fitzRoy - An R Package to Encourage Reproducible Sports Analysis

by Robert Nguyen, James Day, David Warton and Oscar Lane

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. While Australian Rules football (AFL) is the leading commercial sport in Australia, unlike 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. We present the R package **fitzRoy** to provide easy access to AFL statistics.

1 Introduction

Access to data is the key enabling tool for any sports analytics community. Most major international sports have a mechanism to provide free access to match statistics, for example, ballR for NBA, Lahman for baseball and deuce for tennis. Access to sports data can be used by fans, clubs and researchers to better predict match outcomes, to inform decisions and better understand the sport. For example (Romer, 2006) helped change the way teams evaluate 4th down decisions in the NFL, and the way the NBA is played has changed becoming more three point focused (Goldman and Rao, 2013). Sports analytics has also proved a popular avenue for modern data journalism for example, Fivethirtyeight is a popular culture website with a strong analytics following, which publishes models daily across a variety of sports. This sort of product can only be constructed given a publicly available source of sports data.

The Australian Football League (AFL) is the national league of Australia's national winter sport, Australian Rules Football. This is the largest commercial support in Australia with over 1 million club members, a 2.5 billion dollar broadcast rights deal and a participation level of 1.649 million. No current AFL statistics website provides easy access to data for a growing analytical fan base. The Australian Football League (AFL) has an official data provider, Champion Data, which is 49% owned by the AFL. Champion Data have the licence to collect the data for all AFL games and then charge clubs and media organisations fees to access the data. There are two leading websites of publicly available data, afltables and footywire, but data are not available in an easy-to-use form. For example, match statistics are listed on separate web pages for different matches, so hours of time would be required to compile data from over 200 different webpages in order to do an analysis across a single season. Hence, unfortunately, there are significant financial and logistical barriers to prospective analysts and fans studying the game, which stagnates progress advancing our understanding of AFL.

This paper describes the **fitzRoy** package, the first package to provide free and easy access to data on the AFL, with match and player data for the men's competition[1]. Web scraping tools have been developed that provide easy, up-to-date access to AFL match and player box statistics across the history of the game, since 1897, using open source data. The package also provides tools to link match and player data to expert tips from popular websites. For the first time, fans can evaluate the performance of tipsters themselves andcompare them to the betting market.

2 What is fitzRoy?

We developed **fitzRoy**, an R package that allows users to access Australian Rules Football statistics from various websites easily with R. The **fitzRoy** package allows users access to popular AFL statistics websites such as afltables and footywire. These are the two most widely used data repositories in the AFL, which have existed since the late 1990s, and while they are not official repositories, the AFL has not tried to take them offline. However, the data on these websites is not available in an easy-to-use form, e.g. match statistics are stored across different web pages for each match, so compiling season statistics would involve harvesting data from hundreds of webpages. The **fitzRoy** package compiles match or player data into a single frame, and also allows users to access popular bloggers' AFL models via the squiggle website.

A popular website called afltables contains AFL-VFL match, player and coaching stats, records

[1] fitzRoy used to contain access to the AFLW (AFL Womens) data, but unfortunately the data was removed from official AFL media, we are committed to adding AFLW data again, once it comes back

and lists from 1897[2]. The website afltables has been used in research for topics such as umpire racism (Lenten, 2017), umpires assessment of players (Lenten et al., 2019), modelling of the AFL game (Kiley et al., 2016), fixture difficulty (Lenor et al., 2016) and drafting (Lenten et al., 2018). The umpire studies would not have been possible with footywire data as umpire information isn't contained on the game pages.

The footywire website has data back to 1965, but while it does not have as many seasons of data, it has additional game variables not included in afltables. One example is Super Coach score, sometimes used as a proxy for player value (Marshall, 2017). Other examples of variables contained within footywire are tackles inside 50, intercepts and marks inside 50 to name a few.

