Proceedings

14th Australasian Conference on Mathematics and Computers in Sport (ANZIAM Mathsport 2018)

University of the Sunshine Coast

25-28 July 2018

Edited by Ray Stefani and Anthony Bedford

Published by ANZIAM Mathsport.

All abstracts and papers have undergone a peer review.

ISBN: 978-0-646-99402-4

Table of Contents

Topic	Authors	Title	Page
AFL	Daniel T. Hoffman, Andrew J. Simmons, Paul B. Gastin	INVESTIGATING THE RELATIONSHIP BETWEEN INJURY AND MATCH OUTCOME IN AUSTRALIAN FOOTBALL LEAGUE MATCHES	5
AFL	Darren O'Shaughnessy	COMPONENTS OF HOME GROUND ADVANTAGE IN AUSTRALIAN RULES FOOTBALL	6
AFL	Bartholomew Spencer, Tim Bedin, Damian Farrow and Karl Jackson	A METHOD FOR EVALUATING PLAYER DECISION-MAKING IN THE AUSTRALIAN FOOTBALL LEAGUE	7
AFL	Robert Nguyen, Kevin Murray, Berwin Turlach	TAKING HOME CHARLIE	13
AFL	Damien Gattuso, Ian Grundy	USING SHARED EXPERIENCE TO MEASURE THE COHESIVENESS-PERFORMANCE RELATIONSHIP IN THE AFL	19
Basketball	A. Gorman, J. Headrick, I. Renshaw, C. J. McCormack, and K. Topp	CHOOSING AN APPROPRIATEL SIZED BASKETBALL FOR JUNIOR PLAYERS	25
Beach Volleyball	Mattina Koleen T. Yee, Dylan Antonio S.J. Talabis	CONSTRUCTING BEACH VOLLEYBALL STRATEGIES IN EXTENSIVE FORM	26
Cricket	Sam Greer, Ankit K. Patel, Holly Trowland, and Paul J. Bracewell	THE IMPACT OF INJURY ON THE FUTURE PERFORMAMCE RATINGS OF DOMESTIC T20 CRICKETERS	32
Cricket	Zeana Mansell, Ankit K. Patel, Jack T. McIvor and Paul J. Bracewell	MANAGING RUN RATE IN T20 CRICKET TO MAXIMISE THE PROBABILITY OF VICTORY WHEN SETTING A TOTAL	38
Cricket	Jack T. McIvor, Ankit K. Patel, Tamsyn Hilder, Paul J. Bracewell	COMMENTARY SENTIMENT AS A PREDICTOR OF IN-GAME EVENTS IN T20 CRICKET	44
Cricket	Phillip P. Simmonds, Ankit K. Patel, and Paul J. Bracewell	USING NETWORK ANALYSIS TO DETERMINE OPTIMAL BATTING PARTNERSHIPS IN CRICKET	50
Cricket	Ankit K. Patel, Madeleine K. A. Cook, Paul J. Bracewell and Mathew B. West	A FRAMEWORK TO QUANTIFY THE IMPACT OF SOCIAL ENGAGEMENT ON DATA DRIVEN CREATIVE	56
Cricket	Bernard J. Kachoyan , Marc West	DERIVING AN EXACT BATTING SURVIVAL FUNCTION IN CRICKET	62
Cricket	Ankit K. Patel, Paul J. Bracewell and Michael G. Bracewell	ESTIMATING THE EXPECTED TOTAL IN THE FIRST INNINGS OF T20 CRICKET USING GRADIENT BOOSTED LEARNING	68
e-Sports	Niamh McDonald, Dave Matteo, & Minh Huynh	DATA OF DOTA: A PRELIMINARY ANALYSIS INTO FACTORS PREDICTING SUCCESS IN DOTA2	74
e-Sports	Lyn Kee, Minh Huynh	THE FUTURE AND RESEARCH OPPORTUNITIES OF ESPORTS BETTING	78
Gender Bias	Timothy S. McNamara, Tamsyn Hilder, Emma C. Campbell and Paul J. Bracewell	GENDER BIAS AND THE NEW ZEALAND MEDIA'S REPORTING OF ELITE ATHLETES	84
Golf	Ankit K. Patel, Samuel J. Rooney, Paul J. Bracewell and Jason D. Wells	CONSTRUCTING A PREDICTIVE PGA PERFORMANCE RATING USING HIERARCHICAL VARIABLE CLUSTERING	89
Netball	Anthony Bedford, Noeline Taurua and Kylee Byrne	ON THE SUNCORP SUPER NETBALL 2018 LADDER POINTS SYSTEM	95
Performance Measurement	Paul Smith and Anthony Bedford	A FLEXIBLE METHOD OF JUMP AND HIGH INTENSITY EVENT DETECTION	101
Predictive Methods	Ankit K. Patel, Paul J. Bracewell, Jason D. Wells and Patrick Brown	PREDICTING FOOTBALL CROWD ATTENDANCE WITH PUBLIC DATA	107
Predictive Methods	Ankit K. Patel and Paul J. Bracewell	A FRAMEWORK FOR QUANTIFYING THE EFFECTIVENESS OF HUMAN-BASED RATING SYSTEMS	113
Rugby	Emma C. Campbell, Ankit K. Patel and Paul J. Bracewell	OPTIMISING JUNIOR RUGBY WEIGHT LIMITS IN NEW ZEALAND	119
Rugby	Wynton E. Moore, Sam Rooney, Paul J. Bracewell and Ray Stefani	SYSTEMATIC OPTIMIZATION OF THE ELO RATING SYSTEM	125
Rugby	Wynton E. Moore, Paul J. Bracewell, Jack T. McIvor, and Ray Stefani	DERIVING RESULT-DRIVEN RUGBY PLAYER PERFORMANCE RATINGS	131

Topic	Authors	TITLE	Page
Rugby	Phillip P. Simmonds, Timothy S. NcNamara and Paul J. Bracewell	PREDICTING WIN MARGINS WITH SENTIMENT ANALYSIS IN INTERNATIONAL RUGBY	137
Speed Skating	Ray Stefani	UNDERSTANDING OLYMPIC SPEEDSKATING: EFFECTS OF ICE, SKATES, GENDER DIFFERENCES, PERCENT IMPROVEMENTS	143
Tennis	Stephanie Kovalchik	AN APPROACH FOR ASSESSING SERVE PREDICTABILITY	149
Tennis	Martin Ingram	PREDICTING THE OUTCOME OF TENNIS MATCHES USING GAUSSIAN PROCESSES	150

Keynote Speakers

The keynote speaker is Noeline Taurua, former New Zealand Silver Fern netball player and silver medalist at the Commonwealth Games in 1998, former Silver Fern coach and current Sunshine Coast Lightning netball coach. Noeline will give the attendees her perspective as to the effective use of data analysis in shaping player success in her talk Shedding Light on Data in Sport: The Elite Coach.



Invited Speakers

The conference welcomes Scott McLean who will speak on A Systems Approach to Optimizing Performance Analysis in Football and Keith Davies, an expert on Modern Performance Analysis and Nacsport, a video analysis tool.

Sunshine Coast Lightning Analyst

Anthony Bedford serves as Sunshine Coast Lightning Analyst. In addition to his contributed talk On the Suncorp Super Netball 2018 Ladder Points System, he will talk about the challenges of interpreting videos and match data and about The Realities of Probabilistic Models, Visualizations and Expectations on Athletes.

INVESTIGATING THE RELATIONSHIP BETWEEN INJURY AND MATCH OUTCOME IN AUSTRALIAN FOOTBALL LEAGUE MATCHES

Daniel T. Hoffman [a,c] , Andrew J. Simmons [b] , Paul B. Gastin [a]

[a] *Centre for Sport Research, Deakin University, Geelong, Australia*
[b] *Applied Artificial Intelligence Institute, Deakin University, Geelong, Australia*
[c] *Corresponding author: d.hoffman@deakin.edu.au*

Abstract

The aim of this study is to investigate the relationship between injury and match outcome in Australian Football League (AFL) matches. Seven years of match betting odds, match injuries and final margins were analysed in two ways; from the perspective of the home (versus away) and favoured (versus unfavoured) team. Probabilities derived from betting odds at the start of the match were corrected for over-round and used to indicate the predicted match outcome. Final margin was used to represent actual match outcome. Linear regressions and t-tests were performed to examine the nature and strength of the relationships; the influence of predicted match outcome on match injuries, and of match injuries on actual match outcome. Predicted match outcome for the home and favoured team was uncorrelated with match injuries, thus suggesting against the theory that injuries are caused by unbalanced competition. In contrast, match injuries for the home ($r = -0.19$, $p < 0.001$) and favoured ($r = -0.20$, $p < 0.001$) team were negatively correlated with the actual match outcome. Compared to matches where both teams sustained the same number of injuries (home team = 9 point margin, favoured team = 27 point margin), home and favoured teams' average final margin decreased ($p < 0.01$) when they sustained one or more injuries than the away (home team = -2 point margin) and unfavoured (favoured team = 16 point margin) team respectively. A home ground advantage is equivalent to approximately one match injury in the AFL. In conclusion, match injuries can influence the outcome of a match in the AFL, thus clubs that can find ways to reduce their match injuries have the potential to increase their likelihood of a successful match outcome.

Keywords: Australian football, injury, match, team, competition

Acknowledgements

The authors would like to thank the AFL, AFLDA, and the AFLPA for granting research approval and supplying the injury surveillance and Champion Data databases.

COMPONENTS OF HOME GROUND ADVANTAGE IN AUSTRALIAN RULES FOOTBALL

Darren O'Shaughnessy [a,b,c]

[a] *Ranking Software, Melbourne*
[b] *Hawthorn Football Club, Melbourne*
[c] *Corresponding author: darren@rankingsoftware.com*

Abstract

Dozens of papers (e.g. Clarke, 2007) in Australian Rules Football – and hundreds in other sports – have explored the size and psychological causes of home ground advantage (HGA). This paper explores the on-field actions that correlate most strongly with the home field scoring boost over a travelling team. In common with many analyses, it appears that umpires are responsible for approximately one-third of the extra points scored by home teams. In support of theories that posit crowd noise as the primary influence on umpires, those types of free kicks that allow for the most sustained vocal contributions are most biased in favour of the home team.

In general, successful teams are better at both winning the ball from contested situations, and generating successful attacks by creating space in dangerous positions. This paper also compares the nature of HGA-tainted statistics with those from a cohort of similar results at neutral venues. Indicators of "hard effort" such as winning contests are stronger for home winners than for non-home winners, but indicators of "creating space" are weaker than for others. This suggests that the psychological frame of mind for home players is more prominent in reactive behaviour ("see ball, get ball") than in cerebral proactive behaviour ("implement the plan and examine the playing area for space").

Keywords: AFL, Australian Rules Football, home ground advantage, home field advantage

References

Clarke, S.R. (2007). Home advantage in the Australian football league. *Journal of Sports Sciences, 23*, 375-385.

A method for evaluating player decision-making in the Australian Football League

Bartholomew Spencer [a,c] , Tim Bedin [b], Damian Farrow [a] , Karl Jackson [b]
[a] *IHES, Victoria University, Melbourne, Australia*
[b] *Champion Data, Melbourne, Australia*
[c] *Corresponding author: bartholomew.spencer@vu.edu.au*

Abstract

Expected possession value metrics have been a recurring motif in the team sports analytics literature. They provide a means of identifying changes in the expected outcome (EO) as a result of a decision being made and executed by a player. Whilst existing methods identify whether a decision improved the EO, we wish to measure the value of a decision relative to alternative options. Hence, in the presence of multiple positive options, improving the EO is not the only measure of success. In our work on Australian Rules football, the EO of kicking to a teammate is quantified *via* both the probability and value of retaining possession. We measure the former by identifying the theoretical between-player contest that could occur. We model this based on players' velocity, orientation, and the effects of these constraints on re-positioning, as measured from player-tracking data. Unlike traditional metrics such as Voronoi tessellations that limit spatial ownership to a single player, we express the partial (or, contested) dominance of players. We treat players as dynamic objects, capable of repositioning during the ball's trajectory, and consider variation in kicking accuracy. Hence, the optimal receiving location of targets are identified, by searching for local maximums in their vicinity. Analysis of Australian Football League (AFL) matches played at Etihad Stadium in 2017 reveals a trend towards short-range kicks with minor improvements in EO, rather than higher value decisions which are typically to long-range targets. The difference between the theoretical contest of successful and unsuccessful kicks is found to be statistically significant. We present a framework for analysing the decision-making of individuals by quantifying the value of a decision and identifying more valuable alternatives. This metric has applications in player selection and recruitment, performance analysis and predictive analytics.

Keywords: performance analysis, player tracking, spatiotemporal, team sports

1. INTRODUCTION

Quantifying the expected outcome of a possession has been a commonly researched topic in team sports, such as in basketball [1], soccer [2], and Australian Rules football [3, 4]. Results of these studies reveal the importance of space, effects of congestion and pressure, and provide novel insights into player decisions by quantifying the net increase in the expected outcome (EO). While these studies have evaluated the decision-making abilities of players by identifying positive increases in the EO (such as in [1]), we wish to measure the value of a decision relative to alternative options. Hence, in the presence of multiple positive outcomes, improving the EO is not the only measure of a good decision.

In this study we demonstrate a method for quantifying the value of decisions that players make in Australian Rules football. We calculate the EO of a kick by measuring the probability that it will be successful and the value of retaining possession. The probability of retaining possession is calculated by modelling the theoretical contest that could occur based on player positions, orientation and velocity. In doing so, we measure spatial occupancy as a contest that could develop in the time between a kick is made and its arrival (i.e., the travel time of the ball). While Voronoi tessellations have traditionally been used to measure spatial dominance in team sports [5], we believe a contested occupancy model is more appropriate if we consider players as moving objects, capable of repositioning. The model is built upon player-tracking data collected in the Australian Football League (AFL).

2. METHODS

DATA

Player tracking data was collected for all matches played in the 2017 AFL regular season. Matches were limited to those played at Etihad Stadium, ensuring consistent field dimensions and data quality. Tracking data, recorded at 10 Hz, were collected via Catapult's Clearsky local position system (LPS). Matches in which tracking of one or more players was missing for any period were omitted. Detailed match events (e.g. marks, kicks, goals and associated constraints such as pressure type and location), referred to as *transactions*, were manually recorded to

the nearest second by Champion Data. In total, seven matches were used in this study, each consisting of approximately 1.5 million rows of time-stamped coordinates and 3285 rows of transactions. Field equity values were provided by Champion Data. Ball location was extracted from consolidated transaction and tracking datasets by identifying the gain and loss of possession. For model formation and analysis, only possession beginning in a mark and resulting in a kick were included, as these conditions provide the player with adequate opportunity to make an informed decision. Player velocity and orientation were derived from raw positional data. We assume a player's orientation is equal to the deviation of the vector of two consecutive tracking samples.

DECISION MAKING MODEL

While methods exist for measuring the net contribution a decision has made to the expected outcome of a possession chain [1], including transaction-based models in AFL [3], our objective is to identify the value of a decision, relative to available options. If a kick results in an equity improvement of x, while an alternative option existed that would have resulted in an equity gain of y (where $y > x$), the decision behind the decision is sub-optimal despite a positive equity gain. The value of a decision is quantified as the EO of the decision that was made, divided by the EO of the optimal decision. We refer to this metric as the Decision Value (DV).

$$DV = EO_{decision}/EO_{max} \tag{1}$$

The EO of deciding to kick to a spatial location, x, is quantified by measuring the risk and reward components of that decision (2). The reward is equal to the probability of retaining possession when kicking to x (p_x), multiplied by the equity (e) of the attacking team at x, while the risk is the probability of losing possession $(1 - p_x)$, multiplied by the opponent's equity.

$$EO_x = p_x e_{team} - (1 - p_x) e_{opp} \tag{2}$$

We measure the probability of successfully retaining possession by modelling the theoretical contest that could occur at x based on the current position, velocity, and orientation of all players. We refer to the time it would taken an object (i.e., a player or the ball) to reach a spatial location as its *time-to-point*. Ball velocity was considered as fixed at 18.5 m/s as approximated from measuring the travel time of kicks from a match of AFL data. Hence, the players who could contest a kick to x are those with a time-to-point less than the ball's. The probability of retaining possession is equal to the number of teammates with a time-to-point less than the ball's for a location, x, divided by the total number of players who meet the same criteria.

Determining if a player could reach a location, x, in less time than the ball's time-to-point requires measuring the effects of orientation and velocity on repositioning. We record every movement (distance, metres, and angle, degrees) for whole-second integers (≤ 5 s) across four matches of tracking data. These movements are normalised for orientation, grouped into integers, and individual ellipses are fitted in the positive and negative y-axis (equivalent to moving forwards or backwards, relative to current orientation) such that the y- displacements are equal to the maximum and minimum recorded y- values respectively, and the x- displacement for both ellipses is equal to the maximum x- value recorded. The ellipses are joined around the y- axis origin, producing egg-shaped boundaries that represent the maximum distance a player could reach in specified time intervals. We refer to these as *Reachable Regions* (RR). The concept of RR in team sports were introduced in [6] and further explored in [7] where they were fit via a number of methods including convex hulls and motion models. In [7], RR were used in the formation of dominant regions, a variation of Voronoi tessellations that consider player orientation and velocity. Our use of ellipses reduces computation time and produces a smoother bound, at the cost of precision. That is, our method fits ellipses based on the maximum movements, then extrapolates limits between these points, while convex hulls produce unsmoothed bounds that consider all available data. This method represents a novel way of quantifying special dominance.

To calculate the probability of retaining possession at x, RR are produced for all players for the ball's time-to-point. As RR are fit on whole second increments, extrapolation is needed for partial seconds. Any player whose RR contains x is said to be able to contest the kick. We employ the standard ellipse function to check for this (3), which determines that a player could reach the given location (x, y) based on the ellipse width and height (a, b), if the function is satisfied. The dominance, or probability of retaining possession, is equal to the number of teammates divided by the total number of players who could contest. This approach assumes the worst case – that any player who could reach x, based on physical constraints, will do so.

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} \leq 1 \tag{3}$$

As we consider players as able to reposition, rather than as stationary objects, the optimal location to receive a kick need not be the player's current location. We employ a local maximum search to determine the optimal receiving location, based on EO, for a kicking player's 17 teammates.

Kicking Variation

Given that players lack perfect kicking accuracy, we introduce variability in the form of a 2D Gaussian distribution. Player accuracy is difficult to quantify with existing data as we require perfect information on intended targets. For the purpose of this study, we assume variability can be quantified by a Gaussian with covariance equal to 5% of the total kicking distance, suggesting that long-range kicks are more difficult to execute than short-range. With the addition of variation, the modified EO of a kick is equal to the summed product of the Gaussian's probability density function (PDF) and the raw EO values.

Statistical Analysis

The EO, DV and dominance were calculated for approximately 560 kicks in the 2017 AFL season. The explanatory power of our theoretical contest measure was examined by comparing the contest of successful and unsuccessful kicks. Decisions were extracted from the analysed matches and their characteristics summarised, providing insights into the types of decisions that are made in the AFL. We further explore decision making habits of players by summarising the characteristics of decisions that were identified as better options. These include the DV, the Euclidean distance between the kicker and target, and the dominance. Decisions, grouped by team, are compared across quarters to measure the correlation between decisions and score margins, calculated via the Spearman correlation coefficient (ρ).

3. Results

An example of a decision-making output is displayed in Figure 1. In this example, the player, highlighted in red, made a short kick resulting in a DV of 0.48. Local max searches identified two decisions that would achieve a DV of at least 0.55. Note that DV calculations consider variability, hence identified decisions are unlikely to have a DV of 1.0.

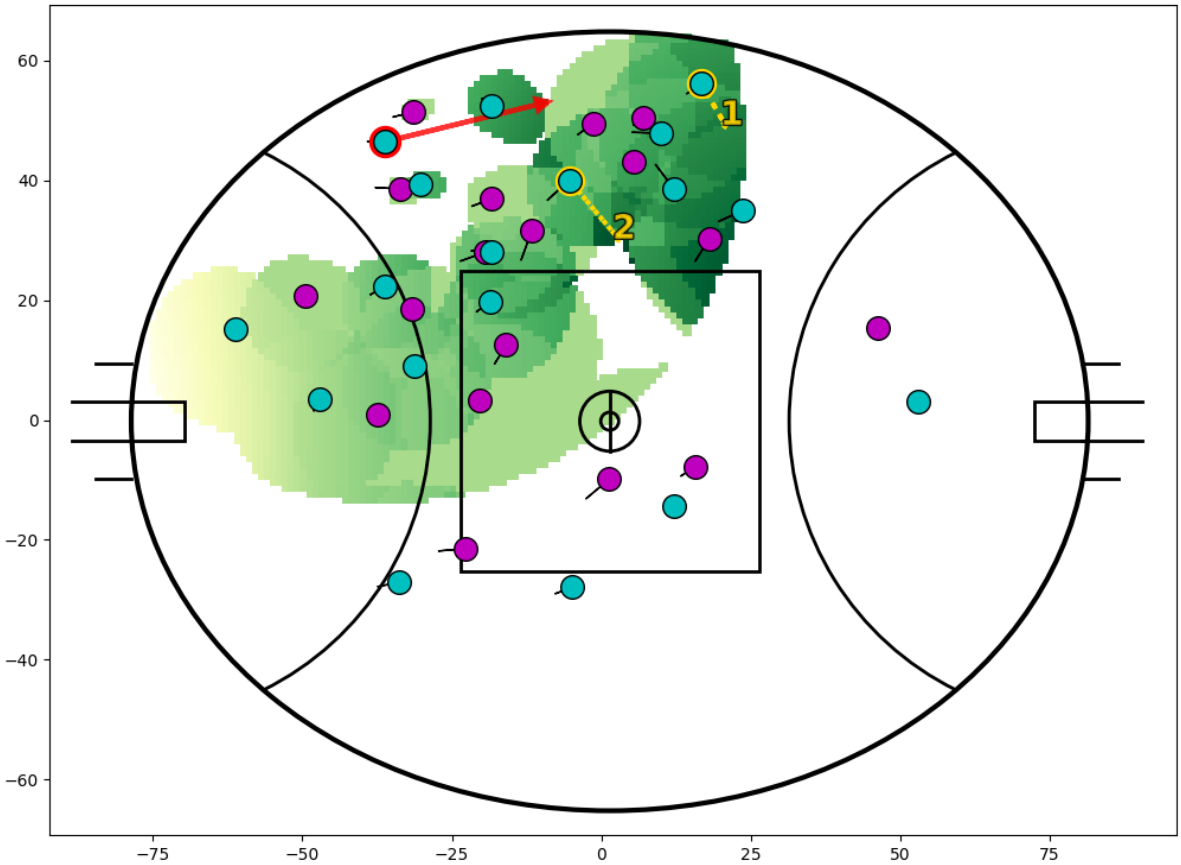


Figure 1. An example output of the decision-making model. The decision that was made by the player in possession is indicated by a red arrow pointing to the final position of the ball. Options identified as having higher EO are highlighted in yellow, with dotted lines representing the optimal receiving location of the identified

players. EO values within 60m of the kicker are visualised, with higher values represented by darker shades of green.

Fitting ellipses to movement data resulted in the RR presented in Figure 2. Due to sampling, data volume, and inaccuracies in wearable technology, some inconsistencies are present in the bands (e.g., subsequent whole-second movements can increase the ellipse boundary by smaller increments than previous bands).

The mean and standard deviation of calculated contests for successful and unsuccessful kicks across all matches was 0.60 ± 0.37 and 0.40 ± 0.25 respectively. There was found to be a significant difference between these two groups ($p < 0.001$).

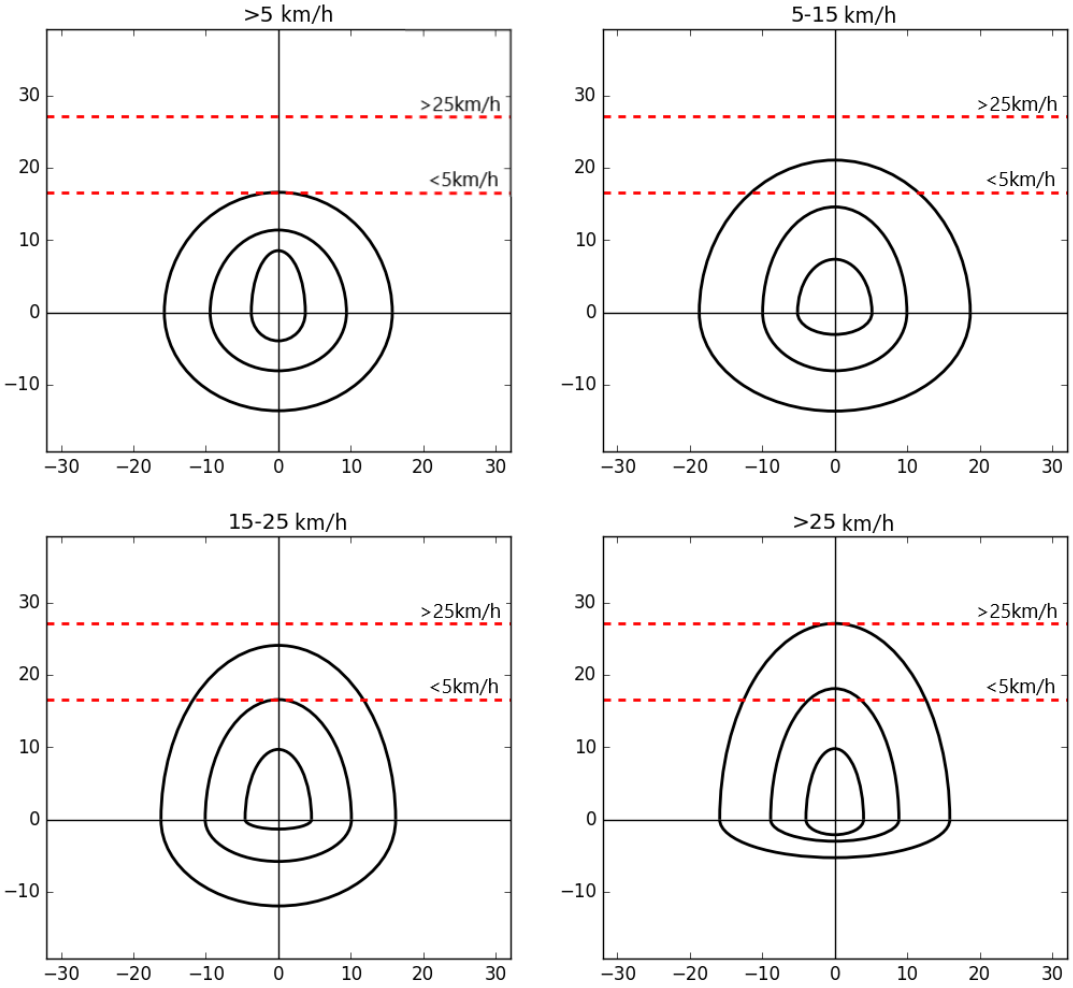


Figure 2. The RR of different speed bands for 1, 2, and 3-second integers. The red dotted lines refer to the 3-second maximum distance for the minimum and maximum speed bands.

Characteristics of kicking options are presented in Table 1. Options are grouped into the decisions that were made and the alternative options that were identified as having higher DV by the decision-making model. There was found to be significant differences between the characteristics of decisions that were made and the alternative options that were identified ($p < 0.001$). On average, kicks that were made are shorter (25.8 m) with lower average dominance but higher variance.

A weak, negative correlation was identified between decision making and score margin ($\rho = -0.14$). Further, there was not a statistically significant difference between the DV of winning and losing teams ($p = 0.85$).

	Distance (m)		Dominance		Decision Value	
	Mean	St.D.	Mean	S.D.	Mean	S.D.
Decisions	25.8	11.1	0.58	0.30	0.24	0.24
Alternatives	36.0	10.0	0.85	0.18	0.55	0.15

Table 1. Summarised decision data.

4. DISCUSSION

Our objective of this research was to develop a method for measuring player decision making in the AFL. To achieve this, we the theoretical spatial dominance of teams derived from player tracking data. Our methods consider the velocity and orientation of individuals. While prior studies have quantified if a decision resulted in an improved possession outcome, the methods detailed in this study value a decision relative to alternatives.

Decision-making results revealed a tendency towards short kicks, while statistical modelling identified long-range targets as having higher EO due to their much higher equity and generally lower theoretical contest. This trend towards shorter kicks is logical due to the lower variability and execution time. Perhaps more importantly, close options are more likely to be identified by players in a shorter period of time due to less visible obstruction. These decisions having a lower theoretical contest may suggest limitations in its calculations. Notably, we assume equal contest ability from players (hence, a player who excels at contesting the ball would be undervalued). Further, we do not consider the effects of interference (e.g., a player standing in front of another). Developing a predictive model to quantify the probability that an individual will commit to a contest, based on spatial features, may address these concerns.

Analysis of player movements revealed minimal differences between speed-bands, particularly in the walking and running categories. At higher velocities, players are not able to cover as much negative space. Inaccuracies in player tracking devices resulted in noisier bands than would be expected in optimal conditions. As we collect more data, it will be possible to produce RR for individuals, allowing for consideration of their maximum velocity and varying ability to accelerate and reorient.

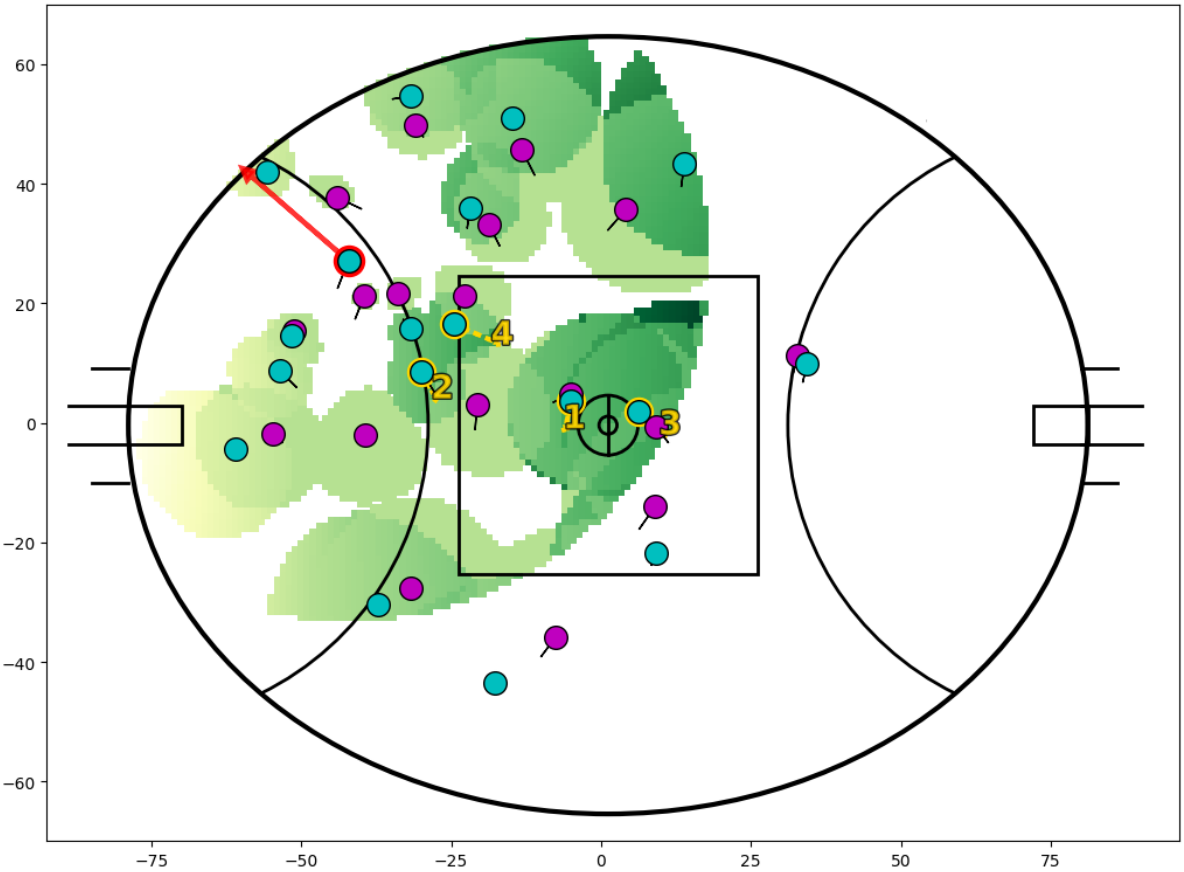


Figure 3. An example output that identifies optimal decisions requiring the ball to be kicked along a path that contains opponents.

While insightful, our decision-making model has limitations due to assumptions that must be made in the absence of additional data. Ball tracking technology does not exist in AFL, hence a reliance on play-by-play possession data to infer its position. If a kick results in a contest, isn't collected cleanly, or goes out-of-bounds, this must be omitted from the analysis as the target location cannot be determined with reasonable accuracy. Without precise locations, it is difficult to record information on kicks that were unsuccessful, as these are less frequently marked by opponents than they are recovered after a contest. Hence, without enough samples to model their relationship, we assume a linear relationship between a team's dominance and retaining possession. Despite that limitation, it was observed that successful kicks had a much higher calculated dominance than those that were unsuccessful. Finally, we assume balls can be kicked to any teammate, which may be unrealistic depending on the trajectory required to pass by opponents. Figure 3 demonstrates optimal decisions that would require the ball to be kicked above opponents. Physics-based modelling of kicks could be added to the calculations to omit unrealistic receiving locations such as in [8].

Should continued work address these limitations, a more robust decision-making metric could be developed. Applications of such a metric would include the ability to measure a player's decision-making capabilities, akin to how we conventionally measure other aspects of performance. This would allow for more informed recruitment, tactics, and match preparation. This need not be limited to the decision-making of players with the ball. For example, movements of players who don't possess the ball may lead to the creation of areas of space, measured by EO, and methods to quantify this could be the subject of future research.

5. CONCLUSION

The primary aim of this research was to develop a method for quantifying decisions made by players in Australian Rules football. Our focus has been on the expression of space as a continuous contest comprising of individuals who are capable of repositioning during the time between possessions. We express the value of a decision as the expected outcome of the kick that was made, divided by the maximum value identified *via* statistical analysis of field equity and possession retention. These methods were exemplified on kicks resulting from marks. Analysis of decisions found a trend towards kicking to close teammates with lower calculated EO than long-range targets. The decoupling of player decision-making from current performance metrics has applications in player selection, recruitment, and performance analysis.

References

[1] Cervone, D., D'Amour, A., Bornn, L., & Goldsberry, K. (2014). POINTWISE: Predicting points and valuing decisions in real time with NBA optical tracking data. In *Proceedings of the 8th MIT Sloan Sports Analytics Conference.*

[2] Green, S. (2012). Assessing the performance of Premier League goalscorers. *OptaPro Blog*. Retrieved 13 June, 2018, from http://www.optasportspro.com/about/optapro-blog/posts/2012/blog-assessing-the-performance-of-premier-league-goalscorers/

[3] O'Shaughnessy, D. M. (2006). Possession versus position: strategic evaluation in AFL. *Journal of sports science & medicine*, *5*(4), 533.

[4] Jackson, K. (2016). *Assessing player performance in Australian football using spatial data*. (Doctoral dissertation, PhD Thesis, Swinburn University of Technology)

[5] Memmert, D., Lemmink, K. A., & Sampaio, J. (2017). Current approaches to tactical performance analyses in soccer using position data. *Sports Medicine, 47*(1), 1-10.

[6] Taki, T., & Hasegawa, J. I. (2000). Visualization of dominant region in team games and its application to teamwork analysis. In *Computer Graphics International, 2000. Proceedings* (pp. 227-235). IEEE.

[7] Horton, M., Gudmundsson, J., Chawla, S., & Estephan, J. (2015, May). Automated classification of passing in football. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 319-330). Springer, Cham.

[8] Spearman, W., Basye, A., Dick, G., Hotovy, R., & Pop, P. (2017). Physics—Based Modeling of Pass Probabilities in Soccer. In *Proceeding of the 11th MIT Sloan Sports Analytics Conference*.

Taking Home Charlie

Robert Nguyen [a,b] , Berwin Turlach [a], Kevin Murray [a]
[a] *University of Western Australia*
[b] *Corresponding author: nguyen.n.robert@gmail.com*

Abstract

Australian Rules Football is the predominant sport on the Australian sporting landscape. Since the first match in 1858, the sport has embedded itself into Australian culture and evolved into a billion dollar industry.

In the spirit of true sportsmanship and fair play, a prestigious award—dubbed the Charles Brownlow Trophy (better known as the Brownlow Medal)—is awarded to an individual player who plays in the Australian Football League (AFL). The winner is determined through an umpire voting system where votes are awarded to the top three players per match. The winner of the Brownlow Medal is the player who has tallied the most points over the season and has also met the 'fairest' criteria. Like most sporting awards, the progress of votes throughout the year is not known by any person due to tight security and secrecy. The votes are only opened and tallied at the annual Brownlow award ceremony. Naturally, interest in the Brownlow winner garners high speculation throughout the year and is facilitated by betting agencies

This paper aims to identify a statistical methodology that may be used to predict the potential Brownlow medallist in a given year. In addition, this paper will use readily available public data in the hopes that it will drive statistical interest in AFL. Previous studies have been completed in this area including work by Bailey and Clarke (2002) which identified leading Brownlow contenders; however, a model predicting the overall winner still remains elusive. This paper hopes to extend the work of Clarke and Bailey using alternative methods to develop a model to accurately predict the overall winner.

Finally, we compare the proposed method (proportional odds ordinal regression) with a probability adjustment method from Champion Data (the official statistics provider of the AFL). We find that our predicted top 10 has less ranking error compared to Champion Data and suggest that this is in turn a great way to promote statistics.

Keywords: interest, statistics, reproducibility

1. INTRODUCTION

Australian Rules Football was originally developed in 1858 by Thomas Wentworth Wills. Wills wanted to keep cricketers fit during winter and subsequently designed a game that was a combination of all known football codes at the time, thus making the sport uniquely Australian. Since then, the sport has evolved and become embedded into Australian culture.

The Brownlow Medal ('Charlie') is the most prestigious individual award on offer in the Australian Football League (AFL). It is awarded to the player who is adjudged the 'best and fairest' across the home and away season. The award was named in honour of Charles 'Chas' Brownlow a former Geelong footballer, secretary and Victoria Football League (VFL) president who passed away in 1924—the year the medal was first awarded.

The Brownlow medallist is determined through a unique voting process where, at the end of each game, the respective games' umpires allocate votes to three players. The umpires allocate votes of three, two and one based on the 'best and fairest' criteria. That is for a single match, three votes to the player that the umpires adjudge to be best, two to the second best and one to the third best. Players who receive more than 100 demerit points from the Match Review Panel for an on-field incident (which most often results in a suspension) during the home and away season are ineligible to receive the medal, as they no longer meet the 'fairest' criteria[1]

This voting system has been criticised in the past due to the level of subjectivity involved with the umpires' decisions. Indeed, even for the umpires themselves, the top three players for a match can be difficult to objectively ascertain and this difficulty can vary from game to game, and umpire to umpire[2] . However, a lot

[1] However, suspended players may still receive votes, but will not be eligible for the overall Brownlow medal.

[2] https://www.heraldsun.com.au/sport/afl/kane-cornes-afl-umpires-shouldnt-give-brownlow-medal-votes-should-focus-on-administering-rules/news-story/4df94899ab764513b6f2bca636fcc522?nk=22e9c4f4f79a8121a1df6ec166854622-1533039517

has changed since the first Brownlow medallist Edward 'Carji' Greeves was announced in 1924. It is hard for many to imagine a time that the Brownlow and its associated gala have not attracted large public interest and fanfare, particularly with the annual televised specials. The award's humble beginnings started with Greeves being simply awarded the medal during a VFL board meeting. Even in 1956, hype surrounding the medal was so low that that year's winner, Peter Box, found out through his neighbour who had just heard the news on the radio.

Over subsequent years the media coverage grew, with the first televised broadcast being made in 1970 on HSV 7 (Slattery, 2010). It has now reached a point where the Herald Sun[3] can spend over four pages on just the dresses of the players' wives and girlfriends – something akin to the coverage of a celebrity red carpet event like the Oscars.

As interest in the Brownlow has increased, so too has the interest in who will win the Brownlow Medal – and as with most modern-day sports so too has the interest in betting. Betting markets have evolved as well, with markets available for who will win outright, who will win amongst sub-sets of players and various other exotic bets.

These betting markets are not always efficient, as can be seen from an example from the end of the 2013 AFL season. Betting odds on star player Sam Mitchell[4] plunged rapidly from 20 : 1 to 6 : 1 after a media punter on The Footy Show[5] backed him to win the medal. This suggests that betting markets are often a function of individual opinion, rather than a sober evaluation of all information available to the market. This plunge occurred two weeks after the home and away season had concluded, and was thus not based on any new information.

As of 2014 there has been only one published study by Bailey and Clarke (2002) on predicting the Brownlow winner; see also Bailey (2005)[6]. The study fitted logistic regression models to current AFL data and came up with leading contenders for the Brownlow Medal. However, this study was unable to predict the overall winner.

This paper identifies a statistical method that may be used to predict the potential Brownlow medallist in a given year with a heavy focus on reproducibility for the general public which we hope will engage the publics interest in modelling AFL Clarke (2003)

To deliver this we encounter 2 constraints:

1. A readily available data source; and
2. A reproducible methodology that works.

2. DATA

There are a number of sources from which to obtain the necessary data, with one of the most direct methods being to acquire the data through a service such as Champion Data, a leader in statistics and graphics services to television broadcasters. This method is typically only available to companies and incurs a substantial cost – one which the general public would probably be unwilling to part with.

On the other end of the spectrum is to use the free and publicly available data sources on AFL websites. The key benefit here is that it is publicly available and free. However, this too comes with some disadvantages due to the nature of how these websites deliver their data – both in terms of format of tables / webpages and completeness of data. Methods of extracting this data can be as extreme as copy and pasting or something a bit more elegant such as writing a web-scraper.

The general public would not be expected to have the expertise to write a web-scraper effortlessly and would need to overcome many issues as described. Thus for the purposes of this paper, a web-scraper was built and packaged for public release called fitzRoy[7].

[3] A prominent Melbourne newspaper

[4] Sam Mitchell finished outside of the top ten players in the 2013 Brownlow

[5] A television football show about the AFL

[6] Bailey was Clarke's PhD student and conducted similar studies as part of his PhD thesis

[7] https://github.com/jimmyday12/fitzRoy

fitzRoy introduces a more direct, less costly and less time-consuming method to acquire the data we need from afltables[8]. It contains AFL data from 1987 to 2017 directly in the R package and consists of a web-scraper for users to pull any data after 2017.

2. METHODOLOGY

To effectively model the results of the Brownlow medal count, it was decided that a simplistic way that would raise interest in statistics should be used. In order to do this we made the following modelling decisions:

1. The model had to be interpretable to fans, this drove an inference over prediction accuracy (black box) kind of approach.
2. The 'obvious' is left out - everyone has an opinion on what matters when it comes to polling in the Brownlow medal. For example characteristics such as hair colour, tattoos, polling history and being captain have been postulated as influences on the ability to obtain points, just to name a few. By leaving out these 'obvious' characterisitcs it encourages others to add them in themselves.

Taking these decisions into consideration the proposed method is as follows:

1. Using the data from fitzRoy[7], we create ratios of each of the game statistics.

   For each player *i* and for each game *j*, we collect their individual statistics (kicks, handballs, marks, etc.) and create ratios where each player *i*'s total statistics for a variable are divided by the total of both teams; in game *j*. For example, player *i*'s kicks are divided by the total kicks by both teams' in game *j*.

   The reasoning behind this is that the *relative* performance amongst their *peers* is of intrinsic interest, rather than the actual number of kicks (or handballs etc.) that is important. In other words, the emphasis is placed on the importance of the ranking of players during a specific game.

2. Standardise all variables in preparation for performing the backwards selection process.

   Gelman (2008) tells us that subtracting the mean improves the interpretation of the main effects in the presence of interactions, which are expected in a dynamic game such as Aussie Rules. In addition, he notes that by dividing by the standard deviation, we put all explanatory variables on a common scale.

3. Fit a full model of all variables and the game margin

   After the variables have been standardised, we use the R package 'ordinal'[9] to fit the full model of all the ratio variables and the game margin. The CLM function of this package fits the following proportional odds model (Christensen, 2010).

$$P(Y_i \leq k) = g(\theta_k - x_i^T \beta)$$

4. Perform backwards elimination and get expected probabilities[10]

   Backwards elimination is then performed based on the AIC criteria using the 'MASS' package in R and the 'stepAIC' function (Venables and Ripley, 2002). This provides predicted probabilities for each player *i* in each game *j* of polling zero votes, one vote, two votes or three votes. By default each player *i*'s probability they will receive k votes will sum to one:

[8] https://afltables.com/afl/afl_index.html

[9] https://cran.r-project.org/web/packages/ordinal/index.html

[10] Where each player *i*'s probabilities sum to 1 across the game

$$1 = \sum_{k=0}^{3} p_{ki}$$

5. Calculate standardised probabilities[11] where p_{ki} is the probability that player *i* receives *k* votes

   As only *one* player can get each of the one votes, two votes and three vote it logically follows that the probability of three votes, two votes and one vote sum to one across *all 44 players* in game *j*. Thus we need to standardize the expected probabilities received from the model in step 4.

$$1 = \sum_{i=0}^{44} p_{ki}$$

   Standardising the probabilities so that this is true implicitly takes into account competition for votes between players in a game.

6. Calculate expected votes and expected standardised votes

   After obtaining both the expected probabilities[10] and the standardised probabilities[11] from steps 4 and 5, respectively, calculate the expected votes and the expected standardised votes for each player *i* for each game. These expected votes are then summed across the season and from this a predicted top ten players for the given years modelled can be obtained and ranked by number of expected votes.

3. RESULTS

This methodology was run with the free data from fitzRoy[7] for 2017 and we were able to compare our prediction to Champion Data the official statistics provider of the AFL.

As the interest in the Brownlow stems from who will win rather than by how much or the total vote count. We will use as our metric ranking error, which simply represents the difference between player *i's* predicted finishing position and their actual finishing position. Using this as our metric and comparing each top 10 we see the results below[12].

Interestingly, the use of free data and a comparatively simple proportional odds regression was able to produce a more accurate prediction than Champion Data's prediction.

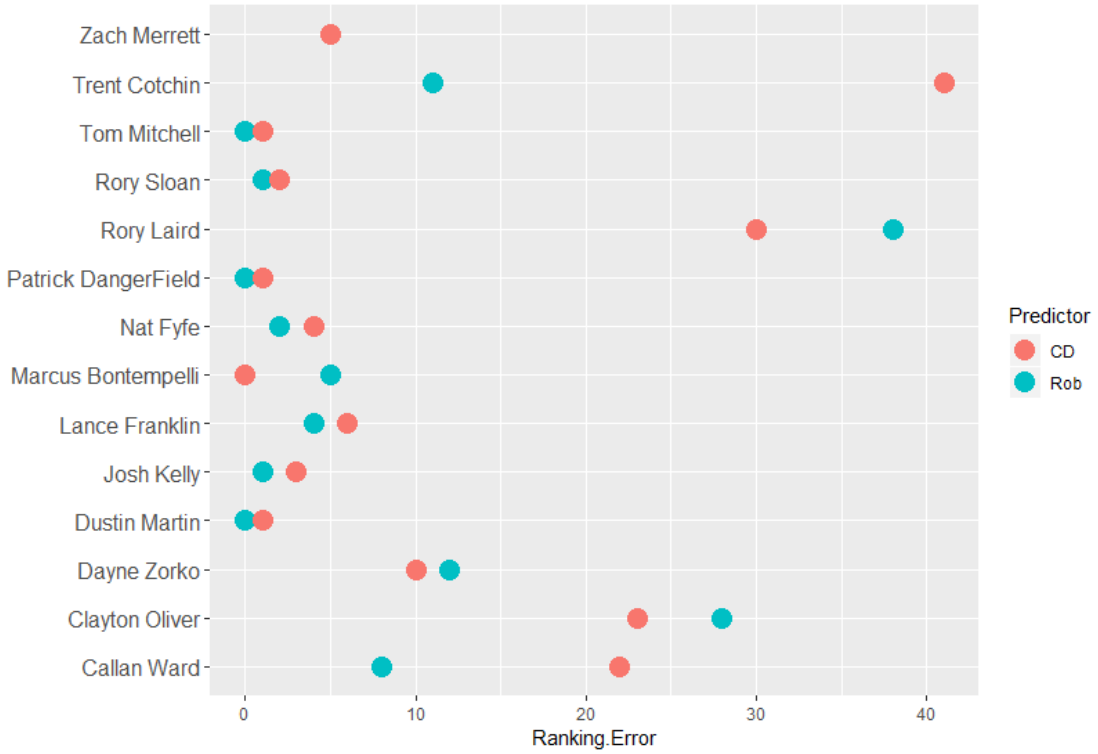
This is converse to general public opinion that Champion Data should have the advantage due to:

- **Better and expensive data:** Access to different and arguably more detailed data; and
- **More advanced modelling techniques:** Creation of 'non-basic' variables such as polling history combined with fitting a more advanced technique (neural nets[13]).

[11] Where the probability of *k* votes sums to 1 *across each game* for the players that played in that game.

[12] Champion data results were taken from https://www.triplem.com.au/sport/afl/news/champion-data-reveal-their-brownlow-predictions to calculate this metric ranking error

[13] As described on this LinkedIn profile: https://www.linkedin.com/in/rob-barone-nugent-8114b5b8/



Ranking Error of Each Player Rob and CD Predicted To Finish Top 10

4. DISCUSSION

The main aim of this work was to see if 'Charlie' can be accurately predicted from 'freely' available data with a method which was reproducible by the general public[14]. We hypothesise that this model can of course be improved with more detailed `robust' statistics. Of particular interest would be the use of data such as:

- **The field location of statistical variables.** Champion Data not only collects statistics on kicks, handballs, contested possessions etc. but also the location on the field that these statistics occurred. One would imagine a kick in the forward half of the ground setting up an attacking play is more 'vote-worthy' than a rushed kick out of defence under pressure.
- **Effect of Possession Chains.** Another statistic collected is 'being in a possession chain', which the AFL website uses as one of its metrics to 'rank' players. It would be likely that such a variable would have an effect in the allocation of umpire votes.
- **Umpire Bias.** While the subjectivity of votes has been mentioned previously our model here does not take into account the actual umpires who vote. This means that predictions do not factor in whether certain umpires are more likely to reward certain game characteristics or favour certain 'players'.

5. CONCLUSIONS

In this paper we show that using more classical statistical methods we were able to accurately predict the Brownlow Medallist in 2017. We can do this using readily available data that is fully reproducible to the general public. By making the prediction fully reproducible to the public, we hope that this encourages more fans to engage in the statistical side of AFL and to give AFL modelling a go.

[14] https://analysisofafl.netlify.com/

Acknowledgements
Robert wishes to thank anyone and everyone who has listened to him rant about statistics and data access. Kevin for first getting him into statistics and Berwin for encouraging freedom of thought.
Robert would also like to thank Dustin Martin for winning so this piece of work seems better than otherwise might be if not the case.
Go eagles

References

Bailey, M. and Clarke, S.R., 2002. Predicting the Brownlow medal winner. In *Proceedings of the sixth Australian conference on mathematics and computers in sport* (pp. 56-62). Gold Coast, University of Technology, Sydney.

Bailey, M.J., 2005. *Predicting sporting outcomes: A statistical approach* (Doctoral dissertation, Faculty of Life and Social Sciences, Swinburne University of Technology).

Clarke, S.R., 2003. Raising Interest in Statistics through Sporting Predictions on the Internet.

USING SHARED EXPERIENCE TO MEASURE THE COHESIVENESS-PERFORMANCE RELATIONSHIP IN THE AFL

Damien Gattuso [a], Ian Grundy [a,b]
[a] *RMIT University*
[b] *Corresponding author: ian.grundy@rmit.edu.au*

Abstract

According to the Harvard Business Review (2014), "team chemistry" is the "new holy grail of performance analytics". In a competition where restrictions have been put in place in attempt to homogenize talent across each club, the understanding of how and when players gel together as a team is becoming ever more important to the task of gaining a winning edge.
Leaving aside psychological factors, it seems obvious that "team cohesion" or "shared experience" must be a large component of team chemistry. The current study aims to draw the link between some simple measures of shared experience and team performance in an AFL context. Linear and logistic regression models are used to quantify the shared experience effect on home-and-away and finals performance. Comparison is made with some other recent approaches.
Overall, the cohesiveness-performance relationship is found to be significant with relatively small effect sizes. Support is found for some commonly held beliefs, e.g. that a settled defence and forward line correlates strongly with team success. However, there is no evidence to suggest that the cohesiveness of the midfield group has any effect on success.

Keywords: Cohesion, Shared Experience, Team Chemistry, AFL

1. INTRODUCTION

The rigorous measures put in place to enhance the 'competitive balance' of the Australian Football League (AFL) have resulted in constant discussion around the evenness of the competition, and the subsequent difficulty in attaining long-periods of success. This is due to the fact that AFL clubs cannot obtain as many elite players as they would like due to wage restrictions such as the salary cap. Successful clubs are further restricted in the pursuit of potential young talent, because the annual draft ensures that the poorest performing clubs get preference in selecting the best eighteen year olds in the country.

In Australian Rules football, player and team skill is largely judged on the accumulation of various statistics (kicks, marks, handballs, etc) during the course of a match or season. With such a plethora of information publicly available to everyone, club recruiters in the AFL have an impetus to investigate new ways of measuring player or team skill that could potentially create a competitive advantage, which can subsequently be exploited through a more targeted list management plan.

This paper will attempt to use the idea of team cohesion to introduce new measures that could be used in the analysis of the strength of a team's playing squad, without taking traditional on-field variables into account. It is a widely accepted notion that the cohesiveness of a sporting team is proportional to its chances of being successful, hence the old adage 'a champion team beats a team of champions'. However, there is a challenge in quantifying cohesion, as it is 'inconsistently defined and measured' (Salas, Grossman, Hughes, & Chris, 2015, p365). Contemporary methods for the measurement of cohesion in other contexts, e.g. business and psychology, involve the use of questionnaires, but this is clearly impractical in a professional sporting setting as access to opposition players is limited. In order to realistically measure the cohesion of opposing sides, new forms of measurement need to be derived. One type of measurement for team cohesion is hypothesised to be shared experience, under the assumption that the two factors are proportionally related to each other. Hence, the primary aim of this report is to determine if there is a significant relationship between various measures of shared experience, and the performance of AFL clubs throughout a season. The resultant findings will provide further evidence as to whether one or more of these measures can act as a proxy in the determination of team cohesion.

2. DATA DESCRIPTION

2.1 Measures of Cohesion – Shared Experience

Shared Experience is calculated here using three derived variables. The first variable is proportional club experience (PCE). PCE measures the total amount of time that a certain club's players spend at the club as a proportion of the total number years that the players have been playing professional football. The amount time that players spend at a club is capped to ten years in order to prevent any one player skewing the results. A possible flaw in this variable is that it will always be high if the total number of years that particular club's players have spent at another professional side is low, raising the possibility of an inexperienced team exhibiting a high PCE value.

$$\text{PCE} = \frac{\sum \text{Yrs At Club}}{\sum \text{Yrs At Club} + \sum \text{Yrs At Another Club}}$$

The second variable is what we will call Team Cohesion. This variable addresses the flaw in the PCE variable, as young, inexperienced teams cannot have a high value. A team cohesion value of zero means that every player at a club has spent at least ten years at another professional club, and zero years at their current club. Conversely, a team cohesion value of one indicates that every player at a club has spent at least ten years at their current club and zero years elsewhere. Therefore, this metric was designed to provide a percentage value on how close a team's squad is to theoretically perfect team cohesion.

$$\text{Team Cohesion} = \frac{(\sum \text{Yrs At Club})^{3.28} - (\sum \text{Yrs At Another Club})^{3.28}}{2 \times (10 \times \text{Number of Players})^{3.28}} + \frac{1}{2}$$

The optimal exponent of 3.28 was found to be the value that minimises the sum of squared errors between the difference of each observations team cohesion value, and its corresponding win percentage across each observation from the 2009 to 2017 seasons.

Our final measure of cohesion is Average Shared Experience. This variable measures the average number of years that players at a club have spent with their teammates. For example, at the start of the 2017 season, Wayne Milera had a shared experience value of 38 as he spent one year at Adelaide with 38 other players. For comparison, eight-year veteran Taylor Walker's shared experience value was 144 at the start the 2017 season. This variable is calculated for the whole team and also separately for the forward, midfield and defensive "teams" within each club.

These measures all use the years (or seasons) that players spend at a club instead of games played. The rationale behind this is that cohesion also encompasses shared experiences off field, such as familiarity with the game plan and assimilation into the team culture. However, games played are an important factor in the influence that players have in annual team performance. Therefore, the data used in this analysis are from players who played eleven or more games during the season in question. This can be referred to as a team's core group of players.

2.2 Measures of Performance

Multiple measures of success were used in the analysis because there are many ways in which the success of an AFL club throughout a season can be judged. For example, although the 2016 Western Bulldogs team finished the home and away season with the seventh highest win percentage (68%), their resultant premiership win that season meant that they were considered to be the most successful team in the competition. Furthermore, the AFL fixture is not fair, as only some teams get to play each other twice. As a result, some clubs receive more difficult draws than others, potentially resulting in end of season win rates and ladder positions that may not exactly represent the quality of a team. Since win percentage is not an exact measure of success, several performance measures were calculated. These were points for (offensive ability), points against (defensive ability), and the binary variables Top 4, Top 8, Bottom 4 and Premiership.
In order to test the relationship between cohesion and finals performance, the number of finals wins by each club who made the top eight was recorded. The number of finals wins for the Western Bulldogs in 2016 was

classified as three (as with every other team who won the premiership within the analysed time frame), in order for the observation to be used in the analysis.

The cohesion values of each team from the start of each season from 2009 – 2017 (n = 157) were compared to end-of-season performance. Data scarcity was due to the fact that Champion Data only provides statistics from the 1999 season, and a ten-year buffer was required to calculate some the cohesion variables.

3. METHODS

The two observations relating to GWS and Gold Coast's first seasons in the AFL, as well as Essendon's 2016 season, in which twelve players were banned for taking prohibited substances, were removed from dataset. These observations were removed because the abnormally low amount of skill in each team's squad clearly played a major role in their lack of success in each particular season. Other outliers in the dataset were detected by finding the Mahalanobis distance of each observation, and those with a distance larger than the critical value of 34.53 ($df=13$, $\alpha = 0.001$) were subsequently removed from the dataset. Outlier detection was performed using the SPSS software (Version 19, SPSS Inc., Armonk, New York, 2010).

Logistic and linear regression was used to measure the relationship between the measures for shared experience and success. The average age and total experience of clubs' players were included as additional explanatory variables in order to compare their effects with the aforementioned cohesion variables. Tests for multicollinearity between the predictors were carried out by measuring the variance inflation factor (VIF) for all predictors, and then removing correlated variables until the VIF of all remaining predictors were below four. Variables found to be highly correlated were separately regressed against the dependent variables. All regression modelling and analysis was performed using RStudio version 3.4.1 (RStudio Inc., Boston, Massachussets, 2017). The *gvmla* function from the *gvlma* package was used to test for linearity and homoscedasticity within the linear models, and the normality of residuals was tested using the Shapiro-Wilk test with $\alpha = 0.05$.

The win percentage and average future win percentage variables were modelled using weighted logistic regression. Baum (2008) suggested that this was a better method for modelling proportional variables than linear regression because predictions are required to stay within [0, 1] bounds. The weights added to these models were the number of home and away games played by the club in each observation (which in most cases was 22). Only the Points For and Points Against variables were modelled using linear regression, and the logistic models were built using the *logit* link function. All variables were taken to be significant at the 5% significance level. Over-dispersion within the logistic models was tested by the determination of a large difference between the residual deviance and residual degrees of freedom (Cox, 2013). Although the evaluation of a large difference is somewhat arbitrary, this analysis used Lindsey's (1999) suggestion that over-dispersion is present when the residual deviance is more than twice the size of the residual degrees of freedom. In order to account for over-dispersion, the *quasibinomial* family with the *logit* link function would be used to fit the generalised models, resulting in the estimation of a dispersion parameter instead of fixing it to a value of one.

The number of finals wins by the clubs that finished in the top eight was regressed using an ordinal probit model, as each outcome can only take four discrete values (0, 1, 2 or 3).
Goodness of fit for the linear models was measured using their adjusted R-squared values. Deviance goodness of fit tests was conducted to measure the fit of the logistic models. The average marginal effects (AME) were calculated for each of the logistic and probit models in order to measure the associations and effect-sizes of the explanatory variables on the various success outcomes.

4. RESULTS

A further four extreme values were removed from the dataset. These were the observations corresponding to Essendon's 2015 season, GWS's 2013 season, Gold Coast's 2012 season and Geelong's 2010 season, leaving a

total of 150 observations. All linear models satisfied the necessary assumption tests of linearity, homoscedasticity and residual normality at the $\alpha = 0.05$ level. However, over-dispersion was exhibited within the win percentage variable, and was subsequently accounted for through the use of a quasi-likelihood approach (Irwin, 2006). Table 1 summarizes the significantly related variables (shown in bold) to each of the performance measures by displaying their respective average marginal effects.

	Win %	Points For	Points Against	Top 4	Top 8	Bottom 4	Premiership
PCE	-7.10 (-42.0 – 27.9)	-134.5 (-622.8 – 353.8)	253.8 (-216.6 – 724.2)	8.2 (-88.7 – 105.1)	81.7 (-28.6 – 192.0)	54.4 (-15.1 - 124)	-13.8 (-75.9 – 48.3)
Team Cohesion	1.4 (-0.5 – 0.33)	**32.05 (6.54 – 57.56)***	4.57 (-19.84 - 28.98)	0.5 (-3.1 – 4.1)	1.3 (-3.8 – 6.4)	-3.3 (-8.5 – 1.9)	0.8 (-1.1 – 2.7)
Average Age	**8.0 (3.8 – 12.2)*****	32.96 (-26.71 – 92.63)	**-156.0 (-213.1 - 98.9)*****	**16.9 (7.9 – 25.9)*****	**21.4 (10.8 – 32.0)*****	**-10.7 (-21.1 – -0.3)***	3.4 (-2.4 – 9.2)
Total Experience	**0.40 (0.2 – 0.6)*****	2.44 (-0.24 – 5.12)	**-6.09 (-8.67 – -3.51)*****	**1.10 (0.6 – 1.5)*****	**1.4 (0.9 – 1.9)*****	-0.4 (-0.8 – 0.0)	0.2 (-0.1 – 0.5)
Ave. Shared Experience (ASE)	0.2 (-0.1 – 0.5)	**4.54 (1.26 – 7.82)***	-2.18 (-5.67 - 1.31)	0.0 (-0.6 – 0.6)	-0.2 (-1.0 – 0.6)	**-0.7 (-1.3 – -0.1)***	0.2 (-0.2 – 0.5)
ASE Midfielders	0.0 (-0.2 – 0.2)	0.98 (-1.52 – 3.48)	1.03 (-1.55 - 3.61)	-0.2 (-0.7 – 0.3)	0.0 (-0.5 – 0.5)	0.1 (-0.3 - 0.5)	-0.1 (-0.3 – 0.2)
ASE Forwards	**0.2 (0.1 – 0.3)****	**2.01 (0.23 – 3.79)***	**-3.39 (-5.23 – -1.55)*****	**0.4 (0.1 – 0.7)***	**0.4 (0.1 – 0.7)***	-0.3 (-0.6 – 0.0)	0.1 (-0.1 – 0.3)
ASE Defenders	**0.3 (0.2 – 0.4)*****	**2.73 (0.62 – 4.84)***	**-3.3 (-5.48 – -1.12)****	**0.6 (0.3 – 0.9)****	**0.5 (0.1 – 0.9)***	**-0.6 (-0.9 – -0.3)*****	**0.2 (0.1 – 0.3)***

Significant at ***0.1%, **1%, *5% levels

Table 1: The average marginal effects and 95% confidence intervals.

