

Big Talk

BadeMian reflects on yesterday, today and tomorrow



Advertisements

Today's Mortgage Rate

3.04% APR

30-Year Fixed	3.71% APR	→
15-Year Fixed	3.04% APR	→
5/1 ARM	3.70% APR	→

lendingtree [Calculate Your Payment](#)

Terms & Conditions apply. NMLS#1136

[Report this ad](#)

Cricket analytics

A few months ago I was invited to give a talk at the Chennai Mathematical Institute. I was free to pick any subject .. but it had to be connected in some way to cricket! And since I was speaking at an elite math institute there also had to be some curves and equations.

I guess I got unduly ambitious and decided that I would do a quick review of everything I know in cricket analytics. What's 'cricket analytics' you might ask. And how is it different from 'cricket statistics'?

Let me explain. I view cricket analytics as something with a little more modeling and a little more data processing. 'Cricket statistics', on the other hand, would just be lists of 'highest', 'lowest', 'first', 'last' cricketing numbers. For example a list that goes Brian Lara 400*, Matthew Hayden 380, Brian Lara 375, Garfield Sobers 365* ...

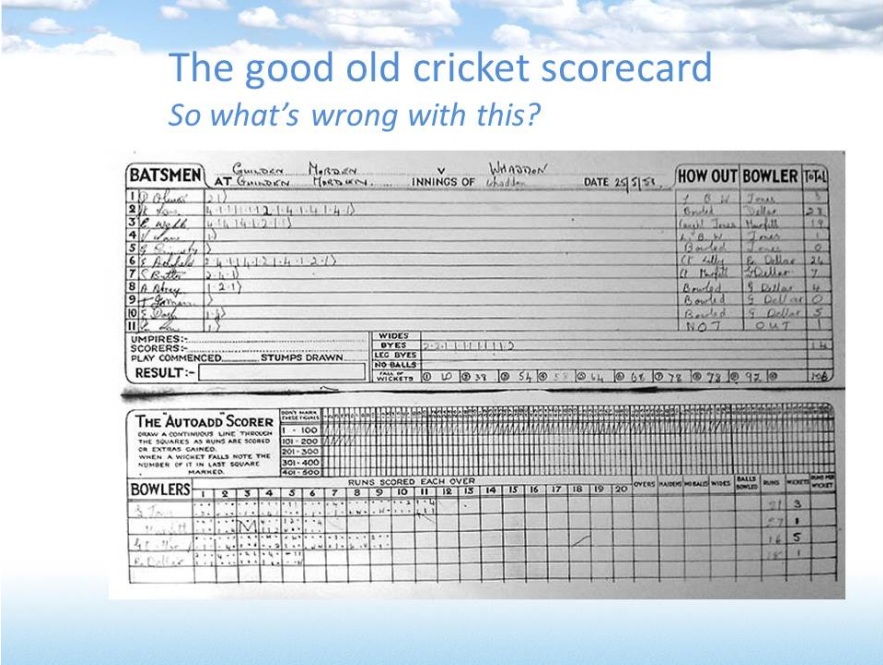
To nobody's surprise my talk went on and on. To my pleasant surprise most of the audience stayed behind till I reached the end. They asked some great questions too ... and occasionally caused me great embarrassment. Here's a sampler: "That photo that you are trying to pass off as Lance Gibbs is actually Joel Garner!".

It was!

In this post I plan to go over my entire talk. There were almost 20 slides, so this narration is going to take some time. I must also add the disclaimer that this is my view of cricket analytics as it was yesterday, is today, and should be tomorrow. I'm sure others can say it better. As Amitabh (Sahir) sang in Kabhi Kabhie:

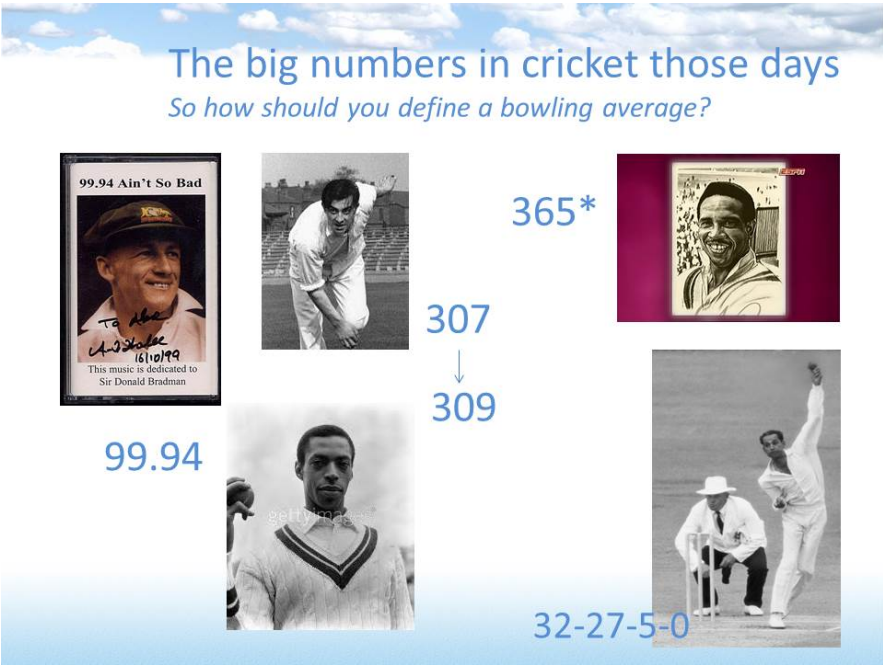
कल और आयेंगे, नग्मों की खिलती कलियाँ चुननेवाले
मुझ से बेहतर कहनेवाले, तुम से बेहतर सुननेवाले

I can't think of a better place to start than with the good old cricket score book. As a little schoolboy I had an intimate association with this book and my happiest childhood moment was when M L Jaisimha leaned over my right shoulder to ask: "How many sixes did Santosh Reddy hit today?"



For a long time I thought that there can't be a better creation than this humble score book but I now realize that it has one really serious weakness: *it doesn't connect the batsman to the bowler on a ball-by-ball basis*. It can tell you that Tendulkar hit a six, it can tell you that Kumble was hit for a six but it cannot confirm that Tendulkar hit that six off Kumble. The only time the printed score card links the batsman to the bowler is when the bowler dismisses the batsman. This weakness has now been corrected by using relational databases.

Let me next elaborate on what I have called 'cricket statistics'. Here are some cricket records that truly excited me when I was a schoolboy and a teenager ... the kind of numbers that my father or uncle passed down to me.

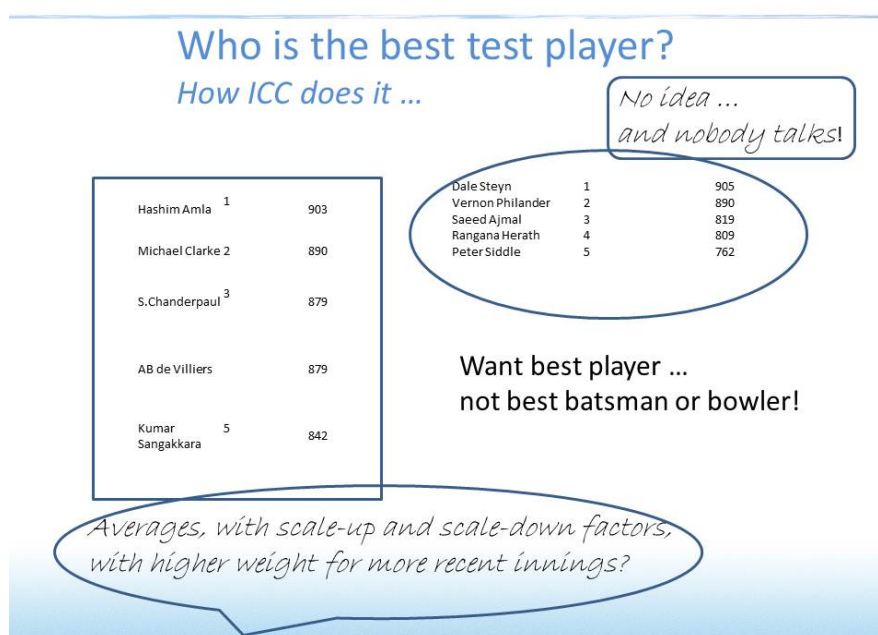


The most romantic story was about how Bradman was bowled by Hollies for a duck ... and how if he had even reached 4 in his last innings, his test average would have been 100.

