waRne - Putting cricket analytics in a spin

Contents

Yeah I forgot I wouldn't have any internet on the plane, so everything I wrote is not referencing anything from a link or pac

Abstract

The importance of reproducibility, and the related issue of open access to data, has received a lot of recent attention. Momentum on these issues is gathering in the sports analytics community, particularly in North America where open access to quality sports data has become the norm. Cricket is the second largest commercial sport in the world, but unlike other popular international sports, there has been no mechanism for the public to access comprehensive statistics on players and teams. Expert commentary currently relies heavily on data that isn't made readily accessible and this produces an unnecessary barrier for the development of an inclusive sports analytics community.

Introduction

Data access is a key enabler for any analytics community. Most major sports have easy access to match statistics, for example nbastatR for NBA, Lahman for baseball, deuce for tennis and nflfastR for NFL. Through access to sports data and reproducible findings and metrics fans clubs and researchers are better able to understand the game, predict match outcomes and rate players. For example the way teams tackled 4th down decisions in the NFL has changed since (Romer 2006) seminial work. As teams have changed their 4th down decision making this allows follow up research such as (Yam and Lopez 2019) which looked at more granular data to see if this happens in practice.

By making data more accessible and more advanced metrics more accessible fans data journalism in sports has grown in recent years. For example in (Horowitz, Yurko, and Ventura 2017) has enabled EPA to enter popular discussion among fans.

Cricket is the second largest sport in the world. However, unfortunetaly there is no easy accessible way to access ball by ball data nor aggregated statistics of teams and players. Data while available on sites like espncricinfo are not in an easy to use form. For example, each match is listed on different webpages so hours upon hours of time would be required to copy and paste a single season. Hence, there are significant logistical barriers for prospective fans and analysts studying the game, which stagnats understanding of cricket. The Duckworth-Lewis-Stern table, used to reset targets for matches that experience rain delays in all of cricket, is maintained by one man in Australia. Not even teams have access to the table. textwhat do they do, call it into the ICC to get a temporary copy or something when there's a delay?

This paper describes the waRne package, the first to provide free and easy access to data for cricket for fans. Web scraping tools are avaiable for fans to easily scrape the play by play commentary data on espncricinfo. For the first time fans can evaluate their favourite teams and players and do so in a reproducible and accessible manner. We have included the data for ODI and T20 from the 2010 archived season on ESPCcricinfo up to the matches on September 8, 2020. If a person wishes to expand this data to include more recent games, this package can scrape and update the dataset.

Why cricket needs reproducibility

Data accessibility enables fans, analysts and researchers to better understand the game. Through being able to reproduce common popular metrics, visualisations and article findings.

Through being able to reproduce, fans are able to make accessible findings for others and importantly they are able to extend and grow concepts. Unfortuntely what we see through leading cricket analytics providers is a track record of confusing output for fans. This can lead to lower engagement and dismissial of cricket analytics.

I'm not sure if we should slam the datajam competition directly. I think the guy tweeting out conflicting news is a good poi

For example in this series of tweets we see the narrative being pushed that Steve Smith is a good player vs pace bowling unfortunetly just a few months prior the same company and journalist published an article which had Steve Smith doing much worse against pace (balls above 140km/h). Unlike a similar sport baseball, fans have no easily accisible way of seeing if Steve Smith vs pace is a strength as alluded to in the original tweet, or a weakness like the same persons published online article. In comparsion, fans are able to get a breakdown of Mike Trout vs fastballs from using baseballr which provides access through statcast data from baseballsavant](https://baseballsavant.mlb.com)).

Unfortunetly this is just one of many examples whereby a relatively simple statistic is provided by media and fans have no mechanism to fact check. Fact checking is an important avenue for fans to not only engage and understand statistics, but having this mechanism also stops analytics companies from putting out misleading conclusions and findings.