The marginal effects of each of the cohesion variables in relation to the number of finals wins were not significantly different from zero at the $\alpha = 0.05$ level. The PCE and average shared experience of midfielders did not have a significant relationship with any performance metric.

4. DISCUSSION

The average future win percentage variable is not significantly related to any metric analysed, indicating that these metrics are poor predictors of performance three and four years into the future. However, in the short-term, the results indicate that there is a significant, but mostly weak, relationship between success and various

measures of team cohesion, thus supporting Mullen and Copper's (1994) noteworthy finding. Out of the cohesion metrics tested, the average shared experience of defenders and the average shared experience of forwards has a significant relationship to the highest amount of success measures, indicating that these are the best cohesion centred variables to use when evaluating the potential for a group of AFL players to achieve success. These cohesion variables exhibit similar effect sizes to average age and total experience, providing further evidence to support the popular notion that more mature teams have a higher likelihood to succeed than young, inexperienced teams.

The cohesion variables were found to be better predictors of performance during the home and away season, rather than club performance during the finals series. This may be due to the fact that there are little to no changes to a club's playing list throughout a finals campaign. Hence, the cohesion of a clubs' players in the finals would be different to the calculated values in this analysis, in which data from the core group of players was used.

Only one variable (defensive shared experience) was found to have a significant effect on the odds of a club winning a premiership, which is viewed as the ultimate (and arguably only) form of success. However, this could be due to the fact that the premiership is decided after only one game, in which luck may have a greater influence on the result. The average shared experience of a clubs' midfield group has no significant relationship to success. This could be due to the fact that midfielders have a lot of influence around stoppage situations, which are effectively random events (O'Shaughnessy, 2016). Hence, a midfielders' ability to impact a game is affected less by their teammates than forwards and defenders, who have a heavier reliance on team structure.

It must be noted that recruitment of players at the end of one season will not necessarily reduce the average shared experience of a club's players by the start of the next season. Furthermore, the effect that the trading of players has on average shared experience values changes in each circumstance. Hence, specific list management recommendations cannot be made, but generally, if a club is planning for near-term success, the recruiting and trading of forwards and defenders should not negatively impact the shared experience of players in those positions.

When making list management decisions, the trade-off between skill and cohesion needs to be taken into account. Since the shared experience of the midfield group was found to have no correlation to team performance, the only effect these players will have on a team is through their playing ability. Therefore, the recruitment of elite midfielders may be a safer option than the recruitment of forwards and defenders in the pursuit of short-term success. The findings from this analysis are particularly useful for clubs who are in or approaching their premiership window, where any small advantage can create success. However, for clubs devoid of talent, the marginal gains from optimising the shared experience of players in specific positions are unlikely to bridge the large gap between the strong and weak teams in the short-term. It may be more beneficial for clubs rebuilding their list to only focus on the optimization of their defenders and forwards shared experience values after they have acquired the appropriate talent to ensure that the list will be competitive in the near future.

The limitations entrenched within the analysis are the small amount of observations and uncorrelated predictors within the dataset. These limitations have hindered the possibility of predictive analysis being performed. Hence, the results form this analysis should not be used to make predictions, but to ascertain the historical effects and observed correlations of the variables under investigation. Furthermore, the lack of predictor variables within the regression models increases the likelihood of endogeneity through omitted selection.

The results presented here are solely based on the sample ranges provided in the dataset. An indefinite increase in the average age or experience of a team will obviously not result in a continued increase in the odds of being successful, as players' bodies begin to deteriorate after a certain time. The range of the average age variable in this dataset is from 22.4 to 27.7 and hence, the results should only be interpreted for these bounds. The fact that intra-team cohesion was only approximated in this study, rather than being psychologically measured through questionnaires, is likely to have contributed to the strength of the cohesion-performance relationship being found to be relatively weak. However, this limitation was built into the current study in an attempt to improve the practicality of measuring team cohesion.

A major criticism on the theoretical relationship between shared experience and success has been the question of causation. Does having a high performing team lead to them exhibiting high shared experience, or does having a high shared experience lead to a high performing team? (Coventry, 2018). Although this paper does not definitively answer this question, a few observations should be made. There should be a distinction made between teams who perform better, or worse than the sum of their parts. In the latter case, the collective skill of a team's players is not necessarily related to their shared experience, and in the signing of a star player/s, this relationship is an inverse one. Hence, if the correlation between shared experience and success was purely because of player skill, then the relationship between these variables would need to be transitive, which it is not. In other words, teams may perform worse than the sum of their parts because of low shared experience measures, but not the other way around. In the case where teams build dynasties in the AFL, these clubs often perform better than the collective abilities of their players. This is not only due to skill, but also to the cohesiveness of the team. Therefore, in the case where teams exhibit high shared experience because of their success, it is not a question as to whether cohesion is an important factor, but rather if shared experience is an appropriate proxy for measuring team cohesion.

5. CONCLUSIONS

The findings from this analysis support the popular notion that the cohesiveness of a team is significantly related to success. The cohesiveness-performance relationship was measured using metrics not currently in mainstream use. The significance of the Average Shared Experience of forwards and the Average Shared Experience of defenders' measures indicate that new methods in the evaluation of the strength of an AFL clubs playing list can be proposed. The effect sizes of these measures on performance are comparable to the average age and total experience of a list, indicating that a slight competitive advantage can be found in their use. In the recruiting of defenders and forwards, an apparent trade-off needs to be considered between the players' ability, and the subsequent effect that these players have on team cohesion. There is no evidence to suggest that the recruitment, or trading, of a midfielder will significantly affect the cohesion of a team. Therefore, only the midfielders' ability needs to be considered when making list management decisions.

Acknowledgements
We wish to thank the recruiters at the Collingwood Football Club for their valuable input and provision of data for this analysis.

References

Baum, C. (2008). Stata tip 63: Modeling proportions. *The Stata Journal* , 299-303.
Coventry, J. (2018). *Footballistics.* Australia: ABC Books.
Cox, S. (2013). *Applied biostatistical analysis using R - Overdispersion*. Retrieved April 28, 2018 from Open Access Textbooks: https://www.otexts.org/node/674.
Irwin, M. (2006). Overdispersion Models. *Statistics 149.* Cambridge, Massachusetts, USA: Harvard University.
Lindsey, J. (1999). On the Use of Corrections for Overdispersion. *Journal of the Royal Statistical Society , 48* (4), 553-561.
Mullen, B., & Copper, C. (1994). The relation between group cohesiveness and performance: An integration. *Psychological Bulletin* , 210-227.
Salas, E., Grossman, R., Hughes, A., & Chris, C. (2015). Measuring Team Cohesion: Observations from the Science. *Human Factors , 57* (3), 365-374.
Schrage, M. (2014, March 5). *Team Chemistry Is the New Holy Grail of Performance Analytics.* Retrieved May 25, 2018 from The Harvard Business Review: https://hbr.org/2014/03/team-chemistry-is-the-new-holy-grail-of-performance-analytics.
O'Shaughnessy, D. (2016). Identification And Measurement Of Luck In Sport. *The Proceedings Of The 13th Australasian Conference On Mathematics And Computers In Sport* (pp. 21-26). Melbourne: ANZIAM MathSport.

Choosing an appropriately sized basketball for junior players: An example of a principled approach towards equipment scaling in sport.

A. D. Gorman[1], J. Headrick[2], I. Renshaw[3], C. J. McCormack[1], & K. Topp[3]

[1]School of Health and Sport Sciences, University of the Sunshine Coast, Sippy Downs, Queensland, Australia.

[2]School of Allied Health Sciences, Griffith University, Gold Coast, Queensland, Australia.

[3]Exercise and Nutrition Sciences, Queensland University of Technology, Brisbane, Queensland, Australia.

In simple terms, the notion of equipment scaling in junior sport involves matching the size of the playing equipment to the size of the child, with the general aim of not only improving skill acquisition, but also improving the child's overall playing experience (see Buszard, Reid, Masters, & Farrow, 2016). The purpose of the current research was to use anthropometric measures of the hand to identify an appropriately sized basketball for junior players. The hand dimensions of adult (*M* age = 28.66 years) and junior (*M* age = 11.55 years) male basketball players were used to create hand-to-ball-size ratios across five different sizes of basketball including size 3 to size 7 (the latter being the regulation size for adults). For the junior players, who were accustomed to using a size 6 ball in their regular competitions, the hand-to-ball-size ratio for the size 3 and 4 basketballs was the closest match to that of the adults' hand-to-ball-size ratio for the size 7 ball. The results from 3-on-3 gameplay also revealed that the size 3 and 4 basketballs elicited a greater number of 3-point shot attempts and more steals/intercepts compared to the larger balls. However, when asked to identify the size of basketball that they would prefer to use in future games, the junior players tended to select the size 5 and 6 balls. Collectively, the results provide evidence of the potential utility of anthropometric measures of the hand for equipment scaling, but highlight the need for equipment changes to be implemented across all age groups to allow sufficient time for children to become familiarised with that equipment.

CONSTRUCTING BEACH VOLLEYBALL STRATEGIES IN EXTENSIVE FORM

Yee, Mattina Koleen T. [a,b] , Talabis, Dylan Antonio S.J. [a]
[a] *University of the Philippines Los Baños*
[b] *Corresponding author: mtyee@up.edu.ph*

Abstract

Beach volleyball is a game wherein two sets of pairs are opposing. The objective of the game is to return the ball to the opponent within three touches under a set of rules. With that, it is necessary to create a strategy especially for games that involve multiple players. Common defense patterns were studied to produce a game tree that can determine an optimal defense pattern. Nash equilibrium existed at 22.11% with the defense patter 2-A-D. When the ball is served, server has leverage over which player to bring the ball to. Given that there are only three touches before returning the ball, the receivers must determine a strategy on how to return the ball to the opponent with imperfect information. It was determined that at 51.14% the Nash equilibrium of the offensive pattern existed by giving the serve to player B and returning the ball in 3 touches.

Keywords: beach volleyball, extensive form, game theory, Nash equilibrium, and strategy

1. INTRODUCTION

Team sports like basketball, football, volleyball, etc., is popular to many individuals. At least once in our lives, we have engaged in at least one team sports [1]. In team sports, strategy is often used in order to perform a successful play. Coaches would usually give a game plan to team members in order to execute a play. Strategies differ in many aspects. It may depend on the players, how crucial the game is or even how the opponents perform.

Volleyball is a famous Olympic sport. Almost all countries in the world have a national team in volleyball. There are 2 kinds of volleyball, indoor volleyball and beach volleyball. Indoor volleyball is played by 12 players with 6 players on each team [2]. On the other hand, beach volleyball is 2 against 2 [3]. Strategy plays a big role in volleyball. The players have to learn about offensive and defensive strategies. In beach volleyball, playing smart improves the chances of winning big [4]. There is a maximum of three touches before returning the ball to the opponent. With three touches, we can maximize the chance of scoring a point [2, 3].

Skills play a big role in beach volleyball but given that there are only two players inside the court, a proper set of strategies would definitely give them the advantage. Pairings in beach volleyball is very crucial. There should be chemistry between teammates. One should know better in spiking or receiving, as this tactic will give the team greater chances of scoring.

Game theory may be defined as the study of mathematical models of conflict and cooperation between decision-makers [5]. In this research, we consider the extensive form of the game. The extensive form contains all the information about a game, by defining who moves when, what each player knows when he moves, what moves are available to him, and where each move leads to [6]. Game trees are used to visualize the sequence of information [9]. A payoff is a number that reflects the desirability of an outcome to a player. When the outcome is random, payoffs are usually weighted with their probabilities [7]. Payoffs for choosing a particular strategy sometimes depend on the previous strategy used, not on who the player is. Game theory is used in an attempt to explain payoffs linked with predictability of individual player's actions, for co-players in his own team, as well as for the opponent team players [8]. A normal form, also known as strategic form, is all the strategies available to the player in which players simultaneously choose their actions [7]. Nash Equilibrium is the best response in which no player would have an incentive to deviate. Sub game Perfect equilibrium is one that induces payoff-maximizing choices in every branch or sub game of its extensive form where it is also Nash Equilibrium.

This study will be a significant endeavor in promoting strategic plays in beach volleyball. This study will also be beneficial to the players and coaches in different levels of competencies. This research will also provide recommendations on how to execute game plays in different situations.

Moreover, this study will be helpful to aspiring beach volleyball players, sports analysts and game officials. It will also serve as a future reference for researchers on the subject of game theory and, sports.

Game theory has been applied in various sports. Through extensive form and strategic form, we aim to determine the Nash Equilibrium in beach volleyball empirically. Specifically, this research aims to:

1. Translate the volleyball exchange into a game and define the components of the games;
2. Identify the plays that usually occurs in beach volleyball and determine the payoffs of the chosen strategies;
3. Create a game tree and game matrix that can show the payoffs of every strategy in beach volleyball;
4. Determine the strategies of players that give the Nash Equilibrium, sub game perfect equilibrium and their corresponding pay-offs.

2. METHODS

In this study, we have considered two models, the defensive strategy and the offensive strategy.

We first consider the defensive strategy. The defensive strategy aims to determine at which positions should the defenders be given the location of the spiker. In this strategy, we have 3 players and we consider 3 plays. Player 1 is the spiker, the offender, and players 2 and 3 are the defenders.

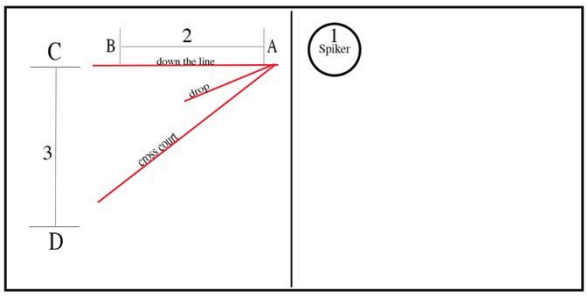


Figure 1. Play 1: Spiker is in opposite position

In play 1, player 1 spikes at the opposite position. Player 2 has an option to block or to receive. Player 3 has the option to go to either ends of the backcourt.

In play 2, player 1 is the spiker and players 2 and 3 are the defenders. Player 1 spikes in the open position. Player 2 has an option to block or to receive. Player 3 has the option to go to either ends of the backcourt.

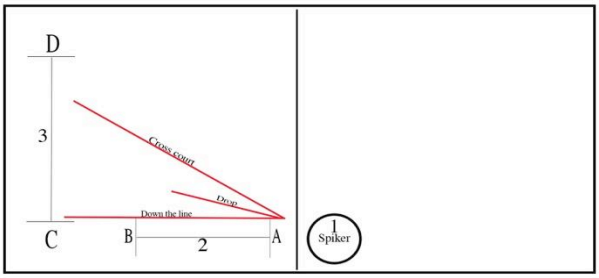


Figure 2. Play 2: Spiker is in Open position.

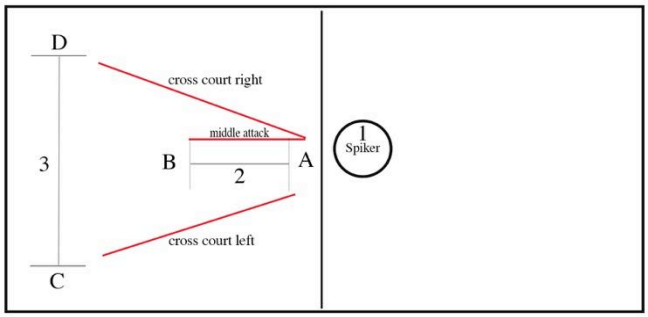


Figure 3. Play 3: Spiker is in middle position.

In play 3, player 1 spikes in the middle position. Player 2 has an option to block or to receive. Player 3 has the option to go to either ends of the backcourt.

In order to determine the Nash equilibrium, we use statistical analysis through excel. We also have to formulate a game tree in order to translate the exchanges into a game and to evidently see the payoffs of the strategies.

In the second model, the Offensive strategy, we also have 3 players. Player 1 is the server. Players 2 and 3 are the receivers. Between players 2 and 3, we separate the players into 2 categories, A and B. A (Player 2) is the dominant offensive player and B (Player 3) is less dominant than player 1. The aim of this strategy is to determine to which receiver should the server give the ball. If the server is given to player 3, at 3 touches, we expect player 3 to return the ball to the opposing team.

A game tree is created to show the exchanges of the strategy. Through statistical analysis, payoff for each possible strategy is calculated.

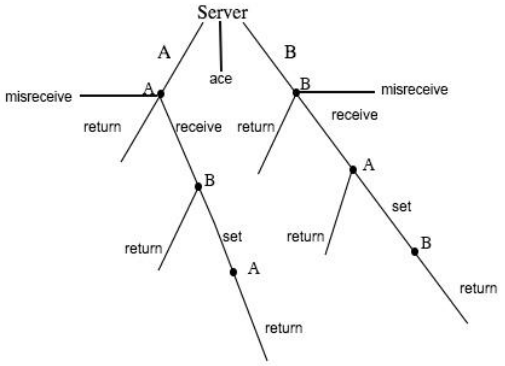


Figure 4. Game tree for the offensive strategy

3. RESULTS

For the first model, the defensive strategy, through statistical analysis, we have determined the payoffs and game tree of the exchanges.

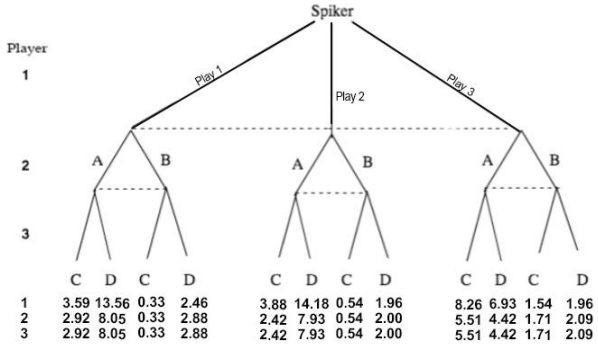


Figure 5. Payoffs of the defensive strategy

With the use of backward induction, a Nash equilibrium for the strategy has been determined.

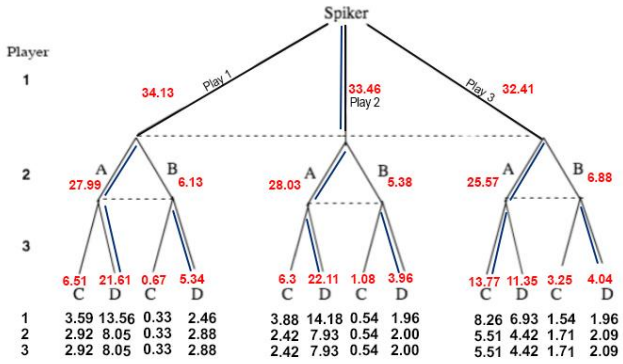


Figure 6. Results of the defensive strategy using backward induction

The results from the offensive strategy are as follows:

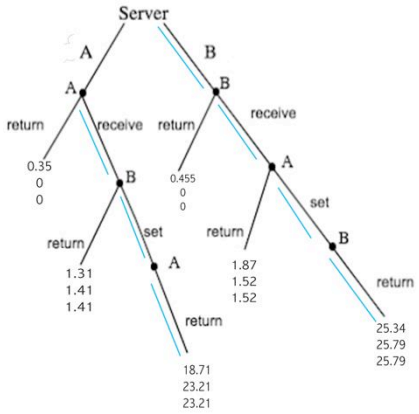


Figure 7. Results of the offensive strategy with payoffs using backward induction

4. DISCUSSION

In the defensive strategy, we have determined through backward induction that the strategy 2-A-D gives us the Nash equilibrium. With that, we can say that this is the best response for each of the players.

But if player 1 chooses play 1, the Nash equilibrium for the game will only happen if $b \geq 0.62$.
Thus, players 2 and 3 will still follow the A-D defense pattern.

Player 1 chooses Play 1

P2/P3	C	D
A	3.59, 2.92, 2.92	13.56+b, 8.05, 8.05
B	0.33, 0.33, 0.33	2.46, 2.88, 2.88

Figure 8. Nash equilibrium if player 1 chooses play 1

Player 1 chooses Play 3

P2/P3	C	D
A	8.26+a, 5.51, 5.51	6.93, 4.42, 4.42
B	1.54, 1.71, 1.71	1.96, 2.09, 2.09

Figure 9. Nash equilibrium if player 1 chooses play 2

If player 1 chooses play 3, the Nash equilibrium for the game will only happen if $a \geq 5.92$. Players 2 and 3 will move the defense pattern to A-C.

With that, Nash equilibrium can be determined depending of what play player 1 executes. This strategy is helpful to create a defensive pattern for the defenders, as this will give the best response at any position of the spiker. Professional players follow a Nash equilibrium at 22.11%.

For the offensive strategy, it was found that the server should give the ball to player B, the less dominant player. This will increase their chances of scoring a point. In defence of the receivers, they should return the ball in 3 touches to have a positive response to the serve. Professional players follow a Nash equilibrium at 51.14%.

5. CONCLUSIONS

Beach volleyball is a game where strategy is commonly used to score a point. With proper strategies on hand, the game exchanges can be translated into a game tree. This will help us in determining the best response each player should do in beach volleyball.

In this paper, we have considered two strategies, offensive and defensive strategies. The aim of the defensive strategy is to determine a defense position at any position of the spiker. We have determined that the Nash equilibrium of the defensive strategy is when player 1 spikes at the open position, player 2 blocks the ball and player 3 receives the ball at the cross court. With that, professional players follow a Nash equilibrium at 22.11%.

As for the offensive strategy, the essence of the offensive strategy is to give the ball to the less dominant player. We know that giving the ball to the less dominant player may resort to a service ace or the ball may not be returned successfully. With that, a strategy has been devised to determine how players should respond. It was determined that the Nash equilibrium at 51.14% is when player 1 gives the ball to player 3 and players 2 and 3 should do three touches. In this way, the game will be at Nash equilibrium.

Acknowledgements

I would like to thank the people who helped in making this research possible. Many thanks to the UPLB Volleyball Club, Issa Ruiz, Yzmarte Doy, Monique Yglesias, Chiki Gonazales and Chinggay Sison. Thanks also to my parents for the constant support and to my adviser, Dylan, for being dedicated and supportive.

References

[1] Volleyverse.com, https://volleyverse.com

[2] Federation Internationale de Volleyball. Official Volleyball Rules 2017- 2020. 2016.

[3] Federation Internationale de Volleyball. Official Beach Volleyball Rules 2017-2020. 2016.

[4] Laura Albert. Beach Volleyball and Game Theory Punk Rock Operations Research. August 12, 2012. https://punkrockor.com/2012/08/01/beach-volleyball-and-game-theory/

[5] Roger B. Myerson. Game Theory. Harvard University Press, 1991

[6] Junki Takeuchi, Rido Ramadan and Hiroyuki Iida. Game-Refinement Theory and Its Application to Volleyball. IPSJ SIG Technical Report, Vol. 2014-GI-31 No. 3. March 17, 2014.

[7] Theodore L. Turocy, Bernhard von Stengel. Game Theory. CDAM Re- search Report LSE-CDAM-2001-09. October 8, 2001.9

[8] Josko Sindik and Nives Vidak. Application of Game Theory in Describ- ing Efficacy of Decision Making in Sportsman's Tactical Performan in Team Sports.Interdisciplinary Description of Complex Systems 6(1), 53-66, 2008.

[9] Steven M. LaValle. Planning Algorithms. Cambridge University Press. 2006.

[10] Yoav Shoham and Kevin Leyton-Brown. MULTIAGENT SYSTEMS: Algorithmic, Game-Theoretic, and

Logical Foundations. Cambridge Uni- versity Press, 2009

[11] TomaszSeweryniak,DariuszMroczekand,L ukaszL ukasik.Analysis and evaluation of defensive team strategies in women's beach volleyball – an efficiency-based approach. Human Movement. 2013, vol. 14 (1), 48–55

[12] Kai Lin. Applying Game Theory to Volleyball Strategy. International Journal of Performance Analysis in Sport 2014, 14, 761-774.December 2014

THE IMPACT OF INJURY ON THE FUTURE PERFORMANCE RATINGS OF DOMESTIC T20 CRICKETERS

Samuel J. Greer [b], Ankit K. Patel [a], Holly E. Trowland [a,b,c] and Paul J. Bracewell [a]
[a] *DOT Loves Data, Wellington*
[b] *Victoria University, Wellington*
[c] *Corresponding author: ankit@dotlovesdata.com*

Abstract

Player absence is common in T20 domestic leagues, and may be caused by a number of factors including injury, international call-ups, enforced rest, or being dropped from the team. This research investigates the effect of absence on player performance, enabling an estimation of expected performance in the return appearance. Non-parametric techniques are used to predict the change in player rating between the last game played and the first game returning from absence. Statistically significant differences are found in return performance, depending on player role and cause of absence. More specifically, injured batters tend to decline in performance, while players absent due to other reasons tend to improve their performance. Similarly, bowlers who are absent for other reasons often return with improved ratings. However highly ranked bowlers who were absent due to other reasons return with a significantly lower rating than high ranked bowlers who were absent due to injury.

Keywords: Randomisation testing, Bootstrapping, ANCOVA

1. INTRODUCTION

The emergence of Twenty20 (T20) as a short form of cricket has put a spotlight on cricketer's skill sets and roles in the team. This is evident in the plethora of academic literature on T20 (Patel, Bracewell and Wells (2017), Verrall, Kalairajah, Slavotinek & Spriggins (2006), Alamar (2013)), popular statistics (Perera & Swartz, 2012) and media attention to events like the IPL player auctions (Gupta, 2011). As well as the attention, the rewards for players are financially lucrative (Subhani, Hasan, Osman, et al., 2012) which has encouraged cricketers like Mitchell McClenaghan (Geenty, 2017) to retire from international cricket and compete solely in domestic T20 leagues around the world.

Each individual player has a role in maximising their team's chances of winning. However, during the course of a season players may not participate in scheduled matches. Reasons for omission could include: injury, rest, tactics, international selection, poor performance or disciplinary matters. For coaches and selectors in T20, a key consideration is maximising the probability of winning by making use of all information at their disposal. Metrics regarding player performance can be used to optimise team selection (Patel et al., 2017), but if a player is missing from a line-up, how does this absence impact their expected performance in the returning game? Moreover, does the reason for omission impact expected team performance?

There has been prior research into sporting injuries as a reason for omission. Verrall, Kalairajah, Slavotinek & Spriggins (2006), aimed to determine whether there is a decrease in player performance coming back off of a hamstring injury in Australian Rules Football (AFL). Using a team from the AFL as a case study, data was collected in the form of a player rating (integer from 1 to 10) for each game in a season. The scope of the data was injured players who returned in the same season, and therefore their pre and post-injury ratings could be analysed. Initially a Friedman Test was used, followed by an Exact Wilcox Signed Rank Test, to identify whether the first test was significant. The authors were able to conclude that player ratings were lower when they returned than before their injury. Additionally, Stretch (2001), investigated the nature of injuries sustained by elite South African cricket players during a season. To complement this, possible risk factors of injuries were also identified. The conclusions of the analyses were that bowling accounts for the largest proportion of injuries (approximately 41%), whilst batting had the smallest proportion, with roughly 21% resulting in injury. First time injuries account for around 64 percent of total injuries, and 63 percent are acute.

Player performance can be quantified. Patel, Bracewell and Wells (2017) used an array of statistical methodologies to identify important variables for predicting player ratings for both batters and bowlers. This player rating system, was followed by an aggregation of the ratings to create an overall team rating for the prediction of match outcomes. The ratings were used for player selection for the 2017 Big Bash fantasy

competition. Overall the researchers finished in the top 1% of participants, and it was concluded that the ratings are a reliable measurement of performance.

This paper investigates how player omission from T20 matches affects future performance (individual and team). This is achieved by examining how the reason for omission affects the expected change of a player's performance once they have missed at least one game within a season of top level domestic T20 Cricket. The expected performance of a team is then quantified and we get an indication of the efficacy of rehabilitation for players returning from injury. Player and team ratings were sourced from Patel *et. al.* (2017), and in light of the Stretch (2001) study, which found over double the number of injuries occur from bowling when compared to batting, these different roles are analysed separately.

2. METHODS

DATA

Player performance was measured using the rating method outlined in Patel *et. al.* (2017). The ratings measure how each respective position dynamically influences the outcome of the match. For example, if a batter scores 20 runs off 30 balls they would have a lower rating than a batter who scores 50 runs off 30 balls. The first batter would have a lesser impact on the outcome of the game than the second batter. The same logic applies for the bowling scores. For example, a bowler conceding more runs off their own deliveries than another, who bowled the same number of balls, the latter would post a higher influence score. These scenarios are a simple example, and additional factors were taken in to account when forming the player influence score, which determines the player rating.

Match summaries that contain match-by-match roster information and articles discussing the match of interest were used to collect information on *matches missed* and the *reason* for an absence. These summaries were extracted from ESPN's Cricinfo (*ESPN Cricinfo*, 2018), a freely available website that contains relevant information on players and matches. Four of the worlds highest level domestic T20 leagues were used to derive the data, over a range of years. These are Australia's "Big Bash League" (BBL) from the 2014-15, 2015-16 and 2016-17 seasons, India's "Indian Premier League" (IPL) from the 2015 and 2016 seasons, the England and Wales's "NatWest T20 League" from 2016, and finally the Caribbean's "Caribbean Premier League" (CPL) from the 2015 season. The data contains 175 batting observations and 215 bowling observations.

MODEL

The dependent variable of the current analysis is *change in rating*, which is the difference between the influence score of the player in the last match played before absence, and their return match. The last match played (referred to as the prior match) rating is also retained as a covariate (*rating prior*). The independent variable of interest is the *reason* for an absence (referred to as *reason* in this project). Given some reasons for absence are rare, *reason* was taken to be a 2-level factor; 'injured' or 'other'.

ANCOVA models evaluate whether the means of a dependent variable are equal across levels of a categorical independent variable, while statistically controlling for a covariate. For this research, the mean *change in rating* across the levels of *reason* was tested in the presence of the covariate, *rating prior*. The standard linear regression assumptions hold, and a further assumption that the slope of the covariate is equal across all treatment groups. This research presents a batter and bowler model.

The models were tested to validate linear model assumptions (i.e. independence of errors, uncorrelated, homoscedasticity and normality). Normality in the batting model was found to be questionable due to the deviations in the upper section of the Q-Q plot but it could be argued that that it was insignificant as it is small, while the bowling model it is violated due to large deviations. Homoscedasticity was satisfied as the residual vs fitted plots were evenly spread around zero, and testing for the independence of errors was undertaken by conducting a Durbin-Watson test. In both models, the residuals produced a test statistic of approximately 2, indicating that the residuals are uncorrelated and therefore independent (Field, 2013).

Parametric tests have a dependence on rigid distribution assumptions, which are not met for the batting model and are violated in the bowling model. Given the violation of the ANCOVA assumptions, it cannot be used directly, but ANCOVA can still be used to frame the models. Non-parametric techniques such as Randomisation Tests may produce suitable power ensuring there is a small chance of a type II error. Randomisation tests are typically used to test the hypothesis that there will be a tendency for a certain type of pattern in the data against a null hypothesis of randomness (Manly, 2006).

Randomisation tests were used in this paper to perform *F-tests* to first assess whether there is a significant interaction effect between *reason* and *rating prior*. This occurs when the regression lines for *reason* and *rating prior* are not parallel. Secondly, the main effect was tested to see whether the expected values of the mean *change in rating* differ between levels of *reason*, regressed against *rating prior*. The observations were permuted to get 1,000 pseudo *F Values*. The interactions are evaluated on statistical significance for the batting and bowling models by calculating the proportion of pseudo *F values* from the randomisations that are greater than the observed F statistic, as these pseudo *F values* are a representation of the null distribution of the statistic. This is equivalent to regular hypothesis testing of the null hypothesis, to determine if there is no significant interaction, against the alternative hypothesis of the presence of a significant interaction.

Following the results of the randomisation testing, the main effects were tested. Non-parametric techniques were also implemented to acquire an estimation of a parameter β. In the batters model where the main effect is tested, β is the difference in *change in rating* between injury and other reason for absence. In the bowlers model, β is the difference in slope of *change in rating* against *rating prior* for the two levels of *reason*. Bootstrapping is a technique in which distributional assumptions are not made when estimating parameters and so the violated assumptions are not an issue (Mooney & Duval, 1993). The approach behind bootstrapping is to implement Monte Carlo style sampling by selecting random samples from the original data with replacement. In this method, for each set of re-sampled data, the corresponding data estimates the parameter of interest, $\{\hat{\beta}_1^*, \hat{\beta}_2^*, ..., \hat{\beta}_A^*\}$ represents an estimate of the sampling distribution of $\hat{\beta}$. One such method of bootstrapping is to re-sample the residuals from the models of interest to get bootstrap estimates and their respective standard errors. The bootstrap estimates $\bar{\beta}^*$ are defined as the mean of the set of A bootstrap estimates. A bias check on $\hat{\beta}$ is then conducted:

$$Bias[\hat{\beta}] = \bar{\beta}^* - \hat{\beta} \tag{1}$$

where $\hat{\beta}$ is the observed estimate for β and $\bar{\beta}^*$ = The mean of the bootstrap estimates.
Bias correction on the estimates is then conducted:

$$\begin{aligned} \hat{\beta}_{BC} &= \hat{\beta} - Bias[\hat{\beta}] \\ &= \hat{\beta} - (\bar{\beta}^* - \hat{\beta}) \\ &= 2\hat{\beta} - \bar{\beta}^* \end{aligned} \tag{2}$$

where $\hat{\beta}_{BC}$ is the bias corrected estimate for β

An appropriate method for constructing confidence intervals for the bias corrected estimate is to utilise the normality of the bootstrap estimates, $\{\hat{\beta}_1^*, \hat{\beta}_2^*, ..., \hat{\beta}_A^*\}$. If the bootstrap estimates presents normality then a standard bootstrap 90% confidence interval will be constructed;

$$\hat{\beta}_{BC} \pm 1.645 \times s.e(\hat{\beta}^*) \tag{3}$$

where $s.e(\hat{\beta}^*)$ is the standard error for the bootstrap estimates $\{\hat{\beta}_1^*, \hat{\beta}_2^*, ..., \hat{\beta}_A^*\}$

3. RESULTS

BATTING RESULTS

Figure 1 reveals a clear negative association between *bat rating prior* vs. *change in rating*, which would be expected, as *change in rating* is partially derived from *bat rating prior*. For example, if an individual has a very high match rating then there would be only so much the rating could realistically increase. As the ratings are unbounded, it is possible to increase after posting a high score, however it is unlikely to be significant. On the other hand, if an individual posts a poor performance it is unlikely their score would decrease much further. This is reflected in Figure 1 showing a negative relationship. The point spread is visually consistent across the plot indicating that for any value of *bat rating prior*, the range for the *change in rating* is equal. Examining the two regression lines based off the ANCOVA model, where each represents a level of *reason*, they appear parallel, suggesting there is no interaction.

Ignoring prior rating, injured players tend to decline in performance (mean *change in rating*=-14.7) while players absent for other reasons tend to improve their performance (21.4). This indicates that injuries have a negative impact on returning batter performance, while absence not related to injury has a positive impact on returning batter performance.

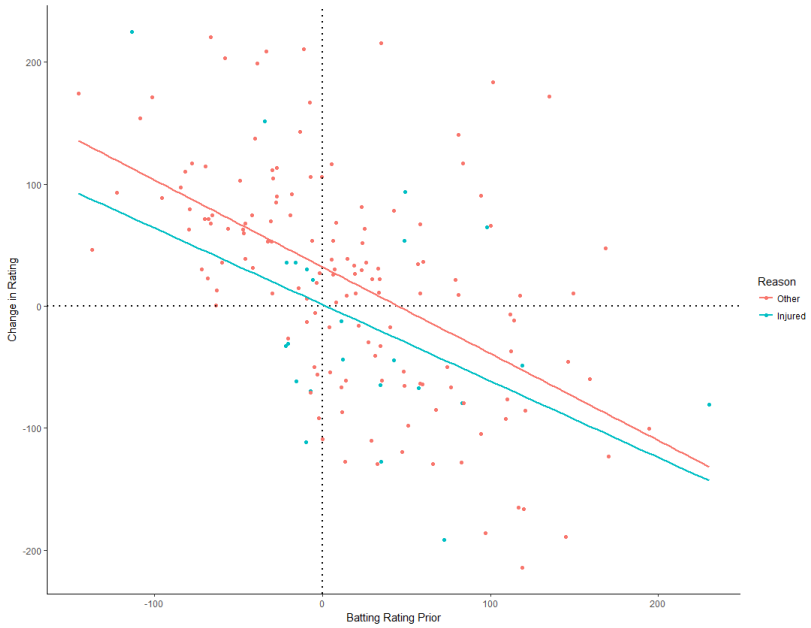


Figure 1: Scatterplot of *bat rating prior* vs *change in rating* with the two levels of *reason* regression lines overlaid

Using the batting ANCOVA model, a Randomisation test is applied to the *F values* to test for interaction effect between *reason* and *bat rating prior*. The following hypothesis tested was conducted:

$$H_0 : \text{No interaction between } \textit{reason} \text{ and } \textit{bat rating prior}$$
$$H_1 : \text{Interaction between } \textit{reason} \text{ and } \textit{bat rating prior}$$

The generated *F values* are greater than observed ($F = 755$), which corresponds to a *p value* of 0.755. This is greater than the 0.1 confidence level and so the null hypothesis is not rejected, showing that there is no statistical significance in the presence of other predictors. Given the lack of interaction, the next step is to test for main effects Using the same logic, the significance between each level of *reason* can be tested when comparing the baseline *Other* using the *t value* as the test statistic. The hypothesis is tested:

$$H_0 : \textit{Reason}\text{: injury is not different from } \textit{reason}\text{: Other}$$
$$H_1 : \textit{Reason}\text{: injury is different from } \textit{reason}\text{: Other}$$

The resulting *p-value* from the randomisation test is 0.082, indicating that *change in rating* of injured players is significantly lower than the players who were absent for other reasons at the 90% confidence level. To estimate the extent of the difference, bootstrapping is applied to generate 1,000 values of the corresponding coefficient. The bias is minimal (-0.18), and is a small fraction of the observed *t value* (-31.03). The bias corrected point estimate is $\hat{\beta}_{BC} = -30.67$ (90% CI; -60.87, -0.48) which shows that on average, the rating for batters who were injured dropped by 30.67 points more than for batters who missed matches for other reasons.

It was found that players who are injured typically have a smaller change in performance when they return to T20 than players who are absent for other reasons. Generally, players who were not injured increased their rating on return. Prior performance has the same effect on change in rating, regardless of reason for absence, and high performing cricketers tend to lower their return performance.

BOWLING RESULTS

Figure 2 is the scatterplot of *bowl rating prior* vs *change in rating* with the two levels of *reason* regressed over the data points. Ignoring prior rating, injured players tend to slightly decline in performance (mean *change in rating*=-3.9) while players absent for other reasons tend to improve their performance (14.6). It can be seen that the regression lines of other and injured reasons for omission are non-parallel, indicating a level of interaction between the two variables.

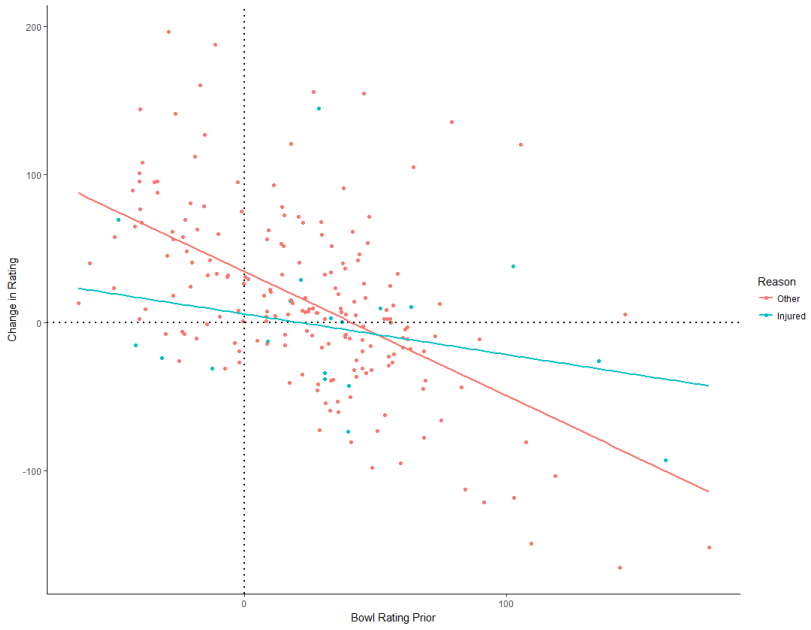


Figure 2: Scatter plot of *Bowl Rating Prior* vs *Change in Rating* with the two levels of *reason* regression lines overlaid

Using the batting ANCOVA model, a Randomisation test is applied to the *F values* to test for interaction effect between. The following hypothesis tested was conducted:

H_0 : No interaction between *reason* and *bowl rating prior*

H_1 : Interaction between *reason* and *bowl rating prior*

The resulting *p value*, 0.018, reveals that the null hypothesis should be rejected and it can be concluded that the interaction in the bowling model is significant. The slope of the regression line for *Injury* is significantly different from the slope representing the regression line for Other.

Given there is a significant interaction effect, the main effect is not tested. Next, the extent of the interaction effect between the levels of reason and prior rating is tested. For the estimate of the coefficient of the interaction between *reason*: Injured and *Bowl Rating Prior*, a bootstrap method is constructed to create estimates of the bias and create the appropriate intervals. The observed value for the coefficient is 0.770 while the bootstrap estimates is 0.766. This shows that the estimate for the coefficient is very close to the observed value, and therefore the bias is assumed non-existent. The bias corrected estimate is 0.57 (90% CI: 0.18, 0.95). This means the effect of the prior rating on the change in rating is higher by 0.56 when a bowler is injured when compared to being out for other reasons.

4. DISCUSSION

Results revealed that ignoring prior rating, injured batters tend to decline in returning performance while players absent for other reasons tend to improve their performance on return. This result may be due to the other reasons of absence such as international call-up and rest. Consequentially, if a player was injured this has a negative impact on their return. This may be the case for injured batters, especially batters who had a high rating before their absence, but players who are absent without injury often return with a higher performance than when they left. Prior performance has the same effect on change in rating for batters, regardless of reason for absence.

A significant interaction effect for bowler was identified. The relationship between change in rating and prior rating is different for bowlers who were absent for injury and bowlers who were absent for other reasons. Like batters, bowlers who were absent for other reasons often returned with increased ratings. However, highly ranked bowlers (> 58) who were absent for other reasons would return with a lower rating than high ranked bowlers who were absent for injury, despite being hindered from their injury. This could be for a number of reasons such as players who are injured generally undertake some type of rehabilitation to

get back to their peak performance level. However, a player who was absent due to rest or form does not go through this process. The reasons why injured bowlers have a smaller drop-off could be the subject for future analysis, especially when compared with batting test results. One reason could be the different pressures put on each positions' mechanics, which is discussed in (Stretch, 2001) where it is shown that bowling makes up the highest proportion of injures in cricket. Moreover, bowler injuries are significantly detrimental or career ending compared to batter injuries.

5. CONCLUSIONS

Overall, interesting insights were found on the impact of the absence on player performance during a season of domestic T20 cricket. From the non-parametric ANCOVA tests, it was found that the reason for absence does have an impact on the change in rating for both batters and bowlers, although to differing extents. It was found for batters that there is no interaction between reason for absence and prior rating, however, the groups do differ with injured players having a lower change in rating. For bowlers it was found that there is an interaction between reason for absence and prior rating. Highly ranked bowlers who were absent for other reasons returned with a lower rating than high ranked bowlers who were absent for injury.

In a real world context these findings present interesting information into how players who miss games perform on return. It also gives coaches and selectors insight into the consequences of absence. For example, a bowler who is injured will not decline as drastically as one who is out for rest. Another conclusion that is presented is that if a "star" bowler is injured, it is inferred that the change in performance will be less drastic than if it was a "star" batter who was injured. This information could be used by a selector to form their team, particularly in terms of depth, as they could adjust their roster to account for these differences. If there is a limited amount of money allocated and these selectors were looking to fill out roster spots with depth in mind, then according to the results they should focus less on greater reserve bowlers. This is because if a key bowler is injured, the decline of the change in rating is not as great as if it were a batter, which is a factor that selectors may need to consider. From the results, further research could be conducted on why there are these observed differences, despite the players participating the same sport. Batters seem to have the most potential for injury, yet not only do bowlers have the largest proportion of injures, their performance is also affected in a different way. The interactions of these two roles make up the essence of cricket yet they impact the game in their own ways, and so the extent of these impacts is an area for potential further analysis.

References

Alamar, B. (2013). Sports analytics: A guide for coaches, managers, and other decision makers. Columbia University Press.

ESPN Cricinfo. (2018). `http://www.espncricinfo.com/`. (Online; accessed 20-January-2018)

Field, A. (2013). Discovering statistics using ibm spss statistics. In (p. 874). Sage.

Geenty, M. (2017). "Hardest decision of my career", as mitchell mcclenaghan hands in nzc contract for t20. Stuff.co.nz.

Gupta, A. (2011). The IPL and the Indian domination of global cricket. In (Vol. 14, pp. 1316–1325). Taylor & Francis.

Manly, B. F. (2006). Randomization, bootstrap and monte carlo methods in biology. In (Vol. 70, pp. 1–3). CRC press.

Mooney, C. Z., & Duval, R. D. (1993). Bootstrapping: A nonparametric approach to statistical inference. In (pp. 9–10). Sage.

Patel, A. K., Bracewell, P. J., & Wells, J. D. (2017). Real time measurement of individual influence in t20 cricket. In *Proceedings of mathsport international 2017 conference* (p. 61).

Perera, H. P., & Swartz, T. B. (2012). Resource estimation in t20 cricket. In (Vol. 24, pp. 337–347). Oxford University Press.

Stretch, R. (2001). Incidence and nature of epidemiological injuries to elite south african cricket players. In (Vol. 91, pp. 336–339). Health and Medical Publishing Group.

Subhani, M. I., Hasan, S. A., Osman, M., et al. (2012). Will t20 clean sweep other formats of cricket in future? In (pp. 98–102).

Verrall, G. M., Kalairajah, Y., Slavotinek, J., & Spriggins, A. (2006). Assessment of player performance following return to sport after hamstring muscle strain injury. In (Vol. 9, pp. 87–90). Elsevier.

MANAGING RUN RATE IN T20 CRICKET TO MAXIMISE THE PROBABILITY OF VICTORY WHEN SETTING A TOTAL

Zeana Mansell [a], Ankit K. Patel [b,c,d], Jack McIvor [b] and Paul J. Bracewell [b,c]
[a] *University of Canterbury, Christchurch*
[b] *DOT Loves Data, Wellington*
[c] *Victoria University, Wellington*
[d] *Corresponding author: ankit@dotlovesdata.com*

Abstract

A framework for optimising a team's run rate in the first innings of a T20 cricket match to increase the probability of winning is introduced. Using ball-by-ball data, several regression models are used to assess how the rate at which runs are scored in the first innings changes with respect to balls and wickets remaining. Comparing the rates of change in the first innings between winning and losing teams enables the optimal scoring rate to be identified. When a team's scoring rate decelerates before 45% of batting resources have been consumed, the likelihood of losing increases.

Keywords: In-game strategy, polynomial regression, optimizing batting aggression

1. INTRODUCTION

T20 cricket is a variation of limited overs cricket designed to boost the games' popularity by delivering a more explosive form of the sport. Each team is given a single innings of a maximum of 20 overs. The primary objective is to score more runs than the opposition. To do so, a team must balance two key resources: balls and wickets. The team that utilizes these resources most effectively is most likely to win.

An important batting strategy to maximize the total runs scored in the first innings of a T20 game is the level of aggression during different stages of an innings. Batting aggression is observed as attempts to hit the ball to the boundary for the maximum rewards of either four or six runs. The first innings is explored as limited information is available to the team batting first regarding a likely winning total. As such, contextual setting of a total assist in improving strategies when batting first. "The optimization exercise in either team's task involves choosing some compromise between scoring fast and hence taking higher risks of losing wickets and playing carefully and hence risking making insufficient runs" (Duckworth & Lewis, 1998, pg. 220). As such, quantifying batting aggression is fundamental in understanding how to optimize the total runs scored with respect to victory. Although batting aggression has not received considered academic attention, the subject has been addressed by Davis, Perera and Swartz (2015), Scarf, Shi and Akhtar (2010) and Clarke (1988). Given aggression is linked to scoring fast, or acceleration in scoring rates, the hypothesis that batting aggression is characterized by the rate of change in run rate is explored. The benefit of this rate of change is quantified by comparing this behaviour between winning and losing team on a ball by ball basis.

New opportunities for sports analytics are due to the increased availability of detailed, machine readable data and greater computational power. These opportunities include monitoring player performance, assessing game tactics in real-time, examining opposition and providing feedback for training. As such, there is an escalating demand for data-driven decisions: such as game strategies, training regimes and player selection. Cricket has vast amounts of data and with the right application, fundamental game strategies and predictions can be made to maximize a team's chances of winning. "During the past decade many academic papers have been published on cricket and performance measures and predictive methods" (Lemmer, 2011, pg. 1).

Arguably, the most significant application of analytics within cricket is the Duckworth Lewis (1998) resource allocation model. Duckworth and Lewis (1998) created a mathematical formulation that is used to reset or recalculate target scores during interrupted one-day cricket matches. The model uses the notion that each team has two resources for scoring runs: balls remaining, and the number of wickets left. A team's run-scoring capability is dependent on the proportion of remaining resources which is used to calculate a team's score during an interrupted match. When the team batting in the second innings has fewer resources at their disposal than the bowling team, their target is adjusted downwards using the ratio of resources available to the two sides.

Similarly, Bhattacharya, Gill and Swartz, (2011), created a modified Duckworth-Lewis system specifically designed for resetting targets in interrupted Twenty20 matches. The method uses a Gibb's sampling scheme related to isotonic regression that is applied to the observed scoring producing a non-parametric resource table.

Clarke (1988) applied a dynamic programming model to one-day cricket to calculate the optimal scoring rate at any stage of an innings. This model was also able to estimate the total runs scored in the first innings and the chance of winning the second innings. The first innings formulation allowed calculations of a team's optimal scoring rate to obtain a given expected total, for any given number of wickets lost and balls remaining. The

second innings formulation developed a probability scoring table that outlines the probability of the second innings batting team scoring the target total, for any number of wickets lost and balls remaining.

Davis, Perera and Swartz (2015) developed a T20 simulator that calculated the probability of first-innings batting outcomes dependent on batter, bowler, and balls bowled and wickets lost. These probabilities were based on an amalgamation of standard classical estimation techniques and a hierarchical empirical Bayes approach, where the probabilities of batting outcomes borrow information from related scenarios (Davis *et. al.*, 2015). Simulation suggested that batting teams were not incrementally increasing aggressiveness when falling behind the required run rate. Aggression was defined as a function of runs and resources remaining; Scarf *et. al*. (2015) also used this definition.

Swartz, Gill, Beaudoin & deSilva (2004) used simulated annealing to conduct a search over a space of permutation of batting orders to find the optimal or near optimal order. Applying a Bayesian latent model, ball-by-ball outcome probabilities were estimated using historical ODI data and were dependent on batter, bowler, total wickets lost, total balls bowled and current match score. The Bayesian log-linear model was applied to the 2003 India World Cup squad and posterior estimates of the parameters were obtained by averaging output from a Markov chain. 71,000 first innings runs using India's 2003 World cup final batting order were simulated; overall a good fit between actual runs and simulated runs was found. The simulation was run 10,000 times and the number of first inning runs were average to produce an estimate. The batting order that corresponding to the maximized first innings runs was found. Overall the study found that the optimized batting order produced 6 more runs than the actual batting order.

The review reveals limited research focusing on the quantification of batting aggression in limited overs cricket. Consequently, the research objective is to determine the optimal or near optimal batting aggression to apply during various stages of the first innings of a T20 cricket match. The expectation is the optimal aggression will lead to a winning total. Only the first innings is considered as the second innings batting strategies are dependent on the number of runs scored in the first innings.

2. METHODS

T20 cricket was chosen to analyze batting aggression because, relative to the other formats, it is a fast transitioning and dynamic game, that requires aggression levels to change quickly to suit match conditions. When selecting an appropriate aggression level, batters must consider a variety of factors, such as: resources remaining and run rate. Here aggression is defined as the rate at which the run rate is modified by the batters, in addition, a new metric, scoring pattern is introduced. This is defined as the change in expected total per percentage of resources used. The concept of T20 batting aggression is inherently tied to scoring patterns and risk associated with changing run rate run rate. It is assumed that in T20 batters should play with a high level of aggression given the shorter format and fewer deliveries to influence a match. The greater the rate of change in scoring pattern the greater the shift in aggression. The optimal change in the scoring pattern was defined in three ways: 1) runs per total balls, 2) runs per percentage resources used and 3) expected runs per percentage resources used. These metrics provide a way to observe the change in scoring pattern throughout an innings as they consider two important factors that influence batting approach - resources and current total. Both metrics were compared against runs scored per percentage of resources used to evaluate the relationship and identify the optimal change in scoring pattern during the first innings. Moreover, it was hypothesized that there exists a strong positive relationship between the rate of change in scoring pattern and expected runs. Resources remaining and expected total were derived using the methodology outlined in Swartz and Perera (2015), and Patel, Bracewell and Bracewell (2018), respectively.

DATA

Ball-by-ball data from the following T20 competitions were used: Indian premier league (IPL; 2015, 2016, 2017 & 2018), Australian Big Bash League (BBL; 2016/17 & 2017/18), English NatWest T20 league (2015 & 2016), South Africa Ram Slam (2016 & 2017), Caribbean Premier league (CPL; 2014, 2015, 2016 & 2017) and NZ Super Smash League (2017/18). Overall the dataset contained 85,700 1st innings observations from 704 matches.

3. RESULTS

Figure 1 represents the approximately linear depletion of resources in T20 cricket indicating that Bhattacharya *et. al.* (2011) correctly considered the consistency of batting aggression in T20 cricket. The slope of losing teams is steeper showing that resources deplete quicker for losing teams compared to winning teams, due to loss of wickets. A polynomial regression was used to determine when in the first innings the rate of change in runs scored changes. The model results, defined as the sum of squared residuals of the least squares fit, were used as a measure of model accuracy. The chosen model for this analysis was the 5th degree polynomial as the model best represented the hypothesised curvature of the data and produced the largest drop in residual error.

To assess the optimal level of aggression at different stages of an innings, the changing nature of runs scored throughout the innings is examined. It was hypothesised that the run rate would change after the first six overs of the first innings, due to lifting fielding restrictions meaning there are more fielders outside the 30-metre circle. Towards the end of the innings, aggression is assumed to increase as there are fewer balls remaining, therefore less opportunity to deplete wicket resources before ball resources.

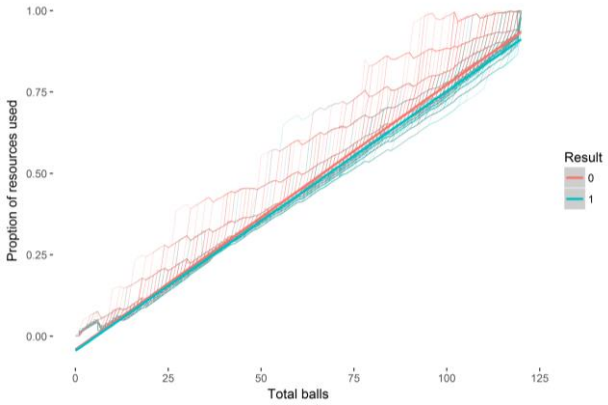


Figure 1: Comparison of resources used by balls consumed for winning and losing teams

Within each stage of an innings linear regression is applied to identify a relationship between runs scored and expected total, and percentage of resources used, for both winning and losing teams. This allows the examination of change in scoring pattern, given the level of percentage resources used. Running total and total balls was used to assess how the runs scored changes, for a winning team, compared to that of a losing team. Unsurprisingly, the winning team scores more runs than the losing team for any given number of balls. The winning team's running total is consistently higher than the losing teams running total over the innings. Essentially, winning teams apply greater batting aggression throughout the first innings, relative to losing teams. At the start of an innings, the running totals are similar, which could be due to the opening batters getting accustomed to the pitch conditions, pace of the ball or game strategy. As the inning progresses, the difference in running total between the winning teams and losing teams steadily increases. Equation (1) represents the polynomial of winning team, while equation (2) represents the polynomial of losing teams:

$$T(x) = (-8.2e-9)\,x^5 + (2.8e-6)\,x^4 - (3.0e-6)\,x^3 + (1.3e-2)\,x^2 + 1.2\,x - 1.2 \quad (1)$$

$$T(x) = (-8.3e-9)\,x^5 + (2.9e-6)\,x^4 - (3.3e-4)\,x^3 + (1.4e-2)\,x^2 + 1.0\,x - 7.0 \quad (2)$$

x represents the total balls during the innings and *T(x)* represents the running total after *x* balls. It was observed from the concavity and inflection points of the polynomial, that at the start of an innings, both teams scores are accelerating, validating the hypothesis that the scoring rate in the first six overs is higher, possibly due to field restrictions. The results showed that losing team's scoring rate decelerates earlier than the winning team. The first inflection point for the winning and losing teams is ~23 and ~20 balls respectively, implying that losing teams tend to decelerate their running total before the winning teams. This suggests that teams should bat aggressively for longer in the initial stages of the first innings to ensure their running total is increasing optimally. Next, the hypothesis that aggression is related to resources available is tested. Linear regression models were applied to various stages of the first innings to understand how run rate changes with resources. Four models representing different stages of an innings were developed which allowed the identification of the level of aggression that should be exercised during different stages of an innings. These stages are based on turning points identified by differentiating equations (1) and (2).

STAGE 1: OVERS 1 - 5

Figure 2 represents expected total regressed on percentage of resources used between overs 1-5.

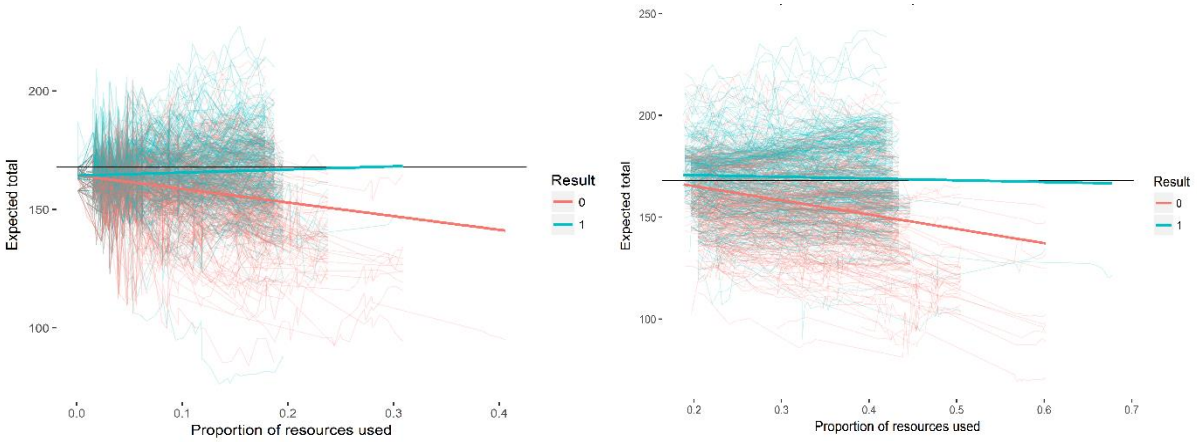
Winning teams are represented by equation:

$$S(r) = 215.8\,r + 1.9 \quad (3) \qquad E(r) = 13.0r + 164.3 \quad (4)$$

Losing teams are represented by equation:

$$S(r) = 163.5\,r + 3.4 \quad (5) \qquad E(r) = -57.6r + 164.4 \quad (6)$$

where *r* is the percentage of resources used, $S(r)$ is the runs scored and $E(r)$ is the expected runs.



Figures 2 & 3: Comparison of deviation of winning and losing teams from an average total of 168 by resources used in overs 1-5 and overs 6-10

For the first five overs, the winning teams utilize the resources available more effectively to accumulate more runs. Figure 2 reveals that winning teams tend towards the average first innings winning total of 168, while losing teams deviate away from this total as soon as the inning begins. By taking the derivative of the above equations, the study can inspect the rate of change in the runs scored per percentage of resources used for winning teams and losing teams and observe how this changes as the innings progresses. Therefore, the run rate and scoring pattern are defined as:

$$run\ rate = \frac{runs\ scored}{\%\ resources\ used} = \frac{dS(r)}{r} \qquad (7)$$

$$scoring\ pattern = \frac{expected\ total}{\%\ resources\ used} = \frac{dE(r)}{r} \qquad (8)$$

Winning teams have the higher run rate during this stage of the innings (i.e. 2.16), however the difference in run rate between the winning teams and losing teams is 0.523 (2.15846-1.6350). Surprisingly, there is a large difference in resource usage on expected runs between the winning and losing teams. Results indicate that winning teams utilize each percentage of resources more effectively (i.e. 13 run increases in expected total) relative to losing teams (i.e. 57 run decreases in expected total). Winning teams use each percentage to positively contribute to expected runs. The proportion of resources used explains 68% and 56% of variation in running total for winning and losing teams, respectively. The proportion of resources used explains 0.3% and 6% of variation in expected runs for winning and losing teams, respectively. These r-squared values indicate that resources used are of greater importance when explaining variation in expected total for losing teams when setting a total. This could be due to other factors such as percentage boundaries and team strike rate having a greater impact on expected total, for the winning team.

STAGE 2: OVERS 6 -10

Figure 3 represents expected total regressed on percentage of resources used between overs 6-10.

Winning teams are represented by equations:

$$S(r) = 157r - 28.2 \quad (9) \qquad E(r) = -8.5r + 172.4 \quad (10)$$

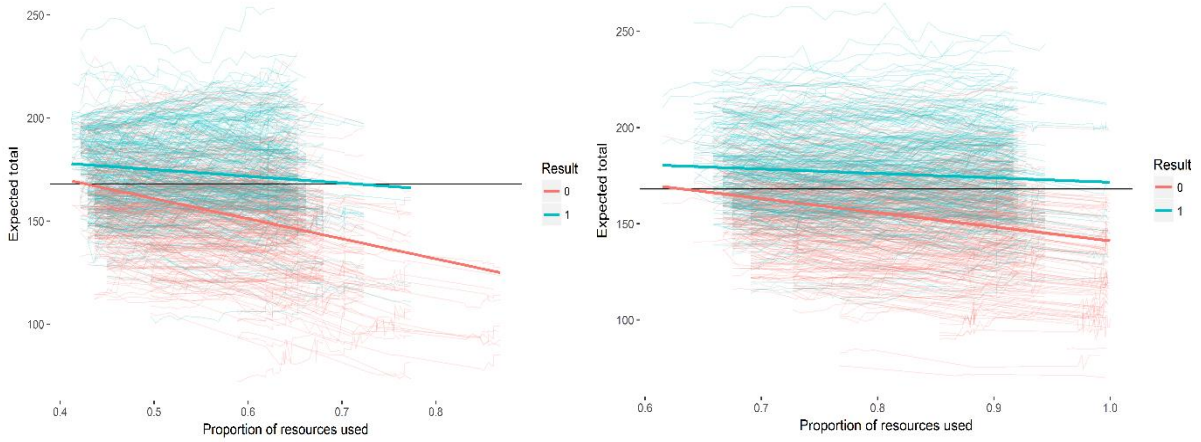
Losing teams are represented by equations:

$$S(r) = 118r - 20 \quad (11) \qquad E(r) = -70r + 179 \quad (12)$$

There is a 28% decrease in run rate for the winning and losing team between overs 1-5 to overs 6-10, revealing a significant drop in aggression for winning and losing teams. Moreover, there is a significantly large decrease in scoring pattern for the winning team, while the decrease in scoring pattern for both the losing team is not so pronounced. This reveals that a decrease in run rate has a greater impact on the expected total for winning teams than for losing teams. However, the winning team better utilizes each percentage of resources towards expected total than the losing team (i.e. -8.49 vs. -70.00). Based on early stages of the first innings teams that ultimately lose are in a position where they need to score more runs with fewer resources because of the run rate in the first five overs. These results suggest that when a team has over half of their resources left, they should slightly reduce run rate and scoring pattern. The proportion of resources used explains 64% and 51% of variation in running total for winning and losing teams, respectively. The proportion of resources used explains 0.09% and

6.6% of variation in expected runs for winning and losing teams, respectively. Again, this indicates that there are other factors, other than resources used that influence the winning team's ability to generate runs, compared to losing teams.