Squiggle is a unique website in the AFL sporting landscape. It contains game analyses but it also aggregates popular AFL bloggers' tips each week. From squiggle users are able to get each models probability of win, margin prediction and a leaderboard based on bits which has been made popular by the Monash probability footy tipping competition[3].

3 Building fitzRoy

The name **fitzRoy** comes from the Old Fitzroy hotel in Sydney where the idea for the package was first conceived. It is also the name of one of the foundation clubs of the Australian Football League, which since merged with another club, and is now called the Brisbane Lions.

We used the R packages **Rvest** (Wickham, 2020), **dplyr** (Wickham et al., 2015), **purrr** (Henry and Wickham, 2019) and **XML** (Temple Lang, 2020) to construct our web-scraper functions that collate data from afltables or footywire into a single data frame. These websites update immediately on completion of each round, hence so does data accessed via **fitzRoy** scraper functions. The key functions accessing afltables player and match statistics are `get_afltables_stats` and `get_afl_match_data`, respectively, and footywire data are accessed via the `get_footywire_stats` function. These functions form the backbone of the package.

4 Applications of fitzRoy

Match Data (Scores)

The `get_match_results` function can be used to obtain match data for any season(s) or team(s). For example to get the game scores for Fitzroy's last AFL season as follows.[4]

```
library(fitzRoy)
library(tidyverse)
library(lubridate)
fitzRoy::get_match_results()%>%
mutate(Season=lubridate::year(Date))%>%
filter(Season==1996)%>%
filter(Home.Team=="Fitzroy" | Away.Team=="Fitzroy")
```

Match Data (Players)

Fans of Australian Football like many major sports like to keep up to date with leaders of statistical categories. One statistic often of interest is goals scored. Users can come up with the leading goalkicker list for Fitzroy in 1996 as follows.

```
library(fitzRoy)
library(tidyverse)
fitzRoy::get_afltables_stats(start_date="1996-01-01",
end_date="1997-01-01")%>%
filter(Playing.for=="Fitzroy")%>%
group_by(ID, First.name, Surname)%>%
summarise(Total_Goals=sum(Goals))%>%
arrange(desc(Total_Goals))
```

[2]The first year of VFL

[3]http://probabilistic-footy.monash.edu/ footy/

[4]fans of Fitzroy Lions might want to avoid this as it only contains one win (Round 8 vs Fremantle Dockers)