With the commentary data on cricinfo, we are not able to get the same resolution on delivery speed and location as the statcast data provides in baseball, but we can often learn the type of delivery the batsman faces through the years. This represents an enormous leap in publically available data, and we hope will lead to the eventual release of the ball tracking data used butnot shared in crickets leagues around the world

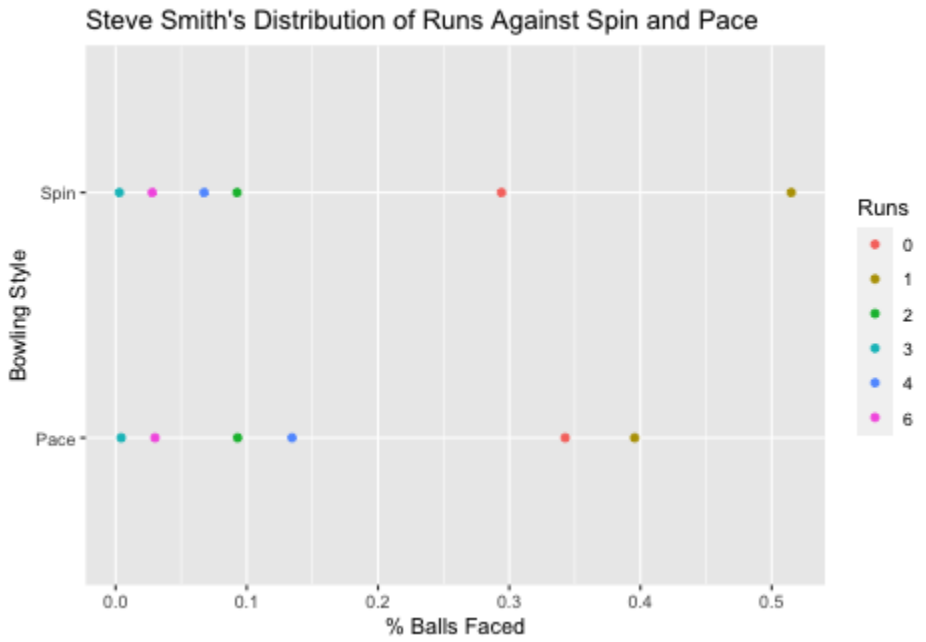
Armed with this new dataset and the waRne package, we are able to not only compare Steve Smiths performance vs bowling type but we are able to easily compare Steve Smiths performance vs bowling types with other players, across multiple seasons, leagues, and cricket formats.

Steve Smith vs Bowling Type Chart

James here need to figure out the code, do we keep it seperate so in the example we still have to join on bowling type to the ball by ball, or will it just be an extra column joined already.

So the way the scraper is set up, I grab the play links when scraping the scorecard and commentary along with the id and r

```
## `summarise()` regrouping output by 'bowling_style' (override with `.groups` argument)
```



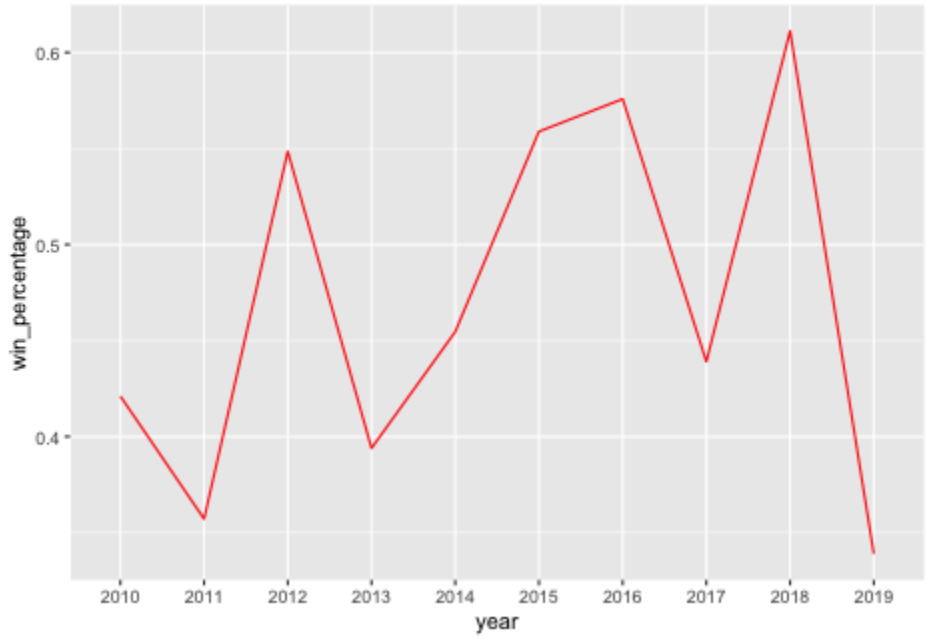
Another reason why is that cricket is a weird sport with a lot of old school rules of thumb that are followed/talked about.