Then there's the romance of Garfield Sobers scoring 365* to overtake Len Hutton's 364, Lance Gibbs overtaking the fiery Freddie Trueman's record of 307 test wickets ... and our own Bapu Nadkarni bowling maiden after maiden in a [Chennai test match](#)! There were other big numbers: Hanif scoring 499 in a first class match and getting run out as he attempted to get to 500, Jim Laker taking all 10 wickets (Anil Kumble would later equal that record) and so on.

Notice that we are talking of just numbers, with practically no 'processing'. The batting average (total runs scored divided by number of completed innings) and the bowling average (total runs conceded divided by total wickets taken) do involve some division, but they have weaknesses; especially the bowling average that doesn't recognize a bowler's run containing ability (Nadkarni, for example, must have hated it).

More interesting questions like 'who is currently the best test player' were rarely asked ... possibly because such questions didn't have easy answers and required much more processing. It wasn't until the mid-1980s that we started asking such questions with the introduction of the Deloitte ratings (the name changed many times over the years: Deloitte -> Cooper Lybrand -> Pricewaterhouse Coopers ... eventually to Reliance ICC rankings). I remember there was quite a flutter when [Dilip Vengsarkar briefly topped the list](#).



Rather surprisingly (or perhaps not so surprisingly!) a description of the ratings formula is not easily available. But based on a study of the numbers, and some inadvertent leaks on the Internet, it would appear that the rating is an intelligent manipulation of the average: start with the average and scale it up or down based on factors such as quality of opposition, state of the match etc. It is also almost surely configured to give greater weight to more recent performances.

There are also apparently different formulae for bowlers and batsmen, although there is an effort to keep them in the same ballpark.

The Reliance ICC team also produces rankings for the best ODI batsmen, bowlers and all-rounders, as always with their hallowed formula shrouded in mystery. I have never liked the idea of hiding the formula. In fact I worked with Professor M J Manohar Rao (who alas passed away 10 years ago) to create the most valuable player index (MVPI) that we continue to [publish](#) on rediff.com.

Who is the best ODI player?

This has more variables!

	Player Name	Runs	Sixes	Fours	Str. Rate	Bat-pts	W	Eco	Bowl-pts	Field-pts	M	MVPI
1	Suresh Raina	543	9	44	84.7	552	3	5.4	69	65	13	686
2	MS Dhoni	486	15	39	88.7	515	0	-	0	113	13	628
3	Ravindra Jadeja	168	6	11	84	169	13	3.3	426	28	7	623
4	Virat Kohli	464	4	41	83.3	464	0	5.9	-6	33	13	491
5	R Ashwin	90	1	8	79.6	86	15	4.8	398	5	13	489
6	G Gambhir	419	0	43	75.9	378	0	-	0	5	13	383
7	B Kumar	53	0	7	91.4	58	9	4.3	280	5	8	343
8	Ishant Sharma	14	0	2	43.8	1.3	12	4.8	316	0	8	317
9	Irfan Pathan	76	0	6	92.7	84	8	5.3	189	5	5	278
10	V Sehwag	183	0	22	91	199	1	4.7	29	23	6	250
11	Yuvraj Singh	160	1	23	90.9	173	0	6.4	-17	35	8	191
12	Manoj Tiwary	86	0	8	81.1	84	4	6.8	80	25	2	189
13	Ashok Dinda	3	0	0	50	1	7	6.3	115	5	6	121
14	Shami Ahmed	1	0	0	9.1	-7	4	4.6	116	0	5	109
15	Rohit Sharma	104	1	13	68	81	0	3.7	12	5	8	98
16	Zaheer Khan	2	0	0	50	0.7	4	5.6	75	15	5	91
17	Ajinkya Rahane	64	0	8	62.1	42	0	-	0	40	5	82
18	Rahul Sharma	0	0	0	-	0	1	5.6	20	0	1	20
19	Umesh Yadav	0	0	0	-3	2	8.1	6	0	2	4	
20	Pragyan Ojha	5	0	0	27.8	-5	1	7.5	8.2	0	2	3

The Castrol index does approximately the same thing!

The Impact Index is a new idea

MVP + II might be a great idea!

MVPI: most valuable player index

Collapse performance to a 'run equivalent'

We've remarked earlier that having different rating schemes for batsmen and bowlers isn't ideal; it would be much nicer if we had a single index that measures overall cricketing ability, instead of measuring batting, bowling and fielding ability *separately*. Our MVPI does this: it collapses a player's batting, bowling and fielding performances into a 'run equivalent': so every performance can be expressed in terms of 'runs' — even if it initially feels awkward to say that Anil Kumble's bowling performance was equivalent to a score of 133 'runs'.

Remember too that in ODI (or T20) matches it isn't enough for a batsman to just score a lot of runs; he must also score them briskly. Likewise it isn't enough for a bowler to take wickets; he must also concede as few runs as possible. So how would we handle this?

Let us suppose that Sehwag scores 50 in 30 balls and Dravid scores 50 in 45 balls. In an ODI context Sehwag's 50 would seem like 75 while Dravid's 50 may seem like 55. The MVPI formula therefore gives Sehwag and Dravid 'bonus' runs; with obviously a bigger bonus for Sehwag because he scored at a faster clip. Of course if someone had taken 80 balls to score 50 then the 50 would have seemed like 35 and there would have been a 'penalty'.

The MVPI therefore adds or subtracts runs to make it conform to a par. An alternative approach, used for example in the [Castrol Index](#), is to multiply or divide the actual score to attain the right par.

The MVPI formulation revolves around such a par criterion. If the par score in an ODI is considered to be 250, then we expect every batsman to score 5 runs in every 6 balls, and every bowler to concede 5 runs per over. We also assume that every wicket a bowler gets is worth 25 'runs'. So if Sehwag scores 50 in 30 balls, when he was expected to score only 25, we give him a bonus of $50 - 25 = 25$ and argue that his 50 is equivalent to 75. Likewise if Kumble has figures of 10-3-17-4, his 4 wickets are worth $25 * 4 = 100$ runs; also he conceded only 17 runs when he was expected to concede 50 ... so we give him a bonus of $50 - 17 = 33$ 'runs' and his overall performance is worth $100 + 33 = 133$ 'runs'.

Let me end by mentioning another very worthy index — the [Impact Index](#) (II) devised by Jaideep Varma and others. This is a relative index in which every performance is evaluated using a 5-point scale. II doesn't really care how many runs a batsman scored, or how many wickets a bowler took; it asks how much *better* a batsman or bowler scored *relative* to other players and therefore how impactful this performance was in the context of the match. I have often imagined how wonderful it would be if we could marry MVPI to II.