STAGE 3: OVERS 11 – 15



Figures 4 & 5: Comparison of deviation of winning and losing teams from an average total of 168 by resources used in overs 11-15 and overs 16-20

Figure 4 represents expected total regressed on percentage of resources used between overs 11-15.
Winning teams are represented by equation:

$$S(r) = 152.2r - 61.4 \quad (13) \quad E(r) = -31.5r + 191.2 \quad (14)$$

Losing teams are represented by equation:

$$S(r) = 96.6r - 36 \quad (15) \quad E(r) = -97.4r + 209.6 \quad (16)$$

The run rate for the winning team is 1.523, while for the losing team it is 0.966. The losing team experiences an 18% decrease in run rate, while the winning team experiences a 3% decrease. This reveals a similar level of batting aggression to stage 2 for the winning team, while losing teams decrease aggression. However, the run rate does not experience a large decrease between stage 2 and 3 for winning and losing teams, the reduction in expected runs for each percentage of resources used is significantly large. This indicates that although run rate does not decrease, other factors such as percentage dots and percentage boundaries decrease indicating a more conservative batting approach. As stage 3 batting approach tends to be more conservative the expected runs experience a decrease per percentage of resources used. This strategy may be due to teams conserving wickets for the final overs to rapidly accelerate run rate. Moreover, the winning team experiences a 270% decrease in scoring pattern, while the losing teams experience a 39.3% decrease. Stage 3 indicates that when more than 75% of resources are depleted, maintaining a very aggressive batting approach is impractical as it increases the risk of losing 10 wickets before the 20 overs have been bowled. The proportion of resources used explains 58% and 41% of variation in running total for winning and losing teams, respectively. The proportion of resources used explain 1.2% and 11% of variation in expected runs for winning and losing teams, respectively. During this stage of the match each percentage of resources used significantly reduces the expected total for winning and losing teams, again indicates reduced aggression levels.

STAGE 4: OVERS 16 - 20

Figure 10 illustrates expected total regressed on percentage of resources used between overs 16-20.
Winning teams are represented by equation:

$$S(r) = 168.18\,r - 108.2 \quad (17) \qquad E(r) = -22.86r + 194.4 \quad (18)$$

Losing teams are represented by equation:

$$S(r) = 115.03r - 71.90 \quad (19) \qquad E(r) = -73.32r + 214.4 \quad (20)$$

The run rate for both the winning and losing team increases. The losing team run rate increases by 19%, while for the winning team, there is a 10.5% increase. Moreover, the winning and losing team experience a 27% and 25% increase in expected total per percentage resources used between stage 3 and 4, respectively. Overall both teams increase aggression during this stage showing that as the inning nears completion, the risk and relative impact of losing all ten wickets is much lower and so batters can afford to play more aggressive shots to hit more

boundaries. The increase in run rate for the losing team is significantly larger than the increase in run rate for the winning team. This could be since the losing team are in a worse position in terms of runs scored and resources remaining, therefore they play more aggressively because the risk of losing all wickets before the end the innings and gaining more runs, outweighs the risk of not making enough runs and having wickets in hand at the end of the 20 overs. However, this increase is not enough to offset the significant decrease experienced for the losing teams between over 1-15. During this stage winning and losing teams experience a significant decrease in expected total for each percentage of resource used, however the resources are used more effectively relative to stage 3. Moreover, winning teams recover their final expected total to slightly greater than 168, while the losing team recover their scoring pattern.

4. DISCUSSION AND CONCLUSIONS

As expected, winning teams score at higher rates for longer, generating higher totals at the end of the first innings. However, teams that tend to set winning totals in T20 cricket have different in-game strategies, based on run rates and scoring patterns, compared to losing teams. To optimise winning, stages during a match when the batting team exercise higher levels of aggression were identified.

The inflection points calculated from the 5th degree polynomial show that the winning team begins with a higher scoring rate than the losing team as soon as the match begins. This is reflected in the difference in the run rate between the winning and losing teams in the first five over period. The fielding restriction in the first 6 overs (where only 2 fielders are allowed outside the 30m circle) allow for batters to play with lower risk of losing their wicket (for example fewer fielders to catch outside the 30m circle). This highlights the importance of having an aggressive opening batter. Importantly, slow starting teams are more likely to lose.

As the innings progresses, the winning and losing teams scoring rate begins to decelerate. However, regressing running total on total balls 5th degree polynomial, showed that the first inflection point of the winning team occurs after that of the losing team. This reveals that the winning team scores at a higher rate for a longer period than the losing team. The changes in runs scored per percentage of resources exhausted for the winning team is much less than the changes experienced by the losing teams. When at least 50% of resources remain, a team should be exercising higher levels of batting aggression. Maintaining a high level of aggression through the first half of the innings is important to increasing the probability of winning. Teams that lost typically decelerated their run rates too early. Once approximately 45% of resources have been exhausted, which is generally around overs 11 - 15, a more conservative batting approach is optimal to ensure the team bats all 20 overs with the better batters facing more balls. Losing wickets in the last 4 -5 overs is not detrimental if a team has remaining resources. If the batting team conserves resources well during overs 11-15 they provide the team with a better opportunity to take riskier but higher rewarding shots that will consequently produce a larger total.

This research indicates the importance of batting aggression in T20 cricket and develops a framework to manage run rate in the first innings to increase the batting team's chances of winning. Other applications could include: team selection, optimising the batting order and evaluation of chasing strategies. This framework serves as a useful way for coaches and players to identify how they should approach batting in the first innings to maximize their resources during different stages of an innings to increase their probability of winning.

References

Bhattacharya, R., Gill, P. S., & Swartz, T. B. (2011). Duckworth-Lewis and twenty20 cricket. *Journal of the Operational Research Society, 62*(11), 1951-1957.

de Silva, R. (2013). A Fair Target Score Calculation Method for Reduced-Over One day and T20 International Cricket Matches. *Journal of Mathematical Sciences & Mathematics Education*, *8*(2), 6-19.

Scarf, P., Shi, X., & Akhtar, S. (2010). Modelling batting strategy in test cricket. In *Progress in Industrial Mathematics at ECMI 2008* (pp. 481-489). Springer Berlin Heidelberg.

Clarke, S.R. (1988). "Dynamic programming in one-day cricket - optimal scoring rates". *Journal of the Operational Research Society*, 39: 331-337.

Perera, G. H. (2015). *Cricket Analytics* (Unpublished master's thesis). Simon Fraser University.

Patel, A. K., Bracewell, P. J., Wells, J. D. (2017). Real Time Measurement of Individual Influence in T20 Cricket. Paper presented at *MathSport International 2017 Conference*. (pp. 62-69).

Patel, A. K., Bracewell, P. J. & Bracewell, M. G. (2018, July 25-28). Estimating Expected Total in the First Innings of T20 Cricket Using Gradient Boosted Learning. Paper presented at *The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. University of the Sunshine Coast, Queensland, Australia: ANZIAM MathSport.

Lemmer, H. H. (2011). The single match approach to strike rate adjustments in batting performance measures in cricket. *Journal of sports science & medicine, 10*(4), 630.

COMMENTARY SENTIMENT AS A PREDICTOR OF IN-GAME EVENTS IN T20 CRICKET

Jack T. McIvor [a,c], Ankit K. Patel [a,b], Tamsyn A. Hilder [a] and Paul J. Bracewell [a,b]
[a] *DOT Loves Data, Wellington*
[b] *Victoria University, Wellington*
[c] *Corresponding author: jack@dotlovesdata.com*

Abstract

Ball-by-ball cricket commentary is a rich source of data with promising applications in statistical models. The cricket commentator describes the most interesting components of each ball as well as expressing an evaluation of player performance in the context of the game. In this paper, we develop a method of player identification and performance analysis using ball-by-ball commentary data. The information is then used for real time event prediction. Applying this method to a large data set of Twenty20 games obtained from ESPNCricinfo we were able to with better accuracy than baseline models: extract discrete phrases associated with one player's action, identify the player who is the subject of commentary and predict in-game events. There is potential to use this methodology across several sports by applying it to play-by-play commentaries.

Keywords*:* Natural language processing, sentiment analysis, entity recognition, dimension reduction

1. INTRODUCTION

Live match reports and commentaries have been around for many years. Recently, practitioners and pundits have picked up on the latent value of these data sources. Sports commentary and match reports are a rich source of in-game and post-game information containing player, team and game level insights. Live text commentary has been characterised by the interactiveness of its language on the one hand and real interaction between the audience and the journalist in the other (Chovance, 2010). Commentary is important to sports broadcasting as it aims to provide information about the game, such as describing the actions of the players and providing statistics. Duncan & Hasbrook (1988) stated that commentary has the effect of drawing the reader's attention to the part of the picture that merit closer attention, an effect called *italicized*. The commentator is a story teller that guides the viewers and readers attention to the most important aspects of the game or a given play. Farrell (2018) stated that commentators, whether written or spoken, convey in-game action as it happens with accuracy, clarity and concision. Three qualities of sports storytelling. Moreover, commentators must fulfil the key purpose of the reporting, i.e. describing the game objectively, fluently, largely error-free and in real-time.

The emergence of live text commentary has produced a huge amount of text commentary data, but there are few studies about utilising this rich data source. However, the literature primarily focuses on commentary synthesis (often post-game summaries) and analysis of fan emotions via social media. Sport-fan comments posted on social media such as Twitter have been found to contain valuable information regarding in-game events, anticipation of results and fan-base emotion towards results; (please see Sinha *et. al.* (2013) and Schumaker *et. al.* (2017)) Other studies that utilise text commentary include: Lareau *et. al.* (2011); Zhang *et. al.* (2016); Lee *et. al.* (2012); Ljaji & Arsic (2015); Minard *et. al.* (2015).

Cricket is a sport in which live commentary adds entertainment value and connects the audience with a shared language of the game. The first radio broadcast in the United Kingdom was by the BBC in 1927 (Baxter, 2007), but Australia had started to cover matches 5 years prior and covered a full test match in 1924. It was not until 1948 that every ball of a Test series was broadcast (Baxter, 2007). Live text commentary has only become available relatively recently, and so there is a lack of research attempting to extract the rich data available. Importantly, the tempo of the game, specifically the time between each bowler-batter interaction is sufficient to enable meaningful written commentaries to be written and published in real time.

This study attempts to extract useful information from live text cricketing commentary and use it to predict in game events, such as boundaries and wickets. To identify the information contained within each comment a pre-trained sentiment analysis model will be applied to determine a sentiment score.

MOTIVATION

Live text commentary provides a more truthful and nuanced view of performance compared to statistical outcomes. For example, the batter may edge a shot resulting in a boundary, while runs have been scored, this may indicate a lack of control and therefore should count negatively towards the batsman. As the commentator is a domain expert, they have a deeper understanding of actions and can provide informed evaluations. The commentator uses rich language to provide a thorough analysis, sharing their expert understanding with the

reader. Therefore, it is hypothesised that a machine which can parse and understand ball-by-ball commentary has an edge over traditional methods at predicting upcoming events within a game.

DATA

We programmatically collected commentary data from 705 games between 2015 and 2017 from ESPNCricinfo. This covers games from a range of domestic leagues played around the world: The Big Bash League (Australia), Super Smash (New Zealand), Caribbean Premier League, Indian Premier League, T20 Blast (United Kingdom) and T20 Challenge (South Africa).

The commentary is separated into distinct comments relating to balls. In total our commentary data covers 94,930 balls. We annotated 900 balls to train and evaluate models. After 900 balls, 1699 unique terms where encountered, accounting for 14.3% of the entire corpus of 11900 words. Importantly, comments share a common sentence structure. A standard example is: *"<Jadeja to Yuvraj Singh>, <no run>, fuller and on middle and leg, worked to short midwicket"*

The first two phrases of the comment are machine generated and always follow the form *<bowler to batter>, <outcome>*. Everything following is the description given by the human commentator. Often this follows a simple pattern, for example describing in order the ball, shot, fielding position, and fielding; sometimes with a following or preceding sentence summarising the action. For example: *"…tossed up and punished. Fuller on off stump, gets on the front foot and lofts it over mid-on fielder."*

However, the commentator is not limited to any predefined structure: *"…steps out and wallops this on the bounce to the long-on boundary. Lovely use of the feet, right to the pitch as he played this nicely…"*

ESPNCricinfo also displays comments made between balls which generally relate to game context. The commentator can curate and select public comments to display, which are often insights or opinions from the fans. This type of commentary is not included in the analysis.

2. METHODOLOGY

Here, we seek to identify evaluations of performance in commentary and then use these to infer sentiment and make predictions. Predictions cover events like wickets, boundaries or runs scored. We expect to have the best chance of predicting events immediately following the ball, with diminishing ability as time passes and play continues. Figure 1 outlines the approach. There are four separate tasks which are applied in chain. Although these models are applied at prediction-time in this order, they will be discussed in the order that they are trained. As will be shown, the phrase extraction task uses the classifier trained for player identification.

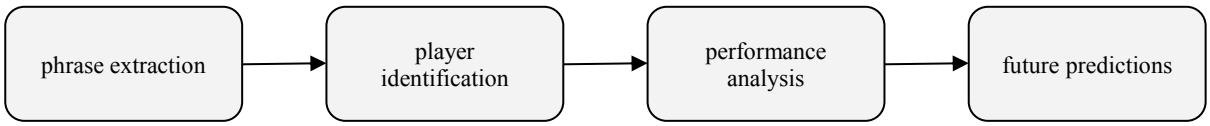


Figure 1: Four stage process for extracting and using key components of speech

Each layer of the framework is dependent on the previous, meaning that errors will snowball, and model performance is tightly tied to the phrase extraction and player identification tasks. These models form the backbone of the pipeline. We will attempt to evaluate the performance on each task both independently of other tasks (i.e. conditional on perfect previous performance), and as a link in the pipeline.

PLAYER IDENTIFICATION

The second link in the chain is player identification, in which we take a comment and seek to identify the category of the player or players to whom the commentary relates.

Task: Use a training set of phrase/player pairs to build a model which takes phrases and attributes (classifies) them to players.

Methodology: We use a bag-of-words model to perform this classification, specifically TFIDF (Term Frequency Inverse Document Frequency) vectors (considering each phrase to be a separate document) and an L2 regularised multinomial logistic regression. The hyperparameters: regularization strength and number of terms are selected using *3*-fold cross validation, this resulted in a unigram model.

Players are classified into four categories: batter, bowler, fielder and wicket-keeper. A fifth category 'meta' is used when the subject of a phrase is not a player but something else (e.g. the weather, the pitch or the umpire). From the labelled set, 50% of phrases are talking about the batter, 41% bowler, 7% fielding and about 1% each refer to the wicket-keeping and meta categories. In some cases, identifying the relevant player is straightforward,

for example when they are referred to by name or as the "batter" or "bowler", whereas in other cases the relevant player is inferred based on the context of the game and the vocabulary used.

Edge cases exist, where a decision about the category was made for the training data when the category is ambiguous. For example, when the bowler fields his own delivery, this is attributed to the *"fielder"* category.

Sometimes a description of the ball is hidden in a phrase about the batter and not given its own phrase. For example: *"...tries to pummel a* ***shortish delivery on off stump*** *over the leg side, but he doesn't get enough power going. Single to long-on..."* (where the bold text relates to the bowler and the remainder the batter). In this example the phrase contains analysis of the batter's actions and a description of the bowler's delivery. Analysing the vocabulary to determine the player category will confuse the model; however, this does not occur often, and the hidden description is unlikely to contain a significant performance evaluation given that the commentator has not given it a separate phrase. In our training set we have attributed the phrase to only one player category, at the discretion of the annotator, in the above example the phrase would likely be attributed to the batter.

Resolving names and pronouns should improve performance. That is, replace *"Kohli drives for four"* with *"<batter> drives for four"*. If we have a model of co-references, we could also replace pronouns. That is: *"Kohli drives for four. He hit that beautifully*" would be replaced with *"<batter> drives for four. <batter> hit that beautifully"*. The simplest heuristic for this may be: if a name is used, replace all subsequent pronouns with this name, then resolve the name to an exogenous list of positions. While players are normally identified in commentary by their last name, it would be beneficial to know nicknames and shortened forms, e.g. *'The Big Show'* for Glenn Maxwell, *'AB'* for AB de Villiers and 'Faf' for Francois du Plessis.

PHRASE EXTRACTION

This step involves separating a comment into chunks, where each chunk includes the phrases relating to the actions of a singular player. Some comments relate to one player only, in which case that comment is also a chunk, other comments relate to multiple players e.g. bowler, batter, and fielder in which case we need to separate the comment into 3 chunks. This is a partitioning of the original text (which we will sometimes call 'chunking')

Task: Using the same training set of phrase/player pairs, build a model which partitions text commentary into phrases which relate to one player.

Methodology: Our method first naively partitions to a fine resolution of pseudo-phrases, then aggregates phrases that relate to the same player. We begin with a naive partitioning method, then enumerate every partition and build a classifier to predict which partitioning is correct. This uses every potential phrase as an observation, with the real phrases used as a target.

From the training set, 17% of comments included a single chunk, 59% two chunks, 18% three chunks and 6% four or more chunks. The most common structure of chunks is *<bowl><bat>*, accounting for 55% of comments. Next most common structures are: 14% of comments as *<bat>*, 8% of comments as *<bat><bowl><bat>*, 6% as *<bowl><bat><field>*.

We used the player classifier to provide features for this model (i.e. probability of each chunk relating to certain player). We want the method to handle chunks that are ambivalent to player, because a word like *"nice"* should have near equal likelihood of relating to a batter or bowler. For example, *"Kohli drives for four. He hit that beautifully.*" contains important information, but it is hard to attribute the second sentence to Kohli without the context of the first sentence.

Naive partitioning can be done on punctuation (the most common are '.', ','), or any potential conjunction (which would include 'and', 'who', 'where'), or on every token (which would be robust to missing punctuation or misspelled words- for instance 'an' in place of 'and').

Accounting for lists of items should improve performance, as a list should always refer to the same player. A very common example are descriptions of the ball, for example: "*full, straight and swinging"*.

PERFORMANCE ANALYSIS

The third link in our chain involves analysing a chunk of commentary to determine how a player has performed, which is achieved by determining the sentiment of the commentary about that player. The structure below shows the intent of the performance analysis to map sentiment to the chunk and player:

"**full and straight**, this is **hit beautifully through cover** and **overrun by the fielder on the boundary**"
(bowler-neutral) (batter-positive) (fielder-negative)

Task: Evaluate how well a player has performed.

Methodology: We use a sentiment analysis as a proxy for performance, so we are in fact measuring how positive the commentator is (towards a player). Sentiment can be an evaluation of performance or a reflection of emotional state (normally game excitement). As a further dimension, it can be targeted towards: the

actions/outcome of a single ball (*'great hit'*), an entire performance (*'great innings'*), historical performance (*'great last year'*), hypothetical performance (*'he wants to be great'*), or sometimes unrelated to performance at all (*'great hairstyle'*).

Comments may be dependent on game context, but they can be understood when standalone. For example, a player who performs well in a high-pressure moment may be rewarded more than an equivalent performance in a low-pressure moment. Certainly, emotional reflections are dependent on game context. We can try normalise against this, although weighting towards exciting periods of play could be beneficial. Predictions (arguably) matter more in these exciting moments and so; we want to be correct more often when it counts.

Most expressions of sentiment are adjectives, adverbs or interjections, for example: "*excellent shot"*, *"played excellently", "excellent!"*. However, this is not necessarily true for evaluating performance more generally, for example: *"fumble"* is a verb that is negative, while *"six"* is a noun that is positive. 45% of observations are completely neutral. For each ball *i*, the sentiment is calculated as follows:

$$sentiment_k = \frac{\sum_j s_{j,k} \cdot p_{j,k}}{n} \qquad (1)$$

$$sentiment_k = \frac{\sum_j s_{j,k} \cdot \mathbf{1}(p_{j,k} \geq 0.5)}{\sum_j \mathbf{1}(p_{j,k} \geq 0.5)} \qquad (2)$$

Where *j* indexes chunks, n is the number of chunks, *k* is the player, $s_{j,k}$ is the sentiment of player *k* in chunk and $p_{j,k}$ is the probability that chunk *j* relates to player *k*. Equation (2) is a binarized version of (1) where $\mathbf{1}(\cdot)$ is the indicator function which evaluates to 1 if the expression is true, otherwise 0. For the remainder of this analysis equation 2 was used.

FUTURE PREDICTIONS

The final link is to use the player performance evaluations to predict likely outcomes. The hypothesis is high sentiment leads to improved future outcomes. For instance, if the batter sentiment is consistently high, the batting team is assumed to be performing relatively better which is expected to lead to more favourable outcomes on average. The challenge of this aspect of the analysis is to prove that the inclusion of this data generates better predictions than models which only account for runs scored on each ball.

Task: Correlate sentiment of the action to the ball outcome, such as: wickets, boundary events or runs scored. Here, it is important to consider the impact over a range of balls in the short term.

Methodology: A suite of time series regression analyses.

To begin with we construct a linear regression with lagged batter and bowler sentiments as predictors (careful not to include information from the current ball). We first target the number of runs scored on each ball with a distributed lag model shown in equation (1):

$$runs_i = \alpha + \sum_p \sum_{j=1}^{n} \beta_{k,j} \cdot sentiment_{k,i-j} + \varepsilon_i \qquad (3)$$

For *p* in {batter, bowler} and where *n* is the maximum number of lags, which is selected by minimising the Akaike information criterion. This results in 3 lags chosen for each player's sentiment (that is, include commentary information from up to and including 3 balls ago). An extension of this model is discussed under results.

3. RESULTS

PLAYER IDENTIFICATION

Given phrases, our player classifier gets 89% accuracy on the test set. This provides an adequate programmatic foundation for utilising natural language processing to derive further insights about a player and proves that vocabulary is sufficient to distinguish players.

PHRASE EXTRACTION

By naively chunking on every potential separator, then aggregating chunks, we obtained 71% accuracy. However, this is a harsh evaluation metric, because every word must be correctly assigned, even if it is not important to the phrase. As most of comments are appropriately classified, the practical impact is assessed by the impact on performance analysis and predictivity.

PERFORMANCE ANALYSIS

Commentary is often descriptive (objective) as opposed to opinionated (subjective). The commentator normally tries to appear unbiased yet interesting, which should mean a lot of neutral sentiment phrases. Here, we found

45% of comments were neutral. Given the nature of the interaction, with the bowler delivering and batsman reacting, it was anticipated that the sentiment of bowler/batter would have a push/pull relationship. However, there is a slightly (but significantly) positive correlation (r = 0.02, p = 1e-7) (see also Figure 2). It could be that in this case the commentators emotions 'crowd out' their evaluation of performance. As further work, it would be useful to decompose sentiment into these emotional and performance components. There is also an expectation that sentiment is higher for the winning team. In this case, batting sentiment was found to be 24% higher for the winning team.

The batters with the highest mean sentiment (limiting to those batter's who've at least 300 balls) are, in descending order: Q de Kock, RV Uthappa, JP Duminy, M Vijay, JC Buttler. The bowlers are: Mohammad Nabi, JD Wildermuth, BJ Dwarshuis, PJ Cummins and AJ Tye.

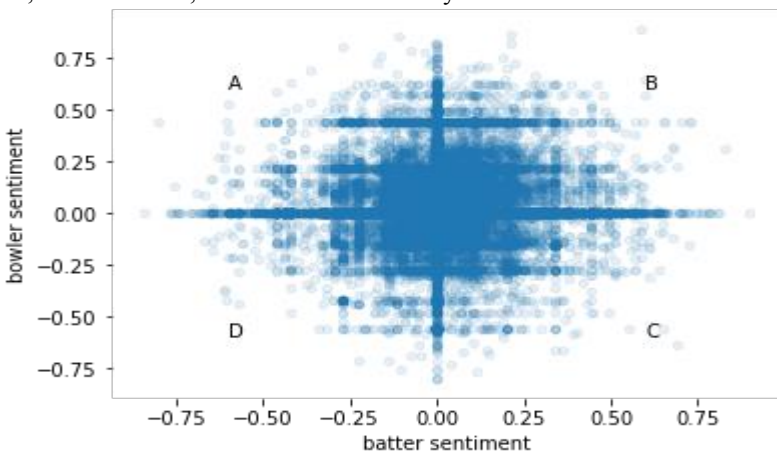


Figure 2: Comparison of bowler and batter sentiment. The striping is due to the use of common words.

The following are examples of commentary found across the quadrants:
A: "looks to drive a good length ball on the middle stump. He gets cut in half as he is beaten. Superb bowling"
B: "nice shape on off stump channel and Hyatt graciously let's go through to the keeper"
C: "back of a length ball but down the wrong line, easy pickings for Walton as he gets a nice tickle past the keeper"
D: "low full toss on off stump, worst all of the over and Malik misses it completely"

FUTURE PREDICTIONS

The F-test statistic reveals the coefficients of this model are jointly significant (p = 1.8e-4). Batter sentiment has positive coefficients while bowler sentiment has negative coefficients, meaning that positive commentary about the batting and negative commentary about the bowling is associated with more runs scored.

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.3268	0.005	241.674	0.000	1.316	1.338
bat-1	0.0677	0.040	1.710	0.087	-0.010	0.145
bat-2	0.0772	0.040	1.950	0.051	-0.000	0.155
bat-3	0.0553	0.040	1.396	0.163	-0.022	0.133
bowl-1	-0.1307	0.042	-3.122	0.002	-0.213	-0.049
bowl-2	-0.0002	0.042	-0.005	0.996	-0.082	0.082
bowl-3	-0.1100	0.042	-2.629	0.009	-0.192	-0.028

Table 1: OLS regression output

A logistic regression targeting whether the ball resulted in a boundary or not is significant (likelihood ratio [LLR] test p = 3.7e-3), while a target whether the ball resulted in a wicket or not is not significant (LLR p = 0.92). Care must be taken because the wicket model is severely unbalanced: only 5% of balls are wickets. This shows that bowlers can build pressure on the batter but still not achieve a wicket, while in the other direction, if the batter is on top they are rewarded more often with boundaries.

To prove the worth of sentiment as an additional set of predictors, we construct an autoregressive distributed lag model as follows:

$$runs_i = \alpha + \sum_{j=1}^{m} \gamma_{i-j} \cdot runs_{i-j} + \sum_{p} \sum_{j=1}^{n} \beta_{k,j} \cdot sentiment_{k,i-j} + \varepsilon_i \qquad (4)$$

For p in {batter, bowler} and where m and n are the maximum number of endogenous and exogenous lags. We then construct a Wald test for joint significance of the sentiment coefficients. The resulting p-value is 0.077, this shows that commentary includes additional information which helps to predict runs scored.

4. DISCUSSION AND CONCLUSIONS

The current approach successfully evaluated in-game performance using ball-by-ball commentary data. However, there are potential improvements which could be made to each underlying task. Phrase extraction and player identification could be improved using syntactical features, such as parts-of-speech tags or a dependency tree. Analysis of performance could be improved by accounting for cricket specific terms and jargon. For example: 'fumble' is always negative for a fielder, 'miscue' always negative for a batsman, and 'overstep' is always negative for a bowler. To improve predictions, we could account for more complicated error structures or use multiple equation prediction to jointly predict runs and wickets. We could also account for exactly which bowler/batter was attributed each sentiment, which would effectively weight match-ups- e.g. when a bowler is doing well against a batter. An alternative approach would be to train a deep learning model to predict game events from terms directly, effectively disintermediating the pipeline described. Our modular approach breaks the problem into successive tasks, where each task produces useful insights and enable auditability.

Understanding the relationship of commentators and the unfolding event has many benefits. Firstly, the efficacy of a commentator or pundit can be objectively assessed. This would also highlight if there is any bias by a commentator to any team or individual. Further expansion could lead to the creation of chat-bots to automatically generate content with limited human intervention. Therefore, the problem described here could be approached in reverse to find terms which appear before interesting events. Importantly these approaches could be applied to postgame reports, analysis and opinion articles.

This work has demonstrated the potential to predict in-game events using natural language processing. Specifically, we found that positive comments regarding batters and negative comments regarding bowlers result in more runs scored and, these predictors provide additional power to scorecard metrics. This study validates the hypothesis that commentators are storytellers with latent game understanding.

References

Baxter, P. (2007). The birth of cricket commentary. In *Test Match Special - 50 Not Out: The Official History of a National Sporting Treasure*. Random House.

Chovanec, J., & Ermida, I. (Eds.). (2012). *Language and Humour in the Media*. Cambridge Scholars Publishing.

Duncan, M. C., & Hasbrook, C. A. (1988). Denial of power in televised women's sports. *Sociology of sport journal*, 5(1), 1-21.

Farrell. (15th May 2018). Pitch perfect: the fine art of live sports commentary. Retrieved from https://www.rte.ie/eile/brainstorm/2018/0515/963663-pitch-perfect-the-fine-art-of-live-sports-commentary/

Lareau, F., Dras, M., & Dale, R. (2011, September). Detecting interesting event sequences for sports reporting. In Proceedings of the *13th European Workshop on Natural Language Generation* (pp. 200-205). Association for Computational Linguistics.

Lee, G., Bulitko, V., & Ludvig, E. A. (2012, October). Sports Commentary Recommendation System (SCoReS): *Machine Learning for Automated Narrative*. In AIIDE.

Ljajić, A., Ljajić, E., Spalević, P., Arsić, B., & Vučković, D. (2015, September). Sentiment analysis of textual comments in field of sport. In *24nd International Electrotechnical and Computer Science Conference* (ERK 2015), IEEE, Slovenia.

Minard, A. L., Speranza, M., Magnini, B., Qwaider, M. R., & Kessler, F. B. (2016). Semantic interpretation of events in live soccer commentaries. *CLiC it*, 205.

Schumaker, R. P., Labedz Jr, C. S., Jarmoszko, A. T., & Brown, L. L. (2017). Prediction from regional angst–A study of NFL sentiment in Twitter using technical stock market charting. *Decision Support Systems*, 98, 80-88.

Sinha, S., Dyer, C., Gimpel, K., & Smith, N. A. (2013). Predicting the NFL using Twitter. *arXiv preprint* arXiv:1310.6998.

Zhang, J., Yao, J. G., & Wan, X. (2016). Towards constructing sports news from live text commentary. In Proceedings of the *54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (Vol. 1, pp. 1361-1371).

USING NETWORK ANALYSIS TO DETERMINE OPTIMAL BATTING PARTNERSHIPS IN T20 CRICKET

Phillip Simmonds [a], Ankit K. Patel [a,b,c] and Paul J. Bracewell [a,b]
[a] *DOT Loves Data*
[b] *Victoria University, Wellington*
[c] *Corresponding author: ankit@dotlovesdata.com*

Abstract

Specific player partnerships in team sports can be the difference between success and failure. A clear example of this is in cricket, where two batters form a partnership and attempt to score as many runs as possible within a limited number of resources. This paper develops a framework to assess the influence any batting partnership has on a T20 match. Ninety ball-by-ball attributes from approximately seventy Big Bash League matches were obtained. These match attributes were analyzed using exploratory data techniques and random forests to identify the most important attributes that influence the number of runs scored, within a partnership. The important partnership attributes were: (1) partnership strike rate, (2) change in expected total, (3) proportion of resources consumed and (4) partnership contribution. These attributes were aggregated to create a partnership match influence (PMI) metric that quantifies the strength of a partnership. Applying the PMI metric as edge weights for Network analysis allows visualization of partnership strength within a team. It was found that the PMI in the second innings of T20 cricket was more indicative of match outcome relative to individual player influence, showing that building strong partnerships must be built to successfully reach the target total

Keywords*:* Network analysis, random forest, feature creation, dimension reduction

1. INTRODUCTION

Cricket is a growing sport, with countries all over the world strengthening their domestic leagues. Internationally, the sport has three formats: One-day cricket, Test cricket and T20 cricket. Subhani, Hasan & Osman (2012) claim that the future of cricket lies within the T20 format and will eventually become the most important format due to its short and exciting nature. In T20 cricket each team has a maximum of 20 overs, each consisting of 6 balls where the primary objective is to score as many runs as possible within the allocated resources. With such significant growth prospects, post, pre and live match analysis will increase in demand. Additionally, metrics such as partnership strength will become valuable property as coaches aspire to optimize their squads, and media outlets strive to market players as effectively as possible. Within 20 over cricket, strong batting partnerships that have a positive effect on their team's batting performance can be identified. Finding important partnerships is an example of how connections as few as two people can have a large impact on the outcome of an event. Given its fast transitioning and highly malleable nature, players have a greater impact on match outcome, therefore partnership influence is no more prevalent than in T20 cricket. This paper introduces a novel method to quantify partnership influence using random forests and neural network to visualize partnership strength.

LITERATURE REVIEW

Here, a brief overview of relevant research emphasizing the use of analytics surrounding limited overs cricket is provided. Brown, Bracewell and Patel (2017a) developed a model to generate the probability of batting partnerships being dismissed in the first innings of limited overs cricket. It was found that the longer a partnership exists, the higher the probability of a wicket falling. Brown *et. al.* (2017a) provided valuable information on individual and batting partnership strategies to maximize the chances of winning when setting a total. Extending this work, Brown, Bracewell, and Patel (2017b) created a model evaluating partnership survival rate using survival analysis to create the optimal batting order in limited overs cricket. Decision trees were applied to illustrate optimal international batting partnerships to produce the best batting order in relation to a strategic partnership from past games. The research concluded that batting order is an important consideration in cricket matches, and it exemplifies the importance of understanding partnerships within cricket.

Brown, Patel, and Bracewell (2017) built a model for predicting real-time dismissal probability for open batsmen in limited overs cricket, using Cox hazard models. The model implements three predictors: the cumulative number of runs scored, the cumulative number of dot balls faced, and the cumulative number of balls faced where less than two runs were scored off the previous four balls. Using this model, Brown *et. al.*

(2017), established that survival probability decreases with time, with Sri Lankan Kumar Sangakkara being the greatest at occupying the crease in these terms.

Prakash, Patvardhan & Lakshmi (2017) applied random forests to develop a team selection strategy for the Indian Premier League. The random forest model identified the following 12 player constraints for T20 teams, crucial in determining the influence of a partnership: 1 captain, 1 wicket keeper, 2 opening batsmen, 3 middle order batsmen, 2 tail order batsmen, at least 1 spin bowler and 2 fast bowlers.

Realizing the expected total at the fall of a wicket would provide a good representation of the effect a partnership has during their innings (Patel, Bracewell, & Wells, 2017). It illustrates whether the batting team is in a positive or negative position, at a given point in the innings. Patel *et. al.* (2017) developed a model to dynamically represent an individual's match influence within a limited over cricket match using ball by ball data. Patel *et. al.* (2017) built four unique models to represent player influence: 1) 1st innings batsmen logistic regression, 2) 1st innings bowler naïve Bayes, 3) 2nd innings logistic regression and 4) 2nd innings naïve Bayes. As well as this, a model was developed for predicting the expected total runs in an innings at any given point. The expected runs model produced a more accurate prediction for the innings total in comparison with the traditional metric, known as projected runs. Bhattacharya, Gill & Swartz (2011) investigated the Duckworth-Lewis system (1998) and its application in T20 Cricket. Bhattacharya et al (2011) discuss their concern surrounding the current Duckworth-Lewis system as it does not consider the more aggressive approach to batting in T20 Cricket. The study complements the ideas underlined by Patel *et. al.* (2017) and reinforced the importance of creating a more appropriate metric for calculating resource consumption, leading to a more appropriate model for setting totals for interrupted T20 matches.

Parera & Swartz (2012) found that the Duckworth-Lewis system was not the most appropriate system for calculating scores for delayed or canceled T20 Cricket matches. The paper investigated the stability of the Duckworth-Lewis system in resetting totals and surmised that the system needs to be improved and thus provided a modified resource table. Shah, Hazarika & Hazarika (2017) applied principle component analysis, to reveal that batting capability dominates bowling capability in limited overs cricket. Shah *et. al.* (2017) found that batting capability is more influential in T20 compared to one day cricket, as the variation due to bowlers is much larger in one day cricket. This implies that bowlers in the one-day format have a greater effect on match outcome in relation to bowlers in T20 cricket. The study also concluded that conditions such as location, weather and pitch condition have a significant effect on match outcome, making it unfair to generalize their findings across the two formats of the game.

Kampakis & Thomas (2015) applied machine learning to predict the outcome of English county T20 cricket. Tasks were approached from two angles; one with player data and one with the player and team data. Random forests, gradient boosted trees, naive Bayes and logistic regression were used to build models capable of predicting the winning team. The random forest model produced the lowest average accuracy with an average of 55.6%, compared to naive Bayes, which was the most accurate with an accuracy of 62.4%. This was due to two sessions where the random forest predicted particularly poorly (2010, 2012). Further analysis of the random forest model showed that on average it performed better at both the player and team level. The optimized naive Bayes model performed consistently above the betting benchmark.

Scarf, Shi and Akhtar (201) fit a negative binomial distribution to partnership scores in test match cricket. Using a non-parametric model that considered run rate as a covariate in the distribution of runs scored, the negative binomial was found to be a good fit. Pollard *et. al.,* (1977) also fitted a negative binomial to partnership scores and found a good fit.

The literature review reveals a lack of research focusing on quantifying partnership performance in limited overs cricket. Even though individual performances are important, a player cannot perform at maximum potential without the support of their team mates, particularly when batting. There is a clear niche for quantitative analysis on batting partnerships in T20 cricket. Therefore, this research attempts to quantify partnership strength in T20 cricket. It is hypothesized that 1) partnerships within T20 cricket can be quantified in terms of their influence on any given match; and 2) the quantified influence values can be applied as weights on an undirected network graph representing players in a batting team, with the edge weights between two batsmen representing a partnerships influence.

2. METHODS

Using various performance metrics, appropriate edge weights are created for a partnership network that represents the strength of all T20 partnerships. Edge weights are defined as the strength of the links between nodes on a network graph (Lawyer, 2015). The weights mapped to each edge within the network represents a given partnerships strength, therefore a measure that quantifies the strength of connection between batters

(i.e. nodes) needs to be identified. The applications of this analysis will be applied to batting partnerships from 70 Australian big bash T20 matches gathered from 2014-2017.

DATA

Ball-by-ball data was programmatically collected from ESPNCricinfo (www.espncricinfo.com) and separated into partnership observations - 770 partnerships were identified. Ball-by-ball resources used were aggregated for every partnership and was defined in terms of wickets lost and balls faced. Moreover, a ball-by-ball player influence score was obtained and averaged for the batting partnership. For each partnership the batters influence scores were summed and divided by the number of balls faced. The ball-by-ball expected total was calculated using the method outlined in Patel, Bracewell & Bracewell (2018). The expected total at the end of the partnership was subtracted from the previous partnerships expected total, producing a measure for the change in expected total as a proxy to partnership contribution. The ball-by-ball resources used was calculated using the method outlined in Perera & Swartz (2012), this produced a measure for the resources consumed by a partnership.

Attributes such as expected runs and resource consumed were used to represent the performance of individual batters as this has been found to be the most descriptive metrics of match outcome (Patel, Bracewell & Wells, 2017). For example, for a given ball, if the expected total experiences a drastic increase during the match, it can be assumed that the performance of the batting team exceeds that of the performance of the bowling team.

MODELLING

A random forest algorithm was applied to identify the most important attributes affecting partnership runs. A random forest is a collection of decision trees, it is combinations of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution as all other trees in the forest (Breiman, 1999). Random forests are an effective tool in prediction due to the rule of large numbers which prevents overfitting. Random forest performs better on large data sets, so in this case it is ideal as the analysis is dealing with many matches.

FEATURE CREATION

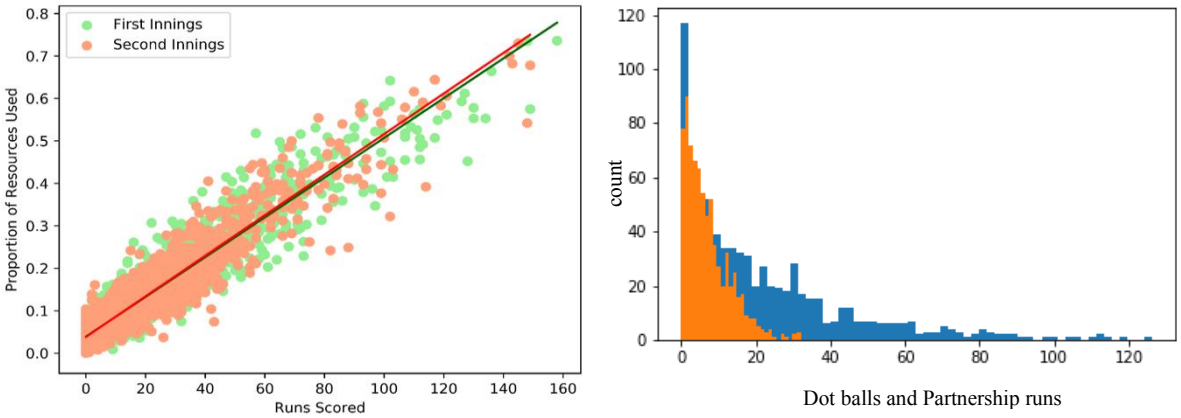
Using the important features identified through the random forest analysis, literature review findings and exploratory plots, a metric quantifying T20 partnership influence was developed. This 'influence' metric was used to create edge weights on a network graph representing the partnership connections within a match, which were later separated into first and second innings batters. Network graphs were created using the *networkX* library in Python.

3. RESULTS

Figure 1 illustrates that partnership runs increases as the proportion of resources used increases, across both innings. The coefficient of determination for both first and second innings partnerships of 0.93 (*p-value* <0.0001), suggests strong evidence for a very strong positive linear relationship between resources used and runs scored. Moreover, there was no statistical difference between the scoring rate per percentage resources used between first and second innings. The bottom left corner of figure 1 reveals a dense cluster of points which highlights that 65% of partnerships score less than 23 runs and consume less than 20% of resources.

Figure 2 shows an approximate Poisson distribution for partnership dot balls and an exponential distribution for partnership runs, respectively, which emphasizes the aggressive nature of T20 cricket. Moreover, it reveals a minimal number of dots implying a low number of balls where the batsmen failed to score. The plot reveals the low proportion of partnership dot balls greater than 15. Figure 2 also reveals the low likelihood of a partnership scoring many runs during an innings due to the quick turnover of partnerships in T20 cricket. Moreover, no difference was found between dot balls and runs scored across both innings.

Dot balls are an unreliable metric when measuring partnership strength in T20 Cricket as a partnership may face a high number of dot balls, however they can still accumulate many runs due to boundaries scored in-between dot balls. Due to the varying effect of dot balls on partnership runs, it is not used to determine partnership influence. This exploratory data analysis revealed the following metrics to be important when evaluating partnership performance: partnership runs, strike rate, expected total and resource consumption. Moreover, other attributes were found via the random forest analysis (next section) should also be included when creating the partnership metric.



Figures 1 and 2: Runs scored plotted against the proportion of resources consumed (left) and histogram of dot balls (orange) and partnership runs (blue) (right)

RANDOM FOREST ALGORITHM

The Random forest algorithms allows the identification of important features for a given response variable. A Random Forest provides a better understanding as to which attributes should be implemented to produce an appropriate partnership influence metric. Here, partnership runs scored were used as the dependent variable, as it accurately represents the partnerships contribution to a given match. The purpose of the random forest in this situation is to identify attributes that should be considered in the partnership performance metric.

Moreover, resources used takes each delivery into account, so if a dot ball occurs, it will still be accounted for within resources used and the number of runs scored. It is worth noting that the influence (i.e. player ratings) attribute is not of interest for the partnership influence, even though it ranked highly. The influence metric defines an individual player's match influence (Patel *et. al*. 2017), however, this study is only interested in the effect of partnership building, therefore including it in the partnership metric may show results towards a single player as opposed to a partnership.

Examining the random forest output and the exploratory analysis, the partnership match influence metric was created with the following variables: Strike rate, proportion of runs, change in expected total and proportion of resources consumed.

PARTNERSHIP MATCH INFLUENCE

The key attributes affecting partnerships in T20 cricket are: Strike rate, proportion of runs, change in expected total and proportion of resources consumed. A partnerships strike rate can be defined as the number of runs scored divided by the number of deliveries faced. Using a weighted combination of these attributes a T20 partnership influence metric (PMI) was created for T20 cricket.

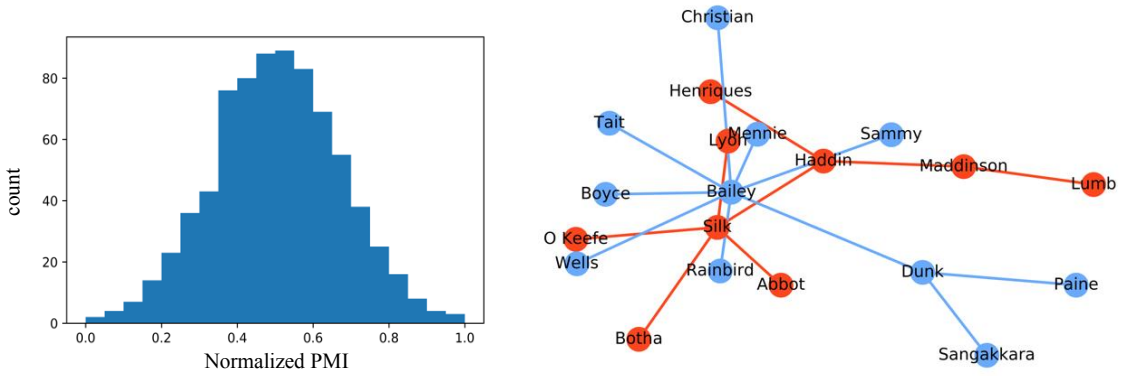
The proportion of resources used is multiplied by the change in expected total as it provides a good indication of a partnerships influence. A partnership must last more than a few deliveries to be considered influential. Strike rate is an important attribute in T20 cricket due to its aggressive and need for efficient contribution, hence it is included with the same weight as the expected total multiplied by the resources consumed. Finally, the proportion of runs scored within the partnership is multiplied to the entire function as it provides a clear representation of the partnerships effect on their team's innings. Equation (1) represents the raw PMI.

$$S_{ij} = \left(n_{ij} + \Delta E_{ij}\varphi_{ij}\right) p_{ij} \qquad (1)$$

S being the raw partnership match influence, η is the average runs per ball, ΔE is the change in expected total during the innings, φ is the resources consumed and ρ is the proportion of runs. Equation (1) was created to evaluate a partnerships strength and influence on a match of T20 cricket. It is implied that the expected runs are set to 0 at the start of an innings so the opening partnership will never have a negative change in expected runs, but they may still achieve a weak score as the metric accounts for resource used. The raw PMI was feed through a normalization (equation 2) process using the Scikit-Learn package 'Quantile Transformer', which calculates the normalized scores using a cumulative density function.

$$\gamma_{ij} = N(S_{ij}) \qquad (2)$$

Figure 3 shows the distribution of the PMI scores post normalization. Normalizing the scores creates a value that is easier to interoperate and compare. Scores above 0.75 are to be noted as they are particularly strong partnerships.



Figures 3 and 4: Distribution of Normalized PMI Scores (left) and Partnership Interactions by team (right) with the Sydney Sixers (orange) and the Hobart Hurricanes (blue).

PARTNERSHIP VS. INDIVIDUAL PERFORMANCE

The PMI metric was measured against individual player ratings (Patel *et. al.* 2017) to test whether partnerships or individuals' performances significantly affect match outcome. It was found that first inning partnerships have an equivalent impact on match outcome relative to individual performances. However, second innings partnerships were found to have a significant impact on match outcome relatively to individual performances. This result suggests that second inning individual batting performances do not significantly affect winning compared to partnerships, and that partnership must be built to successfully reach the target total. In first innings an individual batting performance and partnerships both are equally important to match outcome.

4. DISCUSSION AND CONCLUSIONS

The strongest partnership found in the data was formed between Perth Scorcher batters Marsh and Klinger who achieved a PMI of 1.0 as shown on the right tail of figure 3. The partnership faced 83 balls and scored 127 runs (71% of their teams score), maintaining an average of 1.53 runs per ball.

Figure 4 is a representation of both teams batting performance in the form of a network using a spring layout. The positioning of the nodes is determined based on the edge weights, which are the PMI's for each partnership. Figure 6 shows a match between the Hobart Hurricanes and the Sydney Sixers. In this case, Hobart batted second and lost by 95 runs. Hobart player Bailey can be easily identified as their team's key player by examining the network in figure 4. The blue network is centered around Bailey, showing the multiple connections (partnerships) formed. Aside from Bailey, only one other batter formed more than one connection, with both other connection being very weak partnerships (Dunk with Sangakkara and Paine). The Sydney sixers network shows three batsmen managed to form partnerships with more than one other player, implying a greater team effort in comparison to Hobart who had one stand out player who only formed one reasonable partnership (Bailey, Mennie).

Moreover, to assess whether the PMI was related to match outcome, each teams average PMI and best PMI, across each BBL match, were compared against match outcome. Of the 30 games in which the first innings batting team won 24 times (80%) the [first innings] batting team had the higher average PMI and 25 times (83%) the batting team had the best PMI. And, of the 37 games in which the second innings batting team won 26 times (70%) the [second innings] batting team had the higher average PMI and 19 times (51%) the batting team had the best PMI. These results show that in the second innings resources management (i.e. wickets in hand) and building multiple 'good' partnerships is key. An ANOVA test found a statistically significant difference between winning and losing teams average PMI and best PMI.

A strong partnership can influence the outcome of a cricket match; therefore, it is of significant interest to identify the game-changing partnerships. Exploratory data analysis and random forests were utilized to create a partnership performance metric that identifies the influence a partnership has in T20 cricket. Weights were then applied to the network graphs for batting team in T20 cricket. The weights on the graph are calculated using the partnership influence metric. Using the metric as a representation of performance, the identification of partnerships that show a strong positive connection can be conducted. The research reveals

that partnership influence in T20 cricket is dependent on three factors: 1) the volume of contribution, 2) efficiency of contribution and 3) contribution made under pressure (i.e. fewer resources).

It is important to have metrics such as the partnership match influence as it provides sufficient evidence surrounding the importance of partnerships in team sport. It is too often that sports enthusiasts hear about legendary individual players in team sports. All media outlets and analysts are guilty of over glorifying players that would otherwise not be considered strong players without their fellow team member. Individual players cannot be legends without a strong team behind them, this performance metric exemplifies this and encourages teamwork as a batter will find it very hard to achieve large individual scores if their partners keep getting out.

References

Breiman, Leo (2001). "Random forests." *Machine learning* 45, no. 1: 5-32.

Bhattacharya, Rianka, Paramjit S. Gill, & Tim B. Swartz (2011). "Duckworth–Lewis and twenty20 cricket." *Journal of the Operational Research Society* 62, no. 11: 1951-1957.

Brown, P., Patel, A. K, & Bracewell, P. J. "Optimising a Batting Order in Limited Overs Cricket using Survival Analysis". Paper presented at MathSport International 2017 (2017).

Brown, P., Patel, A. K., & Bracewell, P. J. "Real time prediction of opening batsman dismissal in limited overs cricket." (2017). *The Proceedings of the 13th Australian Conference on Mathematics and Computers in Sport.* (80-85). Melbourne, Victoria, Australia: ANZIAM MathSport.

Brown, P., Bracewell, P. J, and Patel, A. K. "Optimising Batting Partnership Strategy in the First Innings of a Limited Overs Cricket Match." *Journal of Quantitative Analysis in Sports* (in review).

Duckworth, Frank C., & Anthony J. Lewis (1998). "A fair method for resetting the target in interrupted one-day cricket matches." *Journal of the Operational Research Society* 49, no. 3: 220-227.

ESPNCricinfo. ESPN SPORTS MEDIA LTD. Retrieved from http://www.espncricinfo.com/ (Nov14, 2017).

Goldratt, Eliyahu (1999). *Theory of constraints*. North River Press, Great Britain.

Kampakis, Stylianos, & William Thomas (2015). "Using Machine Learning to Predict the Outcome of English County twenty over Cricket Matches." *arXiv preprint.* arXiv:1511.05837.

Lawyer, G. (2015). Understanding the influence of all nodes in a network. Scientific reports, 5, 8665.

Patel, A. K, Bracewell, P. J, & Wells, J. D. (2017). "Dynamic evaluation of player influence in T20 cricket". Paper presented *at MathSport International 2017*.

Patel, A. K., Bracewell, P. J. & Bracewell, M. G. (2018, July 25-28). "Estimating Expected Total in the First Innings of T20 Cricket Using Gradient Boosted Learning". Paper presented at *The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. University of the Sunshine Coast, Queensland, Australia: ANZIAM MathSport.

Perera, H. P., & Swartz, T. B. (2012). Resource estimation in T20 cricket. *IMA Journal of Management Mathematics*, 24(3), 337-347.

Perera, Harsha P., & Tim B. Swartz. "Resource estimation in T20 cricket." *IMA Journal of Management Mathematics* 24, no. 3 (2012): 337-347.

Prakash, C. Deep, C. Patvardhan, & C. Vasantha Lakshmi (2016). "Team Selection Strategy in IPL 9 using Random Forests Algorithm". *International Journal of Computer Applications.* 139, no. 12.

Scarf, P., Shi, X., & Akhtar, S. (2011). On the distribution of runs scored and batting strategy in test cricket. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 174(2), 471-497.

Shah, S., Hazarika, P., & Hazarika, J. (2017). "A Study on Performance of Cricket Players using Factor Analysis Approach." *International Journal* 8, no. 3.

Subhani., Muhammad Imtiaz., Syed Akif Hasan., and Ms Osman. "Will T20 clean sweep other formats of Cricket in future?" *MPRA Paper No. 45144* (2012). 98-102.

Wu, W., Shih, H., & Chan, H., "The analytic network process for partner selection criteria in strategic alliances." *Expert Systems with Applications* 36, no. 3 (2009): 4646-4653.

A FRAMEWORK TO QUANTIFY THE IMPACT OF SOCIAL ENGAGEMENT ON DATA DRIVEN CREATIVE

Ankit K. Patel [a,c,d], Madeleine K. A. Cook [b], Paul J. Bracewell [a,c] and Matthew B. West [a,b]
[a] *DOT Loves Data, Wellington*
[b] *EightyOne, Wellington*
[c] *Victoria University, Wellington*
[d] *Corresponding author: ankit@dotlovesdata.com*

Abstract

There is no literature shortage describing digital engagement and theories on how to measure engagement. However, there exists definitional inconsistencies on what digital engagement is due to a disconnect amongst academia and industry. Moreover, the increasingly cluttered digital media landscape has practitioners demanding a way to quantify the success of digital creative that extends beyond traditional models such as last click and media-mix models. Adopting and extending the digital engagement framework developed by Malthouse *et. al.* (2009) this paper introduces a novel framework to: 1) bridge the definitional gap between academia and industry surrounding digital engagement, 2) develop a methodology that quantifies digital engagement evoked by using creative in a digital environment, specifically Facebook and 3) identifies a more robust and meaningful digital success metric compared to antiquated measures of success. This developed framework and engagement equation are validated through Facebook's real-time ad optimisation platform – Dynamic Creative; and is applied within the sporting domain, specifically T20 cricket. The deployment of this framework and the measurement of content effectiveness can be applied to a range of digital channels outside the realm of social media, specifically Facebook.

Keywords: EdgeRank, digital engagement and gradient boosted machine.

1. INTRODUCTION

Currently there exists no uniform method to quantify digital engagement or an approach that outlines a way digital engagement could be evaluated. This opacity is partially due to the inconsistencies and lack of consensus on what consumer engagement is in a digital paid media context. According to Broadie (2016) there is a lack of consideration for engagement as amongst practitioners and academics there is little agreement on what "engagement" is. This definitional inconsistency has led to many definitions and academic research. Please refer to: Malthouse & Peck (2010); Malthouse *et. al.* (2007); Henderson, 2014; Bowden, 20009; Malthouse & Schaedel (2009), Dessart *et. al.* (2015), Coursaris *et al.* (2013), Ashley & Tuten (2014), Calder *et. al.* (2009)).

Moreover, Mersey, Malthouse and Calder (2015) claimed that the lack of agreement within the industry and academia on engagement has left many vulnerable to the definitions dictated by advertisers, who focus on brand placement, not content, or performance of this content.

From a marketer's point of view, Leander (2010) stated that marketers have different purposes for engaging their digital media audience, specifically on Facebook, but most want to know how engaging their content is and they need engagement to convert their audience to action. Due to the way social media impacts behaviour and behavioural change, and the way audiences consume content, in an increasingly cluttered digital environment, advertisers can no longer rely on traditional digital models, such as last click attribution and media-mix models, to quantify digital success (Nichols, 2013; Slefo, 2017). More specifically, the way consumers interact with media, digital media, has surpassed current measures of success (Villeneuve, 2017). Given the way content on digital advertising and news platforms has changed, there is a need for a measure to quantify authenticity, time effectiveness and aid in the enhanced delivery of this content.

Given the definitional inconsistencies across academia and industry, the lack of a quantitative methodology to effectively measure digital engagement and increasing redundancy of content success metrics, this paper attempts to develop a framework that effectively quantifies content engagement and identifies a more robust measure of success using digital media data, specifically on Facebook, corresponding to creative relating to T20 cricket. More specifically, the paper identifies a framework to quantify the digital engagement associated with creative content (also referred to as creative) and isolates the set of metrics that significantly affect engagement from a statistical and intuitive standpoint. Using the definition of engagement outlined in Malthouse, Calder & Schaedel (2009): a collection of experiences and the consequences of engagement being reactions to an ad, usage and attentiveness, and affective responses - this framework uses the likes, shares, comments and reactions associated with a post as a proxy to the "consequences of engagement". These consequences associated with Facebook content are extracted and aggregated in a meaningful way to derive an engagement score, which measures digital engagement evoked by creative, and therefore quantifying content effectiveness.

2. METHODS

The research objective is to develop a framework that dynamically quantifies digital engagement evoked by creative content and quantify the key metrics that drive engagement. The developed framework will be applied within the sporting domain, specifically, creative relating to T20 cricket. Facebook Dynamic will be used as it implements real-time variables, specifically time-bound content, such as real-time performance statistics, and geo-based content, such as game outcomes targeted to a region.

Engagement will be quantified using an 'engagement' equation derived using the EdgeRank algorithm – Facebook's news feed ranking system. This paper explores how behaviours and reactions relating to a post could be combined to derive a more effective prolonged measure of engagement.

Facebook Dynamic will be used to validate the framework and effectiveness of the input variables, as creative served by Facebook Dynamic is often optimised towards a key metric: *click throughs*. This optimisation is delivered by an optimal combination of variables including messaging and creative assets, subject to constraints such as targeting parameters. Therefore, it is assumed that the creative served through Facebook Dynamic will be the most 'engaging' combination of creative assets, including dynamic variables such as target parameters and messaging. Standard creative served on Facebook is generally optimised towards clicks and it is assumed that an optimal combination of assets and targeting parameters drives higher clicks. At present this is the key measurable metric in lieu of a way to measure true engagement.

However, digital success metrics such as *click throughs* are becoming increasingly irrelevant to advertisers and brands, and marketers are demanding more robust and measurable ways to quantify digital success.

This paper attempts to address two challenges in the way engagement is defined and subsequently operationalised: 1) Academic definition – as stated above there are definitional inconsistencies in engagement. Therefore, this paper attempts to shed light on engagement and the way it is quantified and measured, and the factors and the combination of factors that significantly impact engagement. 2) Commercial – Since Facebook's 2018 media release regarding the change in their algorithm that resulted in the prioritisation of authentic meaningful content on a user's feed. The ability to quantify authentic reactions to content has become increasingly relevant, and the ability create and shape creative to drive more authentic reactions is timely.

For a brand's content to now be prioritised by Facebook's algorithm it needs to be deemed meaningful and authentic to the user. Given Facebook's new algorithm prioritises content based on authenticity and how meaningful it is to the user, it is of paramount importance that advertisers post the most engaging piece of content on a user's feed to draw more insightful interactions such as likes, reactions, comments and shares, not just clicks and likes. Therefore, there has never been a more relevant time to explore a framework to quantify an emotional reaction to a brand's post. Moreover, the new algorithm prioritises active interaction, i.e. sharing, comments and reactions, over likes and click throughs which are deemed passive interactions. Facebook stipulates that actions that require more effort from the user deliver more authentic interactions and therefore are associated with more engaging content (Facebook, 2018). Specifically, a post that delivers a more engaging experience and results in a more emotive response, and therefore a greater level of interaction, is prioritised on a user's feed.

RESEARCH HYPOTHESIS

According to Mersey *et. al* (2009) and Malthouse *et. al.* (2015) the consequences of engagement are a by-product of a collection of experiences. Therefore, to generate an effective measure of engagement two elements must be established: 1) defined consequences of engagement and 2) how these consequences are weighted to represent individual experiences. Establishing these two elements will allow the creation of an engagement metric that correctly weighs and considers each consequence to produce an effective measure of engagement evoked by creative. The consequences of engagement are: *Usage and Attentiveness* – how much attention is on a post? *Affective responses* – the mood, feelings and attitude generated by a post; *Reaction to an ad* – the type of reactions associated with a post. Similarly, Dessart *et. al.* (2015) found three consequences of online media engagement: 1. *Affective* – captures the levels of emotions experienced by a consumer with respect to their engagement focus and relates to content and interactions; 2. *Cognitive* – refers to mental states that a consumer experiences with respect to the object of engagement, and 3. *Behavioural* – refers to consumers' desire to improve their experience and change something of relevancy or learn more to improve or change an outcome.

These consequences of engagement can be inferred through the type of interactions, passive or active, associated with a Facebook post. Using the number and type of reactions associated with a Facebook post as proxies to the consequences of engagement is a valid approach, as Mersey *et. al.* (2010) claimed that people will not attend to messages that have no perceived interest value for them. They will choose among media content offering those items they deem valuable, even if that is only momentary enjoyment. Table 1 outlines the type of actions associated with a post serve as proxies to each consequence. Here, usage and attentiveness metrics refer

to volume of an action, while affective responses and reactions to an ad metrics refer to the context. The emotive and reactionary metric-types provide an additional layer of understanding in terms of context. It is hypothesised that the amount of engagement evoked by creative can be approximated by appropriately weighting and aggregating the consequences of engagement associated with creative.

Usage and attentiveness	Affective responses	Reactions to an ad
No. of comments	comment sentiment	No. of loves
No. of shares	% of loves	No. of laughs
No. of views	% of laughs	No. of surprises
No. of likes	% of surprises	No. of sad
No. of tagged users	% of sad	No. of angry
	% of angry	No. of positive comments
	% of positive comments	No. of negative comments
	% of negative comments	

Table 1: Type of actions associated to each consequence

ENGAGEMENT FRAMEWORK

The developed framework utilised the engagement framework outlined in Malthouse, Calder, and Schadel (2009) (Figure 1). Malthouse *et. al.* (2009) claimed that to understand digital engagement different consumer experiences must be understood. Here, the idea that digital engagement is a collection of experiences is adopted. These experiences lead to consequences of engagement such as usage and attentiveness, affective response and reactions. Further, the framework is extended and introduces an additional layer (3) – engagement scores. This layer aggregates the consequences of engagement, i.e. $f(\alpha, \beta, \gamma)$, to derive an engagement score.