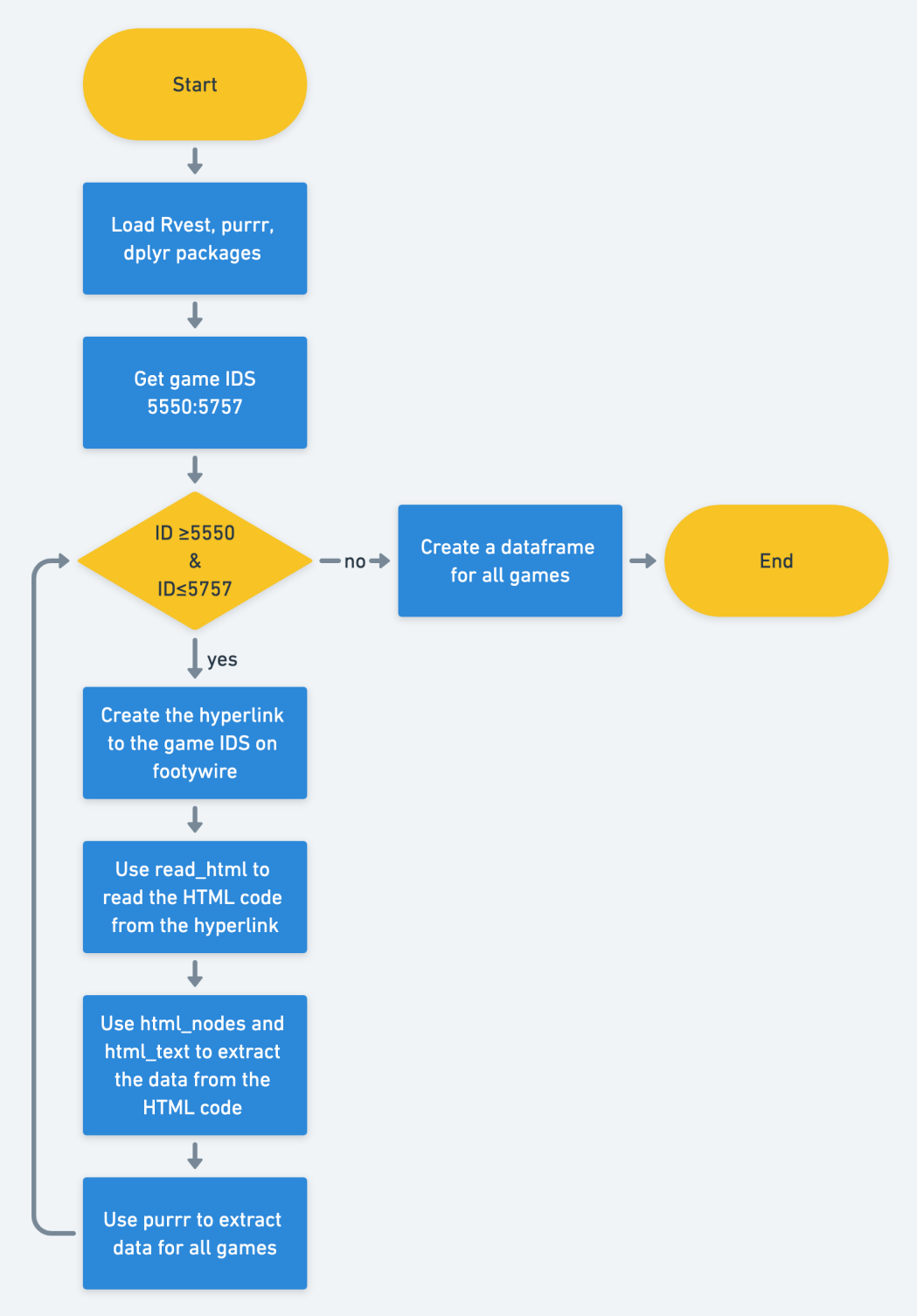

Figure 1: Work flow for **fitzRoy** web-scrapers game IDS 5550 to 5757 refer to the 2013 AFLM Season.

We can see that Anthony Mellington won the goalkicking award for Fitzroy with a modest total of 22 goals for the 1996 season.

Building Sports Models

Sports models are commonly derived using an Elo system which only needs scores (Ryall and Bedford, 2010). The **fitzRoy** package readily provides a data frame of match scores, from which it is straightforward to construct an Elo to predict future match outcomes, as below.

```
library(fitzRoy)
library(tidyverse)
library(elo)
library(lubridate)

# Get data
results <- fitzRoy::get_match_results()
results <- results %>%
mutate(seas_rnd = paste0(Season, ".", Round.Number),
First.Game = ifelse(Round.Number == 1, TRUE, FALSE)
)

fixture <- fitzRoy::get_fixture()
fixture <- fixture %>%
filter(Date > max(results$Date)) %>%
mutate(Date = ymd(format(Date, "%Y-%m-%d"))) %>%
rename(Round.Number = Round)

# Simple ELO
# Set parameters (these should be optimised!)
HGA <- 30
carryOver <- 0.5
B <- 0.03
k_val <- 20

# Create margin function to ensure result is between 0 and 1
map_margin_to_outcome <- function(margin, B) {
1 / (1 + (exp(-B * margin)))
}

# Run ELO
elo.data <- elo.run(
map_margin_to_outcome(Home.Points - Away.Points, B = B) ~
adjust(Home.Team, HGA) +
Away.Team +
group(seas_rnd) +
regress(First.Game, 1500, carryOver),
k = k_val,
data = results
)

as.data.frame(elo.data)
as.matrix(elo.data)
final.elos(elo.data)

# Do predictions
fixture <- fixture %>%
mutate(Prob = predict(elo.data, newdata = fixture))

head(fixture)
```

Building Player Models

Box-score statistics, summary statistics of the involvement of each player in each match, contain a rich history of information. Box score statistics led, for example, to the concept of Value Over Replacement Player (VORP) (Woolner, 2001)[5]. Fantasy teams have gained a lot of interest in recent years, and fantasy scores for players tend to be constructed from box-score statistics.