So much about ranking players. But what about ranking teams?



It seems elementary that teams that win more matches must be ranked higher. If you win with higher margins, or win the series, your team is probably even better. But most of us would agree that to be called a champion team the team must win against stronger opposition and win matches away from home. These two are the more significant criteria — and that's why I have drawn bigger bubbles to represent them.

Many ranking schemes have been proposed, including the [ranking scheme](#) that we ran for a decade on [rediff.com](#). However the official ICC ranking scheme continues to be the one devised by ICC's David Kendix. The ICC [scheme](#), which accords a higher weight for wins against stronger teams, performs reasonably well, but suffers the weakness of not accounting for the *home-away difference*. To explain away this deficiency, many analysts argue that the difference is neutralized because every country plays the other successively at home and away. This explanation is not quite tenable because the rating value decays with time, and even more because at any given point we may not have equal home-away parity between every pair of teams. The Indian test team looked pretty awful after being drubbed 4-0 away in Australia ... till they returned the 4-0 compliment at home!

Our scheme does modestly better than the ICC scheme because it accommodates both the opposition strength and home-away factors, and I am convinced that a vastly superior rating scheme can be devised simply by tweaking our formula. But I also recognize that ICC will not change its ways ... just see how steadfast they are in backing the D/L rule without ever giving a fair trial to V Jayadevan's rain rule!

In fact that's what we are going to talk about next: rain rules!

When it rains ...

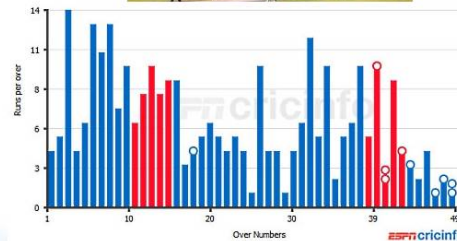
Two 'resources', and the interruption points matter!

India 300/2 in 50 overs
Pakistan 151/9 in 25 overs

Pakistan WIN!

More likely to
win with more
overs and more
wickets. There are
TWO resources!

< 1990



1991-92 WC

Till ODI matches came along, it was okay for cricket matches to end as draws, and indeed a large number of matches were drawn; the very first cricket [series](#) I followed (Pataudi's India versus Mike Smith's England in 1964) ended with five dreary draws! But ODI matches demanded a result *even if the match was interrupted or disrupted by rain*.

Initially the approach was naive (I should perhaps have said 'dumb') and the revised target was *only dependent on the run rate*. But soon captains started getting smart. In particular the wily Arjuna Ranatunga would always opt to field first on a cloudy day if he won the toss. The reason was trivial: it is a lot easier to score 125 in 25 overs than 250 in 50 overs *if all your ten batsmen are allowed to bat*. With the run rate rule, you could encounter absurd situations where Pakistan would win at 151/9 in 25 overs in response to India's 300/2 in 50 overs.

So, before the 1991-92 World Cup in Australia and New Zealand, they decided to come up with a new rain rule — and they apparently requested the venerable Richie Benaud to devise something sensible. Now Benaud was a great captain, and is a greater commentator, but analytics was clearly not his cup of tea. He came up with something that seemed profound: he said that if India scored 300 in 50 overs, and Pakistan had only 25 overs to bat ... then Pakistan must score what India scored in their 25 *most productive overs* out of 50.

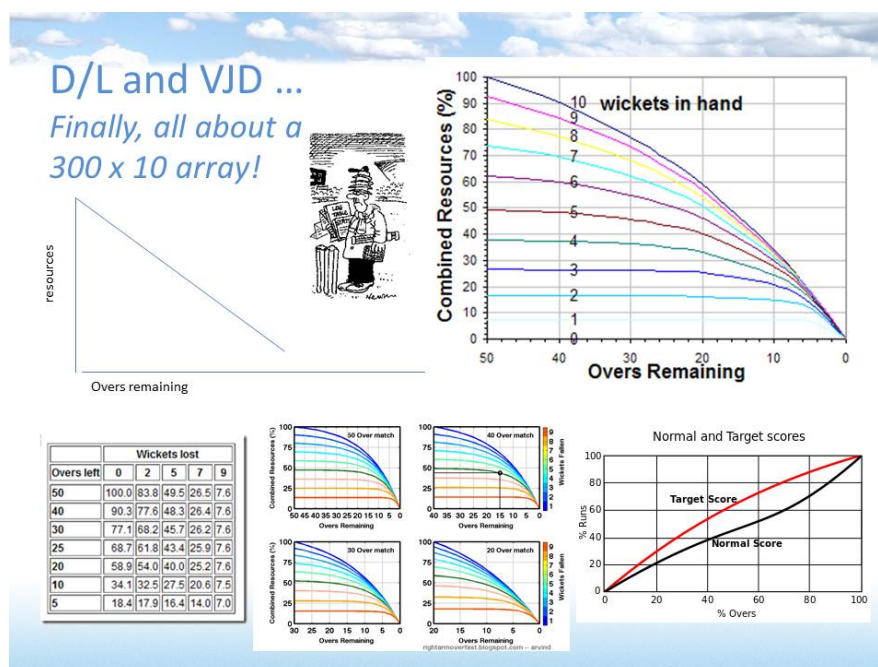
Think of the Manhattan, that plots the runs scored in overs 1,2,3 ... up to 50. Now rearrange the Manhattan so that the tallest building comes first on the left, then the next tallest and so on. Benaud's rule effectively said that if 25 (or more generally 'x') overs were lost, cut off the runs equivalent to the height of the 25 ('x') shortest buildings at the right end of the rearranged Manhattan, and score as much as the height of the 25 (or 50 - 'x') tallest buildings to win.

I hope I didn't confuse the reader. As an example, consider the shocking example of the England-South Africa semi-final in 2002. Batting first, England batted for 45 overs (so England's Manhattan only had 45 buildings). When South Africa were chasing, they needed 22 to win in 13 balls when there was a rain interruption. Two overs were deemed to be lost. So the two shortest buildings at the right end of the Manhattan, respectively of height 0 and 1 runs, were cut off. The balls reduced from 13 to 1, but the target reduced only from 22 to 21!

It was this sorry mishap that cleared the way for the D/L method devised by Frank Duckworth and Tony Lewis. D/L's chief merit was that it recognized that targets must be set based not just on 'overs remaining', but on a combination of the 'overs remaining' and 'wickets remaining' resources. The D/L method could also seamlessly accommodate interruptions at any stage of the match, and multiple interruptions. After the horror of Richie's method, Frank and Tony's method was much much better!

I have [written](#) about the D/L method innumerable times — and even been mentioned in the D/L book — and I have no doubt that D/L constituted a very big step forward in solving the rain rule problem. My only grouse is

that ICC has shut the rain rule door after D/L and not given worthy challengers like Jayadevan's VJD method a fair chance.



What really puzzles me is the way D/L is hyped and considered to be very complicated. I can understand if a cricket commentator with a BA in English, who only studied Chaucer and Shakespeare, throws up his hands in despair. But the average cricket fan mustn't let the method defeat him (as India waited to defeat Pakistan in the Champions Trophy, someone messaged on Cricinfo that he finds his college calculus simpler than D/L!).