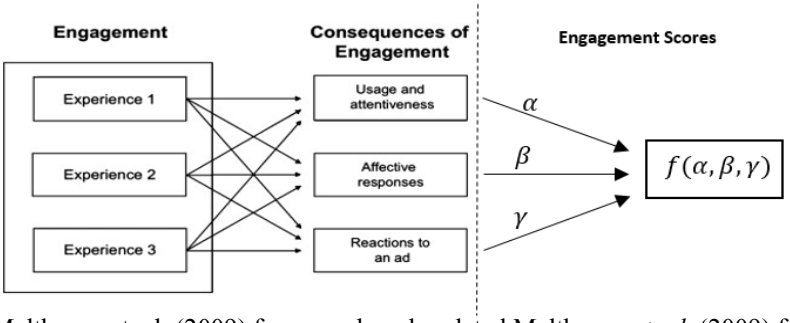


Figure 1 & 2: Malthouse et. al. (2009) framework and updated Malthouse *et. al.* (2009) framework

FACEBOOK DYNAMIC

Facebook's Dynamic Creative Tool automatically delivers the optimal combination of creative assets; data driven creative using mathematical optimisation. It runs different combinations of the creative assets such as images, videos, likes, descriptions and call-to-actions across the target parameters (i.e. constraints) to determine which combination produces the optimal result (i.e. click throughs). The tool automatically randomises ad variations and makes it easier to target the right people with the right message; for more detail please refer to Facebook, 2018; Advertismint 2018; SocialMedia Examiner 2018; Disruptive Advertising 2018.

EDGERANK ALGORITHM

EdgeRank is Facebook's news feed ranking algorithm that decides which stories and posts appear in each user's newsfeed. Equation (1) outlines the EdgeRank equation. The algorithm imposes three elements: 1. Affinity ($\boldsymbol{u}$), 2. Weight ($\boldsymbol{w}$) and 3. Time Decay ($\boldsymbol{d}$). The affinity metric measures the closeness of the relationship between the user, the content and source. The affinity score is derived from a user's actions and factoring in the strength of the action, the closeness of the relationship and time since the action was taken. The weight score measures the type of action that was taken from viewing the content. For example; actions such as a comment are weighted more heavily than a like. The decay score measures the recent and current the content is. The time decay is adjusted based on 1) how long since the user logged into Facebook and 2) how frequently a user logs on.

$$\sum_{edges\ e} \boldsymbol{u}_e \boldsymbol{w}_e \boldsymbol{d}_e \qquad (1)$$

DATA

Reaction and behavioural data relating to a given Facebook post is extracted with the following metrics: post type, reactions, likes, shares, comments and views. Data can only be accessed through administrative access to the account to which the post relates.

The following metrics will be created: Post sentiment (number of positive and negative reactions), ratio of reactions (happy, sad, surprised, love and angry) to *likes,* proportion of each reaction, and the proportion of positive, negative and neutral comments. A sentiment algorithm will be applied to each comment to derive comment sentiment scores.

Corresponding 'delta' metrics will also be created, i.e. change in number of likes, between time *i* and time *i*+1. These delta metrics will introduce a time-component and represent how reactions and behaviours change over-time.

The metrics are classified into three groups: 1) action, 2) context and 3) time. Action metrics refer to *usage and attentiveness* consequences, context metrics refer to *affective responses* and *reactions to ads* consequences, and time metrics refer to *change in rate* metrics.

VARIABLE SELECTION

To establish metric importance and select the appropriate metrics to feed into the engagement equation a variable selection process is necessary. A generalised boosted regression model is applied to identify the most important metrics affecting post effectiveness. First, a dependent variable must be established to identify the relative importance of each metric. Given *click throughs* are an industry standard as a success metric, it will be used as the dependent variables and act as a proxy for 'true' digital success'. Moreover, the boosted model will account for the complex interaction between metrics by taking a sequence of weak learners to construct a complex learner.

WEIGHT ALLOCATION

Given the relative influence score of each metric an Analytical Hierarchy Process (AHP) which generates decision weights through pairwise comparisons and relies on the judgment of experts to derive priority scales (Thomas, 1999). The AHP is applied to three different metric types that construct the engagement equation: action, context and time. The weight allocation process works as follows: 1) conduct a variable importance analysis, 2) identify the top three metrics associated with the three metric-types, 3) normalise the relative influence scores associated with each of the top three metrics - $N(\boldsymbol{\alpha_i}) = \frac{\alpha_{i1} - \min(\boldsymbol{\alpha_i})}{\max(\boldsymbol{\alpha_i}) - \min(\boldsymbol{\alpha_i}}$, and 4) apply the AHP – the normalise weights serve as the initial weights for the pairwise comparison matrix.

It is assumed that the importance analysis will rank context-based metrics (*affective responses* and *reactions to an ad*) higher than action-based metrics (*usage and attentiveness*). This assumption is derived from Mersey *et. al* (2010) - people will not attend to messages and content that have no perceived interest value for them. They will choose among content offering those items they deem valuable, even if that is only momentary enjoyment – therefore the more engaging the post the more active the associated interactions.

Moreover, context-based metrics are more representative of active interactions, while action-based metrics are more representative of passive interactions.

"ENGAUGEMENT" RANK

Bracewell (2003) stated that to generate a meaningful rating system the system must be robust, transparent, reliable and meaningful. Moreover, Patel & Bracewell (2018) claimed that a rating system is an elegant exercise of dimension reduction, and to produce a highly predictive rating system different variables, such as action, context and time, that address different dimensions of the data should be utilised. Therefore, to successfully quantify digital engagement it is proposed that the framework established in Malthouse *et. al.* (2009), should incorporate time-based metrics, in addition to action and context-based metrics. Moreover, the time component serves as a decay measure that accounts for post relevancy and "current-ness" of a post. Equation (2) outlines the engagement (φ) evoked, at time *i*:

$$\varphi_i = \sum_e \boldsymbol{\alpha}_i^{\omega_{i1}} \boldsymbol{\kappa}_i^{\omega_{i2}} \boldsymbol{\tau}_i^{\omega_{i3}} \qquad (2)$$

Where $\boldsymbol{\alpha}_i$ represents the action-based metrics (i.e. usage and attentiveness) at time *i*, $\boldsymbol{\kappa}_i$ represents the context-based metrics (i.e. affective responses and reactions to an ad) at time *i* and $\boldsymbol{\tau}_i$ represents the time-based metrics (i.e. rate of change in $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$) at time *i*. ω_{i1}, ω_{i2} and ω_{i3} represent the [meta] importance weights associated with each metric-type at time *i*. These weights are calculated by the proportion of action, context and time-based metrics that make up the top 10 metrics in the variable importance plot. For example, at time *i*, if context metrics

($\boldsymbol{\kappa}$) make up 50% of top 10 metrics, action ($\boldsymbol{\alpha}$) metrics make-up 20% and the remainder is made up of relevancy metrics than: $\omega_1 = 0.2, \omega_2 = 0.5$ and $\omega_3 = 0.3$.

The following procedure is applied to identify the metrics used in each part of the equation: 1) apply a Gradient Boosted Machine (GBM) importance analysis 2) identify the top 3 metrics associated with each of the three categories, 2) normalise the influences scores, 3) calculate the proportion of importance each metrics makes up in the three metrics. These proportions represent the weights that feed into the metric-type specific equations (3), (4) and (5).

$$\boldsymbol{\alpha_i} = \alpha_{i1}{}^{\gamma_{i1}} + \alpha_{i2}{}^{\gamma_{i2}} + \alpha_{i3}{}^{1-(\gamma_{i1}+\gamma_{i2})} \quad (3)$$

Equation (3) represents the action measure, where α_{i1} represents the most important action metric at time *i*, and γ_{i1} and γ_{i2} represent the AHP weights, at time *i*.

$$\boldsymbol{\kappa_i} = \kappa_{i1}{}^{\gamma_{i1}} + \kappa_{i2}{}^{\gamma_{i2}} + \kappa_{i3}{}^{1-(\gamma_{i1}+\gamma_{i2})} \quad (4)$$

Equation (4) represents the context measure, where κ_{i1} represents the most important context metric at time *i*, and γ_{i1} and γ_{i2} represent the AHP weights, at time *i*.

$$\boldsymbol{\tau_i} = \tau_{i1}{}^{\gamma_1} + \tau_{i1}{}^{\gamma_2} + \tau_{i1}{}^{1-(\gamma_1+\gamma_2)} \quad (5)$$

Equation (5) represents the time-based measure, where τ_{i1} represents the most important time metric at time *i*, and γ_{i1} and γ_{i2} represent the AHP weights, at time *i*.

The raw engagement scores (φ_i), at time *i,* will be transformed into a standard normal variable. Apply a sigmoid function to obtain ratings (0,1).

3. RESULTS

APPLICATION

The framework will be applied within the sporting domain, specifically Big Bash T20 (2017-2018). Creative assets (i.e. text, images, descriptions) relating to a specific match regarding a player performance, will be uploaded to Facebook served in a dynamic fashion, subject to targeting parameters. Facebook Dynamic will then serve the optimal combination of assets, serving the most relevant creative to the right people, optimised towards a given engagement objective.

Content describing individual cricket player influence (Patel et. al., 2017) during the 2018/19 T20 Tri-series between Australia, England and New Zealand was targeted at cricket fans in specific geographies. Subsequent fan engagement was measured using click through rates (CTR). Infographics and line graphs were contrasted using Facebook advertising's split-test function. Split-testing divides the audience into random, non-overlapping groups and enabled the capability to test four variations of each graphic. This randomisation ensured the test was conducted fairly. Overall, line graphs received a CTR 3.0 times higher than the infographics. Highest engagement was with males ages 18 to 30. These results suggest that content delivered in a familiar format is most likely to drive fan engagement. However, to test this conjecture, further work is required.

PROPOSED VALIDATION

To validate the engagement scores, firstly reactions to creative must be tested with a statistically significant sample size, using a sizeable cookie pool in market for a sufficient period. This will be gathered by developing a 'sufficient' number of permutations of creative using standard adserving techniques, i.e. not served dynamically, for creative that is not targeted with the same number of variables to a specified audience. Next, a second phase of creative will be served with a greater number of variable components. Here, the objective will be to measure engagement based on a highly targeted, real-time approach, served through Facebook Dynamic.

The overarching objective of this exercise is to measure the interaction, i.e. the engagement with a piece of content, to interrogate the framework and the validity of the engagement metric. Overall this will assess the hypotheses that the engagement evoked by creative is a function of action, context and time-based metrics, and can be quantified if the relevant metrics are applied with appropriate weightings. An engagement score will be produced at time *i* for each creative within each phase, and assuming the creative produced by Facebook Dynamic is more relevant and authentic, and therefore engaging, the corresponding engagement score should be higher.

Moreover, comparing the engagement scores and click throughs associated with each creative, at time *i,* the hypothesis that the developed metric provides a more robust and meaningful measure of digital success, can be tested. The longer a creative is in market, the greater the number of data points available, making it possible to identify how the metric importance changes over time, how the associated weights change over time, and the type of interactions that differentiate highly engaging pieces of content. This enables the identification of the

metrics and components that affect engagement as creative is market. Moreover, it is possible to track the evolution of metric importance over time, and the way the weightings evolve, allowing time affects to be tested.

4. DISCUSSION AND CONCLUSIONS

Unfortunately, the maturation of the digital environment and increasingly cluttered digital media landscape has been met by a slow evolution in digital success metrics. Given the increasingly difficult task of effectively reaching the target audience, marketers are demanding more robust and measurable ways to quantify digital success.

Adopting and extending the engagement framework developed by Malthouse *et. al.* (2009) this study proposed a possible way to quantify digital engagement evoked by creative, quantify the key metrics that drive engagement and bridge the definitional gap of engagement that exists amongst industry and academia. Upon the development of a robust engagement framework, Facebook Dynamic will be used as the most relevant vehicle in which to test a multitude of creative iterations to measure the performance and effectiveness of the engagement metric with real-time T20 cricket results.

However, the framework has minor flaws, for example the multiplicative nature of the engagement metric means that in any instance in which one of the three metrics-types (i.e. action, context, time-based) has a zero outcome, and the overall engagement score will be zero. This is an unintuitive result from a marketing standpoint as it is not able to be operationalised to drive a more positive result.

References

Advertisemint. (2017). Everything You Need to Know about Dynamic Product Ads. Retrieved from https://www.advertisemint.com/dynamic-product-ads/

Ashley, C., & Tuten, T. (2015). Creative strategies in social media marketing: An exploratory study of branded social content and consumer engagement. *Psychology & Marketing,* 32(1), 15-27.

Bowden, J. L. H. (2009). The process of customer engagement: A conceptual framework. *Journal of Marketing Theory and Practice*, 17(1), 63-74.

Coursaris, C. K., Van Osch, W., & Balogh, B. A. (2013, June). A Social Media Marketing Typology: Classifying Brand Facebook Page Messages for Strategic Consumer Engagement. *In ECIS* (p. 46).

Calder, B. J., Malthouse, E. C., & Schaedel, U. (2009). An experimental study of the relationship between online engagement and advertising effectiveness. *Journal of interactive marketing*, 23(4), 321-331.

Cotter, S. (2002). Taking the measure of e-marketing success. *Journal of Business Strategy*, 23(2), 30-37.

Davis Mersey, R., Malthouse, E. C., & Calder, B. J. (2010). Engagement with online media. *Journal of Media Business Studies*, 7(2), 39-56.

Dessart, L., Veloutsou, C., & Morgan-Thomas, A. (2015). Consumer engagement in online brand communities: a social media perspective. *Journal of Product & Brand Management,* 24(1), 28-42.

Disruptive Advertising. (2018). Automation at its Finest. How to Use Facebook Dynamic Product Ads. Retrieved from https://www.disruptiveadvertising.com/ppc/ecommerce/facebook-dynamic-product-ads/

Facebook. (2018). Set up dynamic ads once, then let them work for you. Retrieved from https://www.facebook.com/business/learn/facebook-create-ad-dynamic-ads

Henderson, C. M., Steinhoff, L., & Palmatier, R. W. (2014). Consequences of customer engagement: how customer engagement alters the effects of habit, dependence, and relationship-based intrinsic loyalty. *Marketing Science Institute Working Papers Series*.

Leander, M. (2017). What is a good Engagement Rate on a FaceBook Page? Here is a benchmark for you. Retrieved from https://www.michaelleander.me/blog/facebook-engagement-rate-benchmark/

Malthouse, E. C., Calder, B. J., & Tamhane, A. (2007). The effects of media context experiences on advertising effectiveness. *Journal of Advertising*, 36(3), 7-18.

Nichols, W. (2013). Advertising Analytics 2.0. *Harvard Business Review*, 91(3), 60-68.

Peck, A., & Malthouse, E. C. (Eds.). (2011). Medill on media engagement. Hampton Press.

Slefo, G. (2017). How Google Plans to Kill 'Last Click Attribution'? Retrieved from http://adage.com/article/digital/google-moves-kill-click-attribution-sf-event/309129/

SocialMedia. (2018). How to Retarget on Facebook and Instagram With Dynamic Product Ads. Retrieved from https://www.socialmediaexaminer.com/how-to-retarget-on-facebook-and-instagram-with-dynamic-product-ads/

Thomas L. Saaty. Decision making for leaders: the analytical hierarchy process for decisions in a complex world. *RWS publications*, 1999.

Villeneuve. N. (2017). Last-Click is Dead, Long-Live Multi-Touch Attribution. Retrieved from https://www.adviso.ca/en/blog/tech-en/last-click-is-dead-long-live-multi-touch-attribution-2/

DERIVING AN EXACT BATTING SURVIVAL FUNCTION IN CRICKET

Bernard J. Kachoyan [a,c] and Marc West [b]
[a] *School of Mathematics and Statistics, University of New South Wales.*
[b] *Affiliation not specified*
[c] *Corresponding author: b.kachoyan@unsw.edu.au*

Abstract
This paper derives a general batting survival function from a minimalist set of assumptions, specifically taking into account that a different number of runs can be scored at each point of an innings. It is shown how this approaches a memoryless survival function at a moderately low score. The paper also shows that an estimate of the value of the underlying parameters can be derived from a batsman's aggregated data summaries available from most common cricket databases. Finally the paper compares the theoretical survival function with those derived from product limit estimates (PLE) to reinforce both that the theoretical distribution is a good estimate of any underlying distribution, and that the whole PLE curve is useful for highlighting specific characteristics in the careers of individual batsmen.

Keywords: Cricket, batting, survival, geometric distribution, memoryless, Product Limit Estimators,

1. INTRODUCTION

This paper derives a theoretical batting survival function from first principles, explicitly taking into account that both the number of runs scored and the probability of dismissal can vary at each stage of a batsman's innings. An exact solution is derived under the assumption that these probabilities are constant, and it is shown that the survival curve converges to a memoryless one. Matching to an individual batsman's aggregate statistics allows the creation of an underlying survival function, which is more directly related to the long term expectations of the batsman than a simple statistical line-of-best-fit, thus providing a more robust baseline to compare deviations from the 'expected' performance at various scoring ranges.

The question of whether a batsman's survival function obeys a geometric and hence conceptually important memoryless distribution has been the subject of analysis since [1]. This pioneering study found a good general fit, although the geometric distribution (GD) underestimated the number of zero scores (known as 'ducks'), and overestimated the number of scores of 100 or more ('centuries'). Reference [2] concluded that at higher scores the survival distribution is "roughly geometric in form" but that again the pure GD underestimates the number of ducks [3]. It also showed that geometric is the only distribution where the traditional calculation of batting average (total career number of runs divided by number of dismissals) is a consistent estimator whatever the censoring mechanism. Hence, the proximity of batting survival functions to a GD is important when considering the validity of the traditional average as a measure of the 'true' average. However, they noted that "the geometric model cannot apply exactly since a score does not increase by a fixed amount at each stage"; it is partly this issue that we seek to address. Most recently, [4] considered a number of generalised geometric models of batting survival and found that models that assume a constant hazard for a large proportion of the run space, typically between 1 and 100 with a different constant thereafter, are the best fit to the 20 batsman they considered, covering different eras.

Reference [5], as far as we can ascertain, is the first to consider the derivation of a geometric survival curve from first principles using a simple set of underlying specific assumptions, namely that the probability that the innings ends with each ball faced, the probability that the batsman makes a scoring shot with each ball faced, and the ratio of the number of runs obtained to the number of scoring shots are all constant. The last was a way around the score not increasing by a fixed amount at each stage and was used as a tuning parameter in the model. This paper extends the work of [5] to explicitly take into account the fact that a different number of runs can be scored at each scoring shot, each with a different probability, and that these probabilities need not be constant throughout the innings. The formulation employs difference equations and leads to an easy generation of the solution. If the probabilities are assumed constant then an analytical solution can be derived, and the solution approaches a GD as a limit after a certain number of runs. The formulation explicitly takes into account the problems at zero [6, 7] using the number of ducks obtained directly from the batsman's data.

Most recent analyses have used Product Limit Estimate (PLE) techniques [4, 8-10]. Although the PLE does have some deficiencies in fully modelling the behaviour of batting survival [4], in this paper it is used to give an indication of the deviation of a batsman's survival characteristics, given by their actual data, from our estimate of their theoretical underlying distribution.

The theoretical formulation of survival will be derived in Section 2. Section 3 uses a large dataset to examine the probabilities of scoring a given quantity of runs at each scoring shot. Section 4 will explore how the quantities needed to create the distribution for each batsman can be derived from aggregated data summaries easily obtainable in public databases. From this, comparisons are made between the purely data derived PLE and the data/analytic estimate of the underlying distribution. For the comparisons to real data, we will restrict ourselves to men's Test (long-form) cricket because, compared to limited overs cricket, Tests are the most conducive to high scoring innings, have a long history, runs tend to be more important than scoring rate, comparison with previous work and to bound the scope of the analysis.

2. THEORETICAL FORMULATION

Begin by defining the batting survival function S_j, the probability that a batsman survives > j runs, as

$$S(j) = S_j = \Pr(r > j) \tag{1}$$

where r is the random variable representing the number of runs scored. This is a discrete function, as only integer number of runs are possible. We now consider μ to be the probability of being dismissed at any particular score. In the general case, this could be a function of the score j already achieved, so that $\mu = \mu_j$. For the present, and in order to explore a closed form solution, we assume that μ is a constant, independent of score.

Let p_k be the probability of scoring k runs at each scoring stroke. In a similar way to μ, the p_k can be a function of the number of runs the batsman has already scored in the most general case, but initially are considered constant through a batsman's innings. Clearly, $\sum_{k=1}^{r_{\max}} p_k = 1$, where $r_{\max}$ is the maximum number of runs that can be scored off one ball. The record number of runs scored off a single ball since 2008 appears to be 8 [11], but in reality the probability of scoring > 6 runs off a ball is negligible. Similarly, the probability of scoring a 5 is negligible, so in practical terms, k can be 1, 2, 3, 4 or 6. The expected value of runs per score, κ, which [5] refers to as the strike constant, is

$$\kappa = \mathbb{E}(r) = \sum_{k=1}^{r_{\max}} k p_k \tag{2}$$

We can construct a survival function by recursively relating S_j to $S_{j-1}, \ldots, S_{j-r_{max}}$, subject to a suitable initialisation of the recursion. In particular, for $j \geq r_{\max}$

$$S_j = \left[\sum_{k=1}^{r_{\max}} \Pr(\text{survived} > j-k \text{ runs}) \times \Pr(\text{scored } k \text{ runs})\right] \times \Pr(\text{survived at } j) = (1-\mu)\sum_{i=1}^{r_{\max}} p_i S_{j-i} \tag{3}$$

The fact that there is more than one path from, say S_{j-4} to S_j, is taken into account through the recursion itself. This can be relatively easily seen by restricting the possibilities to only scoring a 1 or a 2 and will not be shown here.

Being a difference equation with constant coefficients, the recursion (3) leads to a geometric solution

$$S_j \propto \beta^j \tag{4}$$

$$f(\beta) = \beta^{r_{\max}} - (1-\mu)\sum_{i=1}^{r_{\max}} p_i \beta^{r_{\max}-i} = 0 \tag{5}$$

The polynomial (5) has $r_{\max}$ roots and the solution to the recurrence (3) is a linear combination of functions of the form (4), with constants given by the initial conditions. From the nature of the coefficients of $f(\beta)$, it can be shown that there is only one positive real root and that it is ≤ 1. It can also be shown that the other roots (one negative and two complex) are of smaller magnitude than the positive root. Hence the series will converge to the solution given by the positive root. This is a theoretical confirmation that the system will converge to a geometric solution.

Often in survival analysis $S_0 = 1$, but this is obviously not true in cricket as zero scores (ducks) occur with non-negligible regularity. More importantly, it has been well articulated over many studies [1, 2, 6] that any geometric analysis underestimates the number of ducks incurred, and that in general $S_0 < 1 - \mu$, the value expected from a GD. Hence, to improve the fidelity of the analysis we explicitly include S_0 as a normalising parameter to 'anchor' the distribution. An estimate of the value of this quantity will be directly derived from the data for individual batsmen.

We now consider the initialisation of the recurrence. For simplicity of exposition, and because 6s are relatively rare (see Section 3), we will restrict ourselves to $r_{\max} = 4$. However, the numerical results in the subsequent sections will employ the full solution with 6s included. The probability of surviving beyond a score of 1 (S_1) is then given by

$$S_1 = S_0(1-\mu p_1) = S_0\big((1-\mu)p_1 + p_2 + p_3 + p_4\big) = S_0\big((1-\mu)p_1 + (1-p_1)\big) \tag{6}$$

The term in brackets is equivalent to the probability that either the batsman scored a 1 then didn't get out at 1, or that the batsman scored a 2, 3 or 4. By similar reasoning

$$S_2 = S_1 p_1(1-\mu) + S_0 p_2(1-\mu) + S_0(1-p_2-p_1) \tag{7}$$

or, in words

$$\begin{aligned} S_2 = {} & \Pr(\text{survived past 1, scored a 1, and didn't get out}) \\ & + \Pr(\text{scored a 2 from 0 and didn't get out}) \\ & + \Pr(\text{scored a 3 or 4 from 0}) \end{aligned}$$

It can be shown that

$$S_2 = 1 - \left[1 - S_1 + S_0\mu\left(p_2 + p_1^2(1-\mu)\right)\right] \tag{8}$$

or

$$\begin{aligned} S_2 = 1 - [&\Pr(\text{survived to 1 and then got out}) \\ &+ \Pr(\text{scored a 2 from 0 and then got out}) \\ &+ \Pr(\text{scored a 1, survived, scored a 1, then got out})] \end{aligned}$$

Continuing in this manner gives:

$$S_3 = (1-\mu)p_1S_2 + (1-\mu)p_2S_1 + (1-\mu)p_3S_0 + p_4S_0 \tag{9}$$

$$S_4 = (1-\mu)p_1S_3 + (1-\mu)p_2S_2 + (1-\mu)p_3S_1 + (1-\mu)p_4S_0 \tag{10}$$

$$\frac{S_4}{S_0} = 1 - \left[1 - \frac{S_3}{S_0} + \mu(p_4 + 2p_1p_3(1-\mu) + 3p_1^2p_2(1-\mu)^2 + p_1^4(1-\mu)^3)\right]. \tag{11}$$

Thus the final recursion is given by (3) with initial conditions given by (6), (7), (9) and (10). The survival function given by the expression (4) is a memoryless geometric one with constant hazard function, and logarithmic slope, given by

$$h_j = 1 - \frac{S_j}{S_{j-1}} = 1 - \beta \tag{12}$$

This memoryless property would be achieved only for larger values of j where the effects of the initial conditions have decayed. A fully memoryless survival function is only possible with $p_1 = 1$, and all other $p_i = 0$. In a more general case, the recursion is initialised at non geometrically related points $S_0, \ldots, S_4$, and it will take a number of terms to converge to the geometric relationship driven by the positive root of $f(\beta)$. Figure 1 shows the ratio of successive terms for various values of the parameters, p_i and μ. For typical values dominated by the probability of getting a single and low probability of dismissal, the convergence to the analytical geometric solution of (4), $\beta = 0.984$, takes only about 10 runs and the deviation from that value is small even below that. Even in an unrealistically extreme case, where 80% of runs are scored in boundaries and there is a 20% chance of being dismissed at each run, the recurrence still converges ($\beta = 0.937$) by about 50 runs.

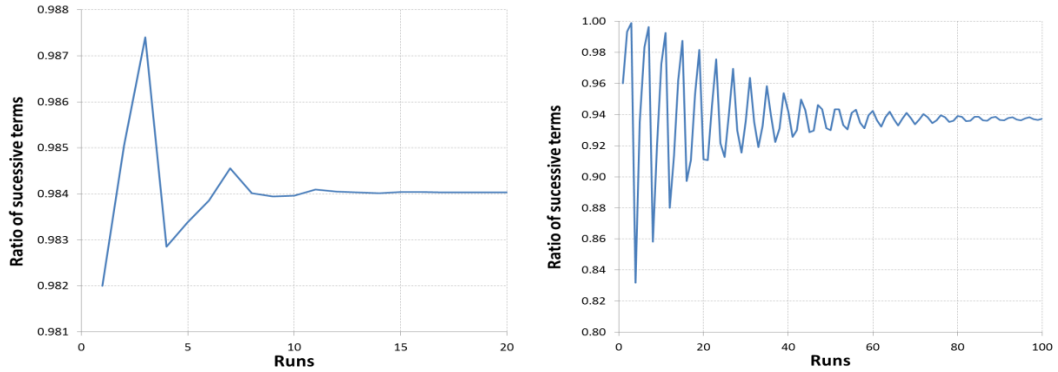


Figure 1: Ratio of successive terms in the recurrence relation (3). Left: $p_1 = 0.60$, $p_2 = 0.14$, $p_3 = 0.04$, $p_4 = 0.22$ and $\mu = 0.03$. Right: $p_1 = 0.2$, $p_2 = 0$, $p_3 = 0$, $p_4 = 0.8$ and $\mu = 0.2$

Equation (5) is equivalent to considering the geometric factor β as a function of the probability of the innings ending μ. In general μ is small, and hence β is close to 1, so it is possible to get an interesting approximate relationship between the two. Specifically, if we let $\beta = 1 - \varepsilon$ and assume $\varepsilon \ll 1$ then (4) and (5) gives:

$$f(\beta) \approx (1 - r_{\max}\epsilon) - (1-\mu)\left(1 - \sum_{i=1}^{r_{\max}-1} p_i\varepsilon(r_{\max} - i)\right) + O(\varepsilon^2) = 0 \tag{13}$$

whence

$$\frac{\mu}{1-\beta} \approx r_{\max} - \sum_{k=1}^{r_{\max}-1} p_k(r_{\max} - k) = \sum_{k=1}^{r_{\max}} kp_k = \kappa + O(\varepsilon) \tag{14}$$

$$\mu \approx \kappa\,\varepsilon \tag{15}$$

where κ is the expected value of runs per score as in (2).

3. DISTRIBUTION OF SCORING SHOTS

As noted previously, although the recurrence relation (3) does not depend on constant values for p_i, the formal analytical solution to that recurrence based on (4) and (5) does. It is therefore worth exploring the validity of those assumptions. Reference [11] provides a source of data available to the authors. This contains 365 men's Test matches from 02 January 2008 (Australia vs. India) to 30 September 2016 (India vs. New Zealand). In those tests, a total of 726,751 balls and 185,056 scoring shots were recorded. Figure 2 plots the percentage of instances a 1, 2, 3, or 4 is scored off each ball faced in all the innings in that database, given a run has been scored off that ball. The raw data is plotted as the dotted curves. This becomes noisy as the number of balls increase, corresponding to a decrease in the number of innings which last that long. A 20-ball moving average is plotted as a solid line to highlight the trends. Only 1.5% of all scoring shots were 6s and these have not been plotted for the sake of clarity. A negligible number of 7s (3 in total) and 5s (108) were scored during this period.

Figure 2 also plots the average number of runs scored off each ball faced and the average number of runs scored off each ball faced if a run is scored, and indicates that the assumption, used by [5] of a constant run rate (expected number of runs per ball) is a reasonable one in the ensemble, despite variations in the details of how those runs are scored.

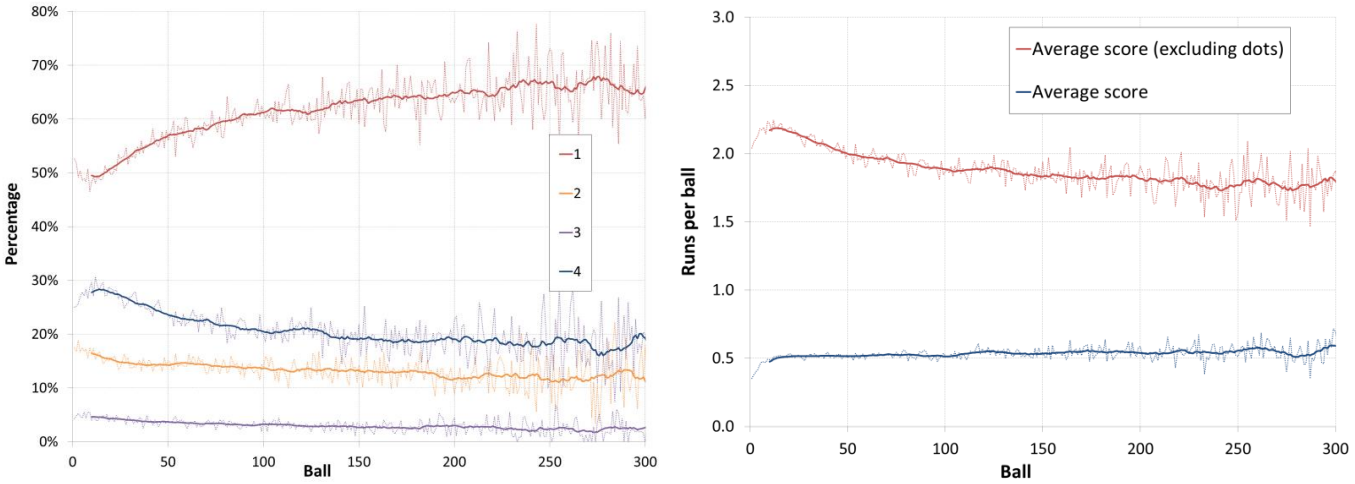


Figure 2: Scoring descriptors per ball. Left: The percentage of instances a 1, 2, 3, or 4 is scored off each ball faced, given that a run has been scored off that ball. The top (red) line shows singles (1s), the next (blue) show 4s, then (orange) 2s and the bottom line (purple) 3s. Right: The average number of runs scored off each ball faced (bottom line) and the average number of runs scored off each ball faced if a run is scored (top line). The darker lines are a moving average of 20 balls.

Figure 2 shows that, unsurprisingly, singles are the most common scoring shot, followed by 4s. Figure 2 also shows that there is some variation in the probability of scoring each type throughout an innings; in particular the proportion of scoring shots that are singles tends to increase with the number of balls faced, and correspondingly the proportion of 4s tends to decrease. Nevertheless, the smoothed values are reasonably stable, especially at > 50-100 balls (equivalent to > 25-50 runs).

In many cricket databases, only the number of boundaries (4s and 6s) and ducks are recorded for individual batsmen. In order to obtain a scoring profile, it is possible to make use of the assumptions about the rate of scoring 2s and 3s, based on the data in Figure 2, to infer the important quantity of fraction of singles. In particular, we know the total number of runs scored R, and the number of 4s and 6s, across a batsman's career. We want to find the total number of scoring shots $S = n_1 + n_2 + \cdots + n_6$ where n_1 refers to the number of 1s scored etc, given estimates of the fraction of scoring shots that are 2s (3s), denoted by a_2 (a_3). Noting that: $a_2 = n_2/S$ and $a_3 = n_3/S$ gives

$$\hat{S} = \frac{R - 3n_4 - 5n_6}{1 + 2a_3 + a_2} \tag{16}$$

where the $\hat{}$ is used to emphasise that this is an estimate of the true value of S.

4. ESTIMATION OF DISMISSAL RATE FROM AGGREGATED SUMMARY DATA

In order to derive the theoretical distribution for a batsman, we need to estimate μ, the probability of being dismissed at each score, from aggregated summary data of a batsman. It is natural to estimate the dismissals per score as

$$\hat{\mu} = \frac{S_0 N}{\hat{S}} \tag{17}$$

where $\hat{S}$ = (estimated) number of scoring shots, N is the number of dismissed innings. We have taken into account the ducks by S_0 = (1 - fraction of ducks), so $\hat{\mu}$ is an estimate of the dismissals per score assuming you don't get a duck, which is important for estimating the slope of the survival curve beyond 0. If only 1s are scored, $R = \hat{S}$. In a more general case we can use the approximation $\mu \approx \kappa\,\varepsilon$ from (15) to find

$$1 - \beta \approx \frac{S_0 N}{\kappa \hat{S}} \approx \frac{S_0 N}{R} \tag{18}$$

under the assumption of constant probability of scoring. Hence, a good first approximation for β can be derived by simply dividing the completed innings of > 0 runs by the total number of runs scored. The β resulting from the rougher approximation (18) approaches the more exact formulation (3) and (5) where dismissal rates are low and the fraction of runs scored in singles is high.

5. EXAMPLES

Table 1 shows example results for a number of batsmen. We have picked a selection of high quality batsmen who have a relatively large number of innings, with the exception of Don Bradman whom we have included as we feel no cricket analysis is complete without him. For the sake of brevity, we will only consider two in detail. As a first example, we consider the Australian batsman Steve Waugh, as he has been extensively studied elsewhere [10]. His career statistics are shown in Table 1 as derived from Cricinfo [12]. His probability of scoring 2 or 3 is unknown and, using Figure 2, is assumed to be 0.14 and 0.04 respectively, from which we estimate the number of scoring shots from (16) as $\hat{S}$ = 5985; hence $p_1 = 0.62$ and $p_4 = 0.20$. From (17) we can also estimate $\hat{\mu} = 0.0327$; that is, he has a probability of being dismissed at each score above zero of about 3.27%. Table 1 shows the PLE survival curve derived from Steve Waugh's actual scores [10] and compares with the survival function using the difference equation (3) and the estimated constants. The chart is plotted on a logarithmic axis to highlight the quasi-log-linear nature of the theoretical distribution.

Name	Innings	Runs	Not outs	Boundaries (4s/6s)	Ducks	Batting Average	p_1	p_4	p_6 (×10)	κ	$\hat{S}$	$\hat{\mu}$ (×100)	*1-β* (×100)
Don Bradman	80	6996	10	681/6	7	99.9	0.65	0.17	0.01	1.73	4035	1.58	0.92
Steve Waugh	260	10927	46	1175 / 20	22	51.1	0.62	0.20	0.03	1.83	5985	3.27	1.80
Gary Sobers	160	8032	21	723 / 31	12	57.8	0.66	0.15	0.07	1.72	4679	2.75	1.61
Adam Gilchrist	137	5570	20	677/100	16	47.6	0.51	0.27	0.40	2.24	2491	4.15	1.87
Jacques Kallis	280	13298	40	1488/97	16	55.4	0.59	0.22	0.14	1.94	6843	3.31	1.71
Mark Waugh	209	8029	17	844/41	20	41.8	0.62	0.19	0.09	1.85	4338	4.00	2.17

Table 1: Summary statistics and calculated quantities for various batsmen. In each case $p_2 = 0.14$ and $p_3 = 0.04$

Solving (5) numerically leads to a geometric slope $1 - \beta$ of 0.0180. Fitting a simple line of best fit to his PLE log curve gives a slope of 0.0146, an approximate 18% difference. A slightly better fit to the slope of 0.0158 is obtained by forcing the 0 run intercept. It is worth noting here that [10] obtained better fits to linearity by considering <100 runs and >100 runs as separate scoring regimes. With these estimates, the expected score per scoring shot for Steve Waugh is $\kappa = 1.83$, (compared to the global estimate used in [5] of 2.16). The estimate for $1 - \beta$ using just the number of runs rather than scoring shots is 0.0179, in this case very close to the full result. Figure 3 is similar to the figure given in [10] except in this case the PLE is compared to an estimate of the underlying theoretical one rather than a line of best fit (LOBF). While proving a good fit to the PLE, it also reinforces the point made in [10] that the whole PLE curve can provide insight into the batting career of a batsman by comparison to the theoretical curve, in this case for example by highlighting the relatively large number of not outs between 100 and 150.

Mark Waugh is an example of a batsman with a significant career but a lower batting average (see Table 1 and Figure 3). Again using $p_2 = 0.14$ and $p_3 = 0.04$, we obtain $\hat{S} = 4.338$ and $\hat{\mu} = 0.040$, leading to a geometric slope $1 - \beta$ of 0.0217. In this case, the LOBF slope is 0.0269, and importantly has a y-intercept > 1. This plot highlights that any memoryless approximation will overestimate the performance at high scores due to its inherent unboundedness. For Mark Waugh, the

estimate for $1-\beta$ using just the number of runs rather than scoring shots is 0.0216, again very close to the full result. In fact for all the players in Table 1, only Adam Gilchrist, who has a relatively large proportion of 4s and 6s, had a >1% difference between the simplified approximation of (18) and the more complete formulation.

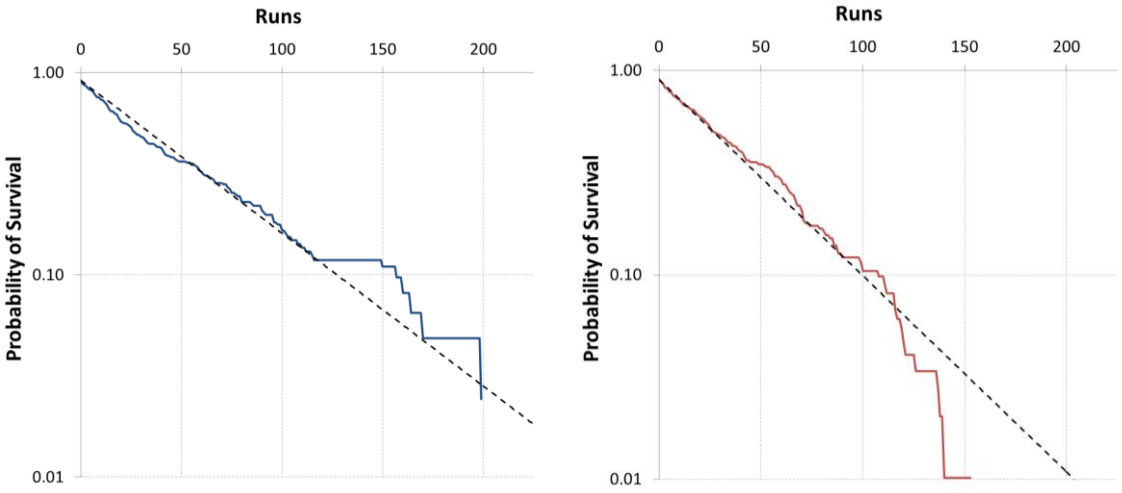


Figure 3: Left: PLE survival function of Steve Waugh (solid line) compared to his theoretical survival function given by the recurrence relation (dotted line). Right: As for the left but for Mark Waugh. Note, for both, the dotted line is not the LOBF.

6. CONCLUSIONS

This paper has derived a recurrence relation for a batsmen's survival function using the probability of being dismissed at each score. The formulation explicitly takes into account the probability of making different scores at each point. The formulation is general enough to cater for these probabilities varying with the number of runs a batsman has already scored as well as the probability of the innings completing varying (that is, being dismissed). The exact solution for the case where these probabilities are constant is derived and show that the survival curve converges to a memoryless curve very quickly. An estimate of those underlying assumptions was derived from a batsman's aggregated data summaries available from most common cricket databases. In particular, the only data needed beyond that to derive their batting average is the number of ducks and number of boundaries. An even simpler estimate based on number of runs and dismissals is also shown to be likely to be a reasonable approximation in most instances. Finally, we compared the theoretical survival function with the PLE derived from a batsman's individual scores.

7. REFERENCES

1. Wood, G. H. (1945) Cricket scores and geometrical progression. *Journal of the Royal Statistical Society* **108** 12-22
2. Kimber, A. C. and Hansford, A. R. (1993) A statistical analysis of batting in cricket. *Journal of the Royal Statistical Society A* **156** 443-455
3. Barr, G. D. I. and van den Hopert, R. (1997) Evaluating batsmen's scores in test cricket. *Journal of South African Statistics* **32** 169-183
4. Das, S. (2016) On generalized geometric distributions and improved estimation of batting average in cricket. . *Communications in Statistics – Theory and Methods* **46** (6) 2736-2750
5. Cohen, G. L. (2002) Cricketing Chances. In: *Sixth Australian conference on mathematics and computers in sport,* University of Technology, Sydney, MathSport (ANZIAM)
6. Bracewell, P. J. and Ruggiero, K. (2009) A parametric control chart for monitoring individual batting performances in cricket. *Journal of Quantitative Analysis in Sport* **5** 1-19
7. Scarf, P., Akhtar, S. and Shi, X. (2011) On the distribution of runs scored and batting strategy in test cricket. *Journal of the Royal Statistical Society A* **174** 471-497
8. Klein, J. P. and Moeschberger, M. L. (2003) *Survival Analysis: Techniques for Censored and Truncated Data*. 2 ed
9. Danaher, J. D. (1989) Estimating a cricketer's batting average using the product limit estimator. *The New Zealand Statistician* **24** (1) 2-5
10. Kachoyan, B. and West, M. (2016) Cricket as life and death. In: *Australasian Conference on Mathematics and Computers in Sport,* Melbourne, ANZIAM MathSport
11. Cricsheet. http://cricsheet.org (2016) [Accessed 14/10/2016].
12. Cricinfo. *Steve Waugh:* http://www.espncricinfo.com/australia/content/player/8192.html. (2016) [Accessed 1/12/2016].

ESTIMATING EXPECTED TOTAL IN THE FIRST INNINGS OF T20 CRICKET USING GRADIENT BOOSTED LEARNING

Ankit K. Patel [a,b,c], Paul J. Bracewell [a,b] and Michael G. Bracewell [b]
[a] *DOT Loves Data, Wellington*
[b] *Victoria University, Wellington*
[c] *Corresponding author: ankit@dotlovesdata.com*

Abstract

Run prediction systems currently utilised within limited overs cricket suffer from two model issues: overly broad match representation metrics and inability to account for contextual match factors. Here, a gradient boosted model (GBM) is developed to account for these two issues. The model outputs are benchmarked against a popular media tool, dynamic programming model (DPM), and actual first innings runs scored. The results show that the developed model converged to actual first innings total faster than the DPM. Importantly, the GBM model outperformed DPM across several statistical accuracy metrics. The proposed model utilises match-level metrics, such as resources remaining, and innings specific metrics that relate to both the batting and bowling teams, such as percentage dot balls and percentage boundaries. Conceptually, these attributes begin to account for environmental and team specific factors. The improvement in accuracy whilst maintaining a simplicity of deployment suggests that maintaining contextual information within an estimated runs model is appropriate for limited overs cricket.

Keywords: Dynamic programming, sports modelling, dimension reduction

1. INTRODUCTION

T20 cricket is a dynamic and fast paced game where the team's prospects of winning can change within a few balls. This allows players to significantly influence match result off fewer deliveries relative to longer formats. Consequently, each ball carries more weight as it represents a greater proportion of the match. However, this introduces a greater level of uncertainty when predicting results as only a small number of balls are necessary to the change match situation. An area of considerable uncertainty is the number of runs the batting team is expected to score in the first innings. It is hypothesised that ball-by-ball predictions of the first innings total can be improved by using match level metrics, such as resources remaining, or shallow metrics such as current total, with team inning metrics, such as percentage dots to produce better first inning run predictions than a model that only considers shallow metrics. The rationale is that some of these within game descriptive actions will encapsulate information about the playing conditions.

Refining estimates using data that is descriptive of actions within the innings is useful for applications in coaching, strategy and entertainment. However, they are not suited to adjusting totals for defining the formal outcome of a match, where the Duckworth-Lewis-Stern is used (Stern, 2016). The now defunct Indian Cricket League used the VJD method, developed by Jayadevan (2002). The outputs from targeting setting and readjusting models are subject to scrutiny and can have a bearing on match and tournament outcomes, thus tremendous rigour must be applied to ensure fair results.

Consequently, forecasting totals is an area that has received considerable attention. Notable research in run prediction in limited overs cricket include: Duckworth and Lewis (1998), Stern, (2016), Jayadevan (2002), Ovens and O'Riley (2006), Brooker & Hogan (2012), Clarke (2000), Scarf, Akhtar & Shi (2010), Kaluarachchi & Varde (2010), Bailey & Clarke (2006), Bandulasiri (2004), Jhawar & Pudi (2016), Asif & McHale (2016), Davis, Perera & Swartz (2015) and Shah, Jha & Vyas (2016).

The Duckworth-Lewis-Stern (DLS) system is the most famous of this research (Duckworth and Lewis, 1998; Stern, 2016) with the primary function is to reset the target total during interrupted matches of limited overs cricket. Importantly, the DLS system can also be used to produce first innings run predictions for uninterrupted matches, with the output embedded in live scorecard publication tools and websites like crichq.com and nzc.nz. This elegant method, which is well entrenched in club, domestic and international cricket due to simplicity of use, is well described in both academic and popular literature (e.g. espncricinfo.com). The premise of the method is that batting teams have two resources to produce runs: balls and wickets. This two-factor relationship is then used to calculate the average number of runs that can be scored given the remaining resources.

Clarke (1988) applied a dynamic programming model to one-day cricket to: 1) calculate the optimal scoring rate, 2) estimate the total number of runs to be scored in the first innings and 3) estimate the probability of winning in the second winnings. The first innings formulation generated a team's optimal scoring rate to obtain a given total, given the number of wickets lost and balls. The second innings formulation generated a probability

scoring table outlining the probability of the second innings batting team achieving the target total, given the number of wickets lost and balls remaining. Ovens and O'Riley (2006) evaluated the ball-by-ball run prediction ability of four models: Average Run Rate, PARAB, Duckworth Lewis (D/L) and Jayadevan. Results showed that the D/L method had the strongest predictive power, predicting 4.50 runs below the actual total, followed by ARR with prediction 17.29 runs below the actual total, Jayadevan with 31.13 runs below the actual total and PARAB 41.60 runs below the actual total. Similarly, Brooker and Hogan (2012) utilised a dynamic programming model to develop a Winning And Score Prediction (WASP) system for limited overs cricket. The system produces predictions using factors such as pitch conditions, weather, boundary size and the quality of the batting team and bowling attack. The WASP works backwards to solve inning specific models. The first innings model produces ball-by-ball prediction of the runs scored, while the second innings model calculates the probability of the batting team reaching the target total and therefore winning.

Swartz, Gill and Muthukumarana (2009) developed a discrete generator simulator, as there is finite no. of outcomes that can occur for any given delivery, for one-day cricket. Applying a Bayesian Latent model, ball-by-ball outcome probabilities were estimated using historical ODI data and were dependent on batter, bowler, total wickets lost, total balls bowled and current match score. It was found that the proposed simulator produced reasonably realistic results, with actual runs and simulated runs revealed an excellent agreement. Ovens & Bukiet (2006) developed a Markov chain approach to predict the runs scored for a given batting line-up. Realising that the interaction between bowler and batter is the primary factor dictating the dynamics of run production, a match was modelled as a sequence of one-on-one interactions, through a multi-dimensional matrix, M, with entries (b, r, w, b_1, b_2) representing the number of balls, runs scored, wickets lost, and the striking and non-string batter, respectively. The probability of being in any given state was calculated, for any given number of balls, by multiplying M, representing the set of probabilities after $b - 1$ balls, by the probability of each event (i.e. number of runs scored off any given ball). Simulation resulted in a runs distribution table and "summing the product of each possible number of runs and its probability of being the result for the match gives the expected number of runs for the batting order considered" (Ovens *et al*., 2009, pg. 497).

Jayadevan (2004) developed a method for resetting the target total during an interrupted limited overs cricket match. A normal score represented a team's general scoring pattern, while the target score represented a team's ideal scoring pattern too achieve the target score. Regressing cumulative percentage runs on cumulative percentage overs it was found that a cubic polynomial equation of order 1 represented a team's scoring pattern (a similar approach was adopted by Mansell, Patel, McIvor and Bracewell, 2018). Moreover, the effect of a wicket was incorporated into the model by examining the pattern of wickets fallen. Applying the model produces a "target runs" percentage table that allocates a proportion of runs that needed to be scored by the batting team during any stage of the second innings.

Swartz, Gill, Beaudoin & deSilva (2004) used simulated annealing to conduct a search over a space of permutation of batting orders to find the optimal or near optimal first innings order. A first innings run simulator was built using a Bayesian log-linear model to generate ball-by-ball outcomes. The model was applied to the 2003 India World cup squad and posterior estimates of the parameters were obtained by averaging output from a Markov chain. Simulating 71,000 first innings runs using India's 2003 World cup final batting order a good fit between actual runs and simulated runs was found. Overall it was found that the optimised batting order produced 6 more runs than that of the actual batting order.

Singh, Singla and Bhatia (2015) developed a first innings run prediction model and a second innings match outcome probability model for one-day cricket by applying linear regression and Naïve Bayes classifiers for each innings applied in 5 over intervals. The first innings model used current run rate and wickets fallen, while the second innings used current run rate, wickets fallen and target score. The error produced by the linear regression classifier were less than a current run rate projection method and the Naïve Bayes classifier had an accuracy of 68% in the 0-5th overs, increasing to 91% between the 40-45th over.

Bracewell *et. al*. (2014) generated team ratings where margin of victory was represented in terms of runs only. Like the approaches outlined previously, the resources available at the end of the second innings were used to determine a likely final total if the innings continued until all resources were consumed. This was used to generate team ratings that outperformed popular opinion for result prediction.

The hypothesis of combining conventional and advanced metrics builds on sabermetrics literature (sabr.org) stating that run prediction models that utilise both conventional and advanced metrics generate better predictions.

2. METHODS

Here, the intent is to show that a model utilising match level metrics (such as resources remaining, pitch conditions) and shallow metrics (such as current total and balls) with team inning metrics, (such as percentage dots and percentage boundaries) produce better first inning run predictions in T20 cricket than models that only

consider shallow metrics. It is anticipated that team specific metrics inherently include information relating to environmental, situational and competitive factors. For example, high percentage boundaries could indicate either poor bowling, good batting, favourable batting conditions or any combination of these factors. Due to the large variations between balls in T20 cricket, ball-by-ball metrics that capture this variation must be utilised to produce predictive outputs. The shallower and less informative the metrics (i.e. covariates) the less informative the model outputs. Therefore, advanced and more informative metrics must be adopted to significantly explain the underlying variation and produce accurate predictions. Current run prediction models do not consider the complex interactions existing between inning specific and match level metrics due to data accessibility, ease of implementation, ability to produce intuitive results that are understandable by players and coaches and ability to implement offline. Therefore, a modelling technique, such as ensembles, that consider these subtle nuisances and capture these interactions must be applied to produce accurate predictions.

It is assumed that models combining match level and team specific metrics produce better run predictions than models that only utilise match level or shallow metrics, such as wicket, current total and balls. This assumption has been extracted from baseball sabermetrics literature focusing on run prediction, which is a well-researched and documented problem within baseball. The literature states - models that utilise both conventional and advanced metrics generate better run predictions. For example, "A mixture of conventional independent variables and sabermetrics independent variables would be broad enough to find models to correlate highly with run production and run preventions" (Benevenatno, Berger & Weinberg, 2012, p. 67). For more readings on baseball run prediction please see Bukiet, Harold & Palacios (1997); Freeze (1974); Passan (2011); Cserepy, Ostrow & Weems (2013). Only first inning run prediction models are considered as there is less information available regarding what is expected to be a winning total (which is known in the second innings).

DATA

Model development utilised ball-by-ball observations from the following T20 competitions: Indian premier league (IPL; 2015, 2016, 2017 and 2018), Australian Big Bash League (BBL; 2016-2017 and 2017-2018), English NatWest T20 league (2015, 2016), South Africa Ram slam (2016 and 2017), Caribbean Premier league (CPL; 2014, 2015, 2016 and 2017) and New Zealand Super Smash League (2017-2018). Overall the dataset contained 85,700 1st inning ball-by-ball observations from 704 matches. Model development utilized 50% of the data for training, 25% for testing and 25% for validation.

ANALYSIS

Figure 1 shows that the underlying distribution for first innings total can be well approximated by a normal distribution. Given this finding a normal distribution will be adopted to during the modelling process. Exploratory analysis found the average first innings runs scored = 160, while the average first innings winning total = 168. Figure 2 illustrates the evolution of first innings total since 2014. There is an upward trend, with runs experiencing an average yearly increase of 4.75.

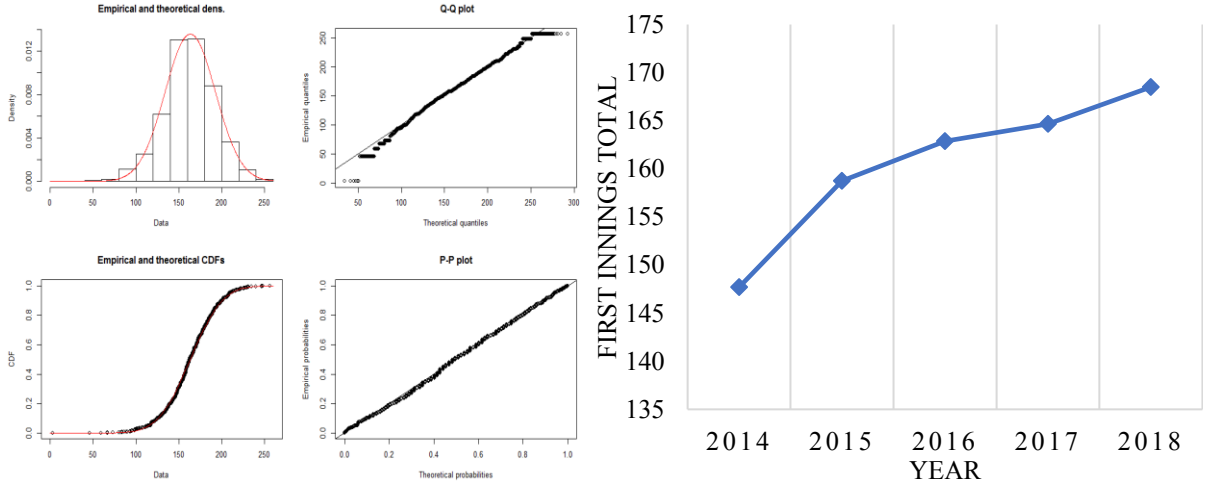


Figure 1 and 2: Distribution of first innings total and Avg. 1st innings total (2014 - 2018)

GENERALISED BOOSTED REGRESSION MODELING

The proposed model uses a gradient boosted regression technique (GBM) to account for the complex interactions between match and inning level metrics by taking a sequence of weak leaners to construct a complex leaner –

increasing model complexity. The initial learners fit simple model and then the weighted combinations can grow more and more complex as learners are added. This produces regression models consisting of a collection of regressors. Learners do so sequentially with earlier stages fitting simple models to the data and analysing the errors. Latter models focus on trying to account for as much error as possible. The models are given weightings and the different models are combined into an overall predictor. Moreover, the gradient boosted method serves as a dimension reduction technique to identify the relative importance of each performance metric, allowing the evaluation of metric importance and elimination of uninformative metrics.

The proposed model was built in *R* using the *gbm* package and incorporates the following parameters: 1) distribution = Poisson – runs scored is a count outcome, 2) n.trees = 20,000 - optimal number of tree for out-of-bag variance, 3) interaction.depth = 5 – 5-way interaction to capture complex variable relationships and 4) shrinkage = 0.0001 – step-size learning rate. The combination of weak-leaners, i.e. a complex learner, that incorporate a 5-way interaction effect will slowly start to reduce the error in first innings total. Ultimately, the new complex learner will account for greater variation and understand the complex interaction between match and inning-specific metrics. The metrics included in the model: projected total $(i.e. current\ total\ /\ resources\ remaing)$, team strike rate, run rate, current runs, wickets, percentage dots, percentage boundaries, resources remaining and balls. These metrics store match level, batting and bowling performance information. Team specific metrics included in the model are: percentage dots, percentage boundaries, strike rate and economy rate. These metrics inherently store information about the interactions between the bowling and batting environment. The gradient boosted model considers the interactions across these 'meta-information' rich metrics to gauge meta (i.e. match-level) understanding.

3. RESULTS

The GBM model was benchmarked against the Dynamic Programming Model (DPM) outlined in Clarke (1988), and the model predictions were evaluated against actual runs scored. The ball-by-ball predictions and were aggregated to an over. A relative importance analysis revealed projected total, team strike rate, percentage boundaries and run rate as the 3 most important metrics. These results show that in T20 cricket the first inning total is heavily dependent on efficient run production. Specifically, run production is dependent on the volume of runs scored per percentage of resources used. Moreover, the analysis reveals the metrics that are utilised by the dynamic model: i.e. current total, balls and wickets are contained within the top 3 important metrics: 1) balls influence team strike rate, 2) current total influences team strike rate and project total and 3) wickets influence resources remaining, which is also present in projected total. This indicates that although current total, balls and wickets are important to evaluate the expected total in the first innings they are more informative when combined with other metrics to explain a greater proportion of variation. Model performance was evaluated using two measures: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). RMSE has the benefit of penalising large errors, while MAE has interpretative power and only describes error. Figure 3 illustrates the

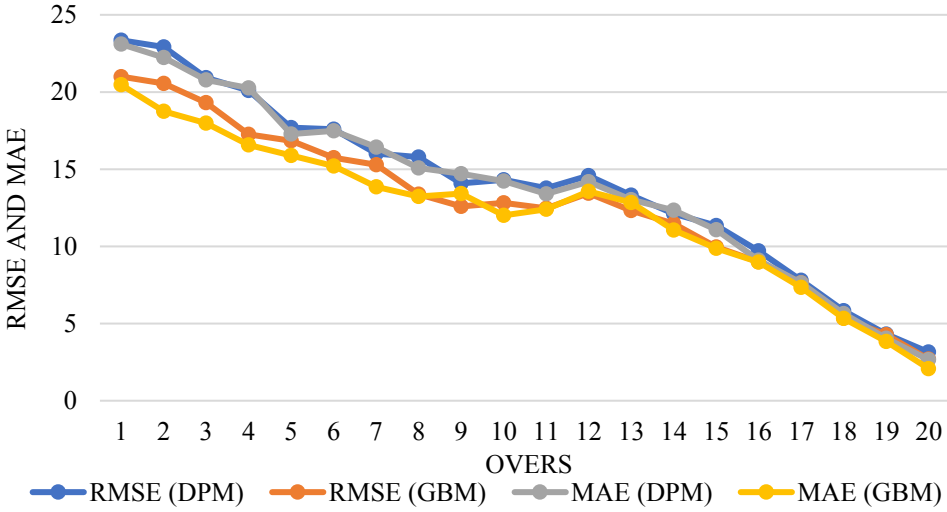


Figure 3: GBM vs. DPM performance measurements

over-by-over predictive accuracy of the two models, measured against RMSE and MAE. On average the proposed model out-performs the dynamic program on an over-by-over basis across both RMSE and MAE. A bootstrapped sample of the over-by-over performance measures created confidence intervals. A statistically significant difference was found between the performance metrics for the two models: GBM and DPM (α ==

5%). This statistically significant difference between the models for the performance measures existed up until the 13th over (~ 78 balls) suggesting that for 65% of the first innings of a T20 match the GBM model produced statistically better results than the DPM. However, examining the performance measures on a match-by-match basis revealed instances where the DPM produced better predictions. On average, the DPM produced better predictions for low scoring matches (i.e. ≤ 158). This could be because in low scoring matches metrics such as current total, balls and wickets metrics have a greater impact on expected total, while efficiency metrics (i.e. percentage dots and percentage boundaries) are of lesser importance. A 30% increase in predictive accuracy is observed between the 5th and 10th over. However, the prediction error during this period is large given that the data does not contain sufficient match information. Surprisingly, model accuracy experiences a decrease in the 12th over. It was found that between overs 6 -10 the batting team the run rate, percentage boundaries and innings strike rate are relatively constant. However, in overs 11-13 these metrics begin to experience a steady increase, indicating that the batters are starting to pick-up and increase aggression. It is assumed that the GBM and DPM fail to effectively account for this sudden increase in batting intensity. The 12th over is where the difference between RMSE and MAE becomes statistically insignificant, indicating that after the 12th-13th over enough match information is known, therefore both models are producing similar results and both models are extracting similar information from the metrics relating to the first innings expected total.

HYBRID MODEL: GBM using DPM

Given the dynamic model generates predictive results and produces better prediction for low scoring first innings, the GBM model was updated using the DPM predictions as an input metric. This hybrid model did not produce prediction improvements as the metrics that are present in the DPM are already present in the GBM. As stated current total, balls and wickets are included in the GBM model in a meaningful manner, such that more information regarding match-state is incorporated. Therefore, the DPM metric is not introducing any new information into the proposed model and introducing confounding issues. An importance analysis revealed DPM metric was the most important metric. This is expected as the DPM combines three conventional metrics in a meaningful way to produce a more informative metric. This suggests that combining weak predictors, in a meaningful manner, creates a stronger predictor that explains a greater proportion of variation.