The AFL runs a fantasy sport competition, and **fitzRoy** could be used to recreate its fantasy points formula, since it is a linear function of box-score statistics.

Box-score AFL data are made readily accessible through **fitzRoy** using the `player_stats` function.

```
library(fitzRoy)
library(tidyverse)

df<-fitzRoy::get_footywire_stats(9721:9927)

eq1<-lm(AF ~ K + HB + M + `T` + FF + FA + HO + G + B, data=df)
summary(eq1)
```

While this might seem like a trivial application, footywire only has fantasy scores going back to 2007, however the statistics used for fantasy go all the way back to 1965, with Tackles being first recorded in 1987.

```
cbPalette <- c("#999999", "#E69F00", "#56B4E9", "#009E73",
"#F0E442", "#0072B2", "#D55E00", "#CC79A7")

fitzRoy::get_afltables_stats(start_date = "1897-01-01",
end_date = "2018-10-10")%>%

group_by(Season)%>%
summarise(
meankicks=mean(Kicks),
meanmarks=mean(Marks),
meantackles=mean(Tackles),
meanfreesfor=mean(Frees.For),
meansfreesagainst=mean(Frees.Against),
meanhitouts=mean(Hit.Outs),
meangoals=mean(Goals),
meanbehinds=mean(Behinds))%>%
gather("variable", "value",-Season) %>%
ggplot(aes(x=Season, y=value, group=variable, colour=variable))+
geom_line()+
scale_colour_manual(values=cbPalette)
```

Goals have been recorded at a player level throughout the history of the game, and the most recent variable that is used in fantasy (tackles) started being recorded in 1987.

The box-score also contains time on ground so users are readily able to compute points per minute which has been a leading indicator for 2 time winner of fantasy sports Moreira Magic.[6]

Champion Data publish Super Coach scores, valued by clubs to inform recruitment decisions and fantasy sport competitions. However the formula for Super Coach scores is propriety and not in the public domain. Following the release of fitzRoy, one well-known blogger attempted to re-create it with a linear model, and managed an R-squared of 91.6[7][8]

Able to Compare Popular Models

Blogging has taken off around the world with popular websites such as fansided, fivethirtyeight and the ringer proving popular among the overseas sporting community. To help promote other people

[5]https://www.theringer.com/mlb/2018/2/20/17030428/sherri-nichols-baseball-sabermetric-movement

[6]https://player.whooshkaa.com/shows/chilling-with-charlie

[7]http://www.matterofstats.com/mafl-stats-journal/2018/10/7/a-first-attempt-at-combining-afl-team-and-player-data-in-a-predictive-model

[8]This is impressive as Champion data uses data not available within **fitzRoy** and its weighted by time and game margin.