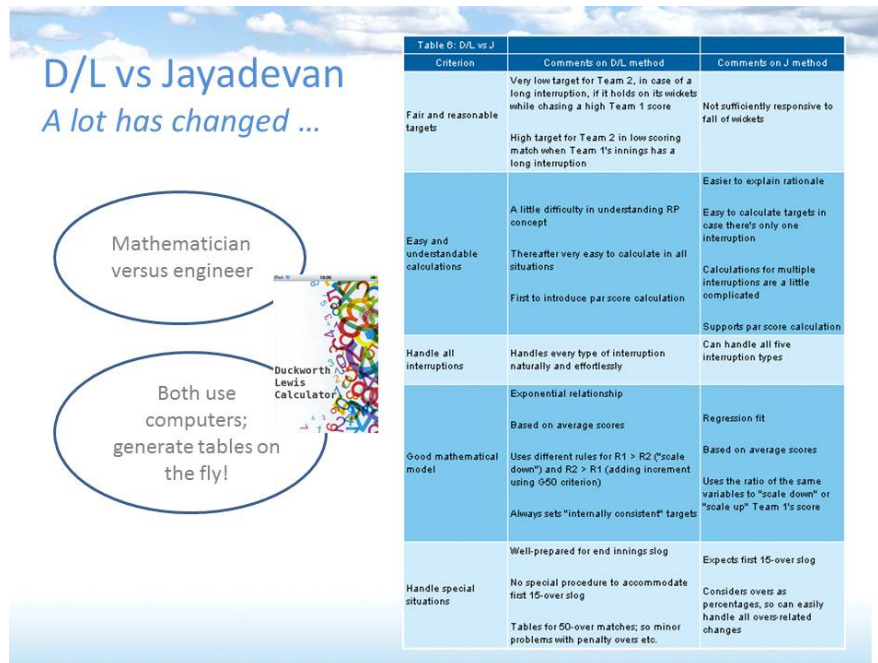
The general idea is to think of a 'resource'. When an ODI innings is starting, the batting side has all its 50 overs and 10 wickets available. You therefore say that it has all its 100 percent resources available. As overs deplete and wickets fall this resource diminishes. D/L merely creates a table that tells you how this resource diminishes from 100% to 0% as the innings advances from 50 overs to 0 overs and from 10 wickets available to 0 wickets available. It is essentially a table with 300 rows (for balls) and 10 columns (for wickets).

That's one part of D/L; the other part is to determine how to reset the target using the resource table if there is an interruption. Remember that in the era of simple run rates we calculated the target by looking at the ratio of overs available to both the teams at the time of interruption. Example: India bats all 50 overs to score 300, and Pakistan are 100/0 after 20 overs when the match is abandoned. Pakistan's par target then would have been: $(20/50) * 300 = 120$ and they would be declared loser. Now, instead of a ratio using overs, we use a ratio using the resource percentage. In our example, while India used up 100% of its resource, Pakistan might just have used up 30% at the time of interruption (remember they have all 10 wickets in hand, and the resource percent judiciously combines overs used and wickets lost) and their par target might just be $(30/100) * 300 = 90$. So they'd be declared winner by D/L.

This is the key idea. Of course things get complicated if the first innings of the ODI match is itself interrupted, or when there are multiple interruptions, or when scores tend to be too high or low ... but the D/L rationale is always to set the target by comparing the resources used up.

To summarize: the D/L method uses (a) a 300 x 10 table of resource percentages and (b) a rule or a formula to reset targets in every interruption situation.

So what does V Jayadevan's method do? Well, Jayadevan proposes his own criterion to populate the 300 x 10 table, and a different rule to reset targets based on what he calls the normal and target curves. I have [described](#) his method elsewhere in considerable detail. Essentially Jayadevan tries to repair Benaud's Manhattan project involving most productive overs. Jayadevan recognized that the most productive overs criterion was perfect if the interruption happened between innings, but misbehaved badly for interruptions within innings. He therefore sought to correct that misbehaviour.

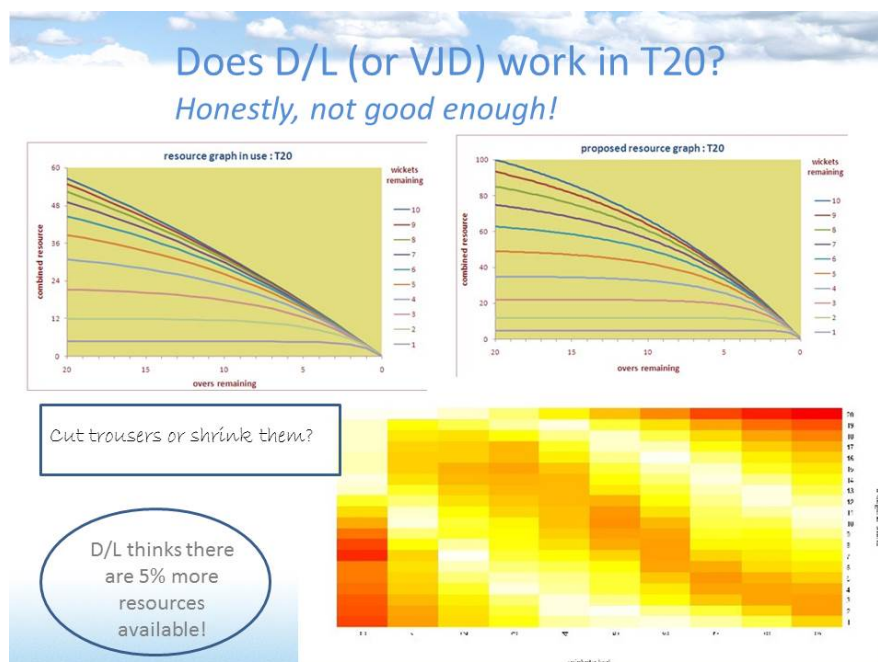


The D/L vs Jayadevan debate has gone on for over a decade, and is frequently seen as an India vs the Rest of the World conflict. I am one of the interlocutors in this conflict and often asked if my technical judgement is clouded by my nationality. Perhaps yes, perhaps no. Back in 2001 I did a [comparison](#) of the two methods and cast my vote in favour of Jayadevan because I thought he was ahead by a whisker.

A lot has changed since; both D/L and Jayadevan have significantly upgraded their methods, and both now need computers to reset targets. This happened after D/L got a big scare during the India-Australia 2003 World Cup final; the match, undeniably Australia's, was briefly going India's way during a 10-12 ball interval when Sehwag was firing all guns. If rain had ended play at that point India could have become undeserving winners and D/L would've got an instant burial.

Where do I stand in the D/L vs Jayadevan debate today? I'm still with Jayadevan — although both methods compute nearly the same target in most 'normal' match situations — chiefly because I think ICC is being completely unfair in denying him an opportunity to demonstrate his rain rule. I still find the debate fascinating, and see this as another example of the [engineer vs mathematician](#) debate that crops up ever so often in information management.

If we look at T20 cricket, however, I can say without hesitation that D/L is simply not good enough. We have to realize and accept that T20 is a very different sort of animal.