4. DISCUSSION AND CONCLUSIONS

Although the proposed model produced better results, there are scenarios where the dynamic programming model produced better predictions. DPM produced better outputs in low scoring matches (i.e. ≤ 163 runs) where the batting team had a 'slow' start. This scenario arises because the proposed model considers metrics that are relatively more important in high scoring matches, such as percentage dots and percentage boundaries and therefore is more sensitive to actions that significantly affected or deviate the slope of the expected total. Figure 2 illustrates the evolution of first innings total since 2014. Overall there is an upward trend across time, although this seems to flatten out. However, recently (2017-2018) there has been a small increase in gradient. Assuming the trend continues it is assumed that the proposed model will continue to outperform the DPM as scores will continue to rise, meaning the latter model will continue to produce more varied predictions over time as it fails to accommodate for highly sensitive metrics that significantly affect expected total and matches where more than 163 are scored in the first innings. The results confirmed the hypothesis that a model that utilises both meta (i.e. match-level) and shallow metrics, and advanced metrics will produce better predictions than a model that only utilise meta and shallow metrics. It reveals that advanced metrics store additional information and combining shallow metrics in a meaningful way creates features that explain additional variation from a given dimension in the data. Moreover, combining weak predictors in a meaningful creates a stronger, more complex predictor that explain a greater proportion of variation than its individual counterparts.

Current models do not consider complex interactions existing between innings specific and match level metrics. Therefore, a modelling technique that incorporates these subtle nuisances and interactions is expected to produce more accurate predictions by capturing these interactions. The literature relies heavily on meta-level metrics such as pitch conditions, boundary size, and shallow team metrics such as batting and bowling characteristics that fail to incorporate inning dynamics and capture the interaction affects existing between players meta and team metrics. This novel method attempts to address these issues to dynamically predict the first innings total. The hypothesis that first innings predictions could be improved by using match-level, team and inning specific data was found to hold true in the first 12 overs over an innings.

Future research will benchmark the GBM system against the WASP model (Brooker *et. al.,* 2011, Shah *et. al.,* 2016*).* Although the WASP utilises team specific metrics such as the average team score, opposition's bowling performance, ground average score, these shallow metrics fail to capture match and inning information at a deeper level. Although metrics such as climate, pitch conditions and boundaries are important when

predicting runs, this meta (i.e. match-level) information can be captured and stored in team and inning-specific metrics; for example, a high percentage of dots and quick depletion of resources indicating strong bowling attack, weak batting performance and/ or poor batting conditions.

Given that the DPM falls-over for high scoring matches (i.e. ≤ 158) and the first innings total is experiencing a 4.75 runs increase year-on-year, it is suggested that future research benchmark the two models year-on-year and observe the period in which the DPM outperforms the GBM. It is assumed that earlier seasons (2014 and 2015) the DPM would outperform GBM due to the low scoring first innings. In addition, reviewing tournament specific model performance will also provide greater insight into the applicability of various models.

More accurate estimations of a first innings total provide interested parties with useful information for both strategic and entertainment purposes highlighting the value of deploying the GBM model in real-time.

References

Asif, M., & McHale, I. G. (2016). In-play forecasting of win probability in one-day international cricket: A dynamic logistic regression model. *International Journal of Forecasting*, *32(1)*, 34-43.

Bailey, M., & Clarke, S. R. (2006). Predicting the match outcome in one day international cricket matches, while the game is in progress. *Journal of sports science & medicine*, *5(4)*, 480.

Beneventano, P., Berger, P. D., & Weinberg, B. D. (2012). Predicting run production and run prevention in baseball: the impact of Sabermetrics. *Int J Bus Humanit Technol*, *2*(4), 67-75.

Bracewell, P. J., Downs, M. C. F, & Sewell, J. W. (2014). The development of a performance-based rating system for limited overs cricket. Brisbane, Australia: ANZIAM Mathsport. 40-47

Brooker, S., & Hogan, S. (2011). A Method for Inferring Batting Conditions in ODI Cricket from Historical Data.

Clark, J. T. (2016). Regression Analysis of Success in Major League Baseball.

Clarke, S. R. (1988). Dynamic programming in one-day cricket-optimal scoring rates. *Journal of the Operational Research Society*, *39*(4), 331-337.

Cserepy, N., Ostrow, R., & Weems, B. (2015) Predicting the Final Score of Major League Baseball Games. (avaialbe at: http://cs229.stanford.edu/proj2015/113_report.pdf. Accessed 1 May 2018).

Shah, A., Jha, D., & Vyas, J. (2016). Winning and Score Predictor (WASP) Tool. *International Journal of Innovative Research in Science and Engineering.*

Davis, J., Perera, H., & Swartz, T. B. (2015). A simulator for Twenty20 cricket. *Australian & New Zealand Journal of Statistics*, *57*(1), 55-71.

Duckworth, F. C., & Lewis, A. J. (1998). A fair method for retting the target in interrupted one-day matches. *Journal of the Operation Research Society, 49*(3), 220-227.

Jayadevan, V. (2004). An improved system for the computation of target scores in interrupted limited over cricket matches adding variations in scoring range as another parameter. *Current Science*, *86*(4), 515-517.

Jayadevan, V. (2002). A new method for the computation of target scores in interrupted, limited-over cricket matches. *Current Science*, *83*(5), 577-586.

Jhanwar, M. G., & Pudi, V. (2016, September). Predicting the Outcome of ODI Cricket Matches: A Team Composition Based Approach. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2016 2016)*.

Kaluarachchi, A., & Aparna, S. V. (2010, December). CricAI: A classification-based tool to predict the outcome in ODI cricket. In *Information and Automation for Sustainability (ICIAFs), 2010 5th International Conference on* (pp. 250-255). IEEE.

Mansell, Z., Patel, A. K., McIvor, J. M., & Bracewell, P. J. (2018, July 25-28). Managing Run Rate in T20 Cricket to Maximise the Probability of Victory when Setting A Total. Paper presented at *The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports*. University of the Sunshine Coast, Queensland, Australia: ANZIAM MathSport.

O'Riley, B. J., & Ovens, M. (2006). Impress Your Friends and Predict the Final Score: An analysis of the psychic ability of four target resetting methods used in One-Day International Cricket. *Journal of sports science & medicine*, *5*(4), 488.

Perera, H. P., & Swartz, T. B. (2012). Resource estimation in T20 cricket. *IMA Journal of Management Mathematics*, *24*(3), 337-347.

Singh, T., Singla, V., & Bhatia, P. (2015, October). Score and winning prediction in cricket through data mining. In *Soft Computing Techniques and Implementations (ICSCTI), 2015 Int.Conference* (pp. 60-66). IEEE.

Stern, S. E. (2016). The Duckworth-Lewis-Stern method: extending the Duckworth-Lewis methodology to deal with modern scoring rates. *Journal of the Operational Research Society*, 67(12), 1469-1480.

Data of DOTA: A preliminary analysis into factors predicting success in DOTA2

Niamh McDonald, Dave Matteo, and Minh Huynh

s3715024@student.rmit.edu.au, 101177497@student.swin.edu.au, mhuynh@swin.edu.au

***Abstract* –** DOTA 2 is a multi-player online battle arena game where two teams of five compete to destroy the other team's base. There are a number of player roles and other variables that can influence success. This research explored the relationship between the variables and the likelihood of team success. Data collected from the 2012 - 2016 DOTA 2 International provided the variables to run a Binary Logistic Regression Model. This was used to reduce the number of variables down to three significant variables, Kills, Deaths, and Gold per Minute. These were used to create a formula for predicting the likelihood of a win. This formula had an accuracy of 95.7% based on the data from the 2017 DOTA 2 International. This research is an important insight for predicting team success and provides a base for future research.

Index terms – eSports, DOTA 2, Binary Logistic Regression

BACKGROUND

Skubida (2016) identifies eSports as "a fusion of sport, media, gaming, and technology". Wagner, (2006) defines eSports at its simplest level as competitive computer gaming. Wagner (2006) identifies the spread of eSports as beginning in the early 90s across Europe and America with the wide acceptance of first person shooting games. It quickly spread to Korea in the mid-90s which eagerly accepted the world of eSports and lead to the development of games which are still common today including StarCraft (Wagner, 2006). Since this time, eSports competitions have grown with broadcasts live across the world, sponsorships, and millions of dollars prize money. In 2017, it was estimated that eSports would have an online audience of approximately 400 million spectators (Global eSports, 2017), with the League of Legends (LoL) World Finals gathering over 36 million unique viewers streaming the event live (Walker, 2016) *More people watched League Of Legends than the NBA finals. Retrieved from: https://www.kotaku.com.au/2016/06/ more -people-watched-league-of-legends-than-the-nba-finals/-*. By comparison, the 2016 National Basketball Association (NBA) finals caught the attention of only 31 million people, with only 1.756 million streaming the event online. This rapid growth of the world of eSports, has allowed for the application of traditional sporting statistics a digital sporting platform.

One such example is Defence of the Ancients 2, more commonly known as DOTA2, is a Multiplayer Online Battle Arena (MOBA). In this gaming format, two teams consisting of 5 players compete in real time to destroy the opposition's central base. In DOTA 2, the players compete on a Map which is geographically balanced for both teams. The map is split into two halves with each team controlling a side which contains their central base (Ancients). The Ancients are surrounded by strategically placed Towers which, every 30 seconds, generate lane creeps which are computer-controlled defenders which fight for each team. Lane creeps are programmed to engage with 'hostile units' which are opposing team lane creeps, and towers.

Each member of the team selects and controls a Hero each of which perform a specific role. There are 113 heroes which players can select. However, there can be no duplicates as once a hero is selected it is no longer available to members of either team. Heroes have specific attributes of Strength, Agility, or Intelligence. These attributes affect which roles a player will serve on the team. A Strength type hero typically performs the "tank" role within a match, generally absorbing damage from enemies to protect their team. By comparison, Agility type heroes are typically responsible for "dealing damage" to the other team and destroying buildings and structures. Finally, Intelligence type heroes are best suited for "supporting" the other roles and disabling the opposing team's heroes in combat.

At the beginning of each match players start off at level one. They attempt to get to max level (25) as quickly as possible. The increase in levels is associated with improved abilities and access to other weapons and skills which in turn increases the players likelihood of killing opposing team members or causing damage to the ancient. Using Gold and experience points (XP) players can "buy" any of 150 items which also contribute to game success.

DOTA 2 is relatively procedural based on: (1) having a team made up of heroes with

complementary skills, (2) earning Gold and XP quickly, (3) levelling-up and purchase increasingly powerful items, and (4) staying alive and kill as many heroes as possible. The table below illustrates the ways in which players are rewarded with gold and XP.

TABLE I
ACTS WHICH REWARD PLAYERS WITH GOLD OR XP

Act	Result
Kills	Number of times player slayed opposing teams heroes.
Deaths	Number of times hero was killed.
Assists	Helping a team mate kill a member of the opposing team.
Last Hits	Being the last player to deal damage to another player before they die.
Denial	Killing a member of same team to prevent opposing team from getting the Last Hit bonus.

Despite the relatively straight forward nature of the game, the various combinations of heroes($2.25x10^{20}$) and items($1.61x10^{10}$) always results in no two games ever being the same, with a total number of possible hero and item combinations equalling 2.62 nonillion ($2.62x10^{30}$). Due to this evolving dynamic, fans, researchers and gamers alike, are continuously striving to seek the ideal combination of hero and item selection, in conjunction with gameplay styles, to optimise their in-game performance. It is perhaps the highly variable, yet fast-paced nature of the game, which has made it so popular. Indeed, this popularity in e-sports over the years has garnered the interest of more than just academic researchers, with professional punters joining in the action. In 2016, betting websites (such as Betfair and Bet365) reported that over 550 million dollars were spent on esports wages (NewZoo, 2017). Of this amount, close to 20% of the wages (99 million dollars) were from DOTA 2 alone

DATA COLLECTION AND STATISTICAL ANALYSIS

Previous research by Pobiedina, Neidhardt, Moreno and Werthner, 2013 suggests that player knowledge of the game has the greatest effect on the likely outcome of a match rather than levelling up, hero selection, or kill death ratio. To account for this variable, this preliminary study only looked at data gathered from the 2012 – 2017 DOTA 2 International which is the highest level of competition. Entry into the International is gained through extensive international competition and takes into account other competitions as well as preliminary competitions. As a result of this it is only professional teams that gain entry. As a result of this we have chosen to exclude hero selection and combat patterns as variables due to the level of familiarity professional player have.

This study aims to look at which variables are indicative of a win and create a preliminary model which can be applied to game data to predict game outcome.

Data was collected from the publicly accessible site Liquipedia (http://liquipedia.net/dota2/Main_Page), with match outcomes for all rounds of the DOTA 2 *International* between 2012 and 2016 being included for the development model. The final model was tested on the 2017 data to assess the effectiveness of predictions.

The data was analysed with a Binominal Logistic Regression on R, due to the dichotomous nature of the outcome variable: win/loss. It was initially run analysing, number of kills, deaths, assists, last hit, gold per minute (GPM), experience per minute (XPM), and denials. Variables were excluded from the model until only significantly related variables remained.

RESULTS

TABLE II
FACTORS ASSOCIATED WITH MATCH OUTCOMES IN DOTA 2

Variables	β	SE	Wald	*p*	OR (95% CI)
Kills	.042	.067	0.395	.529	1.043 (.915 – 1.188)
Deaths	-.198	.028	48.498	< .001	0.821 (.776 - .868)
Assists	.040	.025	2.490	0.115	1.041 (.990 – 1.094)
Last hits	-.001	.001	2.196	0.138	0.999 (.997 – 1.000)
Denies	.019	.014	1.857	0.173	1.019 (.992 – 1.047)
Gold (p/m)	.005	.001	12.875	< .001	1.005 (1.002 – 1.008)
XP (p/m)	< .001	.001	0.023	.0878	1.000 (.998 – 1.003)
Assist	.056	.012	20.771	< .001	1.058 (1.032 – 1.083)
Deaths	-.216	.026	67.167	< .001	0.807 (.767 - .850)
Gold (p/m)	.005	.001	30.539	< .001	1.004 (1.003 – 1.006)

Note: Backwards elimination was utilised to reach the final model

The final model explained approximately 83.2% (*Nagelkerke* R^2) of the variance in match outcomes and correctly classified 93.8% of cases.

$$\hat{p}(win) = \frac{e^{-6.613+.056(assist)-.216(deaths)+.005(gpm)}}{1+e^{-6.613+.056(assist)-.216(deaths)+.005(gpm)}}$$

This model was then applied to the 2017 DOTA international data set. Any prediction value over 0.5 was considered to indicate a win. This was then compared to the actual win/loss results of the tournament. The model accurately predicted 88.3% of results. There are seven games were the model predicted a win for both teams, however in both of these games the team with the higher predicted likelihood of winning came out as the victor of the match. There was only a single game where the model failed to accurately predict the outcome. This was a game in which the winning team had only 12 kills to the losing teams 31. This is a large deviation from traditional play and as such is considered an outlier. However, there was one game that the model predicted wrongly for no explainable reason from the statistics used. The Table below shows the spread of the predictive values (on the x axis) for wins (Diamonds) and loses(squares).

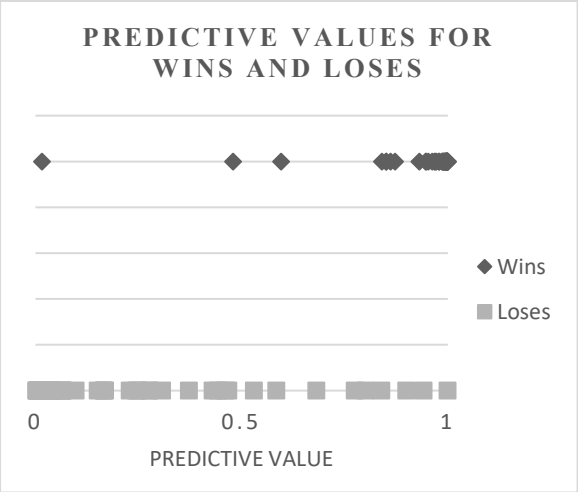


Chart: Predictive values of wins and loses

DISCUSSION

This paper explored a previously understudied area for predicting success in DOTA 2. This paper found a simple equation for predicting the likelihood of a win for teams. The limited number of variables needed for the high return of success in prediction suggests that as a preliminary model, it can be expanded to improve player ability.

Based on previous research into the eSports field the use of a binominal logistic regression did serve to create the most accurate model for predicting success. Somewhat surprisingly, however, the model needed only 3 of the 7 possible variables to predict the outcome of matches. However, this may be due to the fact that kills, deaths, and GPM were the major factors which in themselves represented the other variables.

As expected kills were significantly related to the prediction of success in DOTA2 games. Out of the over 200 matches that were analysed only 17 had games which were won with the winning team having less kills. Furthermore, deaths were also significantly associated with predicting the outcome of the games. Much like kills, teams who has less deaths were significantly more likely to win than those with more. Finally, GPM which is gained through a number of means was also much higher for teams who won the games. One reason for this may be that it takes into account a number of actions players can take, as such, teams who were more successful in kills, farming and attacking the Ancient gain more gold.

Contrary to our expectations, assists, last hits, denies, and XPM were not significantly related to win/loss outcomes and as such were excluded from the final model. One reason for the exclusion of XPM may be due to the fact that XP has a cap to its relevance to game play. XP is earned through being within a 1500 radius of a dying enemy or through farming of computer generated enemies (https://dota2.gamepedia.com/Experience). However, XP is only relevant as players rapidly level up from 0 to the cap of 25. Once this point has been reached XP is no longer as important for the outcome of the match. Furthermore, as XP is gained through the killing of creeps as well as being in the radius of a dying enemy, kills may be representative enough of the information presented by XPM. Similarly, denies and last hits may also not be significant for the same reason. As last hits are representative of the player (or player-controlled creep) which kills an enemy it may be more accurately shown through kills. Finally, denies and assists may be better represented through GPM and as such are not significant. Another possibility is that variables are equally matched in high ranked teams are essentially equally matched and as such do not result in the prediction of a significant difference between teams. Further research needs to be conducted to establish whether the minor changes in these variables may be able to be used as predictors to improve further models.

LIMITATIONS AND FURTHER RESEARCH

This preliminary study was not without limitations. Our model did not consider the players on the teams, or changes within teams between years. The addition of new players and playing techniques could mean that teams could improve over the years. Finally, the games assessed were from one single time point within a year and did not incorporate other major DOTA competitions

that occur. As such, this means there could be some limitations in the generalisability to other competitions as the differences in players, teams, and even locations of tournaments could affect what are the most successful strategies.

Future work needs to further explore the relationships found in this research. An area of interest would be looking at data gathered from lower level tournaments, as well as everyday players to see if the relationship between variables found exists at lower levels of play or whether there are other variables which have a greater influence on success in lower level games. Furthermore, future studies could improve on this model by making real time predictions based on character selections and actions taken during gameplay. In addition, there may be variables which this model did not consider as they were not collected or didn't show up as significant in this analysis which future research could also consider. Another area which would be interesting to assess in future research is the application of an ELO model which could take into account players and player rankings which are not considered in this model. Finally, future research could also analyse player movement and decision making to establish an improved strategy for gold generation.

CONCULUSION

In conclusion, this research is an interesting first step into the creation of a model which players, teams, and coaches could use to improve team performance. Further research can be done, but overall, this paper achieved its aims and established a preliminary model of success for high level teams seeking to play or improve in the DOTA International.

REFERENCES

Pobiedina, Neidhardt, Moreno, & Werthner. 2013. Ranking Factors of Team Success. Proceedings of the 22nd International Conference on World Wide Web, 1185-1194.

Skubida, D. (2016). Can Some Computer Games Be a Sport?: Issues with Legitimization of eSport as a Sporting Activity. *International Journal of Gaming and Computer-Mediated Simulations, 8(4),* 38-52.

Wagner, M. G. (2006). On the Scientific Relevance of eSports. *Proceedings of the 2006 International Conference on Internet Computing and Conference on Computer Game Development,* 437-440.

AUTHOR INFORMATION

Niamh McDonald, Student – Master of Data Science, Department of Science, RMIT.

Dave Matteo, Student – Bachelor of Health Science (Applied Statistics), Swinburne University of Technology.

Dr Minh Huynh, Lecturer – Department of Statistics, Data Science, and Epidemiology, Swinburne University of Technology.

THE FUTURE AND RESEARCH OPPORTUNITIES OF ESPORTS BETTING

Lyn Kee [a,b], Minh Huynh [a]

100993676@student.swin.edu.au, mhuynh@swin.edu.au

[a] *Swinburne University of Technology*
[b] *Corresponding author: Lyn Kee*

Abstract

As of October 2017, it was estimated that around 300 million people worldwide watch eSports. The viewership shows no sign of slowing down and is expected to grow 12% each year. The growing popularity of eSports turns the industry into an excellent revenue-making opportunity. Market research data shows that in 2017 alone, global revenue of eSports was over $1.5 billion. The opportunity for revenue has attracted a range of other interests and investments. Notably, applications of eSports, such as modelling, and betting are at the forefront and are one of the most fruitful areas. According to Pinnacle Sports, one of the eSports betting leader, eSports bets have surged from 100 thousand to five million in just a five-year period. Given the research of eSports betting is still in its infancy, it warrants a promising research potential. This paper will outline the relatively untapped eSports betting industry and discuss opportunities for researchers to collaborate within this field.

Keywords: eSports, modelling, betting

Introduction

[Commentator 1] “oh my god they are going to get control of this, they got to move quick right now”
[Commentator 2] “it is not over yet, they are going to walk right into a giant trap”
[Commentator 1] “oh man oh my god my headphones I don't even know what just happened…”

The converse above was not from a war zone. It is the commentary one would hear in eSports tournaments for computer games such as the *Counter-Strike*: *Global Offensive* (CS:GO). Competitive computer gaming, or eSports, has seen unprecedented growth in recent years. Hamari and Sjöblom (2017) define eSports as: “a form of sport where the primary aspects of the activities are facilitated by electronic systems, through the inputs from the players and teams; as well as the outputs of the electronic systems, which are mediated by human-computer interfaces” (p. 213). Historically, providing competitive computer gaming with the title “sport” has been controversial, and heavily criticised by the media. Despite this, the growth of eSports shows no sign of slowing down, attracting a wide range of interests and investments, and cementing its position as one of the top growing industries of the 21st century. This paper will discuss the history and the relatively untapped betting industry of eSports, then conclude with the discussion of possible opportunities for researchers to collaborate within this field.

A Brief History of eSports

The history of competitive eSports can be traced back to the early 1980s. The Space Invaders Championship held by Atari in 1980 was the first ever documented official major eSports tournament, attracting more than 10,000 participants (Crystal & Smith, 2017). Shortly after the success of Space Invader tournament, Atari announced the $50,000 World Championships tournament. The event, however, ended up being an unmitigated disaster due to the poor projection of levels of participation (Ausretrogamer, 2015). Instead of the expected 10,000 to 15,000 of participants, only 138 players took part. The World Championships was deemed a blight on the history of eSports.

Fortunately, computer gaming took a prosperous turn in the 1990s. Benefiting from the arrival of new consoles and increasing internet connectivity, more and more people became involved in eSports. Games with vital contributions to the growth of eSports in the 90s include *Starcraft* and *Quake*. Several tournaments established within these periods, such as the Cyberathlete Professional League (CPL), QuakeCon, and the Professional Gamers League, became an annual event that attracts hundreds to thousands of attendees.

The start of the 2000s was when eSports became mainstream. The launch of Xbox Live again pushed electronic gaming forward, with players being able to compete while remaining in the comfort of their homes. ESports' popularity gained its biggest surge with the release of *League of Legends* and *Dota 2*. It was suggested that *League of Legends* was the most played computer games in the western countries, with an estimation of over 67 million players per month (Ian, 2014). *Dota 2*, on the other hand, has the highest competition prize pools amongst eSports, totalling millions of U.S. dollars. Both of these games contributed substantially to the growth of eSports tournaments and eSports viewership.

As of October 2017, it was estimated that around 300 million people worldwide watch eSports (SuperData, 2017). The viewership shows no sign of slowing down and is expected to grow 12% each year. According to market research by Newzoo (Pannekeet, 2018) the global eSports audience will reach 380 million by 2018. This influx in viewership can partially be explained by the advances in technologies and the emergence of online streaming services, such as Twitch, was a crucial contributor to eSports viewership (Crystal & Smith, 2017). It was reported that as of 2013, Twitch recorded around 45 million monthly traffic numbers (Popper, 2013).

Motivations of eSports Consumptions

As eSports viewership becomes one of the most rapidly growing form of new media, it has attracted an increasing number of research interests. Although the literature on eSports is still rather rare up to this day, prior studies in eSports research primarily focused on the motivations of eSports consumptions. These include questions such as why people watch eSports and what attracts the participation of competitions. The uses and gratification theory (UGT), a theoretical approach to understanding the underlying reasons for people use of media, is widely adopted to examine media viewing (Hamari & Sjöblom, 2017). Based on the UGT, Hamari and Sjöblom measured eSports consumption motivations and found that escapism was positively correlated with eSports watching frequency. Escapism refers to the experience of mental distraction provided by use of media. Thus, it is argued that watching eSports may afford some levels of gratifications.

Wagner (2006) suggests that whether an activity can be considered as a sports changes as the value system in society changes. In the Industrial Age, physical fitness was considered as one of the most dominant values in society. Therefore, traditional sports mostly aimed at measuring the physical abilities of the athletes. The onset of the Information Age, however, indicate that changes are in place. The mastery of technology by different means is becoming one of the most fundamental values in society. In youth culture, particularly, individuals who feel the need to demonstrate this mastery, may choose to showcase through succeeding in competitions such as computer gaming. The participation of eSports competition can, thus, be interpreted as one of the logical consequences of the transition from Industrial- to Information-based societies.

Contributing to the increasing number of eSports players is the surge in earnings (see Figure 1). When the eSports industry was still in its infancy, it was an incredibly difficult environment for professional players to make a decent income. Through professionalisation, talented eSports players can now earn up to millions of U.S. dollars (Statista, 2018). In *League of Legends* alone, it was suggested there are around 1,000 professional players worldwide, making an average income of $320,000 annually (Heitner, 2018). The top earner amongst eSports player, Kuro Takhasomi, pocketed over three million U.S. dollars (Statista, 2018).

The prize of eSports tournaments has also seen enormous growth in recent years. Prize pools of The International, the world's largest *Dota 2* tournament, have grown from 11 million U.S. dollars in 2014 to 25 million in 2017 (eSports Earning, 2018). This is equivalent to around 140% of growth in just a period of three years. Prizes of tournaments are usually funded by sponsorship. As of 2018, Newzoo (2018) reported 53.2% year-on-year growth in eSports sponsorship, amounted around $350 million of the total revenue. With substantial corporate sponsorships and media coverage, eSports tournaments prize pools are expected to continue its growth. This indicates there will be more players, more tournaments, and enormous opportunities in eSports.

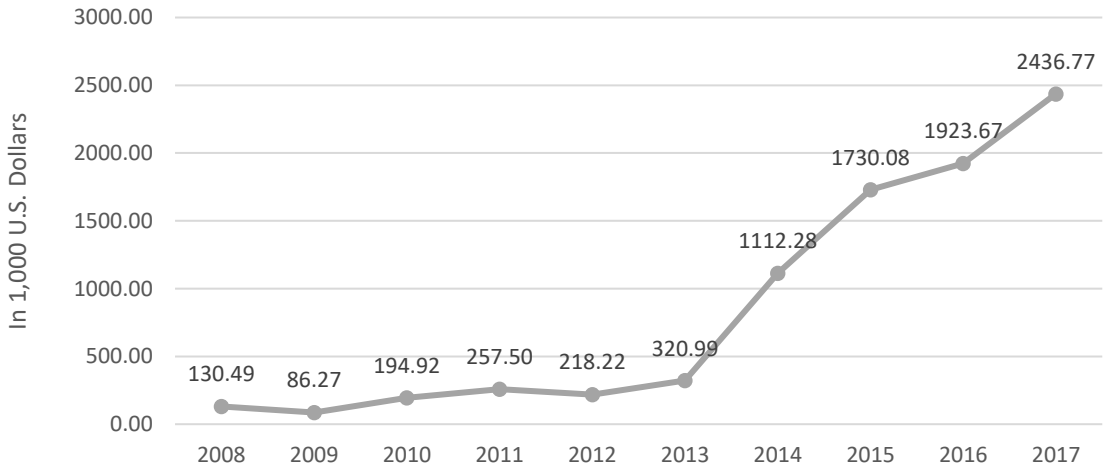


Figure 1: highest earning of eSports players by years.

eSports Betting on the Rise

As with any popular sport, mainstream exposure causes money to follow and eSports is no exception. Betting in eSports began humbly, through a process called skin-betting, where players would wager in-game cosmetic items (called skins) on the outcome of matches. However, it soon became clear that the exponential growth from players, viewers, and sponsors alike would transform this earnest skin-betting process into one based on monetary gains.

Today, cash gambling on eSports occurs through a mixture of traditional sportsbooks (e.g. Bet365, Pinnacle, and Ladbrokes) and eSports-only sportsbooks (e.g. Unikrn). Market research by Eilers and Krejcik Gaming (Grove, 2016) has estimated that the amount fans wagered on eSports in 2016 was close to $5.58 billion (USD). This figure encapsulates both skin and cash gambling. And despite a highly controversial decision by game developer, Valve, to crackdown on unregulated skin gambling, it is expected that the eSports gambling market will reach $12.92 billion (USD) by 2020. However, this growth (an increase of 234%) is not without challenges, one of which being the availability of data for bookmakers.

As with traditional sports, the more data a bookmaker has access to, the more reliable and accurate the odds for eSports can be. Some eSports provide tremendous amounts of data (sometimes so much so that most of it becomes irrelevant), whereas others are rather limited. The gaming company Valve is well known for providing vast amounts of data for their respective eSports (*CS:GO* and *Dota 2*). The data is available through Valve's open Application Program Interface (API), which allows developers, fans and bookmakers to extract the data they want. With over 430,000 players (both public and professional) logging in daily to play at least one game of *Dota 2* per day (based upon the Steam April 2018 Charts), the amount of data that can be generated is immense.

On the other hand, there are also big gaming companies that provide very little data on their competitive games, such as Blizzard Entertainment, who developed several hugely popular eSports titles, such as *StarCraft 2*, *Hearthstone* and *Overwatch*. Because Blizzard provides minimal data on their titles, bookmakers are often taking a significant risk by allowing bets to be placed on these games, before they have had a chance to build an accurate model.

Nonetheless, this has not stopped fans and bookmakers alike from having a go at developing their own prediction models for Blizzard games. For example, numerous attempts have been made over the years to model *StarCraft 2* using a Glicko ratings system. The Glicko model (Glickman, 2013) operates in a similar fashion to an Elo model in the sense that both systems are methods for assessing player (or team) ratings in comparison to how well (or poorly) they performed after a match(s). Unlike Elo however, Glicko focuses on Ratings Deviations (RD), and can be operationalised as:

$$RD' = \sqrt{\left(\frac{1}{RD^2} + \frac{1}{d^2}\right)^{-1}} \quad (1)$$

RD' represents the new ratings deviation after a series of *m* games and *RD* represents the old rating deviation:

$$RD = min\left(\sqrt{{RD_0}^2 + c^2 t}, 350\right) \quad (2)$$

Where *t* represents the amount of time (or *rating periods*) since the last competition or tournament. Players for whom the *RD* is unknown (e.g. they are unrated) are provided an *RD* of 350. The constant *c* represents the uncertainty of a player's skill over a period of time and can usually be estimated by considering the length *t* required before a player's *RD* changes to that or an unrated player.

To determine the new rating *r*, after a series of *m* games, the following formula can be applied:

$$r = r_0 + \frac{q}{\frac{1}{RD^2} + \frac{1}{d^2}} \sum_{i=1}^{m} g(RD_i)(s_i - E(s|r, r_i, RD_i)) \quad (3)$$

Where:

$$g(RD_i) = \frac{1}{\sqrt{1 + \frac{3q^2(RD_i^2)}{\pi^2}}},$$

$$E(s|r, r_i, RD_i) = \frac{1}{1 + 10^{\left(\frac{g(RD_i)(r - r_i)}{-400}\right)}},$$

$$q = \frac{\ln(10)}{400}, \qquad \text{and}$$

$$d^2 = \frac{1}{q^2 \sum_{i=1}^{m} \left(g(RD_i)\right)^2 E(s|r, r_i, RD_i)(1 - E(s|r, r_i, RD_i))},$$

with r_i representing the ratings of the individual opponents, and s_i represent the outcome of individual matches (win = 1, draw = 0.5, loss = 0).

To illustrate Glicko via an eSports example, we will use data sourced from the Starcraft 2 Programming and Predictions website, Aligulac, (http://aligulac.com/). The authors utilise a slightly modified version of the Glickman's (2013) original system, but the underlying principle is still comparable. For example, using the Aligulac data, we have compared the Glicko ratings for the Starcraft 2 player Ty during the Intel Extreme Masters 2017 Championship, to the odds provided through Pinnacle Sports Betting (see Table 1).

Table 1: Comparing Aligulac Glicko ratings to Pinnacle Sports Betting Odds (IEM championship 2017)

Event	Player	Opp	Player r	Opp r	Player	Opp	Player Odds	Opp Odds
Group B M3	Ty	Stats	2451	2280	1	2	1.56	2.45
Group B M5	Ty	jjakji	2632	2065	2	1	1.3	3.6
Group B M9	Ty	Harstem	2451	1961	2	1	1.16	5.33
Group B M11	Ty	aLive	2632	2234	0	2	1.38	3.07
Group B M15	Ty	Neeb	2451	2530	2	0	*u*	*u*
Ro12	Ty	Zest	2451	2418	3	1	1.57	2.42
Quarter Finals	Ty	GuMiho	2632	2340	3	2	1.6	2.37
Semi-finals	Ty	aLive	2632	2234	3	2	1.62	2.33
Gran Finals	Ty	Stats	2451	2280	4	3	1.87	1.95

Note: *u* = unknown

The authors of Aligulac compared the predicted win rate (using the modified Glicko model) to the actual win rate for over 100,000 historic StarCraft matches. Their analyses suggested that predicted and actual win rates were quite comparable, up to about a prediction of 80%, after which, the model tends to overestimate the better rated player (see Figure 2 below).

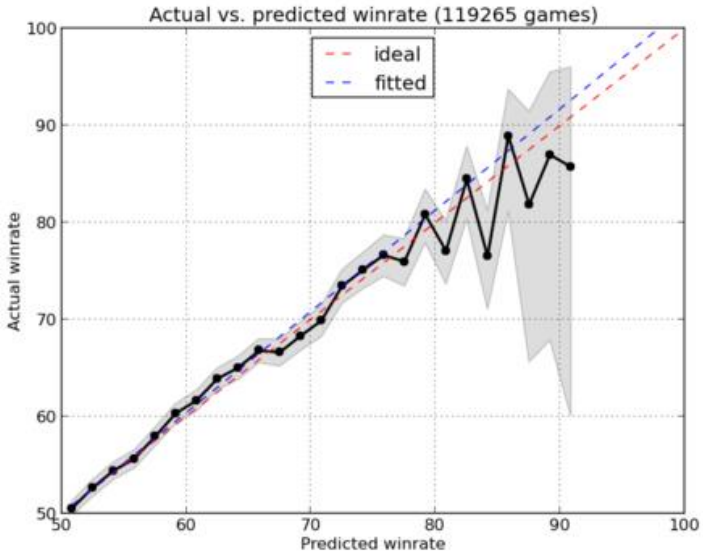


Figure 2: comparing actual to predicted win rates for StarCraft 2 (image sourced from http://aligulac.com)

Conclusion

The opportunity for revenue in eSports has attracted a range of interests and investments. Notably, applications of eSports, such as modelling, and betting are at the forefront and are one of the most fruitful areas. Given the research of eSports betting is still in its infancy, it warrants a promising research potential. A quick search on Scopus returns only 65 eSports related studies, spanning from the year 2005 to 2018. Further, investigation of prediction models for eSports betting, such as the examples provided above, is almost non-existence. The accelerating growth of eSports gambling market and limited research in this area signify exciting opportunities for researchers to collaborate within this field.

References

Aligulac (2018). StarCraft 2 Progaming Statistics and Predictions. Retrieved May 31 2018 from http://aligulac.com/

Ausretrogamer. (2015). *The Atari $50,000 world championships fiasco.* Retrieved May 31, 2018, from http://www.ausretrogamer.com/tag/1980s-gaming-tournaments/

Crystal, S., & Smith, J. (2017). *eSports betting: The past and future,* SCCG Management. Retrieved May 31, 2018, from https://sccgmanagement.com/news/2017/10/20/esports-betting-past-future

ESports Earning. (2018). *Largest overall prize pools in eSports,* ESports Earning. Retrieved May 31, 2018, from https://www.esportsearnings.com/tournaments

Glickman, M. E. (2013). Example of the Glicko-2 system. Boston University. Retrieved May 31 2018 from http://glicko.net/glicko/glicko2.pdf

Grove, C. (2016). Esports & Gambling: Where's the action? *Eilers & Krejcik Gaming*, *2*. Retrieved May 31 2018 from https://www.thelines.com/wp-content/uploads/2018/03/Esports-and-Gambling.pdf

Hamari, J., & Sjöblom, M. (2017). What is eSports and why do people watch it. *Internet Research, 27,* 211–232. doi: 10.1108/IntR-04-2016-0085

Heitner, D. (2018). *A look inside riot games, from $320,000 player salaries to using eSports as a catalyst for sales,* Forbes. Retrieved May 31, 2018, from https://www.forbes.com/sites/darrenheitner/2018/05/02/a-look-inside-riot-games-from-320000-player-salaries-to-using-esports-as-a-catalyst-for-sales/#518f1d0f2c6a

Pannekeet, J. (2018). *Newzoo: Global eSports economy will reach $905.6 million in 2018 as brand investment grows by 48%.* Newzoo. Retrieved May 31, 2018, from https://newzoo.com/insights/articles/newzoo-global-esports-economy-will-reach-905-6-million-2018-brand-investment-grows-48/

Pinnacle. (2017). *The road to five million eSports bets.* Retrieved May 12, 2018, from https://www.pinnacle.com/en/esports/betting-articles/educational/esports-betting-growth-at-pinnacle/ay22gtmplb93agpa

Popper, B. (2013). *Field of streams: How twitch made video games a spectator sport.* The Verge. Retrieved May 31, 2018, from https://www.theverge.com/2013/9/30/4719766/twitch-raises-20-million-esports-market-booming

Sheer, I. (2014). *Player tally for 'League of Legends' surges.* The Wall Street Journal. Retrieved May 31, 2018, from https://blogs.wsj.com/digits/2014/01/27/player-tally-for-league-of-legends-surges/

Statista. (2018). *Leading eSports players worldwide as of January 2018,* Statista. Retrieved May 31, 2018, from https://www.statista.com/statistics/518010/leading-esports-players-worldwide-by-earnings/

Steam Charts (2018). An ongoing analysis of Steam's concurrent players. Retrieved May 31 2018, from http://steamcharts.com/app/570

Super Data Research. (2017). *Esports market report: courtside - Playmakers of 2017.* Retrieved May 12, 2018, from https://www.superdataresearch.com/market-data/esports-market-report/

Wagner, M. (2006). On the scientific relevance of eSport. *Proceedings of the 2006 International Conference on Internet Computing and Conference on Computer Game Development,* CSREA Press, Las Vegas, NV, pp. 437-440.

GENDER BIAS AND THE NEW ZEALAND MEDIA'S REPORTING OF ELITE ATHLETES

Timothy S. McNamara [a], Tamsyn A. Hilder [a], Emma C. Campbell [a,b] and Paul J. Bracewell [a,b,c]
[a] *DOT Loves Data, Wellington*
[b] *Victoria University of Wellington*
[c] *Corresponding author: paul@dotlovesdata.com*

Abstract

Researchers have long recognised that gender influences the visibility of sport in the public. Existing research found that female participation in sport has not resulted in an equal increase in television airtime, however this association has not been explored within a New Zealand context. Here, the effect of gender on an elite athlete's presence in mainstream New Zealand media is quantified. The New Zealand Olympic Committee provides details of every athlete who has competed at the Summer Olympics, Youth Olympics, Winter Olympics and Commonwealth Games on their website (http://www.olympic.org.nz/). New Zealand media articles have been collected by downloading content available via digital editions of major publications. Qualitative evidence that men's achievement at an elite level is over-represented by New Zealand's mainstream media is provided. Counting references to each athlete within the archive of downloaded content provides quantitative evidence that men are over-represented in the media, comparative to female athletes who have achieved the same level of success at an elite level. Furthermore, insight into how media representation changes across sport, career length, career starting point and indicators of peak performance, such as the number of Olympic gold medals achieved within an athlete's sporting career, is discussed.

Keywords: Gender bias, mainstream-media bias, New Zealand sport, text mining, Olympics

1. INTRODUCTION

Society is heavily influenced by the media, and its representation of women is no exception. Over the past 30 years research has shown that women's sport has received less coverage than men's, and that often this coverage is trivialized, stereotyped, devalued and marginalised (Crossman et al., 2007; Fink, 2015; Litchfield & Osborne, 2015). In the United States, articles on female athletes, compared to male athletes, are more frequently placed inside the sports section rather than the front pages, and are more likely to focus on the team rather than a female star (Pratt & Grappendorf, 2008). As well as being underrepresented in the media, female athletes are rarely applauded for their athletic abilities alone. Instead the focus is often placed on their physical appearance, femininity and heterosexuality (Fink, 2015). Researchers have recognised that gender has an impact on the visibility of sport in the public sphere. Unfortunately, despite an increase in female participation in sport, media coverage and marketing of female athletes and women's sport has not progressed at the same rate (Duncan, 2012; Cooky et al., 2013; Fink, 2015).

Much of the research completed to date has focused on the United Kingdom, United States, Canada and Australia, and primarily focuses on print media and television coverage. Existing New Zealand (NZ) research has examined gender bias in television sports coverage. Jones (2013) examined the coverage of the 2008 Olympic Games by four national broadcasting agencies, including TVNZ (the national publicly-funded NZ TV broadcaster), and found that male athletic achievement was given more prominence, and that the content of articles involving female athletes perpetuated gender stereotypes. The worst result for women, reported by Jones (2013), occurred on TVNZ where women featured in just over one third of all stories. Similar results have been found in the coverage of female athletes in NZ newspapers. Research by McGregor and Fountaine (1997) showed that female athletes accounted for only 4.4% of the space devoted to sports news in six of NZ's newspapers.

With regards to the Olympics, research suggests that this situation has been evolving. In coverage of the 2004 Summer Olympics, female athletes were less likely to appear on the front page of *The New York Times* and *The Los Angeles Times* (Pratt & Grappendorf, 2008). This contrasts with later work that could not identify gender-based differences in coverage of the 2012 Summer Olympics (Eagleman, 2014). A gender imbalance has also been found outside of Olympic events. During the 2004 Wimbledon Championships, male players had significantly more total newspaper coverage than female players (Crossman et al., 2007).

Many authors have written about the wider societal impact of the Olympics and other major sporting events. Success impacts the local area by increasing civic pride (Süssmuth, 2010) and the host city's development (Gold and Gold, 2008).

Olympic sponsorship has been investigated from the perspective of the sponsors' share price (Farrell and Frame, 1997; Filis and Spais, 2012). Effective sponsorship defined as sponsoring winning athletes can lead to positive outcomes for the firm (Tripodi, 2001), but the best way to producing effective sponsorship outcomes are often unclear (Wang & Markellos, 2016). Celebrity endorsement is one established mechanism (Ding, 2011). To become a celebrity, athletes require a level of exposure. Understanding the factors that influence how much coverage Olympic athletes receive may provide insight into understanding how to maximise an athlete's sponsorship potential. Does gender affect an athletes' potential for sponsorship deals?

This research aims to explore the effect of gender on a New Zealand athlete's presence in mainstream New Zealand digital media throughout an athletes' career. Understanding these dynamics should provide clarity for many elite athletes. Sports editors will have increased visibility of any unconscious bias. Policy makers will have more information to support their public funding investments. Athletes' managers will have more evidence to support the professional side of their clients' careers.

2. METHODS

SOURCING THE DATA

The New Zealand Olympic Committee (NZOC) maintains a list of every athlete who has competed in the Olympics, the Commonwealth Games and the Youth Olympics. Each athlete record has been downloaded. Variables extracted include name, sport, event, age, height, gender, events participated in and medals won. The NZOC's list of 3,759 (http://www.olympic.org.nz/athletes) is considered authoritative. In places, the variables are incomplete and have been supplemented with information obtained from secondary sources. Resolving missing data is discussed below.

DOT Loves Data maintains an archive of news content from publicly available sources that it calls the Pressroom. To date it incorporates 12.8 million articles spanning from 2005 to 2018. The Pressroom contains an index of content produced on online news publications, including a comprehensive collection of articles published on nzherald.co.nz (NZME) and stuff.co.nz (Fairfax NZ). 41,871 unique articles, comprising 211,446 athlete mentions, were discovered within the Pressroom that relate to NZOC athletes. Of the 3,759 NZOC athletes 3,588 were present in the media. Two thirds of the articles collected were published in the last two years and 83% of articles were published from 2012 onwards. For every athlete, two searches of the archive were made: 1) an exact string match of "FIRSTNAME LASTNAME" and "FIRSTNAME LASTNAME" SPORT, where SPORT is the sporting discipline recorded by NZOC.

DATA PROCESSING

Once the source data were retrieved, some curation was required to enable effective analysis. The NZ Olympic Committee's website does not provide gender and date of birth for all athletes. For these cases, information was added manually based on searching for the athlete online.

It is possible for search results for individual athletes to return the same article twice. Therefore, to minimise the effect of duplicates, results with the same URL were reduced to a single record. This strategy does not eliminate all duplicates however, as news publishers will occasionally publish the same article under multiple URLs. The focus of the current paper is the reporting of New Zealand elite athletes within New Zealand. Therefore, results were restricted to New Zealand-based sources. This might downplay the international profile of some very prominent athletes but reduces the rate of false matches.

Furthermore, it is possible that there may be false matches for some athletes. For example, searching for "Scott Wilson" returns the story "Armed Robbery in North Auckland", that contains the sentence "Detective Scott Wilson said officers hoped someone would recognise the reasonably distinctive hooded-jacket worn by one of the offenders". To reduce the occurrence of false matches the athletes' sport was also included in the search of the archive. Many athletes compete in multiple events over the course of their career. For example, Alethea Boon and Lani Hohepa have both competed in artistic gymnastics as well as weightlifting. This prevents a single categorical variable from representing sport effectively. To enable a fair comparison between sports, each sport is represented as a Boolean variable where a true value represents participation.

DATA ANALYSIS

Linear regression was performed using python inbuilt statsmodels module with a simple ordinary least squares (OLS) model. We examined gender (is_female), year of birth (birthyear), the number of games attended (n_games_attended), the number of gold medals received (n_gold_medals) and the total number of any medal received (n_medals) against the number of media mentions. Note that athletes with no recorded date of birth were excluded from model (1), but not model (2). Some research has suggested that women who compete in so-called 'feminine' sports such as gymnastics and diving receive disproportionate media coverage (Jones,

2013). Therefore, we also controlled for sport by adding athlete's sport into our model. The dataset used in the linear regression is available from Hilder et al. (2018).

3. RESULTS AND DISCUSSION

Initially, we performed an OLS linear regression model including is_female, birthyear, n_games_attended, the n_gold_medals and n_medals. This model, referred to as model (1), is shown in Table 1 and illustrated in Figure 1. The inclusion of birthyear reduced the number of observations from 3,588 to 1,906 as there were several athletes with an unknown year of birth. The results of model (1) are biased towards the present day given that of those athletes with a known year of birth 86% were born after 1970, and 83% of articles were published from 2012 onwards. Similar results were obtained when controlling for sport, with a p value of 0.000 for is_female.

Table 1: Linear regression model (1) results

	coef	std err	t	P>|t|	[0.025	0.975]
is_female	−4.3715	0.993	−4.401	0.000	−6.320	−2.423
birthyear	0.1087	0.035	3.135	0.002	0.041	0.177
n_games_attended	0.5863	0.464	1.263	0.207	−0.324	1.497
n_gold_medals	3.6760	1.398	2.630	0.009	0.935	6.417
n_medals	0.5196	0.935	0.555	0.579	−1.315	2.354

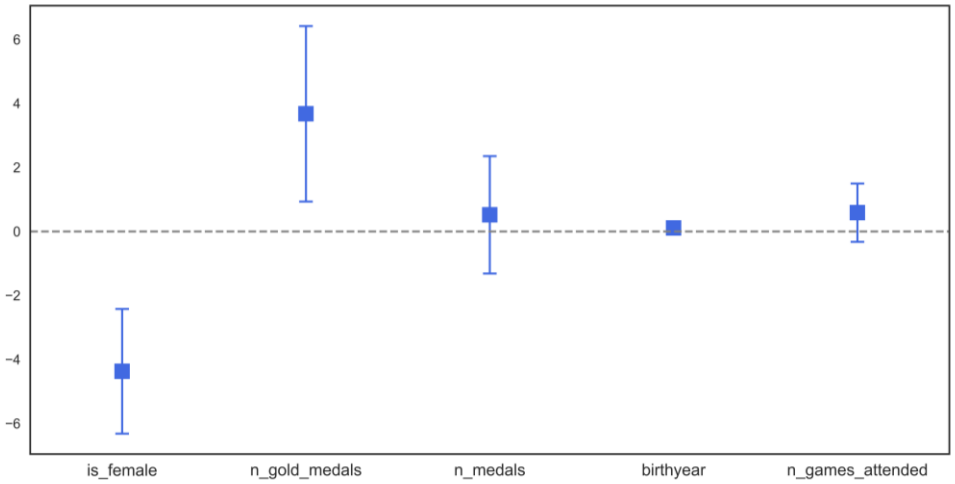


Figure 1: Regression coefficients from model (1) with confidence intervals.

From the results of model (1), n_games_attended, n_medals, and birthyear were not found to be significant indicators of media success (p value greater than 0.1), where we consider 'media success' is to be those athletes who are mentioned more often. Therefore, we performed a OLS linear regression model with these three variables excluded, named model (2). Excluding birthyear from the model increased the number of observations to 3,585. The results of model (2) are shown in Table 2. Both model (1) and (2) demonstrate that female athletes are less likely to be mentioned in the media compared to male athletes (illustrated in Figure 2). As expected, gold medal winning athletes are more likely to be mentioned in the media.

Table 2: Linear regression model (2) results

	coef	std err	t	P>|t|	[0.025	0.975]
is_female	−3.6733	0.713	-5.154	0.000	-5.071	-2.276
n_gold_medals	4.1917	0.933	4.491	0.000	2.362	6.022

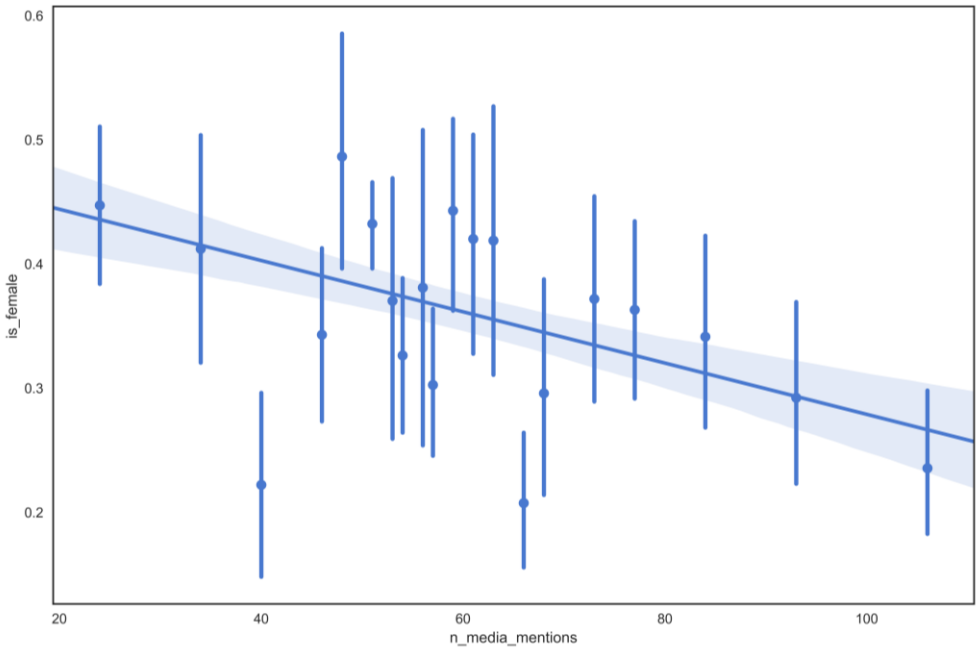


Figure 2: Influence of gender on the total number of media mentions.

Figure 1 and 2 are supported by broader summary statistics. The arithmetic mean and standard deviation of n_media_mentions for all male athletes is 60.44 ± 0.45 and all female athletes is 56.50 ± 0.52. Similarly, the arithmetic mean and standard deviation of n_media_mentions for medal winning male athletes is 62.33 ± 0.78 and medal winning female athletes is 57.56 ± 1.01. For gold medal winning athletes, these values increase to 64.25 ± 1.44 and 59.70 ± 1.87 for males and females, respectively. Even with these simple statistics female athletes are mentioned less than their male counterparts.

4. CONCLUSIONS

This work demonstrates that a gender bias exists in the reporting of New Zealand athletes in New Zealand media. Specifically, female gold medallists receive less coverage on average across their career than male athletes regardless of male athlete success. Importantly, whilst all athletes are considered, the emphasis is on recent reporting with two-thirds of articles published in the last two years. Consequently, these findings are an indication of current reporting practices. In future we hope to extend this research and examine in more detail the type of language used in articles discussing female versus male athletes. Analysis undertaken in National Geographic (Nowakowski, 2017), where it was found that female Disney characters were more likely to be praised for their appearance as opposed to their skills and actions.

References

Cooky, C., Messner, M. A., & Hextrum, R. H. (2013). Women play sport, but not on TV. *Communication & Sport*, *1*, 203-230.

Crossman, J., Vincent, J. & Speed, H. (2007). 'The times they are a-changin': gender comparisons in three national newspapers of the 2004 wimbledon championships. *International Review for the Sociology of Sport*, *42*, 27-41.

Ding, H. (2011). The value of celebrity endorsements: A stock market perspective. *Marketing Letters*, *22*, 147-163.

Duncan, M. C. (2012). Chapter 14 – Gender Warriors in Sport: Women and the Media. *Routledge Online Studies on the Olympic and Paralympic Games*, 1:46, 247-269.

Eagleman, A. (2014). A unified version of London 2012: New-media coverage of gender, nationality, and sport for Olympic consumers in six countries. *Journal of Sport Management*, *28*, 457-470.

Farrell, K. A. & Frame, W. S. (1997). The value of Olympic sponsorships: who is capturing the gold? *Journal of Market Focused Management*, *2*, 171-182.

Fink, J. S. (2015). Female athletes, women's sport, and the sport media commercial complex: Have we really "come a long way, baby"? *Sport Management Review*, *18*, 331-342.

Filis, G. N. & Spais, G. S. (2012). The effect of sport sponsorship programs of various sport events on stock price behaviour during a sport event. *Journal of Promotion Management*, *18*, 3-41.

Gold, J. R. & Gold, M. M. (2008). Olympic Cities: Regeneration, city rebranding and changing urban agendas. *Geography Compass*, *2*, 300-318.

Hilder, T.A., McNamara, T.S. & Campbell, E.C. (2018). Media mentions of New Zealand elite athletes (Version 2018) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1284630.
Jones, D. (2013). Online coverage of the 2008 Olympic games on the ABC, BBC, CBC and TVNZ. *Pacific Journalism Review*, *19*, 244-263.
Litchfield, C. & Osborne, J. (2015). Women in the sports pages: A brief insight into Olympic and Non-Olympic years in Australia. *International Journal of Sport and Society*, *4*, 45-56.
McGregor, J. & Fountaine, S. (1997). Gender equity in retreat: The declining representation of women's sport in the New Zealand print media. *Metro*, *112*, 38-44
New Zealand Olympic Committee (2018). http://www.olympic.org.nz/athletes [Accessed on 8 March 2018].
Nowakowski, K. (2017). Who's the fairest? *National Geographic*, 231, 24-25.
Pratt, J. & Grappendorf, K. (2008). Gender differences in print media coverage of the 2004 summer Olympics in Athens, Greece. Women in Sport & Physical Activity Journal, 17, 34.
RNZ001: Radio New Zealand, Armed robbery in North Auckland, 2014, http://www.radionz.co.nz/news/national/252967/armed-robbery-in-north-auckland
Süssmuth, B. (2010). Induced Civic Pride and Integration. *Oxford Bulletin of Economics and Statistics*, *72*, 202-220.
Tripodi, J. A. (2001). Sponsorship – A confirmed weapon in the promotional agency. *International Journal of Sports Marketing and Sponsorship*, *3*, 82-103.
Wang, J. Y. & Markellos, R. N. (2016). Is there an Olympic gold medal rush in the stock market? *The European Journal of France*, DOI: 10.1080/1351847X.2017.1421245.

CONSTRUCTING A PREDICTIVE PGA PERFORMANCE RATING USING HIERARCHICAL VARIABLE CLUSTERING

Ankit K. Patel [a,b,c] Samuel J. Rooney [a], Paul J. Bracewell [a,b] and Jason D. Wells [a]
[a] *DOT Loves Data*
[b] *Victoria University, Wellington*
[c] *Corresponding author: ankit@dotlovesdata.com*

Abstract

Golf is a high-profile sport. Here, a method for deriving an individual performance-based rating linked to winning is outlined. Player earnings on the PGA tour are used as a proxy for performance as this encapsulates both relative performance and perceived importance of the event. Player data between 2010-2018 with 60 variables was used for analysis. Using a hierarchical clustering technique, four meaningful and expected clusters were found to influence a player's earnings: 1. Short game, 2. Putting, 3. Accuracy and 4. Driving. The most important variable within each cluster was identified by applying a random forest technique. A simple linear model was created with these metrics for each of the 2010-2017 seasons. The models were applied to the following season, explaining no less than 64% of variation associated with player earnings for each season.

Keywords*:* Random Forest, rank-order statistics, dimension reduction, PGA performance

1. INTRODUCTION

Analytically understanding a player's true value and potential is an increasingly important task undertaken by sport team owners, coaches, managers, media and punters. This is primarily for two reasons: 1) Industry growth – revenue within the sporting industry grew by US$145.3 billion over 2010-2015 period (Coopers, 2015); and 2) Large monetary investment, given the large investment of resources and stakes involved, stakeholders cannot rely on subjective views and personal beliefs only. According to Dataconomy (2018), it is expected that big data will increasingly influence the sporting industry in a market worth estimated over $900 billion.

According to Dusek (2017), golf has had little visibility at the MIT Sloan Sports Analytics conference but has recently received more attention, highlighting the increase in interest in golf analytics (Levin, 2017).

With variations in performance, environmental factors and an element of luck on a hole-by-hole basis, predicting tour winners based on performance metrics is challenging. Given the variation in player standings and a growing demand to identify "must watch" players, a predictive PGA rating system is developed. Although there exist methods to rank player performance based on known outcome, such as the Official World Golf Ranking (OWGR), Score-based Skill Estimation (SBSE) and the Sagarin method, these approaches are adaptive rather than predictive. To be successful, a golfer must be proficient in several skills, such as putting, which these systems do not consider. With recent technological advancement in golf clubs and balls, golfers are hitting the ball further with increased accuracy; "Our turning point on doglegs is now 280 yards off the tee, not 250 like it has been traditionally" (Heiny, 2008, p. 2). This change in player capability has led many to criticise the Official World Golf Ranking (OWGF) system. Broadie & Rendleman, 2013; Golf Digest 2015; The New York Times 2017; Golf Channel 2012; The Economists (2017), are among the most notable criticisms.

LITERATURE REVIEW

There is vast literature surrounding the application of analytical techniques within golf; the following section reviews research that attempts to develop predictive and adaptive golf rating systems.

Broadie & Rendleman (2013) investigated whether the OWGR system is prone to bias for the four major tours (PGA Tour, European tour, Japanese Tour and Asian Tour) by comparing the OWGR system with two unbiased methods for estimating golfer performance: 1. Score-based skill estimation (SBSE) method and 2. Sagarin method. The SBSE method provides a player's mean 18-hole score played on a neutral course, statistically removing all intrinsic course difficulties such as course setup and weather. The method does not use tour information to calculate rank and therefore removed any bias for or against any tour. The Sagarin method uses a player's won-lost-tied record against other players when they play on the same course on the same day, and the stroke differential between those players, then links all players to one another based on common opponents. Highly correlated rankings were found between the three methods however a large difference depending on tour affiliation was found which illustrates the existence of bias. There was a clear tendency for OWGR/SBSE ranking pairs to fall below the 45-degree line for non-PGA tour players and above the line for PGA tours. A similar result was found for OWGR/Sagarin relationship. Moreover, regressing OWGR on SBSE (& Sagarin) rankings and PGA tour affiliation indicator found that a golfer's primary tour affiliation is the PGA tour is penalized an average of 37 OWGR rankings positions relative to non-PGA Tour affiliated golfers

(Broadie & Rendleman, 2013). The analysis revealed statistically significant tour bias in the OWGR against PGA tour affiliated golfers and was greater among less skilled players.

Heiny (2013) applied times series analysis (1992-2003) to measure the change on the importance of different golf skills as driving distance increases. It was found that greens in regulation (GIRs), putts per round (PPR), scrambling, driving distance (DrDist), driving accuracy (DrAccu), sand saves and bounce back, experienced the largest change as driving distance increased. The analysis revealed that better performances have higher scores across all categories, except for PPR. GIR was the most highly correlated independent variable with both scoring average and monetary winnings followed by scrambling and DrAccu. Interestingly, correlation between performance measures and scrambling was trending down throughout the 1990's, rebounded in 1999, and then dropped in 2002-2003 (Heiny, 2013). The correlation between DrAccu and performance measures were stable throughout 1992-2002, but then drop in 2003, when DrDist jumped seven yards. The correlation between DrAccu and monetary winnings dropped from 0.22 to 0.

Connolly & Rendleman (2012) investigated what it takes to win on the PGA for Tiger Woods and other professional golfers as a function of individual player skill, random variation in scoring, strength of field and depth of field. The study applied Monte Carlo simulation to estimate the relative difficulty of all PGA tournaments between 2003-2009 and the probability of winning each event for all tournament participants. The scoring model outlined in Connolly & Rendleman (2008) was used to estimate skill and random variation in scoring. The model decomposes observed individual golfer scores into three parts: 1) Time variation in skill, 2) Estimates random effects because of daily round course interactions and 3) Estimates random effect related to separate player-course interactions. Overall player performance was assessed in terms of neutral and normal score, where neural scores are reduced by estimated round-course and player-course effects (Connolly & Rendleman, 2012), while normal scores represent what a player was expected to score under playing conditions, considering estimated skill and potentially autocorrelated component of their scoring. A Monte Carlo simulation was used to determine the mean score per round required to win a PGA Tour event as a function of the number of players participating, the number of rounds, the mean skill level of the participants and the random variation in their scoring. An estimate for a player's probability of winning was calculated by running 10,000 simulations and keeping a running count of number of time each player wins a given tournament. Simulation results reported Tiger Woods with highest winning probabilities, ranging from 34% to 52%, and estimated winning probability to be significantly better over the study period (2003-2009). Moreover, applying the same analysis on a player's normal score and recomputing the player's place in the tournament, showed that playing "normally" Tiger would have won 13 tournaments during 2003-2009.

Broadie (2012) applied an analysis based on strokes gained to assess professional golfer performance in different parts of the game and understand and quantify the contributions of three categories of golf shots – long game, short game and putting – in determining a total golf score for an 18-hole round. This performance attribution analysis is used to rank golfers in various skill categories and examine the relative impact of each skill category on overall score. Broadie (2012) was revealed that Tiger Woods is ranked first in total strokes gained and near the top in each of the three main categories: long game, short game and putting. Tigers long game accounts for about 2/3 of his scoring advantage relative to the average of other PGA Tour golfers. The analysis found long game to be the most important factor in explaining the variability in professional golf scores and that the three skill categories were nearly uncorrelated. A slight negative correlation can be explained by survivorship bias: golfers with a subpar long game need better than average putting to survive on the Tour.

