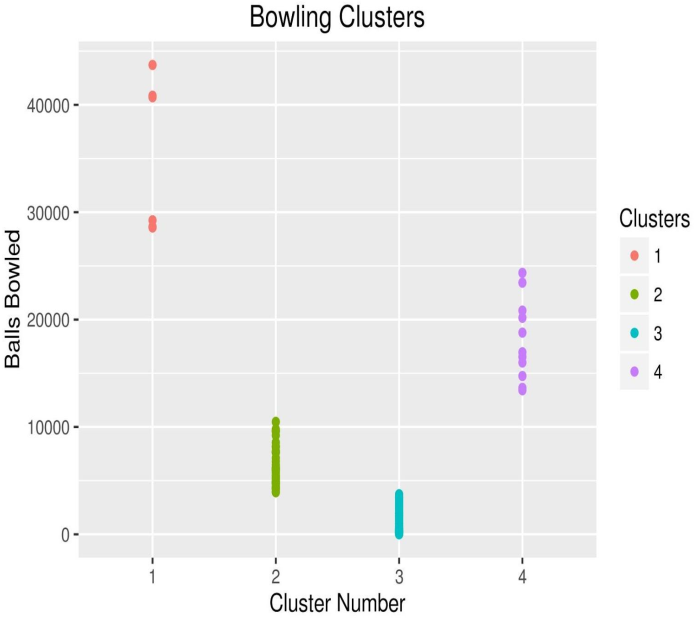
Table T2 : Test Bowling Stats - Showing 9 of 184 records

After a cluster verification process, a first-hand inspection of the data points in each cluster revealed that the overall quality of players in a given cluster were similar, and across clusters this ‘overall quality’ either increased or decreased with variables of the game. In case of batsmen, as observed in plot 1 we notice that cluster 5 is a class apart with respect to runs scored and in case of bowlers, as observed in plot 2 we notice that cluster 1 is a class apart in terms of balls bowled which implies experience. This quality increase or decrease was consistent for the clusters for all the variables of the game such as number of 100s,number of 50s in case of batsmen and wickets taken in case of bowlers. The next logical step was to label these clusters.

In general, as observed, one cluster was a class apart and contained the legends of the game, players who had enjoyed a long and illustrious career. This cluster was labelled the cluster of Legends.



Plot P1 : Cluster Plot-Test Batsmen



Plot P2 : Cluster Plot-Test Bowlers

Among bowlers, the legends were the likes of M Muralitharan, S K Warne, A Kumble and Harbhajan Singh while the batting legends consisted of S R Tendulkar, R T Ponting, J H Kallis,D P M Jayawardene and K C Sangakkara among others.

One cluster below this, was the set of people who were currently in their prime but not yet quite as good. This cluster was labelled as Exceptional.
Similarly, the other clusters were labelled Good, and Average with the players' quality being true to the name of their cluster label.

This whole clustering process was done for batsmen and bowlers across different formats of the game i.e Test Matches, ODIs, T20s (Both Domestic and International)

Name	Matches	Innings	Not Outs	Runs	High Score	Average	# of 100s	# of 50s	Strike_Rate
D P M Jayawardene	443	413	38	12381	144	33.01	18	75	78.67
J H Kallis	323	309	53	11550	139	45.11	17	86	73.14
K C Sangakkara	397	373	40	13975	169	41.96	25	90	78.88
R T Ponting	374	364	39	13589	164	41.81	29	82	80.19
S R Tendulkar	463	452	41	18426	200	44.83	49	96	86.23
S T Jayasuriya	441	429	18	13364	189	32.51	28	68	91.27
S C Ganguly	308	297	23	11221	183	40.95	22	71	73.65

Table T3 : ODI Legends; after k-means clustering of batsmen

Name	Balls	Maidens	Runs	Wickets	Average	Strike_Rate	Economy
A Kumble	40850	1576	18355	619	29.65	65.99	2.69
D L Vettori	28652	1194	12330	361	34.15	79.36	2.58
G D McGrath	29248	1471	12186	563	21.64	51.95	2.49
Harbhajan Singh	28580	870	13537	417	32.46	68.53	2.84
M Muralitharan	43715	1786	18023	795	22.67	54.98	2.47
S K Warne	40704	1761	17995	708	25.41	57.49	2.65

Table T4 : Test Legends; after k-means clustering of bowlers

What resulted from it were about 25 .csv files, each containing the names of players in a particular cluster of a particular format, for either batting or bowling.These .csv files were used to do a complete performance analysis/classification using basic Set Theory.