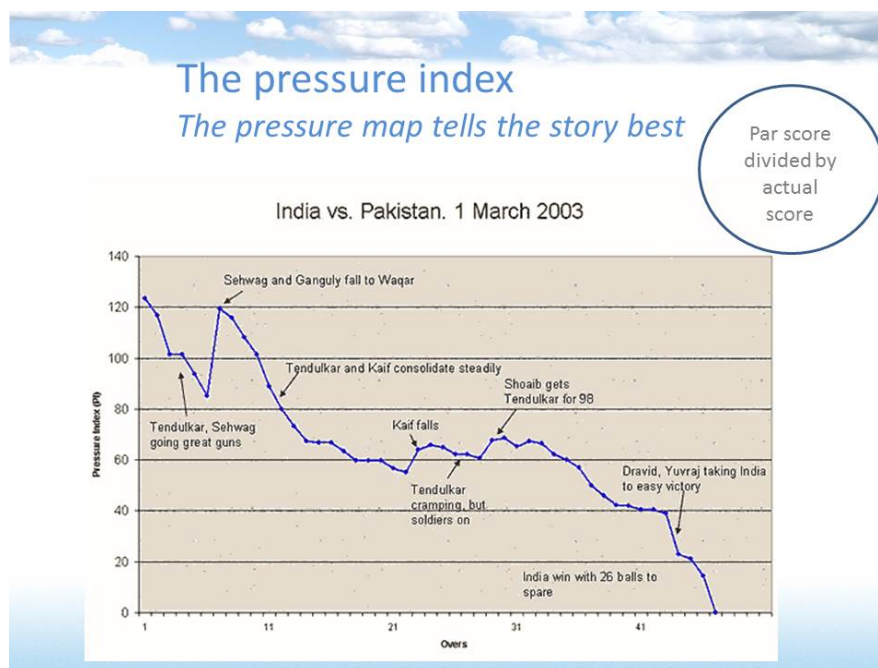
So how does the rain rule go for a T20 match? Shout, scream, smile or gasp when you hear this, but the dumb rule is simply to pretend that a T20 match is an ODI match with the first 30 overs lost for either team! Or simply erase the top 180 rows from our 300 x 10 table. It was this ridiculous construct that got Collingwood hot under the collar and Chris Gayle grinning like a contented cat when WI easily [defeated](#) England in a 2009 World T20 match.

You could view the situation this way: there is a long trouser that doesn't fit you, but you are being forced to wear it. What could you do? Either cut the trouser, or perhaps shrink it. Using the D/L ODI rain rule for T20 is like cutting the trouser. In an [exercise](#) with Rajeeva Karandikar we tried to see if we could 'shrink' the D/L trouser from ODI to T20 size (by 'shrinking' we assume that a T20 game evolves just like an ODI; only everything happens faster). We got mixed results: shrinking wasn't better than cutting, but it wasn't worse either.

In reality, T20 indeed appears to be a different animal. To devise a rain rule for T20 we must return to the drawing board, instead of tinkering with an ODI rain rule. So you might wonder why D/L haven't done this so far? I'm sure they are at it, but their problem appears to be that they don't have enough international T20 match data because not enough matches are played. But what about all that data from six IPLs or the Big Bash? Oh, but IPL is just a silly Indian league that most Englishmen pretend not to notice. And isn't the IPL all fixed?

Actually you must approach the problem differently. The only real requirement is to create that 120 x 10 table such that the resource diminishes after every ball and every wicket. There are so many statistical and probabilistic techniques that'll help you build such a table, and even ensure that your table gets smarter as the months and years roll by. One remarkable exercise at Simon Fraser University actually created such a table. And their big [finding](#) was that D/L over-estimates the amount of resource available in mid-innings by almost 5%. Because there's more resource apparently available, D/L thinks the batting side has the potential to score more and therefore sets a lower-than-expected target for mid-innings interruptions. Now you know why Gayle was grinning when WI easily scampered past England's target.

But let us now move to another truly exciting application of D/L-like resource tables: we call it the 'pressure index'!



Imagine you are in a meeting with your phone switched off and there's an India v Pakistan match happening. As soon as the meeting ends, you check the score. To get an idea of who's winning the chase you need to know three variables: the runs scored, the wickets lost, and the overs remaining.

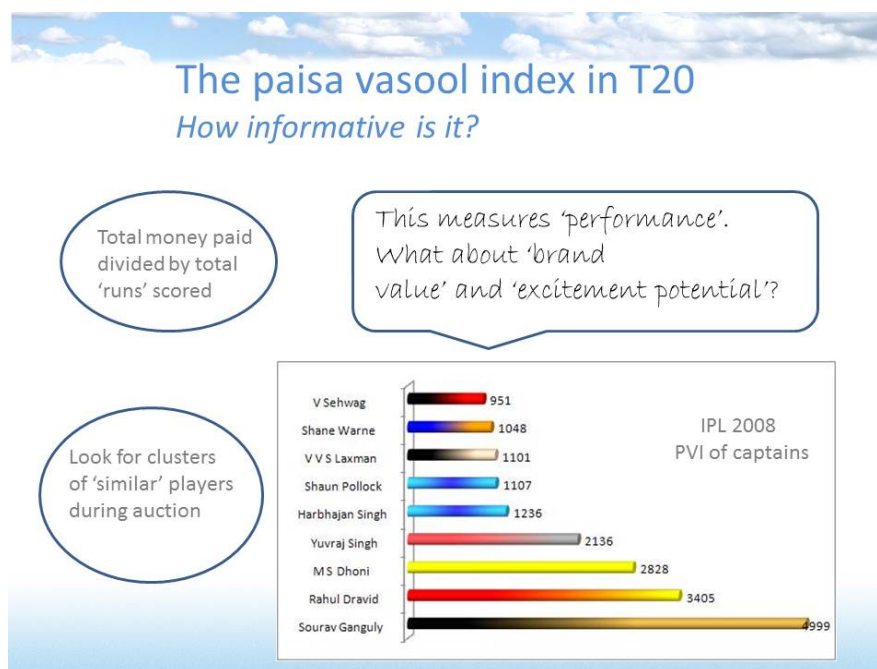
Wouldn't it be wonderful if all this information can be condensed into one number? Well, that number is what we call the [pressure index](#).

We define the pressure index using the idea of a par score. A par score is what the chasing team must score to be on level terms with the bowling team (with all the rain in the ongoing Champions Trophy everyone's talking of the par score). So if the chasing team is exactly at the par score we say that it has a pressure index of exactly 100. If a wicket falls at that point the par score rises and therefore the pressure index goes up to 115 or 120 or whatever (our formula is devised so that the max pressure index value is 200). If, on the other hand, the batsman hits three consecutive boundaries at that point then the chasing team has scored more than the par score and might have a pressure index of 92 or 95 (min pressure index value is 0). In the [WI vs SA](#) Champions Trophy match the WI pressure index was below 100 when Pollard hit that unfortunate shot leading to his dismissal. His dismissal pushed the pressure index up to exactly 100 and the match ended as a tie.

The pressure index keeps changing ball after ball. In a close match it will fluctuate this way and that from 100. If we now plot the ball-by-ball change in the pressure index we obtain what we call the 'pressure map'. In today's age of smartphones and mobile connectivity the best way to report a match could be by using the pressure map. Major events on the map (such as a dismissal, or a high scoring streak) can be hyperlinked so that the map becomes the cricket fan's one-stop cricket match reporting tool.

We carried the pressure index calculation live on rediff.com during the 2007 World Cup. But India's early elimination killed off all interest.

Fortunately, our reporting around the paisa vasool index in IPL 2008 on rediff.com and on *Hindustan Times* was much more successful.



So what then is the paisa vasool index (PVI)? It is really something quite straight-forward, and based on the most valuable player index (MVPI) that I have described earlier in this post. PVI provides a good estimate of a player's value in a professional cricket tournament such as the IPL.

Recall that MVPI collapses a player's performance into a single variable that we can call 'runs' (in quotation marks). The higher the MVPI, the more 'runs' a player is contributing. The PVI is obtained by dividing a player's earning (in US\$) by his MVPI, and is therefore seen to be the amount (in US\$) that the franchise owner pays the player for every 'run' scored.

The [best buys](#) are therefore players with the lowest possible PVI, i.e. players who contribute the most 'runs' at the least cost. In fact an [analysis](#) using PVI even allows you to obtain Moneyball-like inferences.