The Economist's Eagle (Economist Advantage in Golf Likelihood Estimator) rating system measures every player's chance of victory at every point during the tour. The ratings system adopts course and hole difficulty, participants strength, score to par per round and player history. Course difficulty is inferred from player performance. For every round, the actual average score was compared to model expectations. Overperformance implied an easier course, and the opposite for underperformance. A similar approach was adopted to model 'hole' difficultly. A player's world ranking was used estimate the par per round that an average player in any tournament would produce in a major. The system evaluates the probability of all potential outcomes, including extreme ones. Using a multiple linear regression model with OWGR, PAR and distance as covariates, a baseline forecast for any golfer on any hole is created. A non-linear regression is applied to determine the optimal blend of a player's OWGR-based forecast alongside his own historical difficulty-adjusted scores. Finally, a cumulative order logit regression is applied to translate continuous projection to discrete scores. The techniques generate probabilities for 7 outcomes for each golfer on each hole. Using Monte Carlo, the algorithm simulates an expected quality of performance for each golfer on each day in each simulation, as well as a course and hole difficulty, and calculates the outcomes probabilities for each player-hole on expected daily averages.

2. METHODS

Given the amount of literature surrounding golf analytics, there is a lack of research focusing on performance predictions. Here, the type of performance metrics and skill sets required to be a successful PGA golfer are used to construct a meaningful predictive performance rating system. Using $log(season\ money)$ earned as the dependent variable, the study identifies the most statistically significant statistics that impact money earned.

Season-by-season player data was collected from www.pgatour.com between 2010-2018. The data contained 60 variables and approximately 250 player observations per season. Overall 2268 player observations were collected across eight seasons (2010-2018).

Given the large number of performance metrics and the presence of multicollinearity, the first task was dimension reduction. Applying a hierarchical clustering technique four unique clusters were identified. Figure 1 shows four unique clusters for the 2015 PGA tour; these results were consistent across the study period (2010-2018): 1. short game 2. putting 3. accuracy and 4. driving. Next, a random forest algorithm was applied to each cluster to identify the most important features. The dependent variable was $log(season\ money)$. Table 1: outlines the most important features within each cluster for the 2017 and 2018 season.

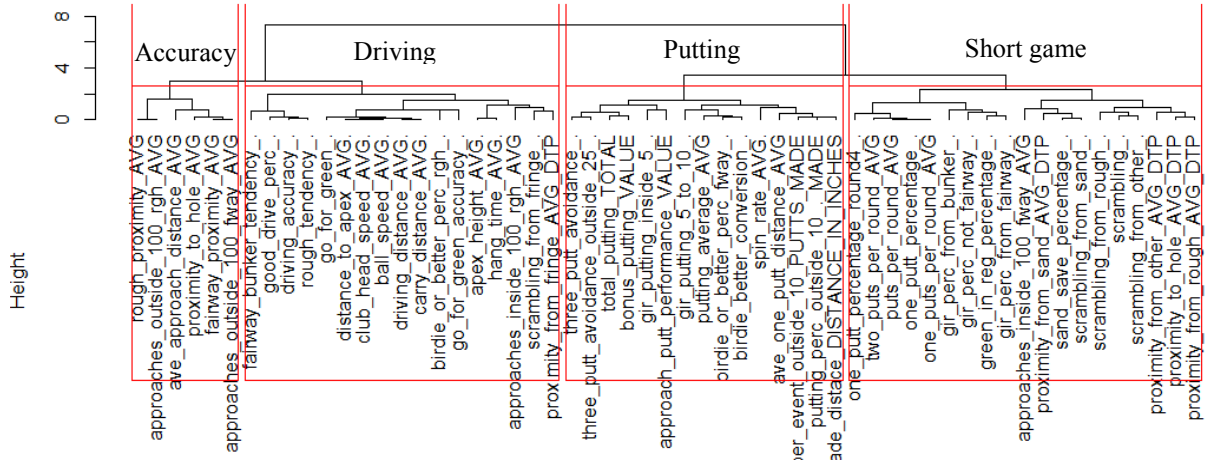


Figure 1: Hierarchical clustering of potential individual performance metrics from the 2015 PGA tour

A correlation analysis was applied to identify the extent to which the most important metric was related. Figure 2 illustrates the relationship between the most important performance metrics across the four clusters for the 2015 PGA data. The correlation values show a weak positive and negative relationship between important metrics across the clusters, while moderate negative relationships between metrics within clusters (i.e. *putts made per event outside 10 putts made* vs. *putting average avg.*; and *driving distance AVG.* vs. *driving accuracy.);* the multicollinearity across clusters has dissipated. This reveals that the issue of multicollinearity has been addressed through variable selection and crudely reducing dimensionality.

Most important features by cluster for regerssion and random forest					
Model	**Year**	**Short game**	**Putting**	**Accuracy**	**Driving**
Regression	2017	scrambling_.	putts_made_per_event_outside_10_PUTTS_MADE birdie_or_better_perc_fway_.	gir_perc_from_fairway_.	ave_approach_distance_AVG
	2018	birdie_better_conversion_. scrambling_.	putts_made_per_event_outside_10_PUTTS_MADE	green_in_reg_percentage_.	ave_approach_distance_AVG
Random Forest	2017	scrambling_.	putts_made_per_event_outside_10_PUTTS_MADE birdie_or_better_perc_fway_. birdie_better_conversion_.	gir_perc_from_fairway_. driving_distance_AVG.	ave_approach_distance_AVG go_for_green_.
	2018	scrambling_.	birdie_better_conversion_. birdie_or_better_perc_fway_. putting_average_AVG putts_made_per_event_outside_10_PUTTS_MADE	gir_perc_not_fairway_. green_in_reg_percentage_.	ave_approach_distance_AVG

Table 1: Important features by cluster (2017 & 2018) - Regression vs. Random forest analysis

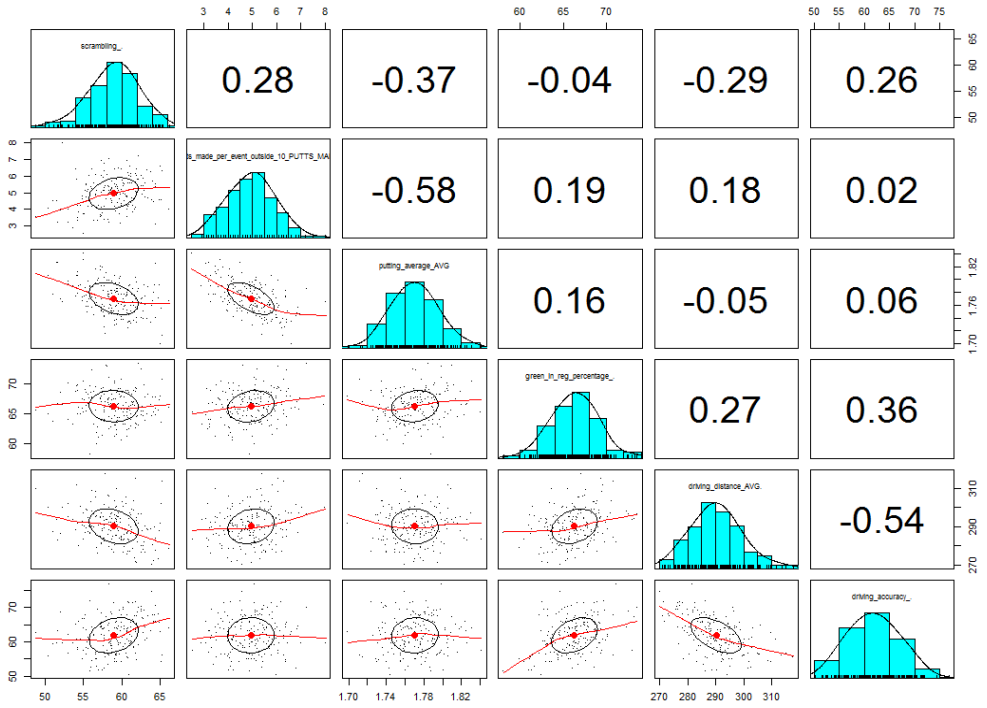


Figure 1: Correlation and distribution plot of 2015 metrics

LINEAR REGRESSION

Season specific regression models were built to investigate the ability of the important metrics, identified by applying a random forest to each cluster, to explain variation in $log(Y_{season\ money_{ij}})$. Equation (1) outlines the linear regression to model logarithmic earnings for player i during season j.

$$log\left(Y_{season\ money_{ij}}\right) \sim \alpha + \beta_1 x_{cluster\ 1} + \beta_2 x_{cluster\ 2} + \beta_3 x_{cluster\ 3} + \beta_4 x_{cluster\ 4} + \varepsilon \quad (1)$$

On average the models explained approximately 67% of the variation in $log(season\ money)$ earned. Table 2 reports the metrics, $p - values$, coefficients and $r - squared$ values for each 2015-2018 models. Each models' predictive power was evaluated by applying the model built off $season_i$ data to $season_{i+1}$ data and using Spearman's correlation rank and Wilcoxon's test. Figure 3 – regression assumptions were not violated.

Applying the models produced interesting player predictions. For example; Aaron Baddeley ranked 110th at the end of the 2010 PGA tour. Applying the 2010 model to 2011 data the model ranked Baddeley 33rd, and his actual rank was 20th. Another example is Rickie Fowler who finished 4th during the 2015 PGA season. Based on the 2015 data the model predicted Fowler to finish 16th during the 2016 season with an actual rank of 30th. However, there were instances where the model predictions were less than favourable; for example, Jimmy Walker ranked 9th at the end of the 2015 PGA tour. Applying the 2015 model to the 2016 data the model ranked Walker 61st, and his actual rank was 10th. These inconsistent results are due to large changes in 1) significant metrics across the season and 2) performance relating to the significant metrics.

Given the moderate predictive power of the regression models, random forest modelling was applied. Instead of incorporating the most important variables from each cluster, this time the 7 most important variables from the entire feature space were identified and fed into a random forest algorithm. Table 1 provides a partial outline of the seven most important variables across the 2017 and 2018 seasons and the cluster to which they relate (i.e. short game, putting, accuracy and driving). Results suggest that *putting, short game* and *accuracy* metrics are the most important areas of a golfer's ability as these metrics appear most frequently in the top seven.

ENSEMBLE MODELLING

The proposed model was built using a Random Forest learner. Random Forest was the model of choice because it averages the predictions of an ensemble of tree fitted to resampled versions of the learning data. This stabilizes the recursive partitions from individual trees and hence better approximates smooth functions. The generalisation error for forecasts converge asymptotically to a limit as the number of trees in the forest become larger (Breiman, 2001). The technique aims to reduce multicollinearity issues by choosing only a subsample of

the features space at each split. An amalgamation of decision tree, it aims to produce de-correlated trees and prune the trees by setting a stopping criterion for node splits. Overall, the goal is to identify the most important metrics that explain the largest amount of variation within each cluster. Producing a collection of uncorrelated and unpruned regression trees.

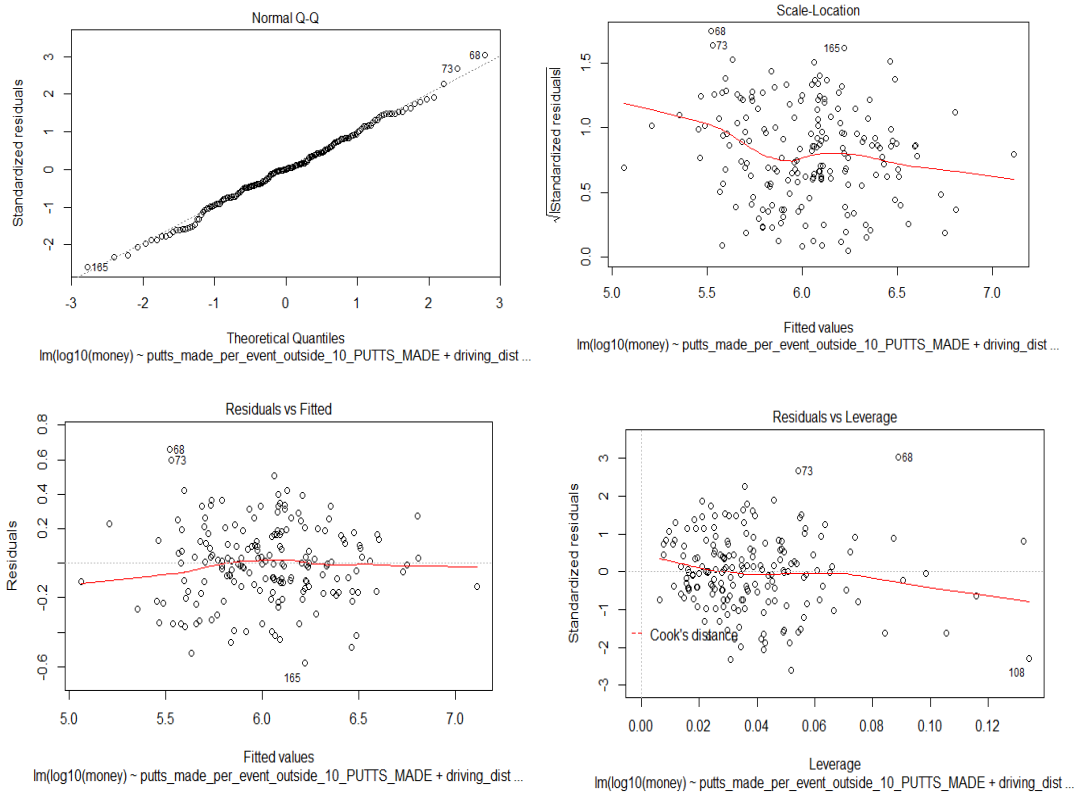


Figure 3: Regression assumptions visuals – PGA 2015 Model

3. RESULTS

Given the ranking nature of the output non-parametric rank statistics were applied to evaluate the models predictive power. The following non-parametric ranking statistics were used: 1) Spearman's rank correlation - a non-parametric measure of rank correlation that assesses how well the relationship between two variables can be described using a monotonic function and 2) Wilcoxon signed-rank - a non-parametric statistical hypotheses test to determine whether two dependent samples were selected from populations having the same distribution (i.e. a paired difference test). The Wilcoxon (table 3) and Spearman's rank (table 4) tests across each of the regression and random forest models (2011-2018) shows no statistical difference between the actual and predicted rank, and strong positive correlation between predicted and actual rank. The p-values suggest that the actual and predicted rank are from identical populations; therefore, have the sample distribution. Unsurprisingly, the largest Spearman correlation exists within the 2013 predictions i.e. applying the 2012 model to 2013 PGA data. This is because the most important metrics across these two seasons are similar, except the driving metric.

The 2012 r-squared value was the third largest (0.6839) of the 7 seasons. After the 2013 season the ranking power of the regression models drops off illustrated, by declining r-squared values, this is because the important metrics across seasons, after 2013, start to differ. The random forest 'importance analysis revealed that putting and accuracy performance metrics were relatively more important than those associated with driving and short game. It was concluded that driving metrics are the least important of the four groups. Table 1 reveals *putts made per event outside 10* and *birdies or better percentage fairway* were most important putting metrics, *scrambling* and *scrambling from rough* were most important short game metrics, *good driving percentage* and *green in regulation* were most important accuracy metrics, while *go for green* were the most important driving metrics. Comparing the predictive performance of the two models it was observed that the ranking accuracy of the models did not significantly change between the regression and random forest models. However, modelling accuracy decreases when applying the random forest model suggesting that interaction affects are not important when ranking a golfer's performance and individual metrics should be treated in isolation. This suggests that a golfer needs an all-round game and can not only focus on a singular aspect.

Year	Regression p- values	Random Forest p- values
2011	0.7374	0.7566
2012	0.9412	0.9412
2013	0.9942	0.8394
2014	0.8663	0.6212
2015	0.8566	0.6837
2016	0.7656	0.8673
2017	0.9465	0.9858
2018	0.4858	0.2344

Year	Spearman's - Regression	Spearman's - Forest
2011	0.7862	0.7663
2012	0.7572	0.7467
2013	0.8086	0.7751
2014	0.7685	0.7891
2015	0.7423	0.7438
2016	0.7634	0.7326
2017	0.7644	0.7694
2018	0.7462	0.7388

Tables 3 & 4: Wilcoxon p-values per year (left), Spearman's rank coefficient (right)

4. DISCUSSION AND CONCLUSIONS

Although no obvious model improvements were observed between the regression and random forest model, there were consistent "important" metrics across seasons: 1) scrambling, 2) scrambling from rough, 3) putts made per event outside 10 putts made, 4) green in regulation percentage and 5) birdies or better percentage fairway. These observations suggest that driving metrics are the least important, while putting metrics are the most important. These findings adhere to the adage: "*Drive for show, putt for dough"*. Moreover, the driving metric coefficients were generally found to be the least significant factor of money earned (table 2). These findings are like those outlined in Heiny (2008). Heiny (2008) found scrambling, green in regulation and driving accuracy to have the greatest relationship with scoring average and $log(season\ money)$. Given the increasing need to understand a player's true value and potential, and the lack of predictive systems ranking golfer performance, this paper introduces a novel and robust ranking system that presents a set of key performance indicators that coaches, pundits and other stakeholders should consider to evaluate player performance.

The research identifies the building blocks to construct a predictive model for a golfers rating. Fundamentally a golfer needs an all-round game, and this work illustrates the type of metrics that have the greatest impact on performance. Model assumptions across the study period were not violated enabling the research to identify a novel, statistically robust PGA performance ratings framework. While the framework is statistically sound, the model assumes all tournaments are equal. Future work will attempt to improve the model by incorporating tour-by-tour data. This will allow the effects of course specific metrics, such as overall course difficulty, tournament money, and how a player's style fits a course, to be included in the model alongside a player's inherent ability. This will also allow us to update player rankings during the season to predict tournament performance, rather than a season ranking. Moreover, models built using tour-by-tour data will produce better predictions as it contains more granular information relative to season data, and accounts for greater variation in player performance.

References

Ayers, R. (2018, January 24). How Big Data Is Revolutionizing Sports. Retrieved from http://dataconomy.com/2018/01/big-data-revolutionizing-favorite-sports-teams/

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Broadie, M., & Rendleman, R. J. (2013). Are the official world golf rankings biased? a. *Journal of Quantitative Analysis in Sports*, *9*(2), 127-140.

Broadie, M. (2012). Assessing golfer performance on the PGA TOUR. *Interfaces*, *42*(2), 146-165.

Price Waterhouse Coopers, (2015). PWC outlook for the global sports market to 2015. Report P-25,

Connolly, R. A., & Rendleman Jr, R. J. (2012). What it takes to win on the PGA TOUR (if your name is "Tiger" or if it isn't). *Interfaces*, *42*(6), 554-576.

Crosue, K (2017, March 4). The No. 1 Ranking in Golf: What Does It Mean to the Players? Retrieved from https://www.nytimes.com/2017/03/04/sports/golf/world-golf-rankings-number-1.html

Dusek, D. (2017, March 3). Golf returns to MIT Sloan Sports Analytics Conference. Retrieved from https://golfweek.com/2017/03/03/golf-returns-to-mit-sloan-sports-analytics-conference/

Heiny, E. L. (2008). PGA tour pro: Long but not so straight. *Chance*, *21*(1), 11-21.

LEVIN, A. (2017). Ranking the Skills of Golfers on the PGA TOUR using Gradient Boosting Machines and Network Analysis. *MIT Sloan Sports Analytics Conference*.

The Economist (2017, August 9). The EAGLE takes flight. Retrieved from https://www.economist.com/game-theory/2017/08/09/the-eagle-takes-flight

Ryan, S (2015, September 14). Golf's World Rankings Are Almost Perfect, And If You Don't Get Them, That's Your Problem. Retrieved from https://www.golfdigest.com/

On the Suncorp Super Netball 2018 Ladder Points System

Anthony Bedford [a,b,c] , Noeline Taurua [b] , Kylee Byrne [b]

[a] *University of the Sunshine Coast*
[b] *Sunshine Coast Lightning*
[c] *Corresponding author: abedford@usc.edu.au*

Abstract

The aim of this paper is to analyse recent rule changes pertaining to bonus points as agreed upon by Netball Australia. The system introduces points for winning a quarter. The analysis suggests a number of scenarios that should be considered so that there is minimal impact to the competition, and to avoid the very high likelihood of action by teams against the competition given contentious scenarios.

There are two key points to the direction of this work: The value of a game – scenarios that can occur when comparing two games' outcomes that are seen as inequitable and the impact on a season – what can occur based upon variations in results that affect the sense of fairness in light of winning games.

There is a multiplicity of possibilities that underlie this work, however for effectiveness we present some of the more probable scenarios foreseen against the actualities of the 2018 season. We look at how behaviours have changed, and how teams affected used the system to their advantage or suffered due to its weaknesses. We consider the effects on the outcomes of a game, and how the season is changed by these in-game point allocations.

Keywords: Points Systems, Fairness, Match fixing

1. INTRODUCTION

To minimise the computational difference with season 2017, the simulations undertaken are extremely constrained, and as such, help bare comparison to that season. The simulation built allows users to modify with equal likelihood a random change of score (+/- a threshold) for any given quarter for all games last season.

However, in this truncated paper, we viewed the least impact on last season's outcomes by varying the actual points scored in the first quarter solely in every game between -2 and 2 points. Notably, it did not take a great deal of variation from last season to yield some results that may agitate teams.

The changes from the league are detailed basically:
Win – Four points; Draw – two points; Losing by five goals or less – one point; Winning a quarter – one point. The authors note that the league dropped the bonus point for a close loss a few weeks before the season started. In this work we shall look at the effect the changes for the seasons from 2009-2016.

2. METHOD

A notable consideration we must make is that teams are ultimately going to change their behaviour based upon this rule. There will be knowledge of the rule before teams take the court, and the ability to manipulate one's own game to favours one's desired outcome is probable. This is a big detail we acknowledge upfront as a constraint. A second constraint is that we consider every team is trying to win, and whilst not covered here, the rules changes and circumstances may allow for the possibility of tanking sections of a game for gain given the bonus points system (i.e. losing 2 quarters, or drawings quarters to promote an opponent, or losing one's own game but receiving reward to fix one's ladder position).

As to not make this too copious, the simplest a smallest change was implemented to see what effect the new rules may have. This is Case A: Change the first quarter score of randomly by either 2 less goal,

no change, or two more goals. Everything else remains exactly the same (i.e. quarter two, three and four scores).

3. RESULTS

3.1: Scenarios for consideration – the outcome of a match

There is now a wide array of possible ladder points outcomes in a given game, yielding an elegant number of possible values, ranging from 0 points to 8 points. This has opened up the interesting scenarios in sometimes an unbalanced nature of outcomes – that is – that not every game will yield the total number of points from its outcome. Running 50,000 simulations, below is a table including the estimated events in season. This is a more hypothetical scenario modelling but does unveil some scenarios for consideration.

Now to some scenarios generated that may cause some concern:

Points	Result		Differential	Comments	Estimated Event in Season
8-0	W	4W:0D:0L-W	8	Winner: Wins4Q, wins by more than 5	1 in 10 games
7-0	W	3W:1D:0L-W	7	Winner: Wins3Q and Draws 1Q, wins by more than 5	1 in 17 games
7-1	W	3W:1D:0L-W	6	Winner: Wins3Q and Draws 1Q, wins by less than 6	1 in 4 games
7-1	W	3W:0D:1L-W	6	Winner: Wins3Q, wins by more than 5	1 in 4 games
6-0	W	2W:2D:0L-W	6	Winner: Wins2Q and Draws 2Q, wins by more than 5	1 in 91 games
7-2	W	3W:0D:1L-W	5	Winner: Wins3Q, wins by less than 6	1 in 15 games
6-1	W	2W:2D:0L-W	5	Winner: Wins2Q and Draws 2Q, wins by less than 6	1 in 13 games
6-1	W	2W:1D:1L-W	5	Winner: Wins2Q and Draws 1Q, wins by more than 5	1 in 13 games
5-0	W	1W:3D:0L-W	5	Winner: Wins1Q and Draws 3Q, wins by more than 5	1 in 1667 games
6-2	W	2W:1D:1L-W	4	Winner: Wins2Q and Draws 1Q, wins by less than 6	1 in 6 games
6-2	W	2W:0D:2L-W	4	Winner: Wins2Q, wins by more than 5	1 in 6 games
5-1	W	1W:3D:0L-W	4	Winner: Wins1Q and Draws 3Q, wins by less than 6	1 in 208 games
5-1	W	1W:2D:1L-W	4	Winner: Wins1Q and Draws 2Q, wins by more than 5	1 in 208 games
6-3	W	2W:0D:2L-W	3	Winner: Wins2Q, wins by less than 6	1 in 7 games
5-2	W	1W:2D:1L-W	3	Winner: Wins1Q and Draws 2Q, wins by less than 6	1 in 70 games
5-2	W	1W:1D:2L-W	3	Winner: Wins1Q and Draws 1Q, wins by more than 5	1 in 70 games
5-3	W	1W:1D:2L-W	2	Winner: Wins1Q and Draws 1Q, wins by less than 6	1 in 29 games
5-3	D	3W:0D:1L-D	2	Draw, however 1 team is rewarded for more Q wins	1 in 29 games
5-4	W	1W:0D:3L-W	1	Winner: Wins1Q and Loses 3Q, wins by less than 6	1 in 63 games
4-3	D	2W:1D:1L-D	1	Draw, however 1 team is rewarded for more Q wins	1 in 156 games
4-4	D	2W:0D:2L-D	0	Draw, with 2Q each	1 in 54 games
3-3	D	1W:2D:1L-D	0	Draw, with 1Q each	1 in 714 games
2-2	D	0W:4D:0L-D	0	Draw, no extra rewards	1 in 5000 games

Table of possible outcomes.

3.2 COMPARISON OF DRAWS

Below we have the five types of Draws, exampled side by side. The largest margin draw is the last game, with the Giants coming back to draw and take 3 points. The closest of all games receives the lowest points (2-2).

	Q1	Q2	Q3	Q4	FINAL	Margin	Quarters		Q1	Q2	Q3	Q4		
Lightning	18	16	12	15	61	0	WDLW	H	1	0	0	1	2	0	4
Magpies	17	16	16	12	61	0		A	0	0	1	0	2	0	3
	W	D	L	W											
	Q1	Q2	Q3	Q4	FINAL	Margin	Quarters		Q1	Q2	Q3	Q4		
Swifts	18	16	14	11	59	0	WDDL	H	1	0	0	0	2	0	3
Giants	17	16	14	12	59	0		A	0	0	0	1	2	0	3
	W	D	D	L											
	Q1	Q2	Q3	Q4	FINAL	Margin	Quarters		Q1	Q2	Q3	Q4		
Firebirds	18	16	12	16	62	0	WLLW	H	1	0	0	1	2	0	4
Thunderbirds	17	17	16	12	62	0		A	0	1	1	0	2	0	4
	W	L	L	W											
	Q1	Q2	Q3	Q4	FINAL	Margin	Quarters		Q1	Q2	Q3	Q4		
Swifts	17	16	14	14	61	0	DDDD	H	0	0	0	0	2	0	2
Giants	17	16	14	14	61	0		A	0	0	0	0	2	0	2
	D	D	D	D											
	Q1	Q2	Q3	Q4	FINAL	Margin	Quarters		Q1	Q2	Q3	Q4		
Swifts	17	16	14	14	61	0	WWWL	H	1	1	1	0	2	0	5
Giants	16	15	11	19	61	0		A	0	0	0	1	2	0	3
	W	W	W	L											

3.3 ONE GOAL WORTH A LOT MORE

In this example, the first game yields the lowest points possible for a game. The second has the Magpies scoring one more goal to win the game – the result a boost of 3 points

	Q1	Q2	Q3	Q4	FINAL	Margin	Quarters		Q1	Q2	Q3	Q4			
Lightning	18	16	15	15	64	0	DDDD	H	0	0	0	0	2	0	2
Magpies	18	16	15	15	64	0		A	0	0	0	0	2	0	2
	D	D	D	D											
	Q1	Q2	Q3	Q4	FINAL	Margin	Quarters		Q1	Q2	Q3	Q4			
Lightning	18	16	15	15	64	-1	DDDL	H	0	0	0	0	0	1	1
Magpies	18	16	15	16	65	1		A	0	0	0	1	4	0	5

3.4 Scenarios for consideration – the outcome of a season

With such variation in total points possible for a team, we should expect more variation in the ladder. Hypothetically, an invincible team could feasibly earn 112 points; its antonym 0 points, in a season. The variance shall certainly create interest throughout the season.

However, this added variance brings more opportunity for a lot of lucky and luckier losers (and unlucky winners!) as some wins are worth more than others.

Below is the given table for the 2017 season as at Round 14, with the new rules in place, as provided. As seen, there is no change in ladder position for any team.

HYPOTHETICAL LADDER																
Actual			Match Results					Bonus Points				Ladder	Goals			
Ladder	#	Team	Played	Win	Loss	Draw	Qtr Wins	WS	Pts	LM	Pts	Points	For	Against	Diff	%
1	1	Melbourne Vixens	14	11	2	1	39	6	0	1	1	86	870	744	+126	116.94%
2	2	Sunshine Coast Lightning	14	11	2	1	33	1	0	1	1	80	808	726	+82	111.29%
3	3	GIANTS Netball	14	10	4	0	28	1	0	3	3	71	773	729	+44	106.04%
4	4	Magpies Netball	14	9	5	0	27	0	0	3	3	66	770	731	+39	105.34%
5	5	Queensland Firebirds	14	7	6	1	28	1	0	3	3	61	780	758	+22	102.90%
6	6	NSW Swifts	14	3	10	1	18	0	0	5	5	37	726	792	-66	91.67%
7	7	West Coast Fever	14	2	12	0	17	0	0	4	4	29	671	777	-106	86.36%
8	8	Adelaide Thunderbirds	14	1	13	0	12	0	0	3	3	19	651	792	-141	82.20%

2017 Ladder under new rules for 2018

One aspect that would have been possible was the Round 13 standings as shown below. There was a chance that the Firebirds could have played finals if the Magpies were to lose to the Thunderbirds in their final match.

It is noted that the Magpies won by 2 points, however if that was a loss, and the Firebirds were to win that game, there was a chance for the Firebirds to make the finals despite winning less games.

HYPOTHETICAL LADDER																
Actual			Match Results					Bonus Points				Ladder	Goals			
Ladder	#	Team	Played	Win	Loss	Draw	Qtr Wins	WS	Pts	LM	Pts	Points	For	Against	Diff	%
1	1	Melbourne Vixens	13	10	2	1	35	5	0	1	1	78	804	702	+102	114.53%
2	2	Sunshine Coast Lightning	13	10	2	1	31	1	0	1	1	74	752	671	+81	112.07%
3	3	GIANTS Netball	13	10	3	0	26	1	0	2	2	68	718	673	+45	106.69%
4	4	Magpies Netball	13	8	5	0	24	0	0	3	3	59	721	684	+37	105.41%
5	5	Queensland Firebirds	13	6	6	1	25	0	0	3	3	54	713	710	+3	100.42%
6	6	NSW Swifts	13	3	9	1	17	0	0	5	5	36	678	725	-47	93.52%
7	7	West Coast Fever	13	2	11	0	17	0	0	4	4	29	629	711	-82	88.47%
8	8	Adelaide Thunderbirds	13	1	12	0	11	0	0	2	2	17	604	743	-139	81.29%

Round 13, 2017 Ladder scenario.

Now from simulation, a number of example emerged that are great cause for concern.

Case A: Firebirds make finals despite lower percentage. This is a product of the rule change – quarters usurp percentage as a first tiebreaker. Magpies miss finals

HYPOTHETICAL LADDER															
Actual			Match Results					Bonus Points				Ladder	Goals		
Ladder	#	Team	Played	Win	Loss	Draw	Qtr Wins	WS	Pts	LM	Pts	Points	For	Against	Diff
1	1	Melbourne Vixens	14	11	2	1	38	6	0	1	1	85	866	743	+123
2	2	Sunshine Coast Lightning	14	11	2	1	34	1	0	1	1	81	806	720	+86
3	3	GIANTS Netball	14	9	4	1	27	1	0	3	3	68	766	737	+29
5	4	Queensland Firebirds	14	7	5	2	29	1	0	2	2	63	784	752	+32
4	5	Magpies Netball	14	7	5	2	26	0	0	3	3	61	768	727	+41

Giants: Lightning lose double chance

HYPOTHETICAL LADDER																
Actual			Match Results					Bonus Points				Ladder	Goals			
Ladder	#	Team	Played	Win	Loss	Draw	Qtr Wins	WS	Pts	LM	Pts	Points	For	Against	Diff	%
1	1	Melbourne Vixens	14	12	2	0	39	5	0	1	1	88	882	737	+145	119.67%
3	2	GIANTS Netball	14	11	3	0	34	1	0	0	0	78	795	711	+84	111.81%
2	3	Sunshine Coast Lightning	14	11	3	0	32	3	0	0	0	76	807	712	+95	113.34%

Case C: Giants finish 3rd due to winning less quarters despite MORE wins than BOTH teams above. GIANTS lose double chance and lose home qualifying final

HYPOTHETICAL LADDER																
Actual			Match Results					Bonus Points				Ladder	Goals			
Ladder	#	Team	Played	Win	Loss	Draw	Qtr Wins	WS	Pts	LM	Pts	Points	For	Against	Diff	%
2	1	Sunshine Coast Lightning	14	11	3	0	35	3	0	2	2	81	827	713	+114	115.99%
3	2	Melbourne Vixens	14	10	4	0	38	6	0	2	2	80	867	748	+119	115.91%
1	3	GIANTS Netball	14	12	2	0	32	1	0	0	0	80	778	707	+71	110.04%

Case D: Giants instead of missing finals get to host the elimination final despite winning less games than Magpies and Firebirds. Firebirds miss finals and Magpies miss home final.

HYPOTHETICAL LADDER																
Actual			Match Results					Bonus Points				Ladder	Goals			
Ladder	#	Team	Played	Win	Loss	Draw	Qtr Wins	WS	Pts	LM	Pts	Points	For	Against	Diff	%
1	1	Melbourne Vixens	14	11	3	0	35	3	0	3	3	82	880	738	+142	119.24%
2	2	Sunshine Coast Lightning	14	11	3	0	34	5	0	2	2	80	820	717	+103	114.37%
5	3	GIANTS Netball	14	7	7	0	30	1	0	5	5	63	745	712	+33	104.63%
3	4	Magpies Netball	14	8	5	1	27	3	0	1	1	62	782	737	+45	106.11%
4	5	Queensland Firebirds	14	7	6	1	26	1	0	2	2	58	773	761	+12	101.58%

Case E: Lightning finish higher due to quarters won. The quarters usurping percentage.

HYPOTHETICAL LADDER																
Actual			Match Results					Bonus Points				Ladder	Goals			
Ladder	#	Team	Played	Win	Loss	Draw	Qtr Wins	WS	Pts	LM	Pts	Points	For	Against	Diff	%
1	1	Melbourne Vixens	14	13	1	0	38	7	0	0	0	90	879	736	+143	119.43%
3	2	Sunshine Coast Lightning	14	9	5	0	34	1	0	2	2	72	799	728	+71	109.75%
2	3	Magpies Netball	14	9	5	0	30	2	0	2	2	68	782	705	+77	110.92%

Case F: Lightning and Giants finish higher due to quarters won with double impact. So Magpies with 9.5 wins yet miss qualifying final, losing double chance and travelling to elimination final. Giants make finals on quarters won and knock out Firebirds who had a superior percentage

HYPOTHETICAL LADDER																
Actual			Match Results					Bonus Points				Ladder	Goals			
Ladder	#	Team	Played	Win	Loss	Draw	Qtr Wins	WS	Pts	LM	Pts	Points	For	Against	Diff	%
1	1	Melbourne Vixens	14	11	3	0	34	5	0	2	2	80	859	749	+110	114.69%
3	2	Sunshine Coast Lightning	14	9	5	0	33	3	0	2	2	71	791	739	+52	107.04%
5	3	GIANTS Netball	14	9	5	0	29	1	0	3	3	68	796	752	+44	105.85%
2	4	Magpies Netball	14	9	4	1	27	1	0	2	2	67	779	735	+44	105.99%
4	5	Queensland Firebirds	14	9	5	0	26	3	0	2	2	64	801	749	+52	106.94%

Case G: Lightning and Giants finish higher due to quarters won with double impact. So Magpies with 10.5 games clear 2nd miss qualifying final, losing double chance. Giants make finals on quarters won despite having less wins than Firebirds below.

HYPOTHETICAL LADDER																
Actual			Match Results					Bonus Points				Ladder	Goals			
Ladder	#	Team	Played	Win	Loss	Draw	Qtr Wins	WS	Pts	LM	Pts	Points	For	Against	Diff	%
1	1	Melbourne Vixens	14	11	3	0	38	5	0	3	3	85	870	740	+130	117.57%
3	2	Sunshine Coast Lightning	14	10	4	0	34	1	0	2	2	76	810	736	+74	110.05%
2	3	Magpies Netball	14	10	3	1	29	1	0	2	2	73	802	740	+62	108.38%
5	4	GIANTS Netball	14	8	6	0	29	1	0	3	3	64	783	753	+30	103.98%
4	5	Queensland Firebirds	14	8	5	1	25	1	0	1	1	60	789	773	+16	102.07%

Case H: Magpies and Firebirds finish higher due to quarters won with double impact. So Lightning with 10 games equal 2nd miss qualifying final, losing double chance, and Giants miss finals despite superior wins to not only 4th placed but 2nd placed team!

HYPOTHETICAL LADDER																
Actual			Match Results					Bonus Points				Ladder	Goals			
Ladder	#	Team	Played	Win	Loss	Draw	Qtr Wins	WS	Pts	LM	Pts	Points	For	Against	Diff	%
1	1	Melbourne Vixens	14	11	3	0	34	3	0	3	3	81	847	743	+104	114.00%
4	2	Magpies Netball	14	9	5	0	33	2	0	4	4	73	786	730	+56	107.67%
2	3	Sunshine Coast Lightning	14	10	4	0	29	2	0	1	1	70	809	733	+76	110.37%
5	4	Queensland Firebirds	14	9	5	0	31	0	0	2	2	69	795	750	+45	106.00%
3	5	GIANTS Netball	14	10	4	0	28	0	0	0	0	68	753	737	+16	102.17%

4. DISCUSSION

A number of notable possibilities occur within the system/s. Importantly, there are a great deal of permutations around the game's outcome and the quarters bonus points, and margin to a lesser extent. It is clear to perceive this may have a large influence on a season's possible result.

It is also worth noting that no assertion has been indicated around tie-breaking rules involving equal points – shall ladder position divert to greater wins, then percentage, then head-to-head?

The purpose here is to highlight a broad array of not so unlikely outcomes that could see a degree of ruckus amongst competing teams. Whilst entertainment and jubilation are at the heart of athletic endeavours, they should not be sacrificed for fairness and equity, and the possibility of criticism or disrepute to the great game. It might be perceived that there is the opportunity for some manipulation, and unfairness, when winning the game is simply not enough.

The initiatives here are noted as incentivising teams to win every quarter they play. Further, a team that might not be winning but going well are rewarded for their efforts through bonus points. In fact, a losing team can receive 4 points for a loss, only one point behind a 'poor' win that would receive 5. (This was in fact the exact scenario of the World Cup final in Sydney).

The notion of fixing is one that is a more serious temptation, however again possible within the rules of the game. A simple example is when two teams are vying for a spot against a given competitor, and one team can 'control' the outcome of the game to benefit themselves or their opponent. Collusion is also a temptation, with 'Quarter fixing' a neat option to veil possible outcome choices (so a team can still win however let their opponent win a quarter/s to gain greater points).

The use of the quarters bonus points seems to yield the largest issue, and one could consider replacing the percentages for the quarters won as a tiebreaker scenario rather than an input into the points. This is akin to other systems that (sadly) exist in netball such as using the head-to-head result to decide ladder when points are tied to decide finals placings.

One other possibility to improve the obvious Draw inequity is extra time. The NHL have a reward for a full-time loss, with the game going into Penalties to decide the winner, however the loser is rewarded. (NHL has 2 points for a win and 1 point for an overtime loss, no draws).

If you look back through the last nine seasons, the finals placings and make-up of the four have come down to the last match and/or round. It is possible that you could actually have a final four confirmed weeks out from the last round due to a large number of points between teams and therefore less interest in the final matches. Additionally, teams won't be playing for places when finishing the season. With the current ladder points a win could determine a jump in places and teams moving in and out of the four.

If we are positioning this competition as world class, best in the world and especially "elite and high performance" where does the value of a win come in with these proposed changes if losing teams can come out of it in a better position than the winning team. In sport you either win, lose or draw and our match is 60 minutes plus potentially extra time, not small matches of 4 x 15 minutes.

From an athlete development point of view, if the focus is on winning quarters you will find teams will run their best line up and make very few changes to chase points for the ladder. Henceforth, less athletes will be exposed to this level of competition and our depth at this level and then International levels will suffer. If these proposed

changes happen, and players 8-10 are getting less time at SSN level, the opportunity to play ANL needs to be available every weekend, and at this stage it's not. The movement of athletes back to ANL for court time needs more thought and integration and there is no time this year for that to happen and changes to both competitions to be made.

Netball as a sport and its coaching are evolving to a space where some athletes could be "impact" athletes for certain amounts of time, which was changing structures on court, and hopefully making it more exciting. The changes in ladder points would again mean limited changes would be made and therefore only showcasing athletes who could run out players for an entire match to chase points again.

5. CONCLUSION

In conclusion, it is noted that whilst this does exist as a theoretical piece, serious consideration of the outcomes is required. Retrospective analysis highlights the possibility of such outcomes, with 2 in 3 seasons analysed yielding significant potential causes for concern. The athlete focus and dynamism of the game are potentially adversely impacted through a desire to maximise points rather than simply maximising a win. Sport is unfortunately riddled with such decisions that have fallen over in the chase for excitement, and with SSN the showcase of the sport, it would be prudent to consider the ramifications of this pending change.

A FLEXIBLE METHOD OF JUMP AND HIGH INTENSITY EVENT DETECTION

Paul Smith and Anthony Bedford.
University of the Sunshine Coast
Corresponding author: paul.smith@research.usc.edu.au

Abstract

With success in competitive sports often relying on players making quick, explosive movements High Intensity Events (HIE's) have become fundamental in assessing athlete performance. Their detection has traditionally relied on video analysis but with many athletes and elite teams now wearing inertial sensors to record their movement there is a demand to develop classification algorithms that find a balance in accuracy between specificity of the athlete, and robustness of detection over a range of athletes or HIEs. In this research we develop a novel approach to detect jumps in a complex sporting environment and apply this method to other sporting applications. Using elite netball as a test case acceleration data from a training session was analysed and key parameters of the event were defined. Code was developed using a single acceleration axis to automatically detect jumps as well as identify the time it occurred and several performance indicators including flight time and maximum acceleration at take-off and landing. Applying the code to an elite netball match it detected 100% of jumps identified by video analysis with no false positives. Time of take-off and landing were within 0.04sec being 1 frame of video taken at 25 frames/sec. The unchanged code was then tested in several sports and successfully classified other HIEs in athletics, and water sports. The specificity and robustness of this method will have wide practical applications in the detection of jumps and the potential to identify other HIEs in many sports. This will enable coaches and sport scientists to classify these events with no pre-conditioning of the data, minimal change to the model, and have the potential to produce results in real time. This method also shows great potential in the wider application of human movement detection in the general fitness and health sectors.

Keywords: High Intensity Events, athletic jumps, IMU, machine learning, sports performance measurement, human movement detection, gait, netball, hurdles, tumble turn.

1. INTRODUCTION

Understanding the physical effort involved in sport, or more specifically the workload of individual positions on a team has become essential to maximise team performance (Rodriguez-Alonso, 2003; Di Salvo 2007). Time-motion studies using video review have become common place in many sports in an attempt to quantify parameters such as distance travelled, time and distance running and sprinting, rest time between high intensity events, or number of tackles (Steele 1991; Adbelkrin 2006; Di Silvo 2007; Davidson 2008; Barris 2008; Gabbett 2008; Hartwig 2011). This technique however is limited due to the time intensive and subjective nature of parameter classification (Barris 2008). Despite advancements in semi-automated video systems to reduce the processing time (Sykes 2009) they still do not allow real-time analysis of athlete performance and have remained price restrictive for its wider application (Randers 2010). Recent developments in microprocessor and battery technology has streamlined the data collection and analysis process by allowing sports scientists to attach electronic data logging devices to athletes. These Inertial Measurement Units (IMUs) containing three axis accelerometers and gyroscopes can now be used to objectively record athlete movement in three dimensional space. By understanding the acceleration along the horizontal, vertical and lateral planes the force produced by the athlete to undergo locomotion can be determined and allow analysis of movement over a range of temporal and spatial scales. This has resulted in the application of IMUs to become commonplace at the elite level in many sports with coaches and training staff now able to quantify player movement including work load in a match or the work rate over a given time period (Di Salvo 2007; Boyd 2011; Chandler 2014, Chambers 2015).

There has been a growing body of research in the use of IMU devices including those contained in smart phones in the field of human activity recognition. With its main application in the health and leisure industries they have been used to automatically detect activities including sleeping, sitting, walking, jogging and climbing stairs (see Xu 2014, Yang 2010). However further development and application of algorithms in sport is required and once developed should be validated for their ability to classify team sport activities (Wundersitz 2015). The acceleration profile of HIEs such as jumping has become well understood (fig 1.) however detecting these in a competitive sport environment is particularly problematic due to the erratic movement of the athlete

throughout the match and digitisation effects of the signal. To resolve this issue automatic activity recognition in the sports domain has been limited to athletes performing a set routine in a laboratory environment (Mitchell 2013; Moran 2015; Wundersitz 2015) or used to identify a high workload event from periods of generally low workload (Ahmadi 2009; McNamara 2015; Ghasemzadeh 2011). These studies have also relied on the use of multiple IMU devices, multiple channels, and pre-processing of the raw data (see Chambers et. al. 2015, Camomilla 2018). To allow for the inherent noise in the signal and variations in execution of the event machine learning techniques have been used with varying success to classify HIEs. Methods such as Random Forest, Support Vector Machines, Logistic Tree Method, and fuzzy logic have been used as the modellers try to find a balance between specificity of the model to achieve the highest event classification accuracy for the athlete, and generality to achieve the highest classification accuracy over a range of athletes.

To resolve the limitations of previous work on human movement classification this research proposes a novel approach that only requires a single channel of raw sensor data and yet provides a model that can be used on a wide range of HIEs. Using jumps as a test case, raw vertical acceleration data will be used to define key parameters for the event. Then implementing a step-wise approach using the Visual Basic coding language, it will test if the given set of parameters for the event is satisfied. The proposed classification method of analysing each sample point gives the model the power to scrutinise the data at the fine scale, yet has the ability to extract trends over multiple time scales. The aims of this study were threefold. Firstly to develop a method that can detect jumps in a complex competitive sports match. Secondly, be robust enough to apply this model to classify jumps in a non-field/court sport. Thirdly, be general enough to detect other HIEs with no change to the model.

2. METHODS

Subjects

Four groups of participants were used for this study. For building the model the subject was a female elite netballer (aged 28 , 167cm tall). For testing the model in a team field/court sport environment data gathered from ten female elite netballers were analysed (age 26.2, S.D 4.4; height 1.81m, S.D. 0.10). For testing the jump detection model in athletics two state-level athletes were recruited, one male age 17, height 1.83m, and one female aged 17, height 1.65m. To test the models ability to detect another type of HIE one female recreational swimmer age 15, height 1.62m was recruited.

Design

Establishment of the jump detection model consisted of video and data analysis taken during a one hour fitness and netball skill session. The participant wore an IMU (Catapult s4, Catapult Sports Australia) between their shoulder blades in a custom made vest over their training top. The IMU contained a three axis accelerometer and gyroscope recording at 100hz on each channel. The session comprised of a series of set routines on a hard-board indoor court and included drills with and without a ball. The video was aligned to the IMU data using Adobe Premiere Pro (CS6, Adobe Systems) and Matlab (2017b, version 9.3, The MathWorks Inc., USA).

Using the video as reference the data was edited into four basic sports activities being walking, jogging, sprinting (see Fox et al. 2013 for definitions) and jumping. For this study jumping was defined as an upward displacement where the athlete makes an explosive movement on the vertical plane resulting in an extended period where both feet are off the ground. Each movement activity was separated into a single event being either a jump, or a step for walking, jogging and sprinting. To facilitate event classification mean and standard deviation were calculated for peak vertical acceleration at take-off, peak vertical acceleration on landing, minimum vertical acceleration between peaks, and time between peaks using a customised Matlab code.

Methodology

To create the jump detection model, time-stamped raw vertical acceleration data from the netball training session was imported into Microsoft Excel (Microsoft Excel 2007, Microsoft). Code was written in Visual Basic (Visual Basic Studio 2007 v6.5, Microsoft) and deployed through a macro. Stating at T=0 the vertical acceleration value was checked. If it was over the minimum take-off threshold for jumping the T+1 to T+n values were checked to confirm a decreasing trend towards the threshold of the flight phase (-1g). If this condition was satisfied the acceleration values were sequentially checked until the next peak in vertical acceleration was located. After recording this local maximum sequential acceleration values were checked to examine if there was a trend towards 0.0g. This movement activity was classified as a jump if both peaks were above the respective thresholds for take-off and landing, and the time between peaks was within the flight time window for a jump. Peak accelerations, flight time, and time of the event were then pasted into a table within Excel. This step-wise process would repeat from T+1 until the end of the data was reached. See Figure 2 for code overview.

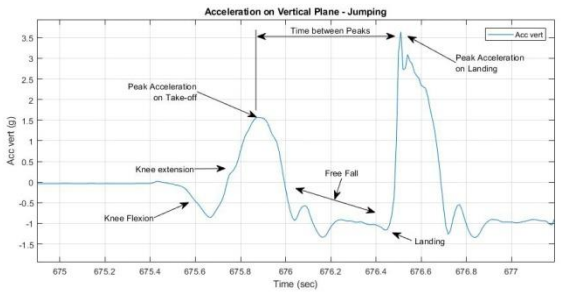


Figure 1. Vertical acceleration profile of a jump showing key phases.

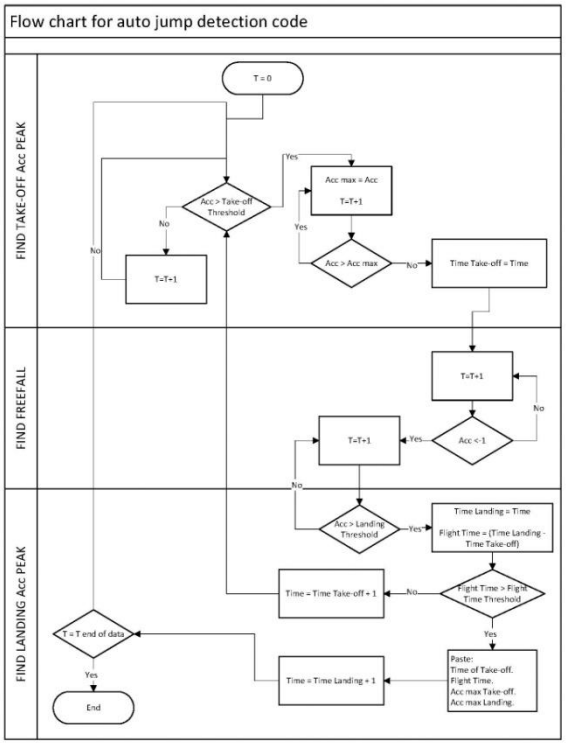


Figure 2. Schematic of code flow to auto-detect High Intensity Events.

Data Analysis

To test the accuracy of the jump detection model the code was run on the training session and an elite level netball match (Sunshine Coast Lightning v Collingwood Magpies). A notational analysis of the video was completed noting each time a player jumped. It was also noted if the player used one or two legs to take-off and land to assess the ability of the code to detect jumps that were performed using different techniques. The raw acceleration data was brought in to Excel and the jump detection macro was run. The video analysis results were then compared to the code output.

To test the generality of the jump detection process the code was tested on two sprint hurdlers. The male participant ran eighteen runs of three hurdles and four runs of five hurdles, and the female ran twenty runs of three hurdles and two runs of five hurdles on a rubberised outdoor athletics track. Hurdle height for the male athlete was 0.91m with a 9.1m separation distance, and for the female athlete hurdle height 0.76m with separation distance of 8.5m for 20 runs, and height 0.50m, separation distance 6.0m for 2 runs. Each athlete wore an IMU (Catapult s4, Catapult Sports Australia) between their shoulder blades in a custom made vest over their training top. An observer was positioned next to the running track with an unobstructed view. The start time of each run was noted along with any other relevant information. The code was run on the vertical acceleration data in Excel with the results compared to the written observations.

To test the ability of the process to detect other HIEs the code was used to detect the movement of a freestyle swimmer making a tumble turn. This event is used to define the start of each lap and calculate the distance swam in a training session. It is also closely analysed by coaching staff to minimise the time taken to change direction and return to pre-turn velocity. A three axis accelerometer (X6-1A, Gulf Coast Data Concepts) logging at 160hz was worn on the lumbar region of the swimmer in a dedicated pocket under their swimsuit. Seventy three minutes of data was analysed that included ten laps as noted by an observer positioned poolside with an unobstructed view. Importing the time and z axis data into Microsoft Excel the unchanged jump detection code was run to detect the tumble turn movement event with the results compared to the notational observations.

3. RESULTS

Indices for peak vertical acceleration on take-off, peak vertical acceleration on landing and time between peaks for each movement activity taken from the netball training session are presented in Table 1a &1b. Significant differences were found for peak vertical acceleration between walking, jogging, sprinting and jumping. As the velocity of the athlete increases from walking to sprinting peak vertical acceleration increased

from 0.47g to 3.39g. Time between peaks was larger for walking (0.50sec) than jogging and sprinting (0.36/0.37sec) for the test data set (Figure 3 & Table1a). Jumping was found to produce the highest values for peak vertical acceleration on take-off and landing, and time between peaks (Table 1b).

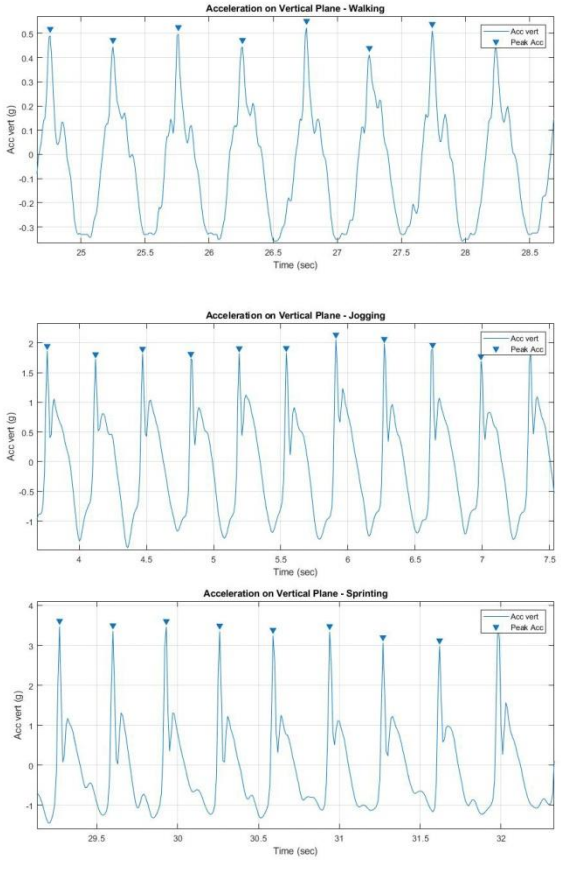


Figure 3. Vertical acceleration profiles of locomotion – walking/jogging/sprinting.

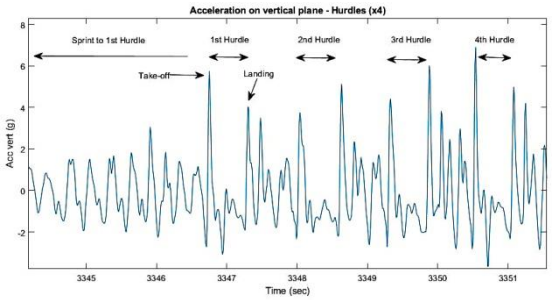


Figure 4. Vertical acceleration profiles of hurdles – Male athlete over 0.91m hurdle.

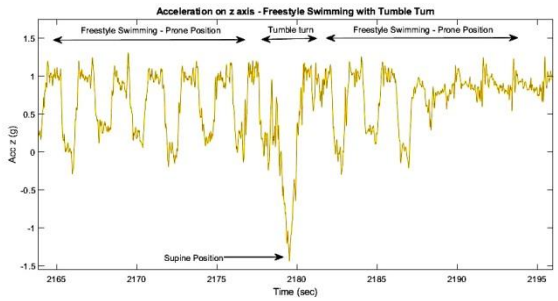


Figure 5. Vertical acceleration profiles of freestyle tumble turn.

Event	Peak Vert Acc/S.D. (g)	Freq./S.D. (sec)	n
Walking	0.47/0.09	0.50/0.023	222
Jogging	1.69/0.27	0.36/0.010	153
Sprinting	3.39/0.54	0.37/0.10	111

Table 1a. Vertical acceleration properties of locomotion.

Peak Vert Acc/S.D (g).		Time between peaks./S.D. (sec)	n
Take-off	Landing		
3.79/0.93	4.17/1.17	0.76/0.21	22

Table 1b. Vertical acceleration properties of jumps taken from training session.

Applying the code to the netball training session and a competitive elite netball match it was successful in identifying all jumps as noted by video analysis (training session 22, match 168). Of the 168 jumps 32 were two leg take-off/two leg land, 28 two leg-take off/1 leg land, 48 one leg take-off/two leg land, 60 one leg take-off/one leg land. No false positive were recorded (Table 2).

The code was successful in identifying 100% of jumps over hurdles (150 from 150). This included 74 jumps over hurdles at a height of 0.91m, 70 jumps over hurdles of height 0.76m, and 6 over 0.5m high (Table 2 & Figure 4). For each hurdle both athletes used their right foot for take-off and left foot for landing. No false positives were identified.

Ten swimming laps including eight tumble turns were completed over seventy three minutes. The code was successful in identifying 100% of tumble turns noted by the pool side observer with no false positives (Table 2 & Figure 5).

High Intensity Event	Number of observed events	
	Observational Analysis	Auto-detection code
Jump (training session)	22	22
Jump* (match conditions)	168	168
Hurdles (0.91m: male athlete)	74	74
Hurdles (0.76m: female athlete)	70	70
Hurdles (0.50m:female athlete)	6	6
Tumble turn	8	8

*Jumps included: 60 1 leg take-off/1 leg landing, 48 1 leg take-off/2 leg landing, 28 2 leg take-off/1 leg landing, 32 2 leg take-off/2 leg landing.

Table 2. Results of Auto-detection code compared to observational recordings.

4. DISCUSSION

The purpose of this research was to present a classification method that was specific enough to detect jumps in the complex environment of an elite sport match and yet had the generality that it could be used to detect other high intensity events in a wide range of sports. While many different methods have been used to automatically detect specific movement activities in sports all to our knowledge have relied on several stages of pre-processing the data, multiple sensors, or multiple channels of sensor data to classify human movement. We are also aware of only one field sports movement, being tackles, which has been conducted in a competitive match environment (Gabbett, 2010, Gastin, 2014). We believe the proposed classification method was successful in achieving these objectives and by analysing the data at the fine scale while identifying trends over multiple time scales shows great potential to be applied in many movement detection applications in the sports and non-sporting environment.

With clear thresholds defined in this study for walking, jogging, sprinting and jumping it was shown that jumping produces the highest accelerations for take-off as the athlete maximises the vertical force applied to the ground to achieve maximum height. This greater height results in the largest velocity on landing, and as athlete resists the upward force of the ground produces the largest peak acceleration with recordings over 8g being observed. It was also shown that jumps that are visually classified as a HIE result in the greatest time period between acceleration peaks. With clear thresholds defined for each activity the step-wise technique developed in this study was able to classify jumps in a competitive match environment with 100% accuracy. Using just the vertical acceleration component of the athlete's motion provided a classification technique independent of the athlete's height or weight. It was also independent of body movement in the lateral, anterior or posterior direction, or rotation during the event. Using this approach it was found that a one or two foot take-off or landing method did not sufficiently change the acceleration profile to the lower values found in sprinting or jogging. It was noted however that landing on one foot did produce a characteristic double spike of negative acceleration as each foot touches the ground. When this landing technique was used it was observed that peak acceleration was reduced due to the increased time that the deceleration of the athletes centre of gravity occurs. However due to the large forces generated in jumping it was still generally above the upper threshold for sprinting, and the time between peak acceleration at take-off and landing were still significantly greater than the other three locomotion events.

The generality of this method was demonstrated by successfully detecting 100% of jumps in a sprint hurdle training session. Sprint hurdlers rely on a jumping technique that minimises both the time spent in the air and height over the hurdles as the try to maximise forward velocity through the run. Despite this the four key parameters of peak vertical acceleration on take-off and landing, the increase in time between peaks from the regular sprinting cadence, and the acceleration transition to -1g from take-off to foot strike were still evident. These patterns were still detected by the code when the female athlete completed two runs of low hurdles set at 0.5m. Despite the athlete not requiring any vertical displacement to clear the hurdle, the action of planting the take-off foot still produced a peak vertical acceleration up to 3.7g, a delay before landing of 0.55 sec as they lifted their legs over the hurdle, and a peak landing acceleration of 5.8g as their landing foot struck the ground. The ability to detect this HIE despite the very specific technique utilised in sprint hurdles and their minimal change in vertical displacement it is proposed that using this detection process should be able to identify other athletics track jumping events seen in 400m hurdles, steeple chase, long jump, triple jump, and pole vault.

The versatility of the described process to detect other HIEs was evident in its ability to successfully classify tumble turns in freestyle swimming without any change to the jump detection code. It was observed that the swimmer undergoes a sudden change of position during the tumble turn as they rotate 180° on the vertical axis from the prone to supine position to touch the wall, then a further 180° rotation back to the prone position to resume swimming. This high intensity movement induced a change in acceleration on the z axis as gravity is subtractive to sensor readings during the freestyle swimming phase, but become additive when the swimmer rotates to the supine position. This creates a spike in z axis acceleration similar to that observed at take-off in the jumping profile. As the swimmer continues to rotate to the freestyle position this produces in a secondary peak

in acceleration due to the change in sensor position in relation to gravity. Although a small sample (8) of tumble turns were detected in this study the ability of the model to analyse the data over a range of scales and the relative effect of gravity to sensor readings each phase of this HIE appears to be clearly identifiable with this method from other movements in freestyle swimming.

5. CONCLUSSION

The findings of this study demonstrate that the described model can be a highly successful method of jump and HIE detection. By defining peak accelerations on a single plane, the time between peaks, and the trajectory of acceleration between peaks the developed code can be used to define events in many sports. It also provides a simple yet powerful analysis tool for movement in sport without the requirement of multiple sensors, any pre-conditioning of the raw data and just a single data channel. These properties make the described method a computationally efficient way to classify HIEs and present itself as an effective method to produce a real-time classification system. It is this power and flexibility that would allow it to be utilised by sports scientists in a wide range of human movement analysis activities, and the simplicity of its application that would enable coaching staff to use the system without any modification to the code. Developing a classification system using this method would have wide applications including enabling performance analysts and physical conditioning staff to quantify and maximise the performance of athletes, coaching staff to more deeply analyse team tactical effects and aid in player recruitment, and it could also become a key tool in tailoring specific rehabilitation programs and objectively assess the fitness of the athlete to return to competition. It is recommended that future research should examine the use of this method to detect other movement events in the sports and non-sports environment. This could be to either identify a specific HIE, or as a full classification system to identify multiple movement events within the data. Further research should also examine what effects sampling rate and filtering of the data has on classification accuracy.

Acknowledgements:

We wish to thank the University of the Sunshine Coast, the Sunshine Coast Lightning netball team and Nick Bennett Performance Coaching.