For example, the list of players who were legends in both ODIs and Tests would be represented as :

ODIandTestBattingLegends □ LegendODIBatting.csv ∩ LegendTestBatting.csv

Many more of these classifications were made, such as :

- ➔ All Round Specialists - Players who were either Exceptional or Legends in all the three formats.
- ➔ Exclusive Specialists - Players who were Legends/Exceptional only in one format but not others.
- ➔ Dual Specialists - Players who were Legends/Exceptional in only two of the three formats.

A combinatorial process was followed for each of these classifications in the sense that, every possible combination of the three formats was considered while building these sets.
The end result was a detailed description of each player's strengths and weaknesses (in terms of format), which in itself provides a good knowledge base for the purposes of team selections.

Main observations that we noted were that K C Sangakkara and D P M Jayawardene were the best of the lot across formats amongst batsmen, and when it came to bowlers Harbhajan Singh and Muttiah Muralitharan were the 'complete' bowlers.

Apart from that, quite a few interesting and sometimes unexplainable observations were made by examining these classifications.

Some notable ones were :

- Only one Asian (V V S Laxman) was a legend in Tests but not in ODIs, everybody else was from outside the sub-continent.To make it even more interesting, when analysis was done to look for legends in ODI but not in Tests, only one cricketer (A C Gilchrist) was not Asian.
- Muttiah Muralitharan was better than Shane Warne (on a purely statistical level). He had performed well across all formats, including T20, whereas Warne had not been exceptional.

- Harbhajan Singh was in the Legends cluster across all formats, which depicts him as a complete bowler (and yet his inclusion is not always guaranteed in the Indian Team.)

- Shikhar Dhawan was found to be one such player who excelled at Domestic T20s, but not on the international level (That can imply that he was selected into the Indian squad purely on the basis of domestic performance)

- The set of exclusive Test Legends was a null set. It is proof of the fact that Tests are considered the hardest format, and that excelling in them suggests being a good player in general (one who can easily do well in other formats too.)

In order to move over to the next section, some additional modifications were made. The clusters were merged into 3, to serve as training data. The legends and exceptional players were put together, since (as the next section will explain) now there was no logical distinction between the two. The second cluster was the one with the good players, who were quality-wise one grade below the top-notch. All the remaining players were put into the third and final cluster as Average players. This was done for both Batting as well as Bowling. This resulted in having three classes, labelled ‘Great’, ‘Good’, and ‘Average’

3. Quantifying the adaptation of a player across multiple cricketing formats and developing a new formula for player evaluation/rating.

We now introduce the concept of a ‘Performance Matrix’ for a player, based on the classifications obtained. This Performance Matrix is used to quantify the adaptation of a player across multiple cricketing formats thereby introducing a metric for the same. This ‘Adaptation Factor’ given to a player is further used to develop a formula for overall player evaluation.

The main reason a classification number of 3 was chosen was because, while developing the adaptation factor for a player it is unfair to distinguish between ‘Legends’ and ‘Exceptional’ players. This is because nearly all the players classified as ‘Legends’ are now retired and most of the ‘Exceptional’ players are on track to become legends of the game.

Developing a Performance Matrix for quantifying player adaptation across multiple cricketing formats.

Performance Matrix for any player is a 3x3 matrix with the columns representing the 3 different formats of cricket which are Test, ODI and T20 and the rows representing the 3 different classifications of players which are ‘Great’, ‘Good’ and ‘Average’. This is consistent for both Bowling as well as Batting and every player has two matrices associated with him.