There is however one weakness in PVI: it estimates the worth of a cricketer *only on the playing field*. Players like Tendulkar or Ganguly are immensely valuable even if they don't perform on the cricket field (Tendulkar sends TV ratings soaring; in his prime Ganguly could single-handedly fill up Eden Gardens). So a more realistic estimate of a player's value must be based not just on on-field performance, but also on his perceived brand value.

Now one of the top talking points in an event like the IPL relates to the points table: which will be the four top teams in the table?

Usually such discussions involve a long series of complex 'if-then' arguments, with the clear picture being elusive till almost the very end. That's because we attempt only deterministic arguments. But what if we used probabilistic arguments instead?



Who makes it to the semi-final?

Try simulation ... or ask Rajeeva

Team	P	W	L	NR	PTS	NRR
Mumbai	9	7	2	0	14	1.11
Delhi	9	6	3	0	12	0.40
Rajasthan	11	6	5	0	12	-0.14
Bangalore	9	5	4	0	10	0.43
Chennai	10	5	5	0	10	0.15
Kolkata	9	4	5	0	8	-0.44
Deccan	9	3	6	0	6	-0.74
Punjab	10	2	8	0	4	-0.70

What I will now describe is an idea from [Rajeeva Karandikar](#). To illustrate the argument let us imagine we are looking at the IPL6 points table. IPL6 had 9 teams, or 36 distinct team pairs such as MI-KKR, KKR-RCB, SRH-CSK etc., etc. For each pair, let us write down our estimate of the win-loss probability. For example, for MI-KKR it could be 0.6-0.4 if we think MI has a 60% probability of winning.

With these probabilities we run a simulation. This means we simply tell the computer to pretend that IPL6 was played over and over, say 10,000 times (it will be impossible to do this in real life, but on a computer it will only take a minute!). The computer therefore ends up with 10,000 possible IPL6 points tables. Looking at these tables a simple counting process will enable us to identify while team is likely to be first, second, third and fourth. It is also clear that we can repeat the same simulation process to identify the likely finalists and the likely winner.

The process can also be easily refined. For example, we could factor in different win probabilities for home-away matches since IPL6 trends indicate significantly higher home win probabilities. It will also be a lot of fun; I can for example visualize a series of attractive contests built around this simulation idea on a cricket portal.

Next we pose an intriguing ODI batting question: If you have to chase a big total would you rather have Adam Gilchrist in your batting line-up or Herschelle Gibbs?

Gilchrist or Gibbs?

Do you get better with overs, or worse?

Herschelle Gibbs	0.93
Adam Gilchrist	1.16
Inzamam-Ul-Haq	1.11
Matthew Hayden	1.19
Desmond Haynes	0.94
Sanath Jayasuriya	0.96
Jacques Kallis	1.01
Brian Lara	1.04
Ricky Ponting	1.14
Kumar Sangakkara	1.09
Virendra Sehwag	1.04
Sachin Tendulkar	0.92

*Beta values of
Weibull
distribution
(difficulty in
getting out)*

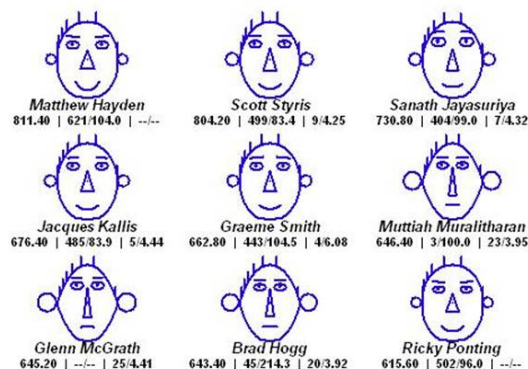
To answer this question, let us think hard about which ODI performance is upper most in our mind when we think of Gilchrist and Gibbs. The Gilchrist knock I remember most is his [172](#) against Zimbabwe in 2004. He had enough overs left to get past 200, but he just threw the opportunity away. My favourite Gibbs knock is his [175](#) at the Wanderers in 2006 as SA chased down Australia's 434/4.

With Gilchrist one feels that he starts strongly and becomes more vulnerable as the innings advances. With Gibbs it is just the opposite; he starts tentatively but looks rock solid as the innings progresses. Can we model this phenomenon? Do some ODI batsmen 'age' well as the innings progresses, and others 'age' poorly? This was the [question](#) that MRLN Panchanana and T Krishnan (who taught me statistics over 35 years ago at Indian Statistical Institute) posed some years ago. Their answer: Yes! They fitted a Weibull distribution and showed how batsmen with a beta value below 1 (like Tendulkar) age well, while batsmen with a beta greater than 1 (like Ponting) age poorly.

Our next chat is about how we can pictorially depict cricketers. Can we caricature their faces so that we can recognize who is a batsman and who is a bowler. Better still, can we use their faces to spot cricketing similarities between two cricketers?

Chernoff faces

How to spot good batsmen and good bowlers?



MVPI | runs / strike rate | wickets / eco rate

This is an old visualization idea: when there are a lot of variables associated with a person or object it becomes difficult to depict all of them holistically. We then use faces to depict them: the roundness of the face may be

linked to 'runs scored', the extent of the smile may be linked to 'strike rate', the curvature of the eyebrows may be linked to 'fielding acumen', the length of the nose to 'economy rate', the loop of the ears to 'wickets taken' and so on. When you do this, every player has a 'face', and the look on the face can instantly tell you the attributes of the player.

Look at 'Hayden' and 'Ponting'. It is easy to see that both have very similar skills (which we recognize to be batting skills). Or look at 'Styris' and 'Jayasuriya'. They look similar because both were batting all-rounders.

We [did this](#) visualization for cricketers right through our 2007 World Cup coverage on rediff.com, it was a fun project (of course it would have been more fun if we had more Indian faces). But if there are thousands of individuals sharing dozens of traits then it is easy to see how these pictures can suddenly become very informative!

I was pleasantly surprised when I saw a 2008 article in [NYT](#) using the same idea to describe traits of baseball coaches.

We will next address a question that has been asked about Sachin Tendulkar all through his glittering cricketing career: Does Tendulkar let you down in a crisis? A lot of folks contend that he isn't a match-winner like Dravid or Laxman. For someone like me watching cricket for close to half a century, this debate has a déjà vu feeling. Back in the 1970s we were saying the same thing about Gavaskar vs Viswanath or Vengsarkar.