References

Ahmadi, A. (2014). Towards automatic activity classification and movement during a sports training session. *IEEE Internet of Things, 2 (1)*. pp. 23-32

Bailey, J., Gastin, P., Mackey, L., Dwyer, D. (2017). The player load associated with typical activities in elite netball. *International Journal of Sports Physiology and Performance. 12*, 1218-1223.

Camomilla, V., Bergamini, E., Fantozzi, S., Vannozzi, G. (2018). Trends Supporting the In-Field Use of Wearable Inertial Sensors for Sport Performance Evaluation: A Systematic Review. *Sensors*, 18, 783.

Chambers, R., Gabbett, T., Cole, M., Beard, A., (2015). The Use of Wearable Microsensors to Quantify Sport-Specific Movements. *Sports Med* 45, 1065-1081.

Choukou, M., Laffaye, G., Talar, R. (2014). Reliability and validity of an accelerometric system for assessing vertical jumping performance. *Biology of Sport: 31:* 55-62.

Harding, J., Mackintosh, C., Hahn, A., James, D. (2008). Classification of aerial acrobatics in elite half-pipe snowboarding using body mounted inertial sensors. *Proceedings of ISEA Conference*.

Korbinian Frank, Nadales, Robertson, Pfeifer (2010). Bayesian Recognition of Motion Related Activities with Inertial Sensors. *UbiComp10 – Adjunct. Proceedings of the 12th ACM international conference adjunct papers on Ubiquitous computing*.

Li, R. (2016). *Wearable Performance Devices in Sports Medicine. Sports Health*., 8(1), 74-78.

Mannini, A. (2010). Machine learning methods for classifying human physical activity from on-body accelerometers. *Sensors*, 10(2), 1154-1175.

Moran, K,. Ahmadi, A. (2015). Automatic detection, extraction, and analysis of landing during a training session, using a wearable sensor system. *Procedia Engineering: 112*, 184-189.

Sadi, F., Klukas, R. (2012). Reliable jump detection for snow sports with low-cost MEMS inertial sensors. *Sports Technology, 4:1-2*, 88-105.

Sadi, F., Klukas, R., Hoskinson, R. (2013). Precise air time determination of athletic jumps with low-cost MEMS inertial sensors using multiple attribute decision making. *Sports Technology, 6:2*, 63-77.

Su, X,. Tong, H., Ji, P. (2014). Activity recognition with smartphone sensors. *Tsinghua Science and Technology: 19, 3*, 235-249.

Roberts-Thomson, C., Lokshin, A., Kuzkin, V. (2014). Jump detection using fuzzy logic. *IEEE Symposium on Computation intelligence for engineering solutions*.

Wundersitz, D. (2014). Validity of a trunk-mounted accelerometer to assess peak accelerations during walking, jogging and running. *European Journal of Sport Science*.

Wundersitz, D. (2015). Classification of team sport activities using a single wearable tracking device. *Journal of Biomechanics 48*, 3975–3981.

Yang, C. (2010). A Review of Accelerometry-Based Wearable Motion Detectors for Physical Activity Monitoring. Sensors, 10(8), 7772-7788.

PREDICTING FOOTBALL CROWD ATTENDANCE WITH PUBLIC DATA

Ankit K. Patel [a,b,c], Paul J. Bracewell [a,b], Jason D. Wells [a] and Patrick Brown [a]
[a] *DOT Loves Data, Wellington*
[b] *Victoria University, Wellington*
[c] *Corresponding author: ankit@dotlovesdata.com*

Abstract

Increasing revenue is important if professional sports teams are to be successful. A particularly important area of revenue is fan support. Demand forecasting is used to construct an operational model to predict crowd attendance at home games for an A-league club using publically available data, based on a variety of potential predictors of demand. Using residual regression, the results indicate a range of environmental and time-based factors are strong determinants of crowd numbers. Appealing weather conditions, strong past performances, family-friendly kick-off times and previous opposition performance result in the highest attendance numbers. Factors associated with winning games increase attendance levels. The model revealed strong predictive power with an r-squared of 0.80.

Keywords: Demand forecasting, AIC, residual regression modelling, variable selection

1. INTRODUCTION

Fan support is critical to the success of any professional sport franchise. For the Wellington Phoenix, a New Zealand-based Football club which competes in the Football Federation Australia (FFA) administered A-League, crowd attendance is a key metric linked to the renewal of their license agreement (Rugari, 2018). An approach to identify and quantify the impact of drivers of crowd attendance is explored using publically available data.

Exploring sport fan engagement is not new. Early research into attendance rates at professional sports games was conducted over forty years ago. Demmert (1973) proposed an econometric model of attendances at US baseball games, while Noll (1974) assessed factors that could predict attendance at games from four US major league sport competitions. Several studies have been carried out to explore the reasons behind differences in attendance rates at sports events (Hart, Hutton & Sharot, 1975).

With increasing professionalism, it is essential that professional sport teams invest equally into customer acquisition and customer retention techniques to maximize revenue (Shin, Sudhir & Yoon, 2012). New customers are important as they frequently share their experiences. Existing customers are more predictable in terms of the products they purchase and are therefore cheaper to target (East, Hammond & Gendall, 2006).

The offering of season tickets is an important strategy in the retention of customers in professional sports. The more season ticket holders a club maintains, the more effective marketing campaigns become, and the amount of income becomes more predictable (McDonald, 2010). Rust & Zahorik (1993), suggested that the loyalty of a customer depends on the customer's satisfaction.

Several studies have investigated aspects of churn at the club level. McDonald, Centra & Vieceli (2003) carried out a survey of lapsed memberships of an Australian professional Rugby League (NRL) club. A survey of 195 questionnaire responses was used for analysis, including the facets of a customer important to churn proposed by Rongstad (1999). These were how and why they were originally acquired, the type of relationship they had with the organisation and their reasons for churning. The results revealed that a customer's reasons for joining were: availability of discounted prices, reserved seats and increased club involvement. Moreover, customers were generally satisfied with their experience whilst being a member and did not renew primarily due to an inability to attend games. A weak relationship was found between customer satisfaction and the likelihood of a previous member re-joining, which suggested that there are other contributing factors.

McDonald (2010) carried out a study of over 4500 season ticket holders, incorporating both survey research and measures of actual behaviour to identify those variables that best explain and predict churn in two Australian Football League teams. Data covered customer satisfaction, complaint handling, length of relationship, number of games attended, and scarcity of season tickets. They found that the length of relationship and number of games attended were the two most important variables in predicting churn. New, low attending season ticket holders were found to be five times more likely to churn than long term, frequent attending season ticket holders. This highlights the importance of customer retention strategies in sports clubs, particularly among new season ticket holders to give them a sense of belonging to the club and an understanding of their value to the success of the club.

García & Rodríguez (2002) proposed a model to estimate attendance at games contested in the Spanish First Division Football League, using data from the 1992/93 and 1995/96 seasons. The authors investigated the effect of four groups of variables on the number of tickets sold at a match: economic variables, variables proxying the expected quality of the match, measures of uncertainty of the result and those capturing the opportunity cost of attending a match. As expected, the results suggested that attendance increases with cheaper tickets, in areas with higher income and more populated areas.

Andreff & Szymanski (2006) constructed a model to estimate the demand at games involving four English first division football clubs. Data was obtained from home games at Leeds United, Newcastle United, Nottingham Forest and Southampton, during three seasons between 1969 and 1972. As explanatory variables, for each game, the authors used the distance travelled by away support, population of the home and away team's catchment area, the league position of the home and away team prior to the match and a measure of time. The results found that the greater the distance travelled by away fans, the lower the attendance numbers, but only at certain grounds. Additionally, a lower quality opposition was associated with lower attendance rates.

Lemke, Leonard & Tlhokwane (2010) used censored normal regression to estimate individual game attendance at games from the 2007 Major League Baseball season. Several factors were assessed, including the occurrence of divisional or interleague rivalries and the probability of winning. The results found that attendance for small market teams was more affected by game characteristics than attendance at large market teams. Additionally, attendance increased steadily as the probability of the home team winning the game increased.

Previous studies show crowd attendance is predictable. This research attempts to extend on these previous studies to predict crowd attendance and identify the key factors affecting home crowd attendance. This developed model utilizes both environmental and time-based factors not identified in the existing academic literature and allow the club to isolate key drivers affecting crowd numbers to better target fans and increase crowd numbers.

DATA

Crowd attendance numbers and data associated with the 14 variables were extracted from various sources (2007-2016). The analysis set contained 98 observations. From www.ultimatealeague.com, the date and time of fixtures were obtained. School holiday information was acquired from www.education.govt.nz. Information regarding weather conditions (wind, rain, temperature and type) were obtained from weather.niwa.co.nz. Detailed information regarding results, table positions were obtained from Wikipedia (e.g. en.wikipedia.org/wiki/2005–06_A-League). Other variables are derived from these downloaded attributes.

2. METHODOLOGY

Several environmental and time-based factors potentially impacting crowd attendance rates were sourced exclusively via the public domain and enables interested parties to develop and monitor initiatives for increasing crowd attendance.

A residual regression modelling approach, as described by Patel *et. al.* (2017) was used with the factors regressed on $log(attendance)$. This method, often used in the development of application credit risk scorecards in the financial services sector, is a simple and effective method for emphasizing interpretability whilst reducing the impact of multicollinearity. The first step in this procedure was to model home game attendance on a set of environmental and time-based factors. This model was defined as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_t X_t + \varepsilon \qquad (1)$$

where Y represents $log(attendance)$ – given that attendance is count data, β_0 represents the intercept term, $X_1, \dots, X_k$ represents environmental and time-based factors, $\beta_1, \dots, \beta_k$ represents the weight for each factor and ε represents the residuals. Residual regression is a two-stage method whereby the first stage considers only the confounding factors in the model and second stage considers the estimated error from the first stage as a dependent variable for a model where only the non-confounding variables will be evaluated. The residuals are $e_j = Y_j - \hat{Y}_j$ with the following properties: 1) zero mean, 2) homoscedasticity, 3) are normally distributed and 4) $Corr\left(e_i, e_j\right) = 0, \forall_{i \neq j}$. The residuals are calculated from the initial model using Pearson residuals $\left(Z = \frac{\hat{Y}-\pi}{\sqrt{\hat{\pi}(1-\hat{\pi})}}\right)$ (Baez-Revueltas, 2009). The model outlined in equation 1 only treats for confounding variables, while the residuals are calculated, using equation:

$$\hat{Y} = \beta_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_t X_t + \varepsilon \qquad (2)$$

The new dependent variable is defined such that it is uncorrelated with the confounding factors and is created with the residuals obtained from the first stage of the analysis (equation 1 and equation 2). The new model fitted is outlined in equation (3 and 4):

$$Y = \beta_0 + \hat{Y} + \beta_{k+1}X_{k+1} + \beta_{k+2}X_{k+2} + \cdots + \beta_t X_t + \varepsilon \quad (3)$$
$$Y - \hat{Y} = \beta_0 + \beta_{k+1}X_{k+1} + \beta_{k+2}X_{k+2} + \cdots + \beta_t X_t + \varepsilon \quad (4)$$

The method enables 'stepwise' variable selection among variables of interest such that the maximum amount of risk explainable by covariates of interest can be estimated while accounting for potential confounding factors (Baez-Revueltas, 2009). Equation (3) outlines stage 2 of the iterative modelling procedure where covariates are regressed on residuals of stage 1 (equation 1, which only treats for the confounding variables). Residuals are calculated from this equation (3). The residuals are considered as the new dependent variable (equation 4) which is equivalent to equation 1. This process iterated through 14 times, introducing 14 factors into the model. The considered factors were: month, day of week, school holidays, weather conditions (i.e. wind, rain and temperature), weather type (sunny, fine overcast, windy, light rain and heavy rain), table position, last game result, point differential, opposition position, opposition, time of day, season opener/season closer, post season and final appearance and time away from home.

DECOMPOSING CROWD ATTENDANCE

Figure 1 illustrates the decomposed time series of crowd attendance. The graph shows a clear declining trend in crowd attendance over time, with random fluctuation occurring counting for approximately 1000-1500 crowd numbers. Given declining crowd numbers it is paramount, to club success, to identify drivers to increase attendance.

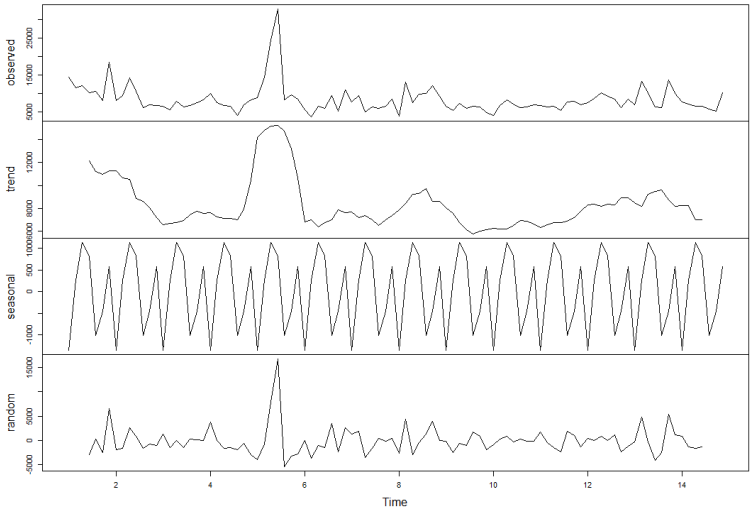


Figure 1: Decomposition of additive time series

VARIABLE SELECTION

A correlation and stepwise regression analysis was conducted to identify the way variables should be stepped into the model. Variables should be stepped-in such that the variable that accounts for the largest variation is *log(attendance)* is fed first, followed by variables of lesser explanatory power.

A correlation analysis between numeric and ordinal variables was conducted using the *hetcor* function in R (polycor package), giving the following results: 1) positively correlated factors with $log(attendance)$: post season games, finals appearances, season opener, season closer, last game result, point differential of last game, opposition position, time of day. 2) negatively correlated factors with *log(attendance)*: table position, weather type, time of day, weather conditions, day and month. No significant relationships were found between other factors.

A stepwise regression analysis applying a forward and backward methodology was conducted. The analysis was conducted using *glmulti* in R to automate model and variable selection, and model averaging. Running a set of 1000 models an information criteria (IC) profile (Figure 2) and variable importance plot (Figure 3) were produced. The red line differentiates between models whose Akaike Information Criteria (AIC) is less than 2 units away from that of the "best" model. The output shows that there were approximately 950 models whose AIC was less than 2 units

away from that of the best model. The importance plot validates the correlation analysis findings, with table position, post season appearance, final appearance and last game result all ranked in the top 5 most important metrics. The model with the lowest AIC included: day of week, rain, last game result, past season appearance, final appearance and table position, as the significant factors affecting $log(attendance)$. This model explained 46% of variation in $log(attendance)$ and the $p - value$ revealed the model to be a good fit. The importance value for a predictor was equal to the sum of the weights/ probabilities for the models in which the variables appear. A variable that appears in models with large weights will receive a higher importance value. The values can be regarded as the overall support for each variable across all models in the candidate set. In Figure 3, the 0.8 (vertical red line) is the cut-off to differentiate between important and not so important variables.

These findings were fed into the residual regression model in the following order: 1. Table position, 2 Post season appearance, 3. Final appearances, 4. Last game result (win, loss and draw), 5. Opposition table position, 6. Rain (mm), 7. Day of week, 8. Season opener, 9. Temperature, 10. Season closer, 11. Wind speed, 12. Weather type, 13, Time of game, 14. Last result differential, 15. School holidays, and 16. Month.

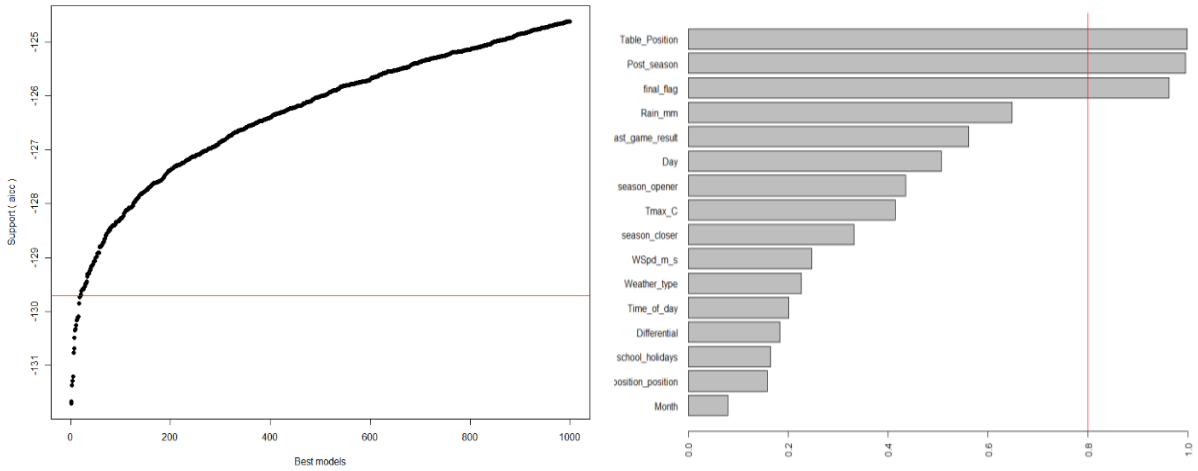


Figure 2 & 3: Stepwise regression AIC profiling and model averaged importance

3. RESULTS

The change in r-squared between each iteration decreases as the factors of lesser explanatory power are fed into the model at latter stages of the modelling process. Overall $the\ log(attendance)$ model explains 65% of variation in log attendance (figure 4). The models explanatory power was also observed without taking the logarithm of attendance (figure 4). This model was found to explain 80% of variation in attendance. Figure 6 illustrates the predicted $log(attendance)$ with actual $log(attendance)$. Figure 2 and 3 illustrates the models predictive power overtime by comparing actual $log(attendance)$ against predicted $log(attendance)$. The model accounts for multicollinearity and confounding issues and produces practical and statically significant results.

The model results reveal that crowd numbers can be explained by a range of environment and time-based factors. Winning, based on table position, last game results, oppositional table position and post-season appearances stimulates numbers. Crowd numbers at the club's home games have remained relatively consistent across 2007-2015, as illustrated in figure 1. The average attendance per home game was 8420. The largest crowd numbers occur in February, March and May. This may be due to favourable weather conditions in the late summer months, while finals series games are contested in May. Fridays and Sundays attract the most crowds, with 728 and 158 more daily attendees on average, respectively. Wednesdays are the quietest days with 2096 fewer daily attendees on average, while daily attendees are 556 below average on Saturdays. Additionally, school holiday periods attract 52 more attendances on average. There is a decrease of 20 in crowd attendance, on average, for every unit increase in wind speed (m/s) and an increase of 22 in crowd attendance for every unit increase in temperature. There is a decrease of 82 in crowd attendance for every mm increase in rain. These results are all practically significant as it is intuitive that bad weather conditions would discourage fans from attending games. Predicted crowd attendance numbers for weather types are practically significant and intuitive. The crowd attendance increases by 603 on fine days, and deceases by 1743 on days with light rain or showers.

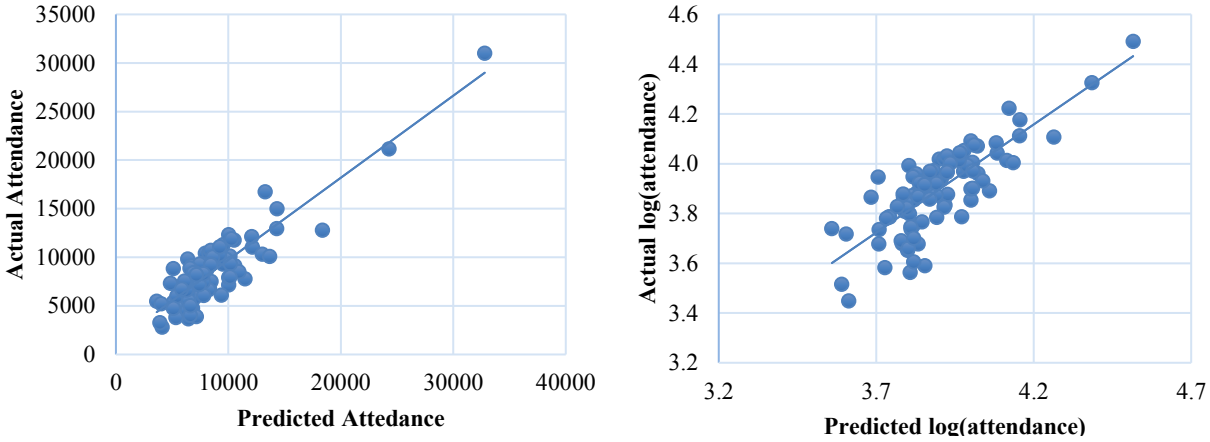


Figure 4 & 5: Model predictive performance vs. $log(attendance)$ predictive power

Typically, crowd attendance increases by 126 for every unit increase in opposition ladder position. This may be due to crowds preferring to attend games involving stronger teams and better players. However, there is an increase of 184, on average, in crowd attendance when the team is playing after a winning performance in their previous game. A drawn result has little impact on crowd numbers, increasing attendance by 1 on average. However, after a loss, crowd attendance decreases by 231 on average. An increase of 5952 in crowd numbers occurs when the team kicks off at 4:00pm which is convenient for families with school aged children. Additionally, Wellington public transport service availability is higher during this time, compared with later kick off times.

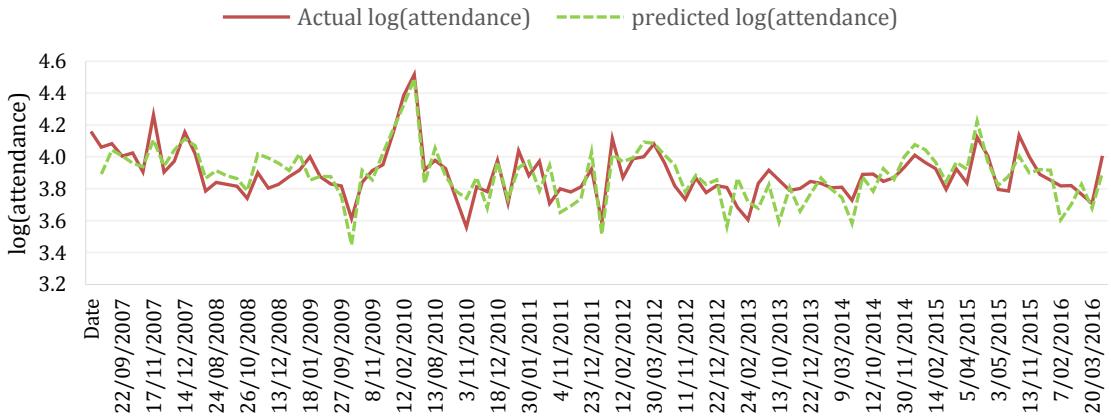


Figure 6: Actual and Predicted Phoenix Home Game Attendance Over Time

Historically, Brisbane Roar attract the highest crowd numbers, with an increase of 788 on the average. The increase in crowd numbers is also high when Melbourne Victory, Sydney FC and Adelaide United are the opposition. These teams have historically appeared in finals giving an indication of their quality. As of June 2018, in eleven out of the previous thirteen seasons, since the A-league commenced, one of these four teams was crowned champions. When the team is contesting a preliminary finals game, there is an increase of 7998, on average, in crowd attendance. However, crowd attendance only increases by 77 during a season closer, which is far less pronounced than the increase associated with finals games. This may provide further insight into fan motivation and worthy of further investigation.

4. DISCUSSION AND CONCLUSIONS

The Wellington Phoenix football club are at risk of their current Hyundai A-league license not being extended beyond 2020. To increase the likelihood of renewal, they need to establish objective data driven strategies to attract as many home game supporters through the gates as possible. The demand models constructed as part of this research suggest that attendance at Wellington Phoenix home games is highest on fair weather game days, game days that

follow a win from the Wellington Phoenix, games against Brisbane Roar, games that kick off at 4:00pm and preliminary finals games. These, and other factors should be considered when adopting strategies to increase supporter numbers.

Future work could involve adding media spend data into the model. Additionally, the current model only adopts attendance data for home games contested at Westpac Stadium. Future iterations of the model could include home game data from other New Zealand venues, which would allow users to understand how schedules and games outside of Wellington should be structured to maximise attendance. The approach could be adopted to forecast demand in other professional sports teams, both domestically and internationally.

References

Andreff, W., & Szymanski, S. (Eds.). (2006). Handbook on the Economics of Sport. Edward Elgar Publishing.

Baez-Revueltas, F. B. (2009). Residual logistic regression. State University of New York at Stony Brook.

Bhattacharya, C. B., & Sen, S. (2003). Consumer-company identification: A framework for understanding consumers' relationships with companies. *Journal of marketing*, 67(2), 76-88.

Coopers, P. W. (2011). Changing the game: Outlook for the global sports market to 2015.

Demmert, H. G. (1973). The economics of professional team sports. *Lexington, Mass:* Lexington Books.

East, R., Hammond, K., & Gendall, P. (2006). Fact and fallacy in retention marketing. *Journal of Marketing Management*, 22(1-2), 5-23.

FFA And Wellington Phoenix Agree to A-League Licence Extension. (2016, Feb 19). Retrieved from https://www.wellingtonphoenix.com/news/ffa-and-wellington-phoenix-agree-hyundai-a-league-licence

García, J., & Rodríguez, P. (2002). The determinants of football match attendance revisited: Empirical evidence from the Spanish football league. *Journal of Sports Economics*, 3(1), 18-38.

Gwinner, K., & Swanson, S. R. (2003). A model of fan identification: Antecedents and sponsorship outcomes. *Journal of services marketing*, 17(3), 275-294.

Hart, R. A., Hutton, J., & Sharot, T. (1975). A statistical analysis of association football attendances. *Applied Statistics*, 17-27.

Juster, F. T. (1966). Consumer buying intentions and purchase probability: An experiment in survey design. *Journal of the American Statistical Association*, 61(315), 658-696.

Lemke, R. J., Leonard, M., & Tlhokwane, K. (2010). Estimating attendance at Major League Baseball games for the 2007 season. *Journal of Sports Economics*, 11(3), 316-348.

McDonald, H. (2010). The factors influencing churn rates among season ticket holders: An empirical analysis. *Journal of sport management*, 24(6), 676-701.

McDonald, H., Karg, A. J., & Leckie, C. (2014). Predicting which season ticket holders will renew and which will not. *European Sport Management Quarterly*, 14(5), 503-520.

McDonald, H., Centra, B., & Vieceli, J. (2003, January). Explaining non-renewal behaviour: an empirical investigation of recently lapsed NRL club members. *In ANZMAC 2003: a celebration of Ehrenberg and Bass: marketing discoveries, knowledge and contribution*. (pp. 2167-2173). University of South Australia.

Noll, R. G. (1974). Attendance and Price Setting. Government and the Sports Business. RG Noll. Washington, DC. *The Brookings Institution*.

Patel, A. K., Bracewell, P. J., Gazley, A. J., & Bracewell, B. P. (2017). Identifying fast bowlers likely to play test cricket based on age-group performances. *International Journal of Sport Science and Coaching*.

Rongstad, N. (1999). Find out how to stop customers from leaving. *Target Marketing*, 22(7), 28-29.

Rust, R. T., & Zahorik, A. J. (1993). Customer satisfaction, customer retention, and market share. *Journal of retailing*, 69(2), 193-215.

Shin, J., Sudhir, K., & Yoon, D. H. (2012). When to "fire" customers: Customer cost-based pricing. *Management Science*, 58(5), 932-947.

Trail, G. T., Fink, J. S., & Anderson, D. F. (2003). Sport spectator consumption behavior. *Sport Marketing Quarterly*, 12(1).

A-League Winners List | Australian Soccer League Past Champions. (2018, March 9). Retrieved from https://skyrockliving.com/a-league-winners-list/

Rugari, V. (2018). Wellington Phoenix could be replaced as A-League formally confirms expansion plans. Stuff. April 3rd, 2018. https://www.stuff.co.nz/sport/football/a-league/102795194/wellington-phoenix-set-to-be-replaced-as-aleague-expansion-looms

A FRAMEWORK FOR QUANTIFYING THE EFFECTIVENESS OF HUMAN-BASED RATING SYSTEMS

Ankit K. Patel [a,b,c] and Paul J. Bracewell [a,b]
[a] *DOT Loves Data, Wellington*
[b] *Victoria University, Wellington*
[c] *Corresponding author: ankit@dotlovesdata.com*

Abstract
The growth in demand for analytics has been experienced across many data rich industries. However, this effect is most evident in three major industry verticals: sport, finance and technology. The common driver in these data-driven and modelling intensive applications is the objective of evaluating, ranking, rating or predicting the performance of an individual or collection of individuals. This common thread dictates that the results must be robust, transparent, reliable and meaningful (Bracewell, 2003). However, these individual-based rating systems may not have an actual outcome, or the observed outcome may take many months to manifest, creating challenges in deriving meaningful conclusions from outputs. A framework is introduced to quantify the effectiveness of ratings that are produced by semi-supervised and unsupervised models. The intent of this framework is a novel scoring method to measure the accuracy, predictive power and assesses the confidence the user can have in model outputs.
Keywords: Proper scoring rules, semi-supervised ratings, dimension reduction

1. INTRODUCTION

The application of analytics in the business environment has experienced tremendous growth (McKinsey & Company, 2016). Business analytics has transformed from a "nice-to-have" to a competitive advantage. "In the past few years, predictive analytics, has gone from a practice applied in a few niches to a competitive weapon with a rapidly expanding range of uses" (CGI: Predictive Analytics, 2013, p.1).

A key factor for the rise in business analytics is the phenomenon of "big data", and its acceptance by senior executives as an important business enabler. The goal of insight and information extraction or revealing hidden patterns within big data is achievable through the application of mathematical and statistical techniques. Importantly, these insights need to be relayed appropriately to the intended audience. Sagiroglu & Sinanc (2013) stated that modern analytics, characterized by improvements in computing power, reduced cost in data storage, greater access to various data sources and cheaper commodity hardware, requires a revolutionary step forward, moving away from traditional data analysis. The Transforming Data with Intelligence survey revealed that the application of advanced analytics creates better aimed marketing, informed decision-making, client-based segmentation and recognition of sales opportunities. This information offers significant potential to generate business value and competitive advantage.

This growth in demand for analytics and data capture has been experienced across many industries, however this effect is most evident in three major applications: 1) Sport, 2) Finance and 3) Technology resulting in considerable academic and commercial attention (e.g. Stefani, 1997; Siddiqi, 2012; Bracewell *et. al.,* 2017, respectively). The consequence is the development of data-driven and modelling intensive applications with an objective of evaluating, ranking and rating or predicting the performance of an individual or collection of individuals. This common thread dictates that the results must be robust, transparent, reliable and meaningful (Bracewell, 2003) to generate trust leading to implementation of the insights derived from such a system.

However, can the characteristics of a good rating system be defined and quantitatively evaluated? To answer this question, the commercial and academic exploration of rating systems within these three industries are explored. Drawing from modelling methodology found across the three industries, this paper develops a ratings framework that allows the evaluation of human performance in a meaningful manner. The resulting framework is analogous to an Associative Neural Network and emphasises the need for interpretable, robust, reliable and meaningful outputs. Essentially, a ratings framework is an *elegant application of dimension reduction.*

SPORTING INDUSTRY

Within the sporting world statistical ranking and rating methodology have been heavily applied in the past two decades at both the individual and team level. Due to the large volume of money resources that is increasingly being invested in teams and individual players, sport analysis and the need for meaningful sports statistics has experienced exponential growth in recent decades. Moreover, the rise in player salary caps over the last 25 years provide ample evidence of the growth in sports analytics, with investors, franchises, clubs and other stakeholders wanting to determine the value of their investment decisions. For example, in the National Football League (NFL) there has been an increase of approximately 950% in player salaries since 1980's, and an increase of

288% in salary cap since 1994 (Vrooman, 2012). Global sports revenue grew by US$145.3billion over the 2010-2015 period (Coopers, 2015), at an annual compound growth rate of 3.7%. Moreover, winning teams earn significantly larger revenue than that of losing teams, there is a strong incentive for coaches and managerial staff of sport teams to succeed. Additionally, "the regulated sports betting market is forecasted to reach $70 billion in 2016, representing a 20% increase from 2012" (Foley-Train, 2014).

Given the large investment of resources, coaches, managers and other stakeholders cannot solely depend on subjective views and personal beliefs to inform team and player selection decisions. Solutions must be augmented with objective approaches by implementing analytical techniques to rank, rate, evaluate and forecast selection decisions. This need for informed data-driven decisions has given rise to the use of sport analytics by managers, coaches, athletes and fans. Forbes (2015) claimed that the popularity of data-driven decision making in sports has trickled down to the fans, which are consuming more analytical content than ever; (see Akhtar, Scarf & Rasool (2014); Alamar & Mehrotra (2010); Annis & Craig (2005); Patel, Bracewell & Rooney (2017)).

FINANCIAL INDUSTRY

Due to the 2007-08 Global Financial Crisis (GFC), the Basel committee on Banking Supervision introduced the Basel III framework in 2010-11, with the intention of strengthening capital requirements by increasing liquidity and decreasing leveraging. This regulation changed the way credit scoring systems are built and the type of applicant attributes each system must incorporate, increasing financial institutions demand for scoring systems that detect subtle changes in an applicant's attributes associated with probability of default. Most credit scoring systems are based on the 12-month view of historical applicant's behaviour and an assumption that a customer's future behaviour is like their past behaviour. However, during the GFC many applicants who were financially stable for many years ended up in financial difficulty. This revealed that the adopted scoring methodology was not necessarily reflective of an applicant's credit worthiness and highlighted flaws in the current scoring methodology. Hand & Henley (1997) stated that the most widely used techniques for building scorecards are linear discriminant analysis, logistic regression, probit analysis, non-parametric methods, Markov chain models, recursive partitioning, expert systems, genetic algorithms, artificial neural networks and conditional independence models. These techniques are used to predict the probability of default in the next 6, 9, 12 or 18 months (Peussa, 2016; Bolton, 2009). There exist subtle nuances in the application of statistical methods within the financial services sector. Specifically, given data is not missing at random, this requires the application of an approach called reject inference. Moreover, to maintain interpretability and minimise the impact of collinearity the data fed into the model in a manner satisfying commercial constraints by iteratively modelling on the residuals. Please see Baez-Revueltas, 2009; Einarsson, 2008; Shad & Rehman 2012; Anagnostopoulos & Abedi 2016; Roy, 2016; Tabagari, 2015; Torosyan, 2017; Patel *et. al.* (2017).

TECHNOLOGY INDUSTRY

Recently there has been a massive rise in the number of organisations focusing on the evaluation of human performance in commercial settings. Organisations such as Codility, Umano and BlueOptima focus on dynamically monitoring computer programmers' performance to provide insight to drive better quality, measures of employee productivity. This stems from increased demand for this type of skill-set in the workforce. This is evident in the technology industry which is experiencing a rapid growth in the demand for developers. According to the U.S. Bureau of Labour, software developer jobs are expected to grow 17% from 2014 till 2024. This rate of growth is much faster than the average rate among other professions. Overall by 2020, employment in all computer occupations is expected to increase by 22% (Thibodeau, 2012). Among these occupations, software developers are expected to experience the highest growth (32%), followed by database administration (21%), network and systems administration (28%), computer systems analysts (22%) and computer and information systems managers (18%). This demand has resulted in a shortage in supply, creating an environment where organizations are competing to identify and place the best candidate against other agencies. Hiring a recruitment agency to identify the talent is expensive and time consuming. These industry trends have led to data-driven technologies identifying, evaluating, rating and ranking developer performance.

2. METHODS

There are two overarching characteristics across the three ratings methodologies: 1) resultant outcome and 2) overall objective. The systems produce a single real number [0,1] representing a human's ability to perform either athletically, financially or technically in their respective environment. Moreover, each system aims to evaluate the performance of the same entity i.e. a human. Here, system definitions, common methods and modelling practices across the three industries are provided.

SPORTS RATING SYSTEM

Formally, a sports rating system assigns each team a single numerical value to represent team or player strength relative to the rest of the league on some predetermined scale (Massey, 1997). Stefani (1997) stated that sport rating systems have three steps: 1) Weigh the observed results to provide competition points - this is the most important factor in determining points for competition *i* for a given competition, 2) Combine the competition points to produce seasonal values, and 3) Aggregate the seasonal value to produce a rating. Generally, sport rating systems fall into two categories: 1) Earned ranking – These systems utilise past performance to provide a suitable method for selecting either a winner or a set if teams that should participate in a play-off, and 2) Predictive ranking – These systems utilises past performance to provide the best prediction of the outcome of future games between two teams. Additionally, Stefani (2011) stated that sport rating systems can be separated into three distinctive types depending on how new ratings are calculated for each rating system: 1) Adjustive, 2) Accumulative and 3) Subjective. A potential drawback of sport rating systems are small sample sizes due to a limited number of contested sporting events. To derive a deeper understanding of the requirements for a meaningful sports rating system, this research builds on work from: Patel, Bracewell & Rooney (2017); Patel, Bracewell & Wells (2017); McIvor, Patel, Hilder & Bracewell (2018); Campbell, Patel & Bracewell (2018).

CREDIT-RISK SCORECARDS

Application and behavioural scorecards incorporate a *binary* or *count* target variable (i.e. approval or non-approval, or a credit rating, respectively). However, unlike the target variable associated with sport rating systems, evaluating the actual 'creditworthiness' of an approved line of credit can take months to observe the true outcome. New scorecard regulations require more robust, dynamic and flexible models capable of accurately measuring an applicant's credit worthiness using a smaller time window of transactional data. However, a smaller time window means a smaller sample size of transactional data, potentially leading to poorer, less predictive credit ratings. There are six key steps involved when developing a scoring method: 1. Data Preparation > 2. Data Cleaning > 3. Variable Selection > 4. Sample Generation > 5. Model Development and Validation > 6. Model Approval. The first 3 steps are data processing. These steps are essential in developing a scoring method; however, the literature predominately focuses on data preparation, model development and validation steps. These three steps have the potential for improving the performance of scorecards. Various model algorithms can be used with different input variables to see which gives the best result. The choice of modelling objective is the primary key to developing scorecards since it defines a full set of technical estimation procedures that are used to select the best model under the objective and defines how to assess its validity. Data preparation and variable selection steps are very important in credit scoring, and it has been found that applying new and more predictive variables can improve the performance of scoring models (Hand & Henley, 1997). The 'model development and validation' step is used to discriminate between 'good' and 'bad' applicants. The better the classifier, the better the performance of the scoring method.

DEVELOPER RATING SYSTEM

Human rating systems are a new application within the technology industry to evaluate the quality of developers and programmers. The limited literature reveals that these rating systems predominately adopt a hierarchical, layered network-based modelling framework such as that suggested by Bracewell, Patel, Blackie & Boys (2017); Bracewell & Blackie, Blain & Boys (2016); Klehe & Anderson (2007) and Silvia (2008). Hierarchical and network-based modelling is used as human ability is a manifestation of skills that rely on other preceding attributes and tasks. Therefore, before a given set of skills and attributes can be quantified a set of preceding tasks and metrics must be defined. For example, the Umano rating system incorporates a network-based, multilevel structure measuring human performance, at any given point in time (Bracewell *et. al.*, 2017). However, unlike the target variables utilized within the sports rating systems and credit risk scorecards, the Umano system has no known target-value to assess predictive power. The appropriate statistical framework for the inability to evaluate predicted outcome to an unknown target value, is unsupervised learning.

LIMITATIONS OF RATING SYSTEMS

The increasing demand for rating systems about human performance heightens the need to measure the performance and validity of the underlying model. However, individual-based rating systems do not necessarily have an observable outcome, creating challenges in deriving meaningful conclusions about model accuracy. For example, the validity of an application scorecard at the time of approval is unachievable given that an applicant's 'actual' credit worthiness can only be assessed after at least a 3-month window. If a 12-month outcome is observed, the true performance of an application credit risk scorecard is only available after 12 months. In this

scenario, predictions can be mapped back to actual outcomes, but only after a minimum time. Furthermore, an approved applicant's credit-worthiness can be 'good' at time t_1 (i.e. 3-months) but bad at time t_2 (i.e. 6-months).

There are many systems across sports, which dynamically assess performance and calculate a single numerical value (or rating). However, unlike the credit rating issues where the effectiveness of the applicant's rating can be assessed after a given period, the issue with sport rating systems are that the rating measure may not tangibly link to the event outcome. For example, if a player rating system produced a rating of 67 (out of 100) during the game, how can the accuracy of such a score be evaluated? Can this rating be mapped to actual in-game events and actions? And is it representative of an intuitive outcome?

BENEFITS AND COSTS FOR EACH OF THE THREE RATING SYSTEMS

Table 1 outlines the benefits and drawbacks associated with each rating methodology. These findings have been determined from the development of rating systems and identifying the communalities of a framework that is applicable across sport, technology and finance. There are two reasons why these three rating systems have been selected to develop a ratings framework: 1) growing demand for predictive and accurate rating systems and 2) the target variable across each system differs; allowing the evaluation of systems with varying dependent variables. Based on these findings and Bracewell (2003) a good rating system must have the following characteristics: 1) Robust – the framework must yield good performance when data is drawn from a wide range of probability distributions that are largely unaffected by outliers, small departures from model assumptions, and small sample sizes, 2) Reliable – Produce accurate and highly informative prediction that speak to real-world events 3) Transparent – Interpretable, easy to communicate and break down, and 4) meaningful – must relate to real-world outcomes, and explain different dimensions of the data and different layers of different dimension.

Rating Systems	Benefits	Costs
Sports	Immediately measurable and observable outcomes Complete data Quantitative and practical metrics	Small sample sizes
Financial	Measurable outcomes Quantitative metrics	Small sample sizes Time lagged variables Incomplete data
Coding Ability	Practical metrics Plethora of metrics across various human characteristics (i.e. behavioural, process and technical)	Small sample sizes No target outcomes Incomplete data

Table 1: Outline of benefits and drawbacks for the three systems

3. RESULTANT FRAMEWORK FOR HUMAN PERFORMANCE

Given the prevalence of human rating systems within major industries, this study attempts to develop a framework that outlines the necessary steps to develop a meaningful rating system. The uniqueness of these rating systems is there: commercial applications, requirement to produce strong predictive power, classification accuracy and intuitive results relating to real world outcomes. Developing rating systems across the three domains identified a set of key steps. As all three systems aim to rate human performance, the overall task is to ensure intuitive ratings that are predictive and informative of human ability.

Rating systems are an elegant form of dimension reduction to produce robust, transparent, reliable and meaningful outputs (Bracewell, 2003). Combining simple and complex metrics in a meaningful manner produces predictive ratings that explain a larger proportion of the variation. A key part of this process is feature engineering and variable selection. This ideology is held throughout credit scoring literature. The developed rating framework is an adaptation of the credit-risk scorecard framework outlined previously.

The framework in Figure 1 represents the different data dimensions and various layers that exist. Given the hierarchical nature, the results of each layer contribute to the resultant rating. Each layer is a modelling exercise with a unique objective to create an intuitive feature that captures sufficient information for a given dimension of the data. Once this is achieved proceed to the succeeding layer. The initial layers combine meta-level and simple metrics to create complex metrics that capture a greater level of information than the simple metrics. The successive layers combine complex metrics with individual metrics, accounting for interactions, to develop metrics capturing information relating to the individual human-level dimension. The latter layers combine complex meta and human-level metrics, accounting for interactions, to create a metric that relate to the actual outcome. Finally, a back-propagation process is utilised where the outcome metrics, i.e. rating are fed into the

proceeding layer to produce a more informative result leading to a more intuitive and predictive outcome or rating. This back-propagation process continues until a non-significant step-size change occurs in ratings or the cost of recursion outweighs the change in ratings.

The layers are analogous to neural network nodes with a set of parameters, weights, thresholds, cost functions or performance metrics. The layer specific objectives must be satisfied to transition to the next layer. The ratings framework is analogous to a feed forward neural network. Given each layer is a modelling exercise, they represent models with their own characteristics: weights, parameters, threshold/ performance measures and cost functions. The definition and objective of each layer is outlined below. The results of each layer need to be predictive, accurate, intuitive (i.e. represent real-world actions) and interpretable, therefore the ratings problem in the field of information theory, feature engineering, variable selection and dimension reduction.

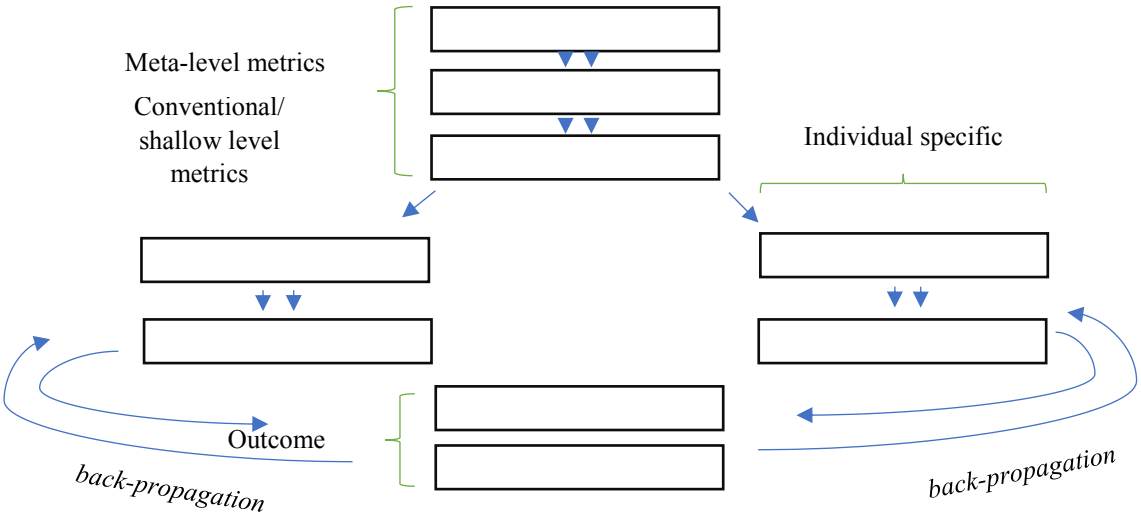


Figure 1: Ratings framework

4. DISCUSSION AND CONCLUSION

As each layer has its own modelling objective, the nature of the layer specific problem dictates the techniques adopted depending if the stage is supervised or unsupervised as well as other modelling assumptions.

For the purposes of dimension reduction, examples of appropriate techniques are: Principle Component Analysis, t-Distributed Stochastic Neighbour Embedding, Stepwise regression, Gradient Boosted Machine, Random Forest analysis, Hierarchical Clustering Analysis. Applicable model validation techniques include: Mean-squared error, 0-1 loss for a categorical outcome, log-loss functions, quadratic loss functions, Expected loss functions, Shannon's Information Criteria, Bayes Information Criteria and Gini impurity.

SCORING RULES

Given the resultant rating is a value between 0-1, it can be interpreted as a probabilistic prediction of a human's performance enabling accuracy to be evaluated. The measures typically applied in metrology, finance and pattern classification, include scoring rules, which "involve the computation of a score based on the probability forecast and, in the event, (or the value of the uncertain quantity) that actually occurs" (Winkler, 1996, p. 1). Scoring rules provide overall measure of "goodness" of probabilities.

In an ex-ante sense strictly, proper scoring rules provide an incentive for careful and honest predictions by the modelling system. In an ex-post sense, they reward accurate predictions and penalise inferior forecasts. The most natural role for scoring rules is simply to provide summary measures to evaluate probabilities considering what happens (ex-ante). Moreover, scoring rules motivate forecaster or forecasting systems to honestly report probabilities, and provide appropriate incentives for systems that do so. De Finetti (1962, p.359) claimed "the scoring rule is constructed according to the basic idea that the resulting device should oblige each participant to express his true feelings, because any departure from his own personal probability results in a diminution of his own average score as he sees it".

References

Anagnostopoulos, Y., & Abedi, M. (2016). Risk pricing in emerging economies: credit scoring and private banking in Iran. *International Journal of Finance & Banking Studies,* 5(1), 51-72.

Annis, D. H., & Craig, B. A. (2005). Hybrid paired comparison analysis, with applications to the ranking of college football teams. *Journal of Quantitative Analysis in Sports*, 1(1**).**

Akhtar, S., Scarf, P., & Rasool, Z. (2015). Rating players in test match cricket. *Journal of the Operational Research Society, 66(4),* 684-695.

Bracewell, P. J., Patel, A. K., Blackie, E. J., & Boys, C. (2017). Using a Predictive Rating System for Computer Programmers to Optimise Recruitment. *Journal of Cases on Information Technology (JCIT)*, *19*(3), 1-14.

Bracewell, P. (2003). Monitoring meaningful rugby ratings. *Journal of Sports Sciences*, *21(8),* 611-620.

Bracewell, P., Blackie. E., Blain, P., & Boys, C. (2016, July 12). Understanding the impact of demand for talent on the observable performance of individuals. Paper presented at The Proceedings of the 13th Australian Conference on Mathematics and Computers in Sports. (pp. 40-45). Melbourne, Victoria, Australia

Bracewell, P. J., Coomes, M., Nash, J., N., Rooney, S. J., Patel, A. K., & Meyer, D. H. (2017). Application of Reject Inference to T20 cricket bowlers: Calculating the probability of taking a wicket using a behavioural credit risk scorecard framework. *Australian & New Zealand Journal of Statistics; (under review)*

Baez-Revueltas, F. B. (2009). Residual logistic regression (Doctoral dissertation, The Graduate School, Stony Brook University: Stony Brook, NY.).

Bolton, C. (2009). Logistic regression and its application in credit scoring (PhD dissertation, U. of Pretoria).

Campbell, E. C., Patel A. K., & Bracewell, P. J. (2018). Optimizing junior rugby weight limits in New Zealand. Paper presented at The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports. Sunshine Coast, Queensland, Australia: ANZIAM MathSport.

De Finetti, B. (1962). Does it make sense to speak of 'good probability appraisers'. The scientist speculates: An anthology of partly-baked ideas, 257-364.

Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. Journal of the Royal Statistical Society: Series A (Statistics in Society), 160(3), 523-541.

Klehe, U. C., & Anderson, N. (2007). Working hard and working smart: Motivation and ability during typical and maximum performance. *Journal of Applied Psychology*, *92*(4), 978.

Massey, K. (1997). Statistical models applied to the rating of sports teams. Bluefield College.

McIvor, J. T, Patel, A. K., Hilder, T.A., & Bracewell, P. J. (2018). Commentary sentiment as a predictor of in-game events in T20 cricket. Paper presented at The Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports. Sunshine Coast, Queensland, Australia.

Patel, A. K., Bracewell, P. J., & Rooney, S. J. (2017). An Individual-Based Team Rating Method for T20 Cricket. *Journal of Sports and Human Performance 5(1): 1-17.*

Peussa, A. (2016). Credit risk scorecard estimation by logistic regression.

Patel, A. K., Bracewell, P. J., Gazley, A. J., Bracewell, B. P. (2017). Identifying fast bowlers likely to play test cricket based on age-group performances. Journal of Sports Science and Coaching 12(3): 328-338.

Patel, A. K., Bracewell, P. J., & Wells, J. D. (2017, June 23). Real-time measurement of individual influence in T20 cricket. Proceedings of the 17th MathSport International Conference. (pp. 61-70). Padua, Italy.

Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In Collaboration Technologies and Systems (CTS), 2013 International Conference on (pp. 42-47). IEEE.

Siddiqi, N. (2012). Credit risk scorecards: developing and implementing intelligent credit scoring (Vol. 3). John Wiley & Sons.

Stefani, R. T. (1997). Survey of the major world sports rating systems. *Journal of Appl. Stat., 24(6),* 635-646.

Stefani, R. (2011). The methodology of officially recognized international sports rating systems. *Journal of Quantitative Analysis in Sports*, *7*(4).

Silvia, P. J. (2008). Another look at creativity and intelligence: Exploring higher-order models and probable confounds. Personality and Individual differences, 44(4), 1012-1021.

Shad, M. Y., Rehman, M. K. (2012). Credit Risk Modelling/ Scorecard. Retrieved from https://pdfs.semanticscholar.org/presentation/e026/4e75803235de2e152668a8637c78983d7cdc.pdf

Thibodeau, P. (2012). IT jobs will grow 22% through 2020, says US Computerworld.

Vrooman, J. (2012). The economic structure of the NFL. In *the Economics of the National Football League* (pp. 7-31). Springer New York.

Winkler, R. L., Munoz, J., Cervera, J. L., Bernardo, J. M., Blattenberger, G., Kadane, J. B., & Ríos-Insua, D. (1996). Scoring rules and the evaluation of probabilities. Test, 5(1), 1-60.

OPTIMISING JUNIOR RUGBY WEIGHT LIMITS IN NEW ZEALAND

Emma C. Campbell [a,b], Ankit K. Patel [a,b,c] and Paul J. Bracewell [a,b]
[a] *DOT Loves Data, Wellington*
[b] *Victoria University, Wellington*
[c] *Corresponding author: ankit@dotlovesdata.com*

Abstract

The New Zealand rugby community is aware of safety issues at junior level (under 13 years of age) and some provinces have applied weight limits tackle grades to minimise injury risk. However, in some of these provinces, the weight-based systems create an unaccommodating situation for heavier participants as they may not be playing within their peer group. Those playing above their age-weight may be competing against individuals up to four school years older. Examining seven years of data from Auckland Rugby Union, Campbell, Bracewell, Blackie & Patel (2018) identified that individuals playing up a grade by weight had a 48% higher chance of not playing in the following year compared to those playing in the correct age band. Here, a framework for developing an optimised age-weight system is introduced. Assuming the importance of playing with peers, the framework accounts for the impact of shifting the current date of birth cut-off (1st January) to more closely align with the New Zealand school system (approximately 1st May) and coincide with the start of the rugby season. Then, various weight quantiles are extracted to isolate near optimal age-weight bands for each grade, given a fixed proportion of displaced participants. Finally, a classification tree is implemented to derive the optimal age-weight categorisation system. The proposed system increased peer-group participation, reduced player displacement, and presents a sensible age-weight grading system.

Keywords: Relative age effect, categorisation system, classification tree

1. INTRODUCTION

To ensure fair competition, youth sports are often categorised according to chronological age. Such categorisation systems are predominantly found in youth sport as relatively older players possess significant assets in terms of their size, weight and strength, which influence their 'perceived potential' (Delorme, 2014). This relative age phenomenon is known as relative age effect (RAE) and refers "both to the immediate participation and long-term attainment in sport, occurring because of chronological age and associated physical differences as well as selection practices in annual age-grouped cohorts" (Cobley, Baker, Wattie & McKenna, 2009, p. 235). The unaccommodating nature of age categorised grading systems, especially for youth level physical sports, can lead to: (1) relatively younger players being dissuaded from participating in sports that favour certain physical attributes, due to their development disadvantage, and (2) relatively younger participants, who play regardless of their physical disadvantage, being more prone to drop out (Delorme, Chalabaev & Raspaud, 2011). Moreover, research suggests that the RAE, an artificial consequence of youth competitions, generates a loss in potentially talented players which, in the long-run, contributes to a decrease in the level of professional and national teams (Delorme, 2014, p.2).

To reduce the effects associated with the RAE, many sports have introduced new categorisation systems that incorporate a player's weight into the grading process to deal with the negative correlates of relative age differences; for example, youth American football (Kerr *et al.* 2015).

Delorme (2014) investigated whether weight categorisation system can help reduce maturational differences and eliminate or reduce RAE in youth athletes (i.e. French amateur boxer). The study results found that a weight categorised system is a possible solution within the RAE phenomenon. Such a system groups according to chronological age and physical attributes (i.e. an age-weight system).

Rugby is a sport whereby youth participants are grouped according to an age-weight categorisation system (also known as a grading system). Categorisation systems based on age and weight are common in New Zealand youth rugby, however, the system currently adopted by several provinces (e.g. Auckland, Wellington, Canterbury and Waikato) is a *weight-age* system rather than an *age-weight* system. The *weight-age* system ignores literature suggesting weight categorisation can counterbalance the RAE. At a youth level it is assumed that physical attributes, such as weight are a by-product of age, and therefore a weight-age system leads to peer-group displacement.

Arguably New Zealand's national sport, rugby is the game of choice for thousands of young players across the country and plays a major part in the early-stage development of youth participants. The experience of being part of a rugby team undoubtedly has an immense impact on the cognitive and psychosocial development of young players (4-13 years).

In some New Zealand (NZ) provinces, junior rugby operates a *weight-age* grading system that was put into effect in 2013. However, this places a greater importance on weight relative to age. NZ is one of the only

countries that implements a junior rugby grading system attempting to account for player weight to reduce RAE when determining team placement. While the intention of this grading structure is to reduce the risk of injury, due to weight disparity, player development is not adequately considered. The current grading system in some regions allow up to a 4-year age gap between players in the same team, therefore it is possible for a child in Year 3 (age 7) to be playing with Year 7's (age 11). A large variation in age results in mismatched sporting abilities. Moreover, contrasting motivational attitudes of players create an unnecessarily difficult job for coaching staff who are required to juggle a wide range of player needs within one team. World Rugby (2013) states that 'creating a competitive atmosphere that is inappropriate for the developmental stage of a child can create negative psychological effects'. If junior rugby participation numbers are to be retained, action must be taken to design a system that caters for a larger proportion of participants to play within their peer group.

Campbell, Bracewell, Blackie & Patel (2018) uncovered that a junior rugby player who participates up one grade because of their weight has a rate of exiting the game in the next season 48% higher than those playing in their appropriate age group. Interestingly, of these players those born in December were 14% more likely to exit than those born in January, indicating that the current date-of-birth cut-off may be influencing player participation. These findings stimulated this review of junior rugby grading systems. In NZ, children can start school as soon as they turn five. Typically, those born in the first few months of the year will be in Year 1. Those born in the latter part of the year will go into Year 0, moving into Year 1 in the following calendar year. This transition is roughly in May, which coincidentally corresponds with the commencement of the junior rugby season.

It is known that cognitive and non-cognitive development is shaped during the early stages of childhood (Heckman & Tremblay, 2006). A study of 'Sports and Child Development' (Felfe, Lechner & Steinmayr, 2011) indicates that participation in sports during childhood enhances the cognitive and non-cognitive development of an individual. Between the ages of 8 and 12 individuals experience the development of psychological factors that influence their perception of sport and sport participation, including the motivational and cognitive readiness of a player (World Rugby, 2013). Motivational attitudes can differ significantly between players who are merely 1 or 2 years apart. Younger participants are driven by notions such as peer group affiliation, team identification and skill development whereas the incentive of winning and playing at an elite level does not influence player participation until 11 or 12 years of age. Cognitive readiness, a player's ability to understand their own performance and skill level, is dependent on adult feedback for 8-9-year olds, however 10-14-year olds assess their performance based on peer group evaluation (World Rugby, 2013).

The RAE shows that elite athletes are more likely to be born in the first 3 months after the cut-off date, in cases where weight bands have not been implemented (Lewis, Morgan & Cooper, 2015). Moreover, this effect is more pronounced in; physical sports, for males and at highly competitive or elite levels. Those born in the first three months following the cut-off date can have up to 11 months more mental and physical development than some of their teammates, an effect that is particularly prominent for younger players (World Rugby, 2013). Applying *weight-age* bands enhances the disparity in player's cognitive development, and further extends the advantage of players born in the three months following the 1st of January cut-off.

These age group disparities highlight the utmost importance of peer-group interactions, achieved by participating in sport with individuals at the same developmental stage as one another. If these interactions are incorrectly managed at such a vulnerable age, the subsequent effect on the child can be detrimental to their enjoyment of playing rugby and their passion for the sport. Consequently, leading to large drop-off rates in youth rugby. As mentioned, the current grading system places a greater priority on weight rather than age. This implies that the current system does not appropriately adjust for the RAE phenomenon.

Therefore, the primary objective of this study is to investigate a framework that can transform and optimise the current weight-age system to an age-weight system. Shifting the system from being weight to age dominant will ensure optimal peer-group participation at appropriate weighting thresholds. With junior rugby playing numbers in decline (Martin, 2017), timely action must be taken to implement a new grading system that strongly emphasises peer group participation.

RESEARCH INTENT

The research attempts to optimise the junior rugby grading system for age and weight. Based on research conducted by Campbell *et. al.* (2018) it is hypothesised that the current grading system operates on a sub-optimal date-of-birth cut-off and weight thresholds. Identifying the optimal *age-weight* bands will enable a larger proportion of participants to play with individuals their own age, foster peer-to-peer relationships, whilst accounting for weight restrictions, due to RAE. It is expected an optimised grading system will cultivate a more enjoyable playing environment, resulting in peer-to-peer relationships extending beyond the field.

The analysis will focus on players playing tackle rugby, Under 8 to Under 12 (U8-U12), as they are the most susceptible to variations in psychological development, as discussed previously. Further, the players within these grades are affected by the weight-age grading system, whereas U6 and U7 players are only subject to age restrictions.

DEFINITIONS

1. Peer group is defined as individuals who are in the same year group at school and therefore have a stronger association to one another than individuals in different year groups.
2. Youth-participants are defined as players between 8 and 12 years old.
3. An optimised system is one which decreases age-weight separation between participants within grades and increases age-weight separation between participants across groups.
4. Player displacement is defined as individuals who do not participate within their peer-group.
5. Correctly categorised is defined as participants whose age falls within their appropriate grade.

ASSUMPTIONS

1. A grading system that places greater importance on weight, relative to age (i.e. a weight-age grading system) produces larger peer-group displacement compared to an age-weight system that places greater importance on age. Given younger participants are driven by peer-group participation, team identifications and the significance of RAE at the youth level, it is assumed that an age-weight system is better than a weight-age system when trying to reduce player displacement.
2. The new age-weight grading system must not allow participants to play more than 1 year above their school year.

DATA

Given the sensitivity of the data and the wider implications for adoption within junior rugby across New Zealand, the emphasis for this research is outlining a framework for defining an optimal age-weight system. The statistics that follow make limited comparisons to demonstrate observed benefits and identify metrics that could be used to characterise improvements.

2. METHODOLOGY

A framework to optimise junior rugby grading systems by addressing the current date-of-birth cut-off and weight bands is proposed. This overall objective has been formalised into two research questions:

a. Identify the effectiveness of the January 1st cut-off, by comparing the entry and exits of players from Junior Rugby across birth months.
   i. Evaluate the proportion of players displaced across all grades, under the current system. This requires a definition of 'correctly categorised' to identify the number of 'displaced' participants (please see above).
   ii. Given a May 1st date-of-birth cut-off, evaluate the proportion of players displaced across all grades.
   iii. Assuming the May 1st cut-off reduces player displacement across the grades, various weight percentiles are tested. This exercise reduces the proportion of displaced players depending on the selected percentile. 10th, 15th and 20th weight percentiles were extracted for each age-grade to allow an 80%, 70% and 60% peer-group participation rate, respectively. This procedure assumes that a fixed proportion of players are playing within their peer group before optimisation.

b. Develop an optimised age-weight grading system that significantly reduces player displacement (i.e. decrease age-weight variation among players within grades and increase variation among players across grades).
   i. A classification tree was used to isolate the optimal age-weight threshold, for each percentile. This generates a classification problem where the optimal age-weight cut-off for each grading system is to be determined. To derive an optimal age-weight system a classification tree is implemented; regressing age and weight, on grade. The resulting system will cater for a larger proportion of players to participate within their peer groups. The tree output will produce optimal age-weight thresholds for each grade, given a fixed proportion of player participation.

HYPOTHESIS

1. The January 1st date-of-birth cut-off is inappropriate as it does not align with the NZ school year nor the NZ rugby season, causing large player displacement.
2. A May 1st date-of-birth cut-off is more appropriate than January 1st, as it more closely aligns with the NZ school year, leading to reduced player displacement.
3. An optimised age-weight system that aligns with the NZ school year, accounting for the RAE and weight, significantly reduces player displacement.