The developed Performance Matrix is given below.

	Test	**Odi**	**T20**
Great			
Good			
Average			

In a matrix P, where i represents the columns which are the formats and j represents the rows which are the category, P[i][j] is 1 if a player belongs to the j^{th} category in the i^{th} format. Also, all P[i][x] = 1 where x > j

In case a player is classified as ‘Great’ in a particular format then it also means the player has passed the barriers of ‘Average’ and ‘Good’ in his career, which is why a 1 is filled for the the ‘Good’ and ‘Average’ row in that format as well.

As an example, we develop the Batting Performance Matrix of Virat Kohli who plays for the Indian National Team. As per previous results, Virat Kohli was a ‘Great’ in ODIs and T20s, and ‘Good’ in Tests.

Virat Kohli’s Batting Performance Matrix is as follows :

	Test	**Odi**	**T20**
Great	0	1	1
Good	1	1	1
Average	1	1	1

Since he falls under the ‘Great’ class in the ODI and T20 format, we fill a 1 for all the respective cells. However in case of the Test format he is classified as ‘Good’ and hence a 0 is first filled in the cell corresponding to the ‘Great’ row.

Using this concept of a Performance Matrix, it is now possible to quantify the adaptation of a player across multiple cricketing formats.

Before doing so, we formally identify the semantics of the Adaptation Factor.

Apart from the primary format in which a game of cricket is being played, secondary and tertiary formats must also be taken into account while evaluating a player. This is due to the fact that situations may arise where the player is required to adapt to one of the other formats. The adaptation factor quantifies this ability of a player to adapt, and is a numerical value between 0 and 1, with a more positive value indicating that a player is likely to change his game and fit in well when a different situation arises.

As an example, we now calculate the Batting Adaptation Factor of Virat Kohli in ODIs.

We first disregard the ODI column in Virat Kohli’s Performance Matrix. The logic behind this step is the fact that we are looking to find how well a player is going to adapt to the ‘Secondary Format’ and the ‘Tertiary Format’ and the player’s skills in the ‘Primary Format’ do not affect this.

Hence, now we are left with two columns as shown.

	Test	**T20**
Great	0	1
Good	1	1
Average	1	1

The adaptation factor is simply defined as :

α = Total Number of 1s / Total Number of possible 1s

We can now find Virat Kohli’s adaptation factor in the ODI format given as:-

$$\alpha_{ODI} = \frac{5}{6} = 0.83.$$

In terms of percentage, when Virat Kohli is playing an ODI match and there is a scenario where he needs to change his game approach to the T20 format or the Test format, there is a 83% chance that he will adapt well to the scenario.

This high Adaptation Factor for Virat Kohli is justifiable considering the numerous occasions he has won the game for Team India when the team faced various match scenarios.

An example of such a situation was the 6th ODI during the Australia tour of India in 2013 wherein Australia posted a mammoth total of 350 in the 50 overs. This meant India had a target of 351 runs which meant the team required the players to adapt to the T20 format of the game. Virat Kohli responded well by strongly adapting which was evident from his 100 of just 61 balls which played a strong factor in India winning the match and making history as one of the biggest run chases of cricket.

Using the aforementioned methodology, we computed the adaptation factor α (both bowling and batting) for all players in our records, across all formats. Next, we used these values in conjunction with a formula we developed to evaluate each player.

Name	α_{Test}	α_{ODI}	α_{T20}
V Kohli	1	0.83	0.83
Yuvraj Singh	1	0.83	0.83
A B de Villiers	1	1	1
K C Sangakkara	1	1	1
A J Finch	0.83	0.50	0.33
A C Gilchrist	0.83	0.67	0.83

Table T5 : Batting Adaptation Factor

Name	α_{Test}	α_{ODI}	α_{T20}
Harbhajan Singh	1	1	1
B Lee	1	0.83	0.83
D W Steyn	0.83	0.83	0.67
M G Johnson	0.83	0.83	0.67
M Morkel	0.50	0.83	0.33
M Muralitharan	0.83	0.83	1