Does Sachin Tendulkar let you down in a crisis?
No!

$$P(\text{Sachin fails} \mid \text{crisis}) = \frac{P(\text{crisis} \mid \text{Sachin fails}) \times P(\text{Sachin fails})}{P(\text{crisis} \mid \text{Sachin fails}) \times P(\text{Sachin fails}) + P(\text{crisis} \mid \text{Sachin does not fail}) \times P(\text{Sachin does not fail})}$$

Easy to check that this probability is well below 1/2



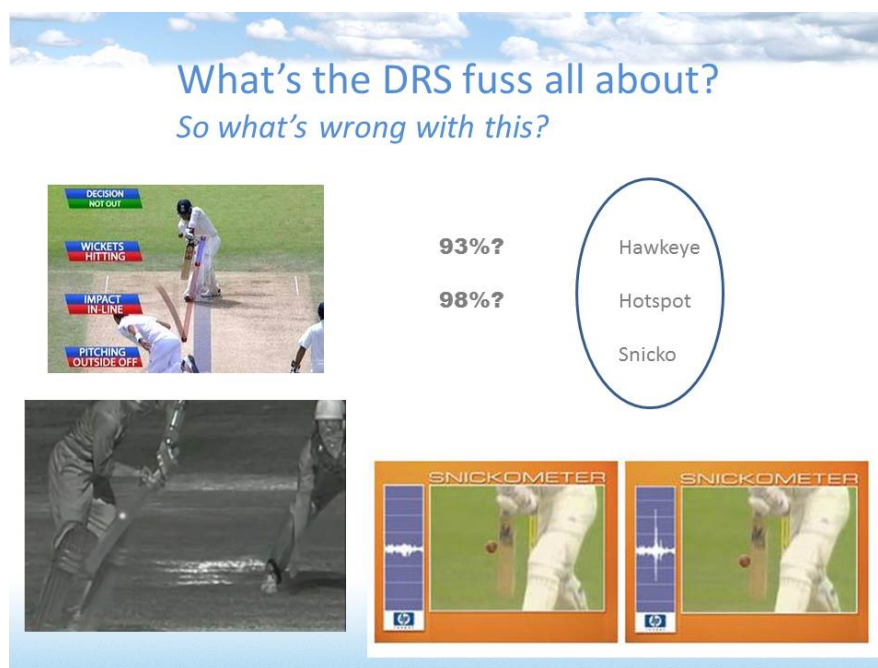
I won't write too much about this because I can scarcely better the compelling writing and [arguments](#) presented by Arunabha Sengupta. Arunabha argues that there is a cognitive fallacy in the reasoning. We are confusing two events ... does Sachin fail in a crisis, or is there a crisis because Sachin fails?

This confusion arises because many of us find it hard to understand conditional probability. If it is indeed true that Sachin fails in a crisis then the probability of the event 'Sachin fails' given the event that 'there is a crisis' must be very high ... say 75% or more. So can we compute the actual probability?

We can ... if readers have ever studied probability in school or college they would recall that we can do this if we use the Bayes theorem!

To compute, we first need to guess some probabilities: what's the probability that Sachin fails? Given his staggering record, very low. Let us say just 0.2. So the probability that Sachin does not fail is 0.8. What's the probability that there is a crisis if Sachin fails? Historically that's pretty high ... say 0.7. And what's the probability that there's a crisis if Sachin does not fail? Pretty low .. I'd say 0.3. If we now do the arithmetic we find that the probability that Sachin fails in a crisis is just under 40%, i.e., in less than 4 cases out of 10!

Let us now gaze into our crystal ball, and see how the future of cricket analytics might look like as we enter the brave new world of high speed communication and big data. I think it is going to be really exciting and enjoyable.



It is easy to see that the first future conflict in cricket analytics will relate to the DRS (for some time it was called UDRS for 'umpire decision review system', but umpires clearly didn't find this amusing). And it is just as easy to see that DRS is eventually here to stay: cameras keep getting better, algorithms keep getting smarter, errors keep getting expensive, and TV viewers and sponsors keep demanding more excitement.

I won't go into too much DRS detail because there are great compilations already available, including the [one](#) by Kartikeya Date that I rate highly, but DRS is all about using great imagery (as in Hot Spot), great gadgets (for example, the Snickometer) and great algorithms (as in Hawk-Eye) to improve the quality of decisions on the cricket field.

Shorn of polemics and controversy a cold-blooded view is that if DRS does better than the umpires we must have it. We already see numbers telling us that umpires succeed 93% of the time while DRS gets it right 98% of the time. It is also evident that in the future the umpires percentage will drop even as the DRS percentage rises (as a parallel, see how dependent doctors are now on medical tests).

I am also amused by the view that no technology should be accepted unless it is 100% accurate; this is often an unattainable ideal; costs zoom as you try to improve by even a fraction of a percentage point. If we want to wait for 100% accuracy, it will be a really long wait!

If pictures from the 2013 Champions Trophy are any indicator, Hot Spot does indeed look improved. The Snicko was always reliable ... so the debate is now really down to how well Hawk-Eye performs. Given the character of technology, we should expect Hawk-Eye to keep getting better, although it might initially also get costlier.

As we debate about how well or how poorly Hawk-Eye performs, I asked Rajeeva Karandikar if there wasn't a simpler way to answer this question. He said there was!

Making sense out of the nonsense

Just devise an old-fashioned experiment!



Bowler bowls, batsman is bowled ...
So what had your Hawk-eye predicted?
Count, count, count ...

What does Hawk-Eye really do? In simple terms it models the trajectory of the ball bowled by the bowler and checks if the ball would have gone on to hit the stumps.

Why not then ask a team of bowlers to actually bowl a few hundred deliveries with the intention of hitting the stumps? For each delivery we ‘freeze’ action as soon as the ball pitches and ask Hawk-Eye to predict its trajectory. We then compare what Hawk-Eye predicts with what actually happens. Did the ball really hit the stumps when Hawk-Eye said it would? Did it really miss the stumps as Hawk-Eye said it would? If there was a mismatch then what was the margin of error?

To make the analysis more robust we could carry out this experiment at different cricket grounds, with different cricket balls, with different pitch wear and tear (we could even deliberately create bowler’s footmarks to the extent umpires would allow) and in different climatic conditions. And then we would count and compute!

But cricket’s biggest future worry is of course ‘fixing’: spot fixing, ball fixing, player fixing, match fixing or whatever! How are we going to fix this?

What do we do about match fixing?

Try big data analytics!

Let us put in all the telephonic conversations logged in at telecom switches, the voices of all suspected bookies, the names of all persons who use a telephone number starting or ending with 111 or 222, the timelines when the bowler bowled an ‘accidental’ beamer, bouncer or wide, and when an experienced batsman failed to ground his bat to get run out, the number of balls taken to score the first 50 runs of an innings, and the next 50, the voice feeds travelling through the Cronje earplugs, the closed-circuit TV footage of all hotel lobbies in Mumbai and Barbados, the number of times there is confusion about which captain won the toss, the frenzy and the lull in the betting markets, and the ups and downs in a cricketer’s bank balance. Also add e-mail transcripts of anyone named John, Shane or Mark, the number of minutes spent by a batsman to tie his Adidas shoe laces and the frequency with which Shahid Afridi bites the cricket ball between the tenth and 15th overs.



Some years ago I wrote a mad [post](#) on a blog I used to write for CastrolCricket asking “What can cricket learn from Google?” It was just wild speculation that we could use big data analytics to uncover evidence of match mixing.

Even today that ramble probably qualifies as senile fantasy ... but I'm not so sure about tomorrow. A couple of weeks ago I was reading the much-acclaimed book [Big Data](#) and was surprised to find a reference to match fixing in Japan's sumo wrestling events. So other folks too are waking up to the big data opportunities in sport.

Essentially big data is all about discovering associations, and match fixing is about deliberately creating associations and correlations. If big data techniques become sufficiently powerful they'll surely find the dirt.

All current approaches to deter match-fixing are based on denial of service (jamming mobile phones in and around a cricket field is a big joke). The principle is: "Make it harder and harder to fix". I rather fancy that this principle should instead be: "Make it easier and easier to find".

Let us end by asking how the cricket analytics story is likely to play itself out tomorrow and the day after.



Long long ago, cricket was played on a cricket ground, and the rest of the world only came to know what's happening via radio or via next morning's newspaper. Today, cricket is 'played' much more on TV screens, and tomorrow it will be played on computer and communication networks worldwide.

The game is adapting to the changing canvas. IPL wouldn't be what it is without TV and Internet; even betting and fixing are far more prevalent now because today's cricket matches can be seen everywhere in realtime.