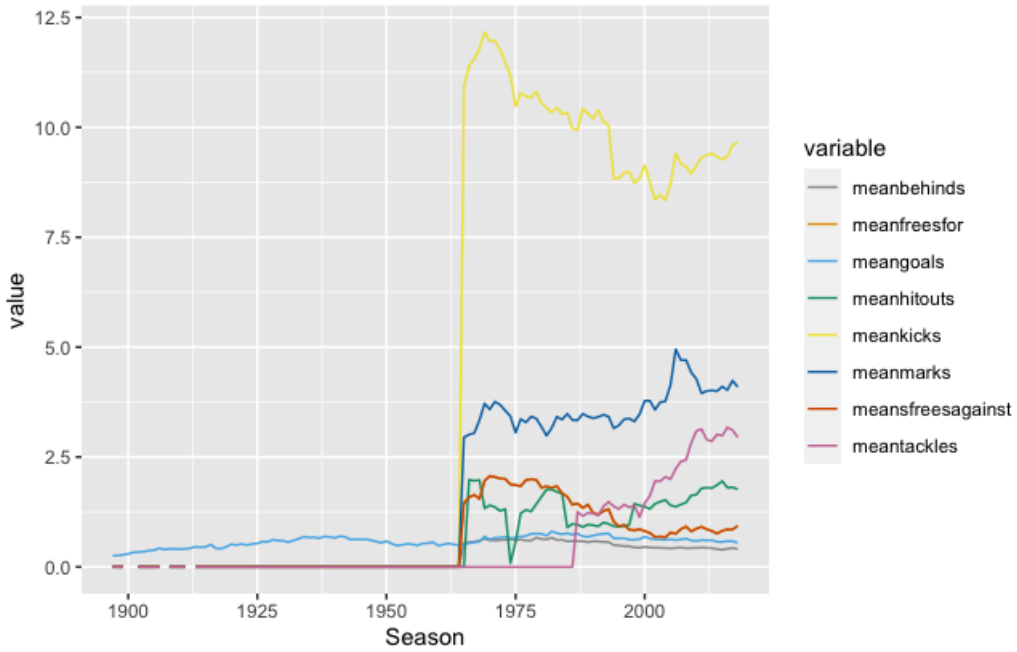

Figure 2: Line graph of mean values of AFLM statistics - By seeing when the line jumps from zero is a quick way to see when a statistic was first collected.

who do modeling work and make their work available online, we provide access to squiggle, the most popular aggregator website in the AFL. This means that the behaviour of different tipsters' models can be analysed easily.

```
library(fitzRoy)
fitzRoy::get_squiggle_data("tips")
```

The above command will enable a user to get the tips from popular blogging sites such as squiggle, matterofstats and liveladders among many. This means that different tipsters models behaviours can be analysed easily. For example studying how they take into account home ground advantages.

5 Future Developments

The developers of **fitzRoy** is committed to giving users the data to analyse the game. In the future this means updating the Womens AFLW data once it becomes available online, updating the scrapers to include the AFL website.

6 Summary

The **fitzRoy** package offers a springboard for sports analytics in the AFL community. It provides easy access to publicly available AFL data, and we have illustrated how this can be used to predict match outcomes and to rank players. In future work we plan to build a statistical model for AFL match outcomes and its key predictors, along the lines of (Yurko et al., 2018; Deshpande and Jensen, 2016). There are endless possibilities: clubs might use it to inform on player recruitment (O'Shaughnessy, 2010) and team style (Greenham et al., 2017); researchers and enthusiasts can use it to better understand the game; there are obvious betting implications (Bailey, 2000); and educators can use it as a teaching tool.

The **fitzRoy** package was only released in 2018 but has already been used by AFL club analysts and bloggers who are now able to access data and develop content they weren't previously able to do (*e.g.* VORP). The AFL analytics community is develops rapidly, and it is exciting to see where it will go over the coming seasons.

Bibliography

M. Bailey. Identifying arbitrage opportunities in afl betting markets through mathematical modelling. In *Proceedings of the Fifth Australian Conference in Mathematics and Computers in Sport*, pages 37–42, 2000. [p160]

S. K. Deshpande and S. T. Jensen. Estimating an nba player's impact on his team's chances of winning. *Journal of Quantitative Analysis in Sports*, 12(2):51–72, 2016. [p160]

M. Goldman and J. M. Rao. Live by the three, die by the three? the price of risk in the nba. In *Submission to the MIT Sloan Sports Analytics Conference*, 2013. [p155]