For example it has become popular conjecture that its better that teams should bat second (chase down a total set by the oppoition). Through waRne we can answer such question for the T20 game.

For example we can explore do teams win more often when chasing in the big bash over the history of the big bash, or we can plot a line chart to see if there is a movement towards batting last and winning.

```
## Adding missing grouping variables: `match_id`
```

```
## `summarise()` regrouping output by 'batting_team_result' (override with `.groups` argument)
```



Using waRne to teach undergraduate statistics

We can also use waRne and other R packages as an easy way to engage students in learning statistical concepts.

For example using the above we can not only look at the answers graphically, but we can run a statistical test. The example we would use is instead of testing is a coin biased (different from 50/50) we can test to see if the winning \% is different from 0.5 for teams chasing.

```
## Adding missing grouping variables: `match_id`
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## [1] 0.2796709
```

James and I entered a reproducibility competition and the winning entry provided a statistic sort of a gateway into the importance of their ''project" was was specifically in 2019 Australian Big Bash T20 season. The statistic given was in the 2019 big bash season teams that batted first won only 43% of games. Thankfully we can try to reproduce that statistic now through the use of waRne

Wikipedia tells us that there were the following:

In the 2018-19 Big Bash League there were 59 matches played. Of which we had 3 DLS results.

Table 1: Games under consideration in 2018-19 Big Bash

Games	Result
21 DEC 2018 - THUNDER VS STARS	THUNDER WON BY 15 RUNS (D/L)
2 FEB 2019 - THUNDER VS SIXERS	SIXERS WON BY 9 WICKETS (D/L)
8 JAN 2019 HEAT VS THUNDER	HEAT BY 15 RUNS (D/L)
17th January 2019 Thunder vs Heat	no result

```
##
##              0  1
##   20102011   8 11
##   20112012  10 18
##   20122013  17 13
##   20132014  13 18
##   20142015  15 17
##   20152016  19 15
##   20162017  19 11
##   20172018  18 20
##   20182019  33 22
##   20192020  20 39
```

‘‘

So here we can see that we in the big bash season 2018-19 we have a current winning percentage batting first of 22/(22+33) =0.4 but this doesn't include the duckworth lewis games or the washed out completely game.

If we were to add in the duckworth lewis games we get 24/(24+34) which is equal to 0.414. Which isn't the result.

We can also clearly see that in the 20192020 big bash season the team batting first wins 39/(39+20) or 0.66 so what could be going on? Did Quantium mean only games played in the actual year 2019?

In that case from the 2018-2019 season we would remove the following games

Table 2: Games in 2018 of the 2018-19 Big Bash Season

Date	Game	Result	Winning Team
19 December 2018	Heat Vs Strickers	Strickers 5 Wickets	Chased
20th December 2018	Scorchers Vs Renegades	Renegades by 4 wickets	Chased
21 December 2018	Thunder vs Stars	Thunder by 15 runs (D/L)	Batted First
22 December 2018	Sixers vs Scorchers	Sixers by 17 runs	Batted First
22 December	Hurricanes vs Heat	Hurricanes by 15 runs	Batted First
23 December 2018	Strikers vs Renegades	Renegades by 5 wickets	Chased
24 December 2018	Stars vs Hurricanes	Hurricanes by 6 wickets	Chased
24th December 2018	Thunder vs sixers	Thunder by 21 runs	Batted First
26th December 2018	Strickers vs scorchers	Scorchers by 7 wickets	Chased
27th December	Sixers vs Stars	stars by 5 wickets	Chased
28th December	thunder vs hurricanes	hurricanes by 7 wickets	Chased
29th december	sixers vs renegades	sixers by 33 runs	Batted First
30th decemeber 2018	scorchers vs hurricanes	hurricanes by 6 wickets	Chased
31st decemember 2018	strickers vs thunder	strickers by 20 runs	Batted First

i.e. we remove 14 games here of which 6 teams won batting first 16/(22+33-14)

Then from here we have to add on games in the 2019-2020 big bash season that were played in 2019