Our cricket contests are now much more data-centric and almost certain to use Twitter or Facebook (does anyone even recall those spot-the-ball contests of the 1960s and 1970s on *Sport & Pastime* or *Sportsweek*?). We are now seeing ads on TV inviting us to watch matches on the computer, instead of on TV! Today's cricket nostalgia would involve browsing on YouTube, instead of old cricket books and magazines featuring Neville Cardus, Ray Robinson or Jack Fingleton.

Surprisingly, cricket websites still aren't embracing this new world. Cricinfo is playing out a huge nostalgic trip as it looks back to 20 years ago. Have they thought enough about how Cricinfo will be 20 years *later*? Are they even aware of the phenomenal value of their cricket statistics? Have they realised that text mining of their ball-by-ball summaries are likely to be the richest source of cricketing information? When I created pressure maps during the 2007 World Cup I depended heavily on Cricinfo's summaries. So why isn't Cricinfo itself preparing pressure maps and selling them?

I see this as a big opportunity. Someone should create that ultimate one-stop cricket portal; this post itself contains several ideas and suggestions for such a portal: instant par scores and run scoring strategies after every ball; what-if scenarios after every ball (Dwayne Bravo wished that the umpires had given his team just one more ball in that WI vs SA match ... so what could've happened if WI had that extra ball?); ongoing pressure index and pressure map, Chernoff faces, quizzes and contests about the IPL points table; drama and

discussion accompanying DRS; round-up of the latest player ratings; sale of cricket merchandise ... and I could go on, but I guess it now time to stop!

Advertisements



[Report this ad](#)

Share this:

Twitter

Facebook 6

Like

2 bloggers like this.

Related

[The D/L story so far](#)
In "Analytics"

[The new face](#)
In "Analytics"

[Lies, damn lies and analytics?](#)
In "Analytics"

June 19, 2013

9 Replies

« Previous

Next »

Leave a Reply

Your email address will not be published. Required fields are marked *

Comment

Name



Email



Website

Post Comment

☐ Notify me of new comments via email.



Kapil on [June 21, 2013 at 8:31 am](#)

Absolutely awesome article. And completely agree with your amazement over the hyping of the D/L method. Calculating the target based on D/L is easier than the average 8th grade math problem – don't understand why fans are so scared of it. And yes, it is completely useless in T20 – basically because T20 is not a "normal" ODI match. How many ODIs have we seen where both teams manage to keep all 10 wickets intact for as many as 30 overs – which is essentially what DRS assumes when being applied to a T20 match.

And love the pressure map, though I don't believe it can be a one-stop point of information as it basically tells you who is ahead based on a statistical calculation which you don't understand – as opposed to letting you decide for yourself who is ahead, based on your own cricket viewing experience which provides you with an empirical understanding of runs, wickets and overs. However, I think it would be a great companion to the regular Manhattan.

In addition to the map, I also feel that broadcasters should regularly show the D/L projected score during an innings as opposed to only showing projected scores based on run rates like current rate, 6 an over, 8 an over etc.

And as far as DRS goes, even if ICC doesn't want to do an actual experiment for whatever reasons (like cost and possible unavailability of top level bowlers), somebody definitely needs to do a similar analysis just using existing footage that we have. DRS cameras have now been employed all over the world (including in India in WC 2011). Why not use all that footage and test DRS on deliveries which were

either left alone or missed by the batsmen. These deliveries either hit the stumps or missed them narrowly (balls clearly far from the stumps can be excluded from the experiment) and by freezing the balls when they passed the batsmen, we can find out the true accuracy of Hawk-eye.

↩ Reply

Advertisements

Golf Video of the Day



#1 Reason the Average Golfer Can't Hit 200+ Yards...

[Watch Video »](#)

[Report this ad](#)

Pingback: [The Best Sporting Blogs from the last seven days | Talking Sports |](#)



Rahul Tyagi on June 22, 2013 at 1:51 am

One thing that seems to be a problem with methods like D/L (and even VJD) is that they are more than likely to be unfair to a team that is higher ranked/stronger. Basically, the method is based on an assumption that most teams' resources look like the average value of those resources over the training data history. so, essentially, they are answering the question "how many runs an average team is likely to make with these resources". Now, if one of the teams is much above average than the other, this assumption is likely to penalize it by underestimating its strength (and overestimating the strength of the below average team). Basically, if we know that one of the 4 wickets remaining is Adam Gilchrist, we know that Australia is in a much stronger position than D/L's estimation.

This is also why it is better for a captain to complete as many overs of its best bowlers as he can when there is an almost certain weather interruption coming. D/L doesn't know that the real bowling resources have already been spent when rain arrives!

One way of tackling this is to incorporate the player strength ratings (just batting avg/SR and bowling avg./SR if you want to have something simple but perhaps more appropriately Reliance ratings, or whatever improved version we trust) in our calculation of resources available. The problem with that is how to treat really new players.. perhaps a constant prior score can be attached to batsmen/bowlers in their first few matches, after which Reliance ratings can take over...

↩ Reply



Rahul Tyagi on June 22, 2013 at 1:56 am

By the way, I don't think the example you used in order to tell people about usefulness of simulation is actually not very appropriate. If you already have estimates of pairwise probabilities of results between

all the teams (based on your prior guesses or past years' performances etc.) then you don't really have to simulate 10000 different IPLs in order to calculate the probabilities of various teams' final standings. it is a simple matter of multiplying and adding those probabilities for the matches. This is true because the assumed probabilities are static and do not have any "memory" i.e. they do not depend on the results in the current tournament.

The simulation will be useful (and indispensable) if you treat the probabilities as dynamically changing.. e.g. probability of team A winning over team B is some reasonable combination of its strength in the previous year, its current win/loss ratio in this year and the same numbers for team B. THEN, you really need to simulate and see what the numbers turn up at the end.

[↩ Reply](#)



Rahul Tyagi on [June 22, 2013 at 1:57 am](#)

I should've said "I think the example..." rather than "I don't think..."

[↩ Reply](#)

Advertisements

Wayfair: Official Site



New Day, New Deals on
Everything Home

[Report this ad](#)



kidney function on [July 12, 2013 at 10:22 pm](#)

I'm impressed, I must say. Really rarely do I encounter a blog that's both educative and entertaining, and let me tell you, you have hit the nail on the head. Your idea is outstanding; the issue is something that not enough people are speaking intelligently about. I am very happy that I stumbled across this in my search for something relating to this.

[↩ Reply](#)



srinivasbhogle on [July 29, 2013 at 9:32 am](#)

Since I wrote this post, the CastrolCricket site has unfortunately closed down. So links pointing to my posts there can no longer be accessed.

[↩ Reply](#)



srinivasbhogle on [July 29, 2013 at 9:35 am](#)



I hear that Nate Silver has ended his association with NYT and will now work with ESPN. Given that Nate first made a name for himself with 'baseball analytics', my friend S Amarnath speculates that ESPN might persuade him to get into cricket analytics.

[↩ Reply](#)

Advertisements

**You don't need to go to an office
to write code. Work with us!**

AUTOMATTIC

[APPLY](#)

[Report this ad](#)

Pingback: [Cricket Analytics – what should be the future? | Number Sports](#)



[srinivasbhogle](#)

Srinivas Bhogle, 60, is an information scientist and manager.

[View Full Profile →](#)

Wayfair: Official Site



New Day, New Deals on
Everything Home

[Report this ad](#)

About me

[About](#)

[My life after ISI](#)

[View Full Site](#)

[Create a free website or blog at WordPress.com.](#)