G. Greenham, A. Hewitt, and K. Norton. A pilot study to measure game style within australian football. *International Journal of Performance Analysis in Sport*, 17(4):576–585, 2017. URL https://doi.org/10.1080/24748668.2017.1372163. [p160]

L. Henry and H. Wickham. *purrr: Functional Programming Tools*, 2019. URL https://CRAN.R-project.org/package=purrr. R package version 0.3.2. [p156]

D. P. Kiley, A. J. Reagan, L. Mitchell, C. M. Danforth, and P. S. Dodds. Game story space of professional sports: Australian rules football. *Physical Review E*, 93(5):052314, 2016. URL https://doi.org/10.1103/PhysRevE.93.052314. [p156]

S. Lenor, L. J. Lenten, and J. McKenzie. Rivalry effects and unbalanced schedule optimisation in the australian football league. *Review of Industrial Organization*, 49(1):43–69, 2016. doi: https://doi.org/10.1007/s11151-015-9495-7. [p156]

L. J. Lenten. Racial discrimination in umpire voting: an (arguably) unexpected result. *Applied Economics*, 49(37):3751–3757, 2017. URL https://doi.org/10.1080/00036846.2016.1267848. [p156]

L. J. Lenten, A. C. Smith, and N. Boys. Evaluating an alternative draft pick allocation policy to reduce 'tanking'in the australian football league. *European Journal of Operational Research*, 267(1):315–320, 2018. URL https://doi.org/10.1016/j.ejor.2017.11.029. [p156]

L. J. Lenten, P. Crosby, and J. McKenzie. Sentiment and bias in performance evaluation by impartial arbitrators. *Economic Modelling*, 76:128–134, 2019. URL https://doi.org/10.1016/j.econmod.2018.07.026. [p156]

K. Marshall. The effect of leadership on afl team performance. In *Proceedings of MathSport International 2017 Conference*, page 255, 2017. [p156]

D. O'Shaughnessy. On the value of afl player draft picks. In *10th MathSport Conference, Darwin, Australia*, 2010. [p160]

D. Romer. Do firms maximize? evidence from professional football. *Journal of Political Economy*, 114(2):340–365, 2006. URL https://doi.org/10.1086/501171. [p155]

R. Ryall and A. Bedford. An optimized ratings-based model for forecasting australian rules football. *International Journal of Forecasting*, 26(3):511–517, 2010. URL https://doi.org/10.1016/j.ijforecast.2010.01.001. [p158]

D. Temple Lang. *XML: Tools for Parsing and Generating XML Within R and S-Plus*, 2020. URL https://CRAN.R-project.org/package=XML. R package version 3.99-0.5. [p156]

H. Wickham. *rvest: Easily Harvest (Scrape) Web Pages*, 2020. URL https://CRAN.R-project.org/package=rvest. R package version 0.3.6. [p156]

H. Wickham, R. Francois, L. Henry, K. Müller, et al. dplyr: A grammar of data manipulation. *R package version 0.4*, 3, 2015. [p156]

K. Woolner. Introduction to vorp: Value over replacement player. *Retrieved from Stathead. com: https://web. archive. org/web/20070928064958/http://www. stathead. com/bbeng/woolner/v orpdescnew. htm*, 2001. [p159]

R. Yurko, S. Ventura, and M. Horowitz. nflwar: A reproducible method for offensive player evaluation in football. *arXiv preprint arXiv:1802.00998*, 2018. URL https://doi.org/10.1515/jqas-2018-0010. [p160]

Robert N. Nguyen
School of Mathematics and Statistics
University of New South Wales
Sydney,NSW 2052 Australia
robert.nguyen@unsw.edu.au

James T. Day
Fusion Sport
Australia
jamesthomasday@gmail.com

David I. Warton
School of Mathematics and Statistics and
Evolution & Ecology Research Centre
University of New South Wales
Sydney,NSW 2052 Australia
David.Warton@unsw.edu.au

Oscar Lane
lane.oscar@gmail.com