Table T6 : Bowling Adaptation Factor

Showing 6 out of 487 records in each table

We now define and introduce this formula to evaluate a player as :

$$\boldsymbol{\beta_X = k_1 * (PFR) + \alpha_X * (k_2 * SFR + k_3 * TFR)}$$

Where,

- β_X is the final rating value assigned to the player, after considering his adaptability, for a format X.
- k_1, k_2, k_3 are constants which act as weights given for each format.
- PFR is the **P**rimary **F**ormat **R**atings, as given by the International Cricket Council (ICC) for that year. Similarly, SFR and TFR are Secondary and Tertiary format ratings for that year.
- α_X is the adaptation factor as obtained from the prior computations, for the format X.

The optimal value for the weights were decided to be :

- 0.55 for the PFR
- 0.40 for the SFR
- 0.05 for the TFR

These weights were decided as a result of logical reasoning, and a detailed discussion about the magnitudes of influence of each format.

To add more clarity to the Ratings, we use the official ICC Player Ratings which is a moving average calculated after every match and made public by the ICC. This was web-scraped from www.relianceiccrankings.com. Since the ratings of the player change throughout the year, the values of July were used for every year as it is approximately mid-year and hence serves as a good estimate of the overall year's performance of the player. The player ratings for the top 100 players from 2007 - 2016 were web-scraped for the three formats of the game for both batsmen as well as bowlers.

Name	Rating
V Kohli	897
A J Finch	871
A D Hales	860
F du Plessis	795
C H Gayle	732
M D K J Perera	707
S K Raina	677

Table T7 : 2015 International T20 ICC Ratings For Batsmen

Name	Rating
N W Bracken	786
D L Vettori	765
S E Bond	714
W P U Vaas	693
M Muralitharan	687
B Lee	677
A Nel	677

Table T8 : 2008 International ODI ICC Ratings For Bowlers

Name	Rating
J H Kallis	883
S R Tendulkar	874
K C Sangakkara	838
I J L Trott	833
A N Cook	821
V Sehwag	732
T T Samaraweera	766

Table T9 : 2011 International Test ICC Ratings For Batsmen

Showing 7 of 100 records in tables T7,T8, and T9

In the next section we establish a proof of concept where we evaluate the legitimacy of this formula based on the complete team's performance.

4. Finding the correlation between average team rating and team performance.

It is quite intuitive that a team with a higher average player rating is bound to perform better. For the sake of pure statistics and analysis, we can disregard external factors such as emotions, pressure, team morale among other things which may influence the team's performance.

In order to justify the impact of the player rating formula described earlier, it was decided that a good test would be to statistically correlate the team's rating along with its margin of win or loss (in terms of runs). If the correlation was found to be positively strong, it shows that the ratings are a good judge of the team's performance, which is intuitively expected.

To put it in a concise form, we asked the question "Does having a better rating imply better performance?"

Another round of web-scraping was done, this time from www.howstat.com, from where we accumulated details of all the ODI matches India have played since 2007. For each match recorded, ratings of each player of the India XI were calculated (using the formula described previously) and then the average team rating was computed. When these values were correlated against the margin of win or loss (A loss by 100 runs is considered as -100 while a win by 100 runs was considered as +100) a shocking negative correlation was obtained. This meant that better the team rating, worse the team performance was; which was extremely non-intuitive. The player evaluation formula had failed the test.

It so happens in cricket, that teams tend to rest their key players while competing against lowly opponents. We deduced that this was the flaw in the test of legitimacy. The Indian team is usually devoid of its main players while touring minnows like Bangladesh or Zimbabwe. This results in a lower rating than normal as a lot of youngsters are given a chance, and their ratings are not even close to that of the star players they're replacing. But, this weakened national team still ends up defeating the minnows, hence leading us towards an inconclusive correlation coefficient. Also, on the other hand, even the best possible team sometimes ends up losing to a team that is simply better than they are. It was understood that the average player rating was not a good metric to test out the formula.