Table 3: Games played in 2019 of the 2019-20 Big Bash

Date	Teams	Result	Winning Team
17 December	Thunder vs Heat	Thunder by 29 runs	Batted First
18th December	Scorchers vs Sixers	Sixers by 8 wickets	Chased
19th december	Renegades vs thunder	thunder by 6 wickets	chased
20th decemeber	hurricanes vs sixers	hurricanes by 25 runs	batted first
20th decemember	stars vs heat	stars by 22 runs	batting first
21st decemember	strickers vs thunder	no result	no result
21st december	scorchers vs renegades	scorchers by 11 runs	batting first
22 deceember	stars vs hurricanes	stars by 52 runs	batting first
22 decemember	heat vs sixers	heat by 48 runs	batting first
23 dec	strickers vs scorchers	strickers by 15 (DLS)	batting first
24th dec	renegades vs hurricanes	hurricanes by 7 wickets	chased
26 decemember	sixers vs scorchers	sixers by 48 runs	batting first
27 dec	strickers vs stars	strikers by 5 runs	batting first
28th december	thunder vs sixers	sixers won super over (batting first)	sixers by superover or sixers batting second
29th dec	strikers vs renegades	strikers by 18 runs	batting first
30th dec	hurricanes vs stars	stars by 4 runs (DLS)	batting first
31st december 2019	thunder vs strikers	thunder by 3 runs	batting first

12 batting first and 3 batting second

(16+12)/(12+3+22+33-14)

So over the past 2 years (not including DLS) we have

(22+39)/(20+39+33+22)

i.e. >50% of teams win chasing so what does that mean?

Easier engagement of fans

To the surprise of many, its hard to engage cricket fans into the analytics behind the game. Reasons for this are generally centred around reproducibility and explainability to fans.

Without an easy accessible medium how can crickets version of an analytics community grow.

something something look towards how EPA has changed the way fans are engaged in NFL analytics - so maybe the change in run rate can be like EPA?

Cool cricket examples

- Dhoni spin/pace/yorkr
  - shows off joins to player information and extracting information from play by play
- espncricinfo articles but instead of ipl do same tables for big bash
- recreate stats from espncricinfo like this page

Things to do	person
clean scrapers/csv for waRne	James
data dictionary	Robert
examples	Robert - start (assume same csv being used) - cool cricket example stuff

Dhoni Dilemma

With this new dataset, we allow fans to put on their analyst thats and to be able to ask themselves, what exactly makes a good closer and what exactly is the Dhoni Dilemma.

To understand the Dhoni Deilemma, we need to see some interesting things to do with the change in RRR.

Cricket is an obvious sort of sport, to win in the end you have to get one run more than your opposition. A commonly used statistic during the broadcast is the required run rate, for example if you need 120 to win off 20 overs your RRR is 6 runs an over

What follows is then obvious is the chasing team wins if they are able to lower the RRR that is score quicker.

Espncricinfo did article on teams that win games get a higher proportion of their runs in the powerplay and first 10 overs. But proportion of team runs in T20 games doesn't really make sense as a batting metric as we are not just concerned with how many runs they get, but how quickly they get them. So we introduce a new way to measure batsman change in RRR.

This graph makes sense as we can see the best openers in the world lower the required run rate so whats the Dhoni Dilemma?

Dhoni is renowed for being one of the worlds best closers but we can see it seems as though in the middle he is letting the game get out of hand. How can one of the worlds best closers

notes for James

So I don't think its like a complex paper, because of the way cricket is at the moment I think the biggest influence this stuff can make is

- being able to fact check
- having batsman metrics that are beyond average, total runs and SR
- What we can maybe mention? is that change in RRR while simple seems to be quite effective, a better measurement would be the DLS which we know can be derived from the ball by ball data or alternatively if someone has a resource table they can instead of change in RRR use change in resources or whatever

For the examples, I think its pretty easy to get the graphs to look nice, its more about whats the dataset look like? i.e. do we need the first few lines for reproducibility to be joins or is it sort of like done already

Horowitz, Maksim, Ron Yurko, and Samuel L Ventura. 2017. "NflscrapR: Compiling the Nfl Play-by-Play Api for Easy Use in R." *URL Https://Github. Com/Maksimhorowitz/nflscrapR, R Package Version* 1 (0).

Romer, David. 2006. "Do Firms Maximize? Evidence from Professional Football." *Journal of Political Economy* 114 (2): 340–65.

Yam, Derrick R, and Michael J Lopez. 2019. "What Was Lost? A Causal Estimate of Fourth down Behavior in the National Football League." *Journal of Sports Analytics* 5 (3): 153–67.