In order to fix this defect, we came up with a new solution. Instead of taking into consideration just the Indian team's rating for that match, the *difference* of the two team's ratings was taken. In the cases where India's opposition was stronger than them, it would result in a negative value and a loss would be justified.

A few more modifications were done to the data set :

- Only the top 8 teams were considered. Any matches against Bangladesh, Ireland, Zimbabwe, Netherlands, Kenya were disregarded, purely because most of these players were inexperienced and didn't have too much exposure to International Cricket at the highest level. They are almost never a part of the yearly ICC ratings, and it was a hard task to give them a fair value in order to evaluate them.

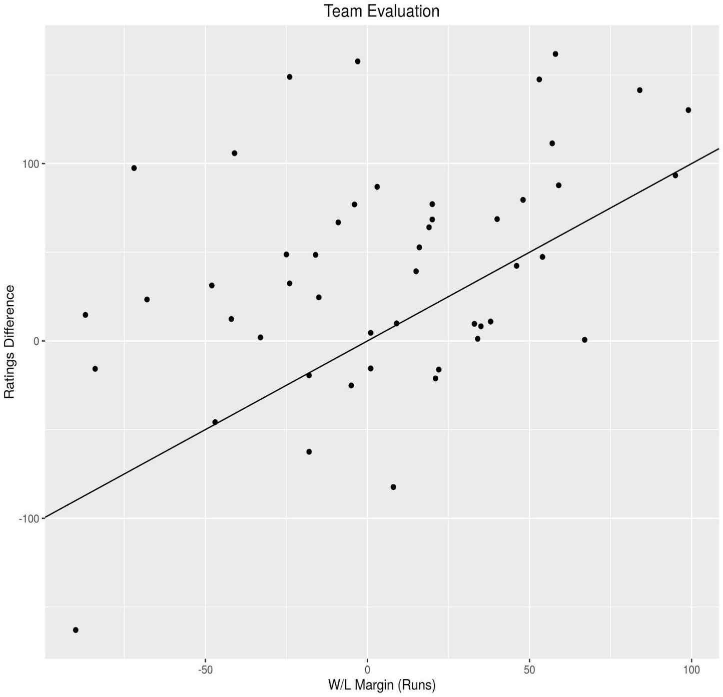
- Marquee tournaments such as the World Cup and the Champions Trophy were not considered since they have a huge public appeal and are high pressure matches.

- Matches where India had won or lost by more than a margin of 100 runs, were removed. This was accounted for, by the explanation that a loss of more than a 100 runs is considered very rare for a top team and cannot be explained by pure statistics. Reasons could include fatigue, bad luck, or just the fact that it was not their day. Cricket is not a very predictable game.

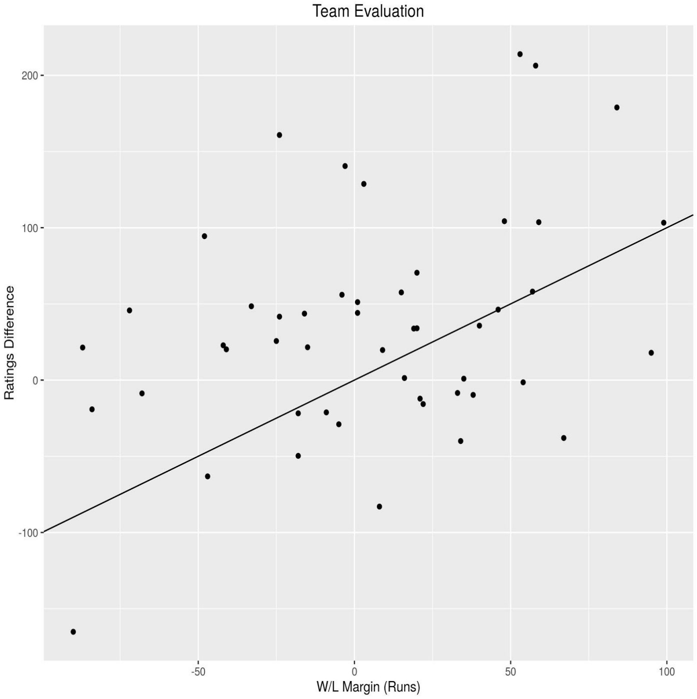
- India vs Pakistan matches are always electrifying experiences. There is always a lot of emotion involved since the competitive rivalry goes back a long time (and not just in cricket). We presumed these emotions had the potential to both enhance and diminish an individual's performance. Hence, these matches were not counted either.

The same test was again performed, this time plotting the difference of ratings and also with the use of this new and more practical data set. The logical changes that were made turned out to make a strong difference.Initially the test was done using ICC's player rating of the year,excluding the adaptation factor and the formula developed in the project. A correlation coefficient of 0.36 was obtained, which is positive and indicates a relation between ratings and the team performance.

The test was run again,this time using the developed Adaptation Formula that was defined in the project. In order to justify the significance of the Adaptation Factor and validate our hypothesis, we needed a stronger correlation coefficient than the existing ICC rating. On performing the test, the correlation coefficient was found to be 0.43. We conclude that the influence of the adaptation factor is a significant metric in the domain of cricket player evaluation.



Plot P3 : Pearson's product-moment correlation of **0.4341**
p-value = 0.0018
95 percent confidence interval:
0.1742 - 0.6375
(Using developed formula)



Plot P4 : Pearson's product-moment correlation of **0.3649**
p-value = 0.0099
95 percent confidence interval:
0.0932 - 0.5859
(Using pure ICC Ratings)

As the plots show, the slope of the best-fit line is more positive when the formula involved the adaptation factor. When the formula was removed, and the players were evaluated only on their ICC rankings for the year, plot P4 was obtained. It is understandable that a more positive slope of the best-fit line is more desirable.

A statistically significant correlation coefficient of 0.43 was a good measure to legitimize the hypothesis, but to further strengthen it, we observed that removal of a couple of outliers pushed up the value to 0.6. However, there was no logical backing to completely discard these data points from the analysis as there was no common connection between them. But the mere fact that there was a jump in rise in the value leads us to an exciting finding that on some level it is possible that the outcome of a game of cricket (which is considered unpredictable by many) can be mathematically modeled.

Inspecting the impact of each team on the correlation coefficient

In order to observe the impact of each team's contribution to the overall correlation, the same test was run but with different data sets, each without the matches played against a particular team.

It was observed that on some occasions the value was higher than the default, while in other cases in dropped lower. But usage of the adaptation formula always resulted in a higher correlation than when the ICC ratings were used thereby showing the strong significance and importance of the adaptation factor.

Team Excluded	**Correlation (Using ICC Ratings)** All teams : 0.3649	**Correlation (Using developed formula)** All teams : 0.4341
England	0.4296	0.4677
South Africa	0.3228	0.3505
Australia	0.3914	0.5214
Sri Lanka	0.4227	0.4944
West Indies	0.3586	0.4299
New Zealand	0.2424	0.3413

Table T10 : Correlation coefficients with modified data sets

5. Conclusion

While there exist several methodologies of player evaluation in cricket, we are not aware of any system that takes into account a player’s adaptation in multiple formats and this project introduced this Adaptation Factor metric for all three formats of the game. The clustering approach towards developing a Performance Matrix as well as a set theory approach for finding ‘Single Format Specialists’ and ‘Multi Format Specialists’ which was introduced in this project also improves player evaluation and assists in finding major player competition. A formula for overall player evaluation was developed as well which lead to an exciting finding which showed that there is a prominent positive correlation between team rating and team performance. The fact that all the teams taken into consideration for the analysis have all been ICC World Cup winners,finalists or semifinalists made the finding more valuable.The strong impact of the Adaptation Factor was shown when the developed Adaptation Formula gave a stronger correlation as compared to the Reliance ICC ratings which does not take into account the Adaptation Factor of a player.

In spite of the various intangible entities that play a role in cricket such as player mindset, team strategy and team morale, the positive correlation result and the developed graph showed that the performance outcome in the game of cricket can be reduced to a mathematical model at some level.