3. RESULTS

The findings from Campbell *et. al.* (2018) suggest that aligning the junior rugby date-of-birth cut-off with the school year is could increase peer group participation. Here, using the data available, a review of the proportion of displaced participants by grade highlighted that the proportion of players displaced increases with age. This is not unexpected given the different growth rates of children. Youth growth charts reveals the different rates of change (for example: https://www.cdc.gov/growthcharts/clinical_charts.htm). This preliminary analysis suggests that there are two issues with the current grading system: 1) date of birth cut-off to adjust for the RAE and 2) weights limits to adjust for physiological disparity between participants within grades. It is suggested that shifting date-of-birth to align with the NZ school year will increase peer-group interactions and create greater alignment of rugby experiences.

ADJUSTING AGE CUT-OFF

Applying a May 1st cut-off produces a tighter grading system meaning there are fewer observations at tails of the distribution, across each grade. The proportion of displaced players, given a May 1st cut-off leads to a 7.4% reduction in players displaced compared to the current system (January 1st cut-off). A decrease in the standard deviation of ages and weights is also observed across all grades.

The proposed cut-off produces reduced age variation across grades. There is a 10.0% reduction in weight variability and a 12.4% reduction in the variability of age. This analysis suggests that a May 1st cut-off is more appropriate than January 1st to address player displacement, as it reduces weight variability between participants within grades, after accounting for age, and reduces overall player displacement.

INITIAL WEIGHT LIMIT ADJUSTMENTS

Given that the May cut-off produces 'better' age-weight bands and reduces the proportion of player displacement across all grades, the next steps are: 1. further reduce player displacement and 2. Identify the optimal age-weight threshold for each grade to reduce age separation across grades. This section addresses (1) by examining varying levels of participation and the affect each level (10th, 15th and 20th percentiles) has on the mean age, weight and standard deviation within each grade.

Examining the average age and weight standard deviation for each grade across these three percentiles, it is observed that the average age separation across the grades decreases as the percentiles decrease. This is because as the percentile decreases the system is reducing age disparity across the grades. However, as the percentile decreases and increasingly becomes age accommodating, the average weight standard deviation increases. Interestingly, compared to the current system (i.e. January 1st), the May cut off, decreases average age and weight standard deviation for all grades, across all percentiles. Given these results the study adopts the 15th percentile threshold, because: 1) it leads to a greater proportion of participants playing within their peer group, producing a reduction in participant displacement of approximately 20%, 2) it reduces age separation across all grades relative to the current system; and 3) it reduces weight separation across all grades relative to the current system.

MODELLING – CLASSIFICATION TREE

Applying a classification tree enables identification of optimal, grade specific, age-weight thresholds. The classification model will generate an optimised classification tree with each branch representing the ideal age-weight attributes necessary to participate within a given grade.

Classification trees, also known as Decision Trees, are a supervised classification learning technique made-up of 'decision nodes' with each decision node containing an individual test function of discrete outcome. A decision tree is a hierarchically organised structure, with each node splitting the data space into pieces on value of a feature, in this case, either age or weight. CART models employ a categorical target variable and the tree is used to identify the "class" within which a target variable would most likely fall into. The main elements of CART models are: 1) rules for splitting data at a node based on one variable (i.e. age or weight), 2) stopping rules for deciding when a branch is terminal and can be split no more; and 3) a prediction for the target variable (i.e. grade) in each terminal node.

"Regression trees organise nodes in a recursive, unidirectional, hierarchical fashion by repeated application of the test function" (Coomes, 2015, p. 16). Tree 'induction' (i.e. training) starts with all data set observations at the 'root' node and corresponding test function. The function splits observations into subsets that are input, via 'branches', to subordinate 'leaf' nodes, which in turn split observations to lower nodes.

The technique is a robust non-parametric alternative to classical parametric models. These models are robust to the distorting influences of complex variable interactions and interrelationships that would produce an unreliable model. CART models are "immune to the potential model-defeating characteristics of these effects and are a useful tool in identifying terms for the regression model to help models perform better" (De Ville & Neville, 2013, p. 55). Binary recursive partitioning is applied to the sample space which minimises the training error to improve fit. The recursive technique is a partitioning method "whereby the data are successively split along coordinates axes of the explanatory variables so that, at any point, the split maximally distinguishes the response variable in the left and right branches is selected" (Crawley, 2012, p. 686), these sequences of splits define a binary tree. The optimal split is found over all variables and all possible split points that bring about the largest drop in the residual sum of squares.

A Gini impurity measure ($i(t) = 1 - \sum_{j=1}^{k} p^2(j|t)$) was applied as the splitting rule, the branch splits are created at each node such that the variable that creates the greatest separation in the target variable is selected. This procedure further decreased age-weight variability between observations within a branch and increases age-weight separation between observations across branches. This criterion is ideal for this type of problem as pure classes (i.e. age-weight grades) must be identified. Overall the metric measures the impurity represent at each node which reveals the probability of obtaining two different outputs, which is an "impurity measure".

The classification model produced the age-weight categorisation system outlined in table 1. Table 2 shows an example system from the 2017 season).

DOB	Weight	Grade
1/05/2011 - 31/04/2012	No weight limit	Grade 1
1/05/2010 - 31/04/2011	No weight limit	Grade 2
1/05/2009 - 31/04/2010	if weight < 23.5kg	Grade 2
	if 23.5kg ≤ weight ≤ 35kg	Grade 3
	if weight > 35kg	Grade 4
1/05/2008 - 31/04/2009	if weight < 26.5kg	Grade 3
	if 26.5kg ≤ weight ≤ 40kg	Grade 4
	if weight > 40kg	Grade 5
1/05/2009 - 31/04/2008	if weight < 29kg	Grade 4
	if 29kg ≤ weight ≤ 45kg	Grade 5
	if weight > 45kg	Grade 6
1/05/2008 - 31/04/2007	if weight < 32kg	Grade 5
	if 32kg ≤ weight ≤ 51kg	Grade 6
	if weight > 51kg	Grade 7
1/05/2007 - 31/04/2006	if weight < 35kg	Grade 6
	if 35kg ≤ weight ≤ 57kg	Grade 7
	if weight > 57kg	Grade 8
1/05/2004 - 31/04/2005	if weight < 39kg	Grade 7
	if weight > 39kg	Grade 8

YOB	Weight	Grade	Grade Base Weight
2011 or after	Non-tackle, no weight limit	Under 6	
2010	Non-tackle, no weight limit	Under 7	
2009	Under 33kg	Under 8	Under 24 kg; Under 7
	33 kg or more	Under9	
2008	Under 36kg	Under 9	Under 26 kg; Under 8
	36kg or more but under 45kg	Under 10	
	54kg or more	Under 11	
2007	Under 40kg	Under 10	Under 27kg; Under 9
	40kg or more nut uder 60kg	Under 11	
	60kg or more	Under 12	
2006	Under 46kg	Under 11	Under 31kg; Under 10
	46kg or more but under 65kg	Under 12	
	65kg or more	Under 13	
2005	Under 55kg	Under 12	Under 37kg; Under 11
	55kg or more	Under 13	
2004	Open weight	Under 13	Under 46kg; Under 12

Tables 1 & 2. Comparison of proposed age-weight system (left) and a current weight age-system for 2017 (right)

4. DISCUSSION AND CONCLUSIONS

The proportions of junior rugby players displaced across all grades indicates that the current grading systems adopted in various NZ provinces are sub-optimal. In addition, as players get older and weights have increased variance, the proportion of players displaced also increases. The results indicate that the January 1st date-of-birth cut-off is sub-optimal. The conjecture is that this is due to misalignment with the NZ school year and commencement of the NZ rugby season. This finding is based on the reduced variance in both age (10.0%) and weight (12.4%) when a May 1st date-of-birth cut-off is implemented. This leads to reduced player displacement (7.4%). Adjusting weight limits using a classification tree in conjunction with the May 1st cut-off creates a system with a 20% reduction in player displacement.

Therefore, using the framework outlined here, an optimised age-weight system that aligns with the NZ school year, accounting for the Relative Age Effect (RAE) and weight, significantly reduces player displacement. This decrease in peer group displacement could lead to increase playing numbers and cultivate a more enjoyable playing environment, resulting in peer-to-peer relationships extending beyond the field. Future research could investigate the validity of applying a beta distribution to weights by grades and see how the shape parameters change across grades, dictating the rate of change in weight across grades.

References

Campbell, E. C., Bracewell, P. J., Blackie, E. & Patel, A. K. (2018). The Impact of Auckland Junior Rugby Weight Limits on Player Retention. *Journal of Sport and Health Research*. In press (accepted January 3, 2018).

Cobley, S., Baker, J., Wattie, N., & McKenna, J. (2009). Annual age-grouping and athlete development. *Sports medicine*, *39*(3), 235-256.

Coomes, M. (2014). Comparison of reject inference methods on complete data with gradient boosting machine variable selection.

Crawley, M. J. (2012). *The R books*. John Wiley & Sons.

Delorme, N., Chalabaev, A., & Raspaud, M. (2011). Relative age is associated with sport dropout: evidence from youth categories of French basketball. *Scandinavian journal of medicine & science in sports*, *21*(1), 120-128.

Delorme, N. (2014). Do weight categories prevent athletes from relative age effect? *Journal of sports sciences*, *32*(1), 16-21.

De Ville, B., & Neville, P. (2013). *Decision trees for analytics using SAS Enterprise Miner*. SAS Institute.

Felfe, C., Lechner, M., & Steinmayr, A. (2016). Sports and child development. *PloS one*, *11*(5), e0151729.

Heckman, J., & Tremblay, R. (2006). The case for investing in early childhood.

Lewis, J., Morgan, K., & Cooper, S. M. (2015). Relative age effects in Welsh age grade rugby union. *International Journal of Sports Science & Coaching*, *10*(5), 797-813.

Martin, R. (2017 July 10). Sports clubs tackle declined in player numbers. Retrieved from https://www.radionz.co.nz/news/national/334813/sports-clubs-tackle-decline-in-player-numbers

Patel, A. (2016). Roster-Based Optimisation for Limited Overs Cricket.

PlayRugby (2017 June 06). Junior Rugby. Retrieved from http://www.playrugby.co.nz/

Sabato, T. M., Walch, T. J., & Caine, D. J. (2016). The elite young athlete: strategies to ensure physical and emotional health. *Open access journal of sports medicine*, *7*, 99.

World Rugby (2013). *Putting players first: Weight Consideration Guideline.* Retrieved from http://playerwelfare.worldrugby.org/?subsection=64

Kerr, Z. Y., Marshall, S. W., Simon, J. E., Hayden, R., Snook, E. M., Dodge, T., ... & Nittoli, V. C. (2015). Injury rates in age-only versus age-and-weight playing standard conditions in American youth football. Orthopaedic journal of sports medicine, 3(9), 2325967115603979.

Stracciolini, A., Friedman, H. L., Casciano, R., Howell, D., Sugimoto, D., & Micheli, L. J. (2016). The relative age effect on youth sports injuries. *Medicine & Science in Sports & Exercise*, *48*(6), 1068-1074.

SYSTEMATIC OPTIMISATION OF THE ELO RATING SYSTEM

Wynton E. Moore [a,d], Samuel J. Rooney [a], Paul J. Bracewell [a], Ray Stefani [b]
[a] *DOT loves data, Wellington, New Zealand*
[b] *California State University, Long Beach, USA*
[d] *Corresponding author: paul@dotlovesdata.com*

Abstract

The Elo rating system, originally developed for chess, is increasingly being applied to mainstream sports. Beginning from first principles, we derive a generic framework for applying the Elo model to a wide range of sporting competitions. We show how to determine all model parameters systematically, utilising optimisation techniques to achieve maximum predictive power. Accurate initial ratings and home ground advantage are efficiently determined by constrained optimisation on a small training set. Once ratings trajectories have been computed, a logistic model is trained to predict win/loss outcomes from ratings differences. The Elo learning rate k is optimised for predictive performance, using a cross-validation scheme which respects the time structure of the match schedule. As an application, we derive team ratings and predict match outcomes for international and domestic rugby union tournaments.

Keywords: Elo ratings, learning rate, predictive modelling, cross-validation, rugby union

1. INTRODUCTION

Forecasting the outcome of sporting events is an ongoing challenge. Sport provides a unique setting for constructing predictive models, due to the accessibility of data and the range of factors, both quantifiable and not, that drive the outcome of competition. Accurate forecasting of sporting results is more important than ever, with the increasing sophistication of analytics used by professional sporting organisations, and the proliferation of online sports betting markets.

Early sports modelling work was based on ratings methodologies Stefani (2011). The most well-known of these are the Bradley-Terry Bradley and Terry (1952) and Elo (1978) models. Other modelling approaches have looked to incorporate predictive features into a modelling framework. Match outcomes have been directly modelled using various explanatory variables within supervised machine learning frameworks. For instance, football has seen the application of ordered probit regression models Goddard and Asimakopoulos (2004); Goddard (2005), and other machine learning classifiers Hucaljuk and Rakipovic (2011); Odachowski and Grekow (2013). Score distributions have also been studied, using parametric approaches; see for instance Maher (1982); Dixon and Coles (1997), where Poisson distributions are applied to goals scored in football.

Originally developed for rating chess players, Elo ratings can track changing abilities of players or teams (for the purpose of the Elo ratings system, teams are considered as single entities) as reflected in one-on-one matchups, which are not required to follow a schedule. In recent years they have been adapted to a wide variety of sports, both individual (such as tennis United (2018); Morris and Bialik (2015); Abstract (2018)) and team (football da Silva Curiel (2018); Hvattum and Arntzen (2009); Christoph Letiner and Hornik (2009); Lasek et al. (2013), American football Silver (2014), basketball Silver and Fischer-Baum (2015), and baseball Silver (2006), among others). See Aldous (2017); Király and Qian (2017); Stefani (2011) for recent mathematical reviews.

The Elo framework provides an intuitive model of team ratings, where the ratings of each entity are updated based on which team had a superior performance. The standard model relies on three parameters: (i) the initial rating r_0, (ii) B, which represents the scale for ratings differences, (iii) k, which controls the sensitivity of ratings to new match outcomes. A shift in ratings to account for home ground advantage is often used as an additional feature in Elo models of team sports. While these parameters are intuitive and provide parsimony, they are often chosen in an ad-hoc manner relying on domain experts to validate the model results.

In this paper we systematically apply the Elo rating system to the sport of rugby union. We approach the problem of applying the Elo model from first principles, by reviewing the foundations of the system to provide insight into the model parameters. We then adopt a systematic approach to determine all model parameters. Initial ratings and home ground advantage are determined by constrained optimisation of static ratings over a training set. In this approach the initial ratings are completely decoupled from the Elo learning rate k. We then optimise k by cross-validation of a logistic model which predicts match outcome as a function of ratings difference. We note an analogy between the Elo update algorithm and stochastic gradient descent, which provides a theoretical motivation for the optimisation of k. Our approach provides a generic framework for completely determining the Elo model for a wide range of sporting competitions.

2. METHODS

ELO RATING SYSTEM: REVIEW

The Elo rating system assigns each team i a dynamic rating r_i. Each match between two teams i,j causes the ratings r_i, r_j to be updated according to which team performed better. The ratings updates are zero-sum: if team i increases their rating r_i, then r_j will decrease by an equal amount. Hence the mean of all ratings is conserved and may be freely chosen at the outset; conventionally it is set to 1500.

The value of the ratings update is obtained by comparing the actual match outcome with the expected outcome based on prior ratings. Specifically, the update to r_i after a match between teams i and j is

$$r_i \;\rightarrow\; r_i + k\left[y_i - L\left(\frac{r_i - r_j}{B}\right)\right] \qquad (1)$$

and the corresponding update to r_j is obtained by switching labels $i \leftrightarrow j$. Here y_i represents the match outcome in favour of team i:

$$y_i = 1\,(win)\; or\; 0\,(loss) \qquad (2)$$

In the case of a draw $y_i = 1/2$ may be used, but this is not important for what follows. The second term in (1) represents an expected outcome between zero and one, where $L(x)$ denotes the logistic function

$$L(x) = \frac{1}{1+e^{-x}} \qquad (3)$$

Only the difference in ratings $(r_i - r_j)$ affects the expected outcome in (1). The constant B sets a scale for ratings differences. It is conventionally set to $B = 400/ln10$, so that an increase of 400 in the ratings difference corresponds to a tenfold increase in odds. Note that the sum of ratings $(r_i + r_j)$ is conserved by $L(x)+L(-x) = 1$.

The constant k in (1) controls the sensitivity of ratings to new match results. The value of k strongly affects accuracy of the ratings and has received much attention in the literature and applications. Various schemes have been devised for adapting the value of k to reflect winning margins or the significance of playoff matches. We will instead content ourselves with constant k and determine the optimal value by cross-validation.

It is worth mentioning there exists an upper bound on k above which the model does not make sense Aldous (2017). The rating r_i after a match must be an increasing function of the prior rating. If k is too large, the negative logistic term in (1) will have a stronger gradient than the prior term r_i. This implies an upper bound of:

$$k < 4B \qquad (4)$$

This ensures the post-match value of r_i is always an increasing function of the pre-match value, regardless of the opponent's rating r_j. The value of B quoted above leads to a bound $k \lesssim 690$, which is much larger than typical values encountered in practice.

CONTINUOUS MEASURE OF MATCH OUTCOME

The update rule (1) does not consider the margin of victory; it cares only about the binary win/loss outcome y_i. The modern game of rugby is a high-scoring game which gives ample opportunity to replace y_i with a continuous outcome variable valued on [0,1]. Here we adopt the points ratio:

$$v_i = \frac{points(i)}{points(i) + points(j)} \qquad (5)$$

to measure the outcome of a match between teams i and j. This measure is suited to rugby, where v_i explores the range [0,1] with reasonable support, and should also be well-adapted to sports such as American football. In other sports such as basketball, v_i is probably too close to 1/2 to be of use.

The modified update rule for our ratings system is simply:

$$r_i \;\rightarrow\; r_i + k\left[v_i - L\left(\frac{r_i - r_j}{B}\right)\right] \qquad (6)$$

The logistic term now acts as an expected value for the points ratio v_i. The ratings will change when the points ratio is above or below this expected value. It is possible for a winning team's rating to decrease if they did not win convincingly (meaning the points ratio was less than expected). However, we insist that a winning team cannot fall behind their opponent (assuming they had a prior ratings advantage). That is, only in an upset should it be possible for the ratings difference $(r_i - r_j)$ to change sign. This condition results in a slightly more stringent bound on k:

$$k \leq 2B \qquad (7)$$

Again, this bound $(k \lesssim 345)$ is far more than typical values found in practice. Another interesting feature of the continuous outcome variable is seen when considering the interplay between skill and chance. In a hypothetical sport which involves no element of chance, the original Elo update rule (1) will drive ratings differences to infinity. This is because the logistic function can never saturate the binary outcome variable, except at its asymptotes at $\pm\infty$. The continuous-outcome rule (6) instead attracts ratings differences to finite values which reproduce the points ratio. The dominant team's rating is not pushed to infinity. This behaviour is appealing because finite ratings contain richer information and are more human-interpretable. In realistic

scenarios with an element of chance, the continuous-outcome ratings are more stable and less sensitive to match scheduling.

INITIAL RATINGS AND HOME GROUND ADVANTAGE

Conventionally the Elo system has been initialised by setting all ratings to the mean value (1500) before evolving them over some training set of historical matches. Eventually the ratings should equilibrate, but this typically requires a large quantity of training data. We have found it more expedient to set the initial ratings equal to a set of *static* ratings $\left\{r_i^{(0)}\right\}$ which minimise a loss function over the training set, Opisthokonta (2016) used a similar method. This is a constrained optimisation problem, where the mean rating is constrained to be 1500. A large historical training set is neither necessary nor advantageous, because we want the static ratings to be accurate near the end of the training period. Another advantage of the static initial ratings is their independence from k; this will be crucial when we study the optimisation of k in Section **Error! Reference source not found.**.

Home ground advantage is an important factor in rugby union. To allow for this we apply a constant rating shift h to the home team in any match. This constant h is optimised simultaneously with the initial ratings. The loss function for the optimisation is:

$$SS\left(\left\{r_i^{(0)}\right\}; h\right) \;=\; \sum_R \left[v_i \;-\; L\left(\frac{r_i^{(0)} \,-\, r_j^{(0)} \,\pm h}{B}\right)\right]^2 \qquad (8)$$

subject to the mean rating constraint. We have chosen to define the loss function using the continuous outcome variable v_i and hence a squared loss. One could also use the binary outcome y_i and log-loss. Our intention is for the logistic function $L((r_i - r_j \pm h)/B)$ to predict v_i; a separate logistic function with latent parameter will later be trained (using log-loss) to predict y_i. Here R denotes the set of training matches, and the home ratings shift h is added (subtracted) if team i is at home (away). Typical values are h~40, which corresponds to a contribution of $10^{\frac{1}{10}} \approx 54$ to the ratio $v_H/(1 - v_H) = points(home)/\, points(away)$. In the binary Elo system this would be a contribution to the odds in favour of the home team (but the value of h may be different). Once h is determined during initialisation, it is fixed and does not change as ratings evolve. The ratings update rule (6) is adjusted to take account of h in the expected outcome, becoming:

$$r_i \;\rightarrow\; r_i \,+\, k\,\left[v_i - L\left(\frac{r_i - r_j \,\pm h}{B}\right)\right] \qquad (9)$$

BALANCE RELATION

An under-appreciated aspect of the original Elo system, which was pointed out in Aldous (2017), is the following: if one wishes to predict match outcomes solely as a function of ratings differences, then the probability distribution on match outcomes is determined by the update rule (1). If the ratings are well-fitted and represent the true strengths of the teams (and if these true strengths are static), then the *expected* change in ratings should be zero. Of course, in reality the true strengths are dynamic, which is what makes sports interesting. The point is that, on average, any change in ratings should be due to varying strengths only. To see what this means for the distribution of match outcomes, take the expectation of the ratings update (1):

$$E[_i] \;=\; k\,\left[E[y_i] - L\left(\frac{r_i - r_j \,\pm h}{B}\right)\right] \qquad (10)$$

The *balance relation* states that (10) is required to vanish. This implies that the binary match outcome y_i follows the distribution with $p(y_i = 1)$ equal to the logistic term.

In the continuous-outcome system the points ratio v_i instead follows an unknown distribution. It would be interesting to derive this distribution by assuming, for instance, a Poisson process for point-scoring. The balance relation places a constraint only on the mean of the distribution (which is distinct from the win/loss probability). A separate logistic model will be fitted to predict win/loss outcome as a function of ratings difference.

3. RESULTS AND DISUCSSION

Elo ratings timeseries for the 2017 season of Super Rugby are displayed in Figure 1, and details of the playoff rounds are given in Table 1. Initial ratings were determined by fitting static ratings to the entire 2016 season, and ratings were evolved using the points ratio update rule (6). The points ratio in favour of the home team is denoted by v_H. The root-mean-squared error in the points ratio predictions is 0.14, compared with a standard deviation of 0.18. Particularly noticeable is the large ratings jump for the Crusaders from their 17-0 victory over the Highlanders in the quarter-final, which achieved the maximum points ratio of one.

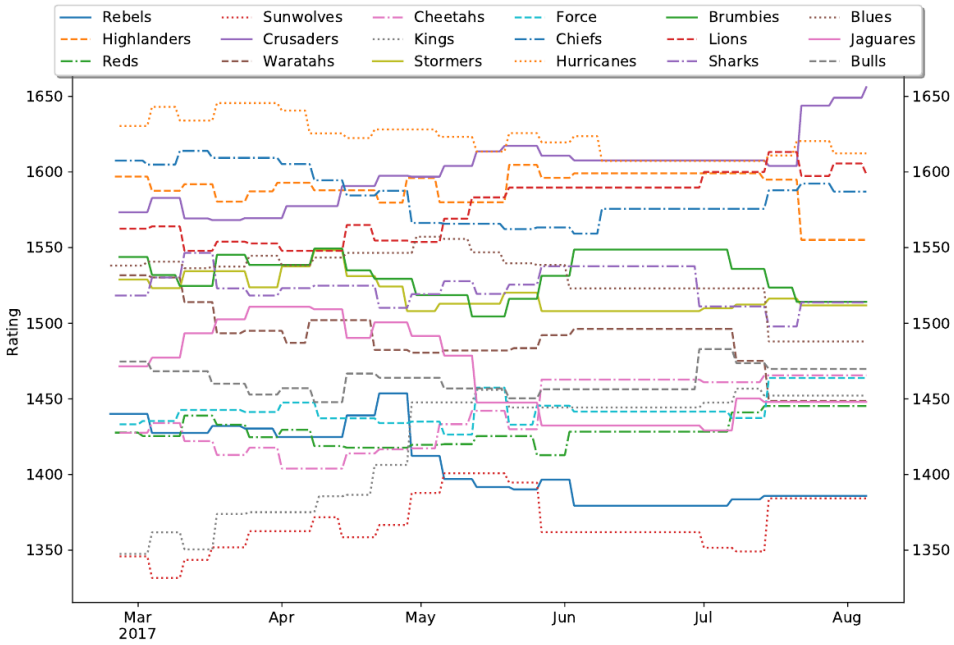


Figure 1: Elo ratings timeseries for the 2017 season of Super Rugby, with $k = 90$. Initial ratings were determined by fitting static ratings to the entire 2016 season. The home ground advantage is $h \approx 31$.

Date	Home score	Away score	V_H	Predicted v_H	Home ratings gain
July 21	Brumbies 16	Hurricanes 35	0.31	0.42	-9.6
July 22	Lions 23	Sharks 21	0.52	0.70	-15.9
July 22	Crusaders 17	Highlanders 0	1.0	0.56	39.8
July 22	Stormers 11	Chiefs 17	0.39	0.44	-4.4
July 29	Lions 44	Hurricanes 29	0.60	0.51	8.2
July 29	Crusaders 27	Chiefs 13	0.68	0.62	5.2
August 5	Lions 17	Crusaders 25	0.40	0.48	-7.0

Table 1: Predictions and results for the playoff rounds of Super Rugby 2017.

A separate logistic model was trained to predict win/loss outcomes, with latent parameter $B \approx 59$. This is much less than $B \approx 174$ in the points ratio model, so the win/loss model is much more responsive to ratings differences. This results from the use of a log-loss penalty to train the win/loss model; ratings differences are a reliable predictor, so stronger predictions are preferred. The mean accuracy of the win/loss model is 77%, compared to a baseline of 56% home team wins.

Figure 2 shows static initial ratings for international teams, trained on matches from the 2017 calendar year. In this example there are two subgroups of teams which play a regular annual tournament (The Rugby Championship: New Zealand, Australia, South Africa, Argentina; and the Six Nations: England, Ireland, Scotland, Wales, France, Italy). The mean ratings in these two groups are 1652 and 1608 respectively; the relative difference is set by interactions between the groups and their common interactions with other teams. It corresponds to an average points ratio of $points(i)/points(j) \approx 1.29$ between teams i (belonging to Rugby Championship) and j (Six Nations). The home ground advantage is $h \approx 41$.

OPTIMISATION OF K: CROSS-VALIDATION

The Elo ratings model (6) contains a parameter k which plays the role of a learning rate. This parameter can be tuned to achieve optimal predictive performance on a test set of unseen matches. The optimal value strikes a balance between bias (low k, static ratings) and variance (high k, ratings fluctuate strongly). Note that the optimisation of k is necessarily separate from that of the initial ratings discussed in Section Initial Ratings and home ground advantage. If one attempts to simultaneously optimise k with the initial ratings, the solution will be $k = 0$ (static ratings) and the initial ratings will reflect the entire data set.

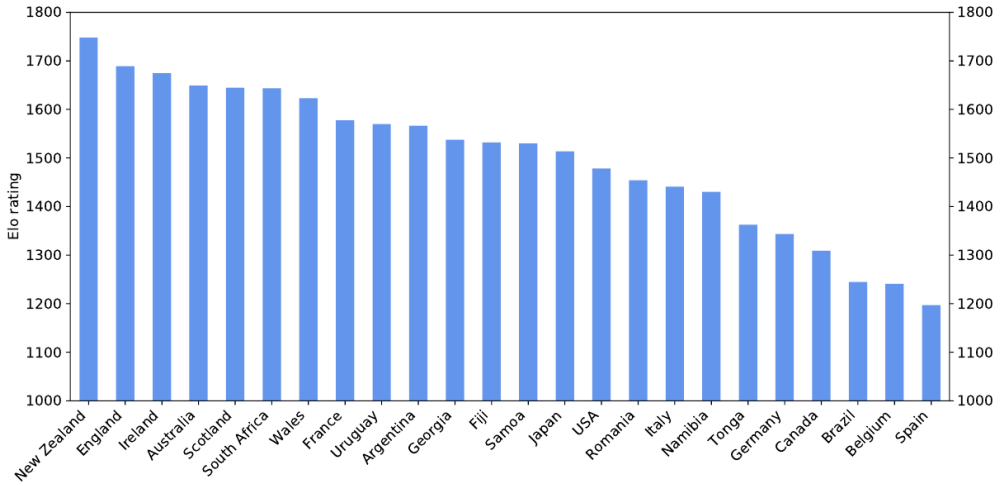


Figure 2: Static initial ratings for international rugby teams trained on matches from the 2017 calendar year.

Predictive performance of the Elo ratings is measured by training a logistic model to predict win/loss outcomes as a function of ratings differences. Lasso regularisation is used to prevent overfitting, and the inverse regularisation strength C is varied. Hence, we are optimising predictive performance as a function of k and C simultaneously.

Three-fold cross-validation was performed using the Python package *Scikit-learn* Pedregosa et al. (2011), and a cross-validation scheme (TimeSeriesSplit) which respects the time structure of the data. Seasons 2016 and 2017 were combined to ensure the volume of data was sufficient for cross-validation. Both parameters exhibit a smooth trade-off between bias and variance, and there is a clear minimum loss at $(k = 90,\ C = 6 \times 10^{-3})$.

BIAS-VARIANCE DECOMPOSITION

The above determination of k by cross-validation is motivated by a bias-variance decomposition of the loss function. The form of this decomposition may be seen through an analogy with stochastic gradient descent, where k plays the role of the learning rate. Consider a simplified scenario of just two teams with a true ratings difference x_{true}, and corresponding win probability $p = L(x_{true}/B)$. The bias-variance decomposition for stochastic gradient descent was given in Schaul et al. (2013); following similar arguments, we arrive at

$$\langle L(x) \rangle \;=\; L(x_{true}) + \frac{p(1-p)}{2B^2}\left[\left(1 - \frac{2kp(1-p)}{B}\right)^2 (x - x_{true})^2 \;+\; 4k^2\sigma^2\right] \;+\; O(x - x_{true})^3 \qquad (11)$$

Here $\langle L(x) \rangle$ is the expected log-loss for the logistic model which predicts match outcomes, x is the current estimate of the ratings difference, and the variance is $\sigma^2 = p(1-p)$. The precise details of this expression are not important; what is significant is the separation into irreducible randomness, bias, and variance (the first three terms respectively). The expansion (11) holds when x is close to the true value x_{true}, and in that case the optimal learning rate is $k_* = (x - x_{true})^2/B$. The bias-variance decomposition (11) motivates the existence of an optimal value for k, which is correctly determined by cross-validation. In realistic scenarios with more than two teams, there is a tension between different pairs of teams which reduces the optimal learning rate; however, a qualitatively similar decomposition still exists.

CORRELATION BETWEEN OPTIMAL K AND UNEVENNESS

When the cross-validation analysis is repeated for different rugby tournaments, the optimal value k_* of the Elo learning rate shows significant variation. This variation is not random but appears to depend on the "unevenness" of the tournament: tournaments with a clear hierarchy between good and bad teams tend to have a higher value of k_* (Aldous (2017) observed a similar correlation), to quantify this "unevenness" it is desirable to use a measure which is determined independently of k, from training data only. The standard deviation of the initial ratings $\{r^0\}$ fulfils these requirements. Hence, we investigate the relationship *(k* scales like an Elo rating, by the update rule (9)):

$$k_* \;=\; \alpha + \beta\ s(\{r^0\}) \qquad (12)$$

The optimal learning rate k_* was computed by cross-validation for nine domestic and international club rugby tournaments, using data from 2016 and 2017. We focus on club tournaments (rather than international test matches) because they provide a sufficient volume of data and matches follow a weekly schedule. While the Elo system does not require matches to follow a schedule, infrequent matches may lead to a higher value of k*

simply because team strengths are likely to change over time. By focussing on club tournaments this possibility is avoided. The coefficient $\beta \approx 2.26$ is statistically significant at the 95% level. Because the predictor $s(\{r^0\})$ is generated from training data only, the correlation (12) allows us to estimate the optimal k for future predictive performance in real-life scenarios.

4. CONCLUSIONS

The Elo ratings model provides a simple and intuitive approach for describing the dynamics of team strength based on match day performance. We have presented a systematic approach to determining the parameters of the Elo model. Our approach maximises predictive power and does not depend upon validation by domain experts. The key points are the efficient optimisation of initial ratings, the decoupling of initial ratings from k, and the use of cross-validation to determine k itself. There is a significant relationship between k and unevenness of competition which is suitable for future investigation.

References

Abstract, T. (2018). Tennis Abstract: ATP Elo ratings. tennisabstract.com/reports/atp_elo_ratings.html.

Aldous, D. (2017). Elo ratings and the sports model: A neglected topic in applied probability? *Statist. Sci.*, 32(4):616–629.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Christoph Letiner, A. Z. and Hornik, K. (2009). Forecasting sports tournaments by ratings of (prob)abilities: A comparison for the euro 2008. *International Journal of Forecasting*, 26(3):471–481.

da Silva Curiel, R. S. (2018). World football Elo ratings. eloratings.net.

Dixon, M. J. and Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265–280.

Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco Pub.

Goddard, J. (2005). Regression models for forecasting goals and match results in association football. *International Journal of Forecasting*, 21:331 – 340.

Goddard, J. and Asimakopoulos, I. (2004). Modelling football match results and the efficiency of fixed-odds betting. *Journal of Forecasting*, 23:51 – 66.

Hucaljuk, J. and Rakipovic, A. (2011). Predicting football scores using machine learning techniques. In *Proceedings of the 34th International Convention MIPRO*, pages 1623–1627.

Hvattum, L. M. and Arntzen, H. (2009). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, 26(3):460 – 470. Sports Forecasting.

Király, F. J. and Qian, Z. (2017). Modelling competitive sports: Bradley-Terry-Élő models for supervised and on-line learning of paired competition outcomes. *CoRR*. arxiv.org/abs/1701.08055.

Lasek, J., Szlávik, Z., and Bhulai, S. (2013). The predictive power of ranking systems in association football. *International Journal of Applied Pattern Recognition*, 1(1):27–46.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118.

Morris, B. and Bialik, C. (2015). Serena Williams and the difference between all-time great and greatest of all time. fivethirtyeight.com

Odachowski, K. and Grekow, J. (2013). Using bookmaker odds to predict the final result of football matches. In *Knowledge Engineering, Machine Learning and Lattice Computing with Applications*, pages 196–205, Berlin, Heidelberg. Springer Berlin Heidelberg.

Opisthokonta (2016). Tuning the Elo ratings: Initial ratings and inter-league matches. opisthokonta.net/?p=1412.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Schaul, T., Zhang, S., and LeCun, Y. (2013). No more pesky learning rates. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 343–351, Atlanta, Georgia, USA. PMLR.

Silver, N. (2006). Lies, damned lies: We are Elo? baseballprospectus.com

Silver, N. (2014). Introducing NFL Elo ratings. fivethirtyeight.com/features/introducing-nfl-elo-ratings.

Silver, N. and Fischer-Baum, R. (2015). How we calculate NBA Elo ratings. fivethirtyeight.com

Stefani, R. (2011). The methodology of officially recognized international sports rating systems. *Journal of Quantitative Analysis in Sports*, 7(4).

United, O. (2018). Weekly tennis ELO rankings. tenniseloranking.blogspot.com.

DERIVING RESULT-DRIVEN RUGBY PLAYER RATINGS

Paul J. Bracewell[a,b,c] , Jack McIvor [a,] , Wynton E. Moore [a], and Ray Stefani
[a] *DOT Loves Data, Wellington*
[b] *Victoria University, Wellington*
[c] *Corresponding author:* paul@dotlovesdata.com

Abstract

A result-driven player rating system is developed for rugby union, using a methodology outlined by Patel *et. al*. (2017) for deriving cricket player ratings. Random forest regression (combined with expert domain knowledge) is used to identify the most important player attributes for each position on a rugby team. A dynamic probability of winning score is used as the target for the regression model, as binary match outcome is too coarse to be an effective target. This probability of winning is derived by training an ensemble of logistic models to predict match outcome from time-dependent match statistics. In addition to its role as a target, the probability of winning model enables each player's accumulated influence on probable match outcomes to be directly quantified. Finally, a novel team rating is derived from the constituent player ratings. This team rating allows natural comparison between teams across different tournaments. It is found to predict match outcomes with greater accuracy than the Elo rating system (Moore et. al., 2018).

Keywords: Random forest, Ensemble, Real-time estimation

1. INTRODUCTION

A core component of media coverage of sports events is the provision of player ratings. Major New Zealand media outlets: NZME (www.nzherald.co.nz) and Fairfax (www.stuff.co.nz) regularly generate rugby ratings following a match. This provides an indication of the perceived demand for this type of information.

This commercial interest motivated the development of a meaningful rating system that is both informative and intuitive. For any rating system to be successful in popular media, it must withstand substantial scrutiny. As a consequence, we leverage insights from tools that are popular within cricket. For example, the real time Winning and Score Predictor has had both academic (Shah *et. al*., 2016) and public scrutiny (Bayer, 2015). Past experience with commercial rugby rating systems (Bracewell, 2003) and investigation of the key features of a meaningful human-based rating system (Patel *et. al*., 2018) highlighted key elements within the process that need to be resolved. Substantial research has been undertaken to track and quantify individual player involvement in rugby games (e.g. Deutch *et. al*. (2007), Hughes (2004), James *et. al.* (2005), Quarrie *et. al.,* (2007), Quarrie *et. al*. (2012). These studies typically identify the key characteristics that differentiate either the physical or physiological demands of different positions. Not surprisingly, the outcomes of this research have tremendous value for developing skill and conditioning programmes for athletes. However, these studies typically do not extend the research further to quantify the impact of the action within the context of the match.

In order to construct a meaningful rating system that is robust and transparent, the approach adopted is focused on winning. As the entities involved are professional sports people, match outcomes and the contribution to that outcome are critically important. Patel *et. al*. (2017) demonstrated that by constructing a team rating model composed of individual rating systems, better predictions of match outcome are obtained (13% increase in predictivity). If betting agencies are used as a proxy for public opinion that analysis also demonstrated the combined ratings outperform head-to-head odds as a predictor of match outcome.

Intuitively, victory requires winning key moments within a game. It is further hypothesised that quantifying this impact provides an even richer perspective of player behaviour. The intent then is that this approach enables meaningful individual comparisons, such as: *Player V* is the most valuable player now, with an individual rating of *W*, because they rate high on *X* and *Y* which improves their team's chances of winning by *Z*%.

As we usher rugby audiences into a new era of appreciating the game through informed analytic insight, we need to build trust. Our framework, geared around moments (Brown *et. al*. 2016) in real time lends itself to a transparent framework. Importantly, as we expand on team ratings using individuals (Bracewell *et. al*, 2016) creates a strong position to understand and compare competitions (Bracewell *et. al*. 2009; Jowett *et. al*. 2016), due to relative performance and depth of talent. Ultimately, what is described here is an expert system. That is, using a combination of machine learning, shaped by winning outcomes and guided by human observers, we are creating a dynamic system that will output meaningful, rugby orientated output that will stimulate, engage and challenge thinking of those interested in rugby across all levels.

3. DATA

A rating system which scores specific player attributes requires a source of rich match data. Acquisition of this type of data often requires a commercial relationship due to the effort and cost required to capture the data. This creates a barrier to accessing this type of information. Opta Sports (www.optasports.com) is a leading supplier of detailed sports data. Since 1996, they have collected a large amount of rugby data on a match-by-match basis. Using notational analysis on recorded match footage (Hughes, 1993), approximately 1600 coded actions are captured per match. This is not live data; the actions are coded manually following completion of the match from recorded footage. Each row in the dataset consists of the following fields: time, pitch coordinates, sequence and phase identifiers, player and team identifiers, and three levels of action descriptors. This data is received as part of a commercial relationship and accessed via an API.

4. METHODOLOGY

At its heart, a player rating system should reflect each player's contribution to winning matches. Simply, for each game, we have player actions and the subsequent match outcomes. Thus it is a relatively trivial exercise to build a descriptive model which aligns the independent variables (player actions) with the dependent variable (match outcome). However, as discussed below, this is insufficient. As a consequence, we evolve this to a live odds model, which provides a sensitive measure of contributions to winning. This in turn becomes the dependent variable in constructing individual rugby player ratings.

Our first attempt at identifying winning contributions was to construct a model with the match outcome (win or loss) as the target variable and player actions as the features. A number of different models were tried, including logistic regression and random forests. Linear models using points difference as the target were also attempted. The problem with these approaches is a low signal to noise ratio. There is only one outcome per match, but over a thousand actions. Which team has the upper hand may change a number of times during a match, and may depend on various factors other than the current score. There are at least two possible approaches for increasing the signal to noise ratio. The simplest would be to aggregate the actions to match level before modelling. Alternatively one can go in the other direction, attempting to define a new target variable which captures the ebbs and flows of the match. Here we adopt the second approach; our new target variable is the “live odds," ie. the estimated probability of victory P(Win) for a chosen reference team (the home team). The live odds model is trained on macroscopic features such as current score, possession and territory. It is not trained on individual player actions; rather, it acts as a target for modelling the effects of these. The two approaches described in the previous paragraph are not mutually exclusive. Once every observation in a match has been scored with the live odds model, we may aggregate to any time step of our choice before modelling the effect of player actions. Aggregating over the full duration of the match would recover the first approach.

LIVE ODDS

The live odds model must generate a probability P(Win) at each moment of a match, based on macroscopic features. A natural and simple way to accomplish this is to define a family of logistic regression models labelled by time. The time ranges from 0-80 minutes in increments of a chosen time step, which we take to be one minute. The choice of time step is a balance between ensuring sufficient training data (longer time step) and retaining as much time structure as possible (shorter time step). We trained the live odds model on a set of over 600 matches from a wide variety of rugby tournaments, at both international and club level. The training set for the logistic model labelled by time t consists of one observation per match. Each observation consists of the “state" of the match at time t as represented by the macroscopic features, and the eventual outcome of the match (win or loss). The features are given by the home team's advantage in each of: pre-match team rating, current score, and rolling averages of possession and territory taken over a five minute window. We use the Elo team ratings discussed in (Moore *et. al*., 2018) as the pre-match ratings. This system has a few key advantages over the original methodology: it targets the points ratio (which holds more information than simply win/loss), accounts for home team advantage and optimises the speed of update for each competition.

Figure 1 shows how the coefficients evolve over time with respect to predicting match outcome. Not surprisingly, the most important element in the early stages of the match are the ratings of the competing teams. This is trivial, as it is not unreasonable to expect better teams to have a greater chance of winning. As the game unfolds, the score begins to hold more information about the likely match outcome. Around half time the score difference becomes the most important element for predicting final outcome. Interestingly, the impact of the team ratings decays almost linearly until the final ten minutes of the match, where there is a rapid decay. This provides interesting insight into professional rugby. An interpretation of this evolution of the rating coefficient

is that more favoured teams have the potential to claw their way back into the game at any stage in the first 70 minutes. However, when it comes to the last ten minutes there may be insufficient time to structure scoring opportunities to deliver the win.

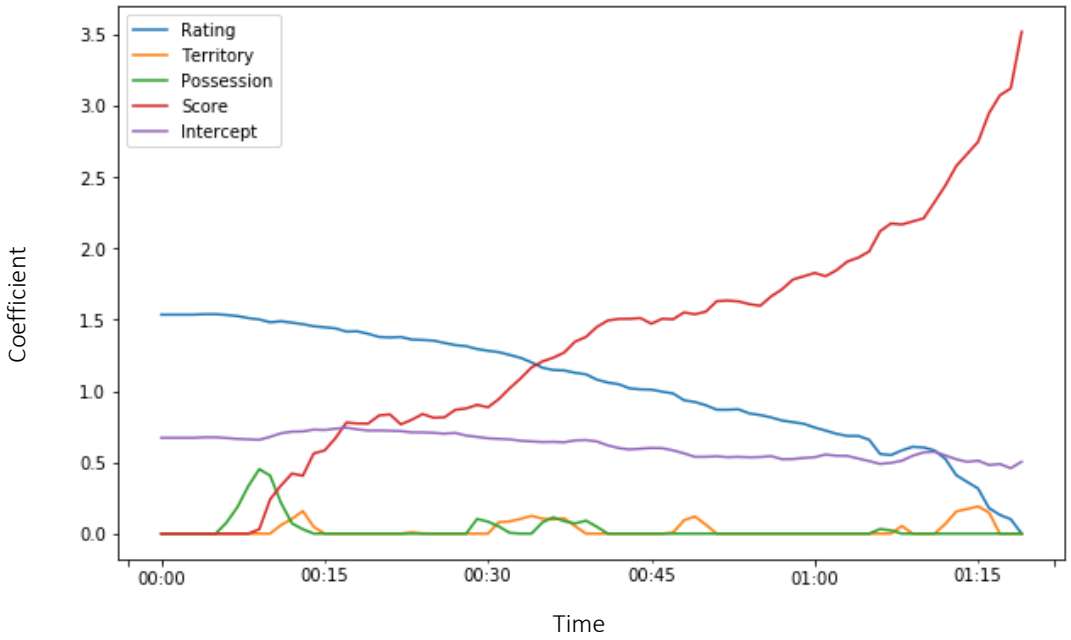


Figure 1: Coefficients vs time in the live odds model

As expected, the accuracy of the model improves over the course of the game. This is shown in Figure 2, where the log-loss of the model drops consistently over the course of the match. The decay of the log-loss is approximately linear over the first 60 minutes of the match, and then tends to the minimum much more quickly in the last 20 minutes.

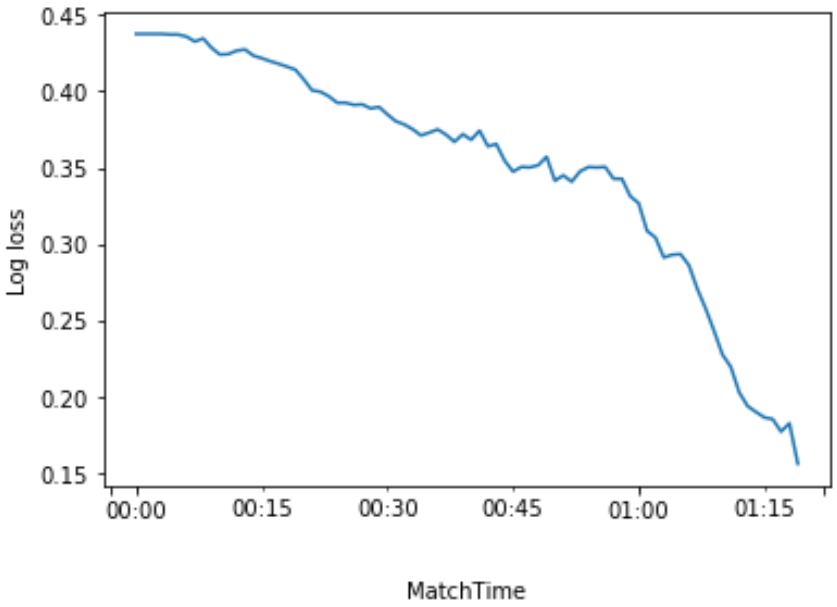


Figure 4: Log loss for the live odds model vs time

INDIVIDUAL METRICS

This probability of winning is derived by training an ensemble of logistic models to predict match outcome from time-dependent match statistics. In addition to its role as a target, the probability of winning model enables each player's accumulated influence on probable match outcomes to be directly quantified. This enables a framework to meaningfully convert actions recorded as player statistics to player ratings. This is firstly achieved by identifying positional groupings and defining position-specific attributes. In addition, each interaction an individual has is assigned a value based on the shift in probability of winning over the next two minutes. This enables a wider range of behaviours to be included. Furthermore, it provides some resistance to match context and reduces the impact of the interpretation of the coders who record the game activities.

There are some interactions within a game where it is difficult to isolate individual impacts. Scrums are an example of this behaviour. To handle this situation, a group score is calculated and distributed to the individuals. Each scrum a forward player is involved in counts toward their scrummaging score. Each player is a member of a forward pack at the scrum, and that forward pack can either win or lose the scrum, whether by winning possession directly or via a penalty. Scrum wins add one point to each player's scrummaging score, and losses subtract one point. These scores are aggregated for each player across all scrums they take part in, in all their match appearances. To validate this statistic, it was found that teams that had the strongest scrum rating were 32% more likely to win.

Positional specific metrics are identified through regression analysis to identify attributes that are statistically and practically related to within game changes to the probability of winning. For example, in addition to the influence measure, midfield backs (second five eighth and centre) are characterised by the following key metrics: Line Breaks, Defenders Beaten, Offload in Tackle, Carry Metres, Tackles. Missed tackles negatively contribute to their rating. Openside flankers have a different profile. In addition to influence, the ability to turn the ball over, via Jackal, Snaffle or Turnover Tackle is favourable. Other positive factors are tackles, pick and go metres, carry metres and the scrum score outlined previously. Missed tackles negatively contribute to their rating.

INDIVIDUAL MATCH RATING

Then, to develop the rating, the following process is deployed. Firstly the player features are aggregated to match level and normalised by the minutes played. Then these features are grouped by position. For each position a linear transformation of the aggregated information is learnt which sets the variance of each feature to unity. This enables cross position comparison, and is similar to the process that Bracewell (2003) deployed using factor analysis. This is further extended to form an overall rating as a linear combination of features, where the weights in the linear combination are defined by position and are provided by a consensus of rugby experts. Again, grouping by position, a set of quantile transformations is learnt which transform the match ratings into standard normal variables. This raw rating is theoretically non-bounded. Consequently, a sigmoid function is applied to obtain match ratings in (0; 1).

In order to derive player ratings based on a series of matches, the following extension occurs. For each player, their current rating is calculated as an exponentially weighted mean of past ratings. Similar to deriving the match ratings, a single quantile transformation is trained which maps the current player ratings into a standard normal variable. This transformation is blind to position which is important for cross position comparison. Finally, a sigmoid function is applied to obtain current player ratings in (0; 1). This transformation is order preserving and has properties that may be adjusted by the exact specification of the sigmoid. For this application, we choose a logistic function which transforms ratings to follow a logit-normal distribution. This allows the ratings to follow a pre-specified distribution which accounts for how they are perceived. For example, the public may expect only the top few players to be in the (0.9, 1) band, with a mode of about 65.

To externally validate the efficacy of individual ratings combined to create a team rating, as described by Patel *et. al.* (2017), an expected probability of victory was obtained by training the ratings against historical outcomes. Odds provided by the New Zealand TAB were converted to probability of victory. The assumption is that the odds will be a proxy for public opinion regarding likely outcome. These two statistics were compared against the ratio of victory, defined as the winning team points divided by the total points scored in the match. Contrasting the results from the late rounds of the 2017 Super Rugby Tournament against the prediction revealed that the individually-derived team ratings out-performed the TAB. As a specific example, in week 14 these new ratings outperformed the TAB odds by 24.4%. Interestingly, only one game in that round was beyond what was expected. The Sunwolves decimation by the Cheetahs (7-47) was the only result more than 2 standard deviations of the expected result. That is, there was less than a 1 in 20 chance of observing that result. All other games were within 1 standard deviation.

5. CONCLUSION

To win games in an elite rugby environment, key moments need to be won. The emphasis on winning and predictivity ensures we obtain parsimonious models that are aligned with perception and are more readily interpretable and as a consequence, defendable. The motivation for the methodology adopted was inspired by commentary regarding commercial systems in cricket and previous experiences with rugby systems. For instance, when we compared a team rating system for teams combining individual ratings (Patel *et. al.,* 2017) with a team rating based only on team performances (Bracewell *et. al*., 2014) we obtained a 13% improvement in predictivity.

As a consequence, a result-driven player rating system is developed for rugby union, using the methodology outlined by Patel *et. al*. (2017) for deriving cricket player ratings. Random forest regression (combined with expert domain knowledge) was used to identify the most important player attributes for each position on a rugby team. A dynamic probability of winning score is used as the target for the regression model, as binary match outcome is too coarse to be an effective target. This probability of winning was derived by training an ensemble of logistic models to predict match outcome from time-dependent match statistics. In addition to its role as a target, the probability of winning model enables each player's accumulated influence on probable match outcomes to be directly quantified.

Importantly, due to the standardisation and scaling that occurs between the positional groupings, individuals can be compared across positions. As such, we can also expand this research to comparing across generations, positions, partnerships and coaches to name a few issues that consumers would find interesting.

References

Bayer, K. (2015, January 24). WASP has sting for diehard ODI fans. New Zealand Herald. Retrieved from https://www.nzherald.co.nz.

Bracewell, P.J. (2003). Monitoring Meaningful Rugby Ratings. Journal of Sports Sciences, 21 (8), pp. 611-620.

Bracewell, P.J., Blackie, E., & Boys, C. (2016). Understanding the Impact of Demand for Talent on the Observable Performance of Individuals. Proceedings of the 13th Australian Conference on Mathematics and Computers in Sports. (pp. 39-44). Melbourne, Australia: ANZIAM Mathsport.

Bracewell, P.J., Downs, M.C.F, & Sewell, J. W. (2014). The Development of a Performance Based Rating System for Limited Overs Cricket. Darwin, Australia: ANZIAM Mathsport

Bracewell, P.J., Forbes, D.G.R., Jowett, C.A., & Kitson, H.I.J. (2009). Determining the Evenness of Domestic Sporting Competition Using a Generic Rating Engine. Journal of Quantitative Analysis in Sports. 5(1).

Brown, P., Patel, A.K., & Bracewell, P.J. (2016). Real Time Prediction of Opening Batsmen Dismissal in Limited Overs Cricket. Proceedings of the 13th Australian Conference on Mathematics and Computers in Sports. (pp. 80-85). Melbourne, Australia: ANZIAM Mathsport.

Deutsch, M.U., Kearney, G.A., & Rehrer, N.J. (2007). Time-motion analysis of professional rugby union players during match-play. J Sports Sci; 25(4): pp. 461–472.

Hughes, M. (1993). Notation analysis in football. In Science and Football II (edited by T. Reilly, J. Clarys and A. Stibbe), pp. 151–159. London: E & FN Spon.

Hughes, M. (2004). Performance analysis—a 2004 perspective. Int J Perform Anal Sport; 4(1): pp. 103–109.

James, N., Mellalieu, S.D., & Jones, N.M.P. (2005). The development of position-specific performance indicators in professional rugby union. J Sports Sci; 23(1): pp. 63–72.

Jowett, C.A., Rooney, S.J., & Bracewell, P.J. (2016). Team Performance and Conference Calibration. Proceedings of the 13th Australian Conference on Mathematics and Computers in Sports. (pp. 50-55). Melbourne, Australia: ANZIAM Mathsport.

Patel, A.K., & Bracewell, P.J., (2018) A Framework for Quantifying the Effectiveness of Human-bnased Rating Systems. Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports. Sunshine Coast, Australia: ANZIAM Mathsport.

Patel A.K., Bracewell P.J., & Rooney S.J. (2017). An Individual-based Team Rating Method for T20 Cricket. Journal of Sport & Human Performance 2017; 5(1), pp. 1-17.

Quarrie K.L., & Hopkins W.G. (2007). Changes in player characteristics and match activities in Bledisloe Cup rugby union from 1972 to 2004. J Sports Sci; 25(8):pp. 895–903.

Quarrie, K.L., Hopkins, W.G., Anthony, M.J., & Gill, N.D. (2013). Positional demands of international rugby union: Evaluation of player actions and movements. Journal of Sport Science and Medicine in Sport; 16(4), pp. 353-359.

Moore, W.E., Rooney, S.J., Bracewell, P.J., & Stefani, R. (2018) Systematic Optimisation of the Elo Rating System. Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports. Sunshine Coast, Australia: ANZIAM Mathsport.

Shah, A., Jha, D., & Vyas, J. (2016). Winning and Score Predictor (WASP) Tool. International Journal of Innovative Research in Science and Engineering.

PREDICTING WIN MARGINS WITH SENTIMENT ANALYSIS IN INTERNATIONAL RUGBY UNION

Phillip Simmonds [a], Timothy S. McNamara [a] and Paul Bracewell [a,b]
[a] *DOT Loves Data, Wellington,*
[b] *Corresponding author: paul@dotlovesdata.com*

Abstract

Sentiment analysis quantifies the emotional polarity of natural language. Using this quantified sentiment, the respective opinions of the writer surrounding a given topic can be identified. In sport, it is assumed that writers will not all share the same opinion and in fact, some writers pre-match sentiment will be more correlated with match outcome. This paper uses sentiment analysis of pre-match reporting to predict the final margin for New Zealand's national rugby union team, the All Blacks, in international test matches using matches contested from 2011 to 2017. Collectively, the sentiment of New Zealand mainstream journalists is not related to the final margin ($p = 0.11$). However, by considering individual reporters, the most predictive writer is identified. The pre-match article sentiment from author Gregor Paul of the New Zealand Herald is significantly and meaningfully related (R-Sq of 26%). When the sentiment of Paul's articles is negative, the margin tends to be wider in the All Blacks favour. As such, this approach serves as a mechanism for rating the quality of journalistic opinion.

Keywords: Natural language processing, opinion mining

1. INTRODUCTION

This research finds that commentators' personal style within their pre-match articles can be used to predict match outcomes in international Rugby Union matches between the New Zealand All Blacks and their opponents.

Sentiment analysis is a family of text mining techniques that assign polarity scores to natural language. Typically, it is treated as a supervised machine learning problem. Example sentences are supplied that have been labelled as "positive" and "negative". Given sufficient training data, learning algorithms are able to distinguish positive from negative language. Positive language use scores above 0.0, and negative language scores below 0.0.

Digital delivery of news reporting and sports commentary provides a wealth of accurately time-stamped textual data that can be easily indexed via technological means. Our team has created an index of 22 million news articles, providing us with the ability to create time series of sentiment by author and by topic.

Rugby union is a team sport played by 15 players per team on a field 100m long and approximately 70m wide. Players score points by placing the ball on the ground within the in-goal area at the opposition's end (a try is worth 5 points) or by kicking the ball within the bounds of two vertical bars and over a horizontal bar at the opposition's end (a conversion following a try is worth 2 points, a dropped goal and penalty goal are both worth 3 points). The game is governed globally by the International Rugby Board (IRB). Teams seek to advance down the playing field to put themselves in a position to score points by running, kicking or passing. Importantly, passing forward is not permitted.

Arguably, rugby union is the national sport of New Zealand and the national representative team, the All Blacks, have dominated the international game. The All Blacks are the first team to win three World Cups. Within New Zealand, rugby is widely reported with both mainstream news media and specialist publications. This work seeks to understand the relationship between these reports and the actual match outcomes. Of interest is the relationship between the sentiment of pre-match reporting and match outcome.

PREVIOUS WORK

Shiladitya *et. al.* (2013) used post-match sentiment analysis gathered from Twitter to predict NFL match results in future games. The sentiment analysis was used within a logistic regression to predict match outcomes and sports betting outcomes. The authors discovered that using simple features over many tweets can outperform or at least match that of traditional features which use game statistics. 42 million tweets where gathered during the 2010-2012 seasons.

Turney (2002) carried out sentiment analysis on reviews across the film industry, vacation destinations and other areas to create a "thumbs up" or "thumbs down" system via a simple unsupervised classification algorithm. After classifying the reviews, the author found that movie reviews were very hard to classify.

Gratch *et. al.* (2014) explored the emotions of the public using sentiment analysis of Twitter posts during the 2014 FIFA World Cup. The purpose of this research was to identify exactly what makes a football match

exciting. The research found that rate of tweets per minute increased with the difference in scores (one team wins by a significant margin) with 99% confidence. A sentiment analysis was also carried out in which the authors categorised each tweet as: positive, neutral or negative. Research found that, as the rate of tweets increase, the proportion of which are negative increases with a correlation of 0.26. The percentage of neutral and positive tweets showed no relationship with the volume of tweets. This indicates that people tweet more negatively and more often when a team is beaten by a significant margin.

Opinion has been widely analysis outside of sport. Li and Wu (2009) used text mining and sentiment analysis for online forums hotspots detection and forecasting. K-means clustering and SVM machine learning techniques where used to identify forum hotspots. Both methods proved to be highly predictive with both methods concluding with the same top 4 hotspot forums. The purpose of this research was to aid the decision-making process for Internet social network users in detecting hotspots.

More traditional source than social media are also fruitful sources of sentiment data. Zhai *et. al.* (2011) used sentiment analysis on headlines of *New York Times* articles to predict the daily market trend. The classifier was inaccurate on many occasions and often predicted the inverse to what happened. Godbole *et. al.* (2007) carried out large scale sentiment analysis of news articles and blogs. The authors analysed the sentiment surrounding the following topics: business, crime, health, media, politics and sport. They found that sports people often blog positivity most often. It was also found that American politicians speak positively in blogs, but negatively in news articles. Ljajic *et. al.* (2015) used specialized sports specialized dictionaries with hard-coded weights established beforehand. Sentiment polarity could best be classified using a logarithmic difference of ± 2 when comparing counts of positive and negative words, where the difference in term frequency multiplied by inverse document frequency for positive over negative terms was logged (DifLog$t$). DifLog$t > 2$ implies a positive term classification.

Here, we will evaluate the relationship between the sentiment of pre-match reports from mainstream media and compare against the actual match outcome. The intent is to understand if this sentiment is predictive of outcome, and if there is meaningful, collective consistency of sentiment across journalists. Finally, the ability to evaluate the pre-match insights supplied by journalists is investigated.

2. METHODS

DATA COLLECTION

Articles were collected from an index of news media articles maintained by DOT Loves Data. We extracted all articles containing the phrase 'all blacks' within the headline or body text, then filtered results further to those articles published on the day before All Blacks games during time period of 2011 to 2017. This research archive 22M news articles from sources such as www.stuff.co.nz and www.nzherald.co.nz with material since 2005. Authoritative match result details were programmatically extracted from the All Blacks' website (www.allblacks.com/).

DATA MANIPULATION AND ANALYSIS

We manually curated the downloaded content to ensure that only relevant pre-match articles about the given games where included in the analysis. 43 games contained at least three pre-match articles, with reports on other matches omitted. Further filtering was conducted to enable a comparison between writers. Articles of writers who shared their pre-match opinions on more than three matches were retained.

To obtain stronger results, the difference in results was binned in order to create a more appropriate response variable for future use. All modeling and statistical work was undertaken using R and its core libraries (www.r-project.org).

3. RESULTS

ANALYSIS

All Blacks performances from 2011 to 2017 resulted in an 88%-win rate with the average difference in final score being +19.88 in favour of the All Blacks. Their results include 8 loses, 2 draws and 86 wins, with a 95.6% (1 loss, 1 draw) home win rate and an 82.3% away win rate (7 losses, 1 draw). The rarity of an All Black loss during this period means that there is insufficient data to perform an analysis using a binary match outcome.

Opposing Team	Number of Matches	Average Point Difference (+)
South Africa	14	15
Australia	21	12.5
France	10	16.2
Argentina	13	21.46
England	6	3.8
Ireland	6	16.34
Wales	6	21.34

Table 1: Summary statistics for All Blacks performances between 2011-17

For this period, the All Blacks toughest opponents in terms of points difference was England's national team. The All Blacks beat them by only 3.8 points on average. This was followed by Australia who, on average, lose by 12.5 points to the All Blacks. Table 1 shows the average difference in scores against each respective international team. For consistency, the margin is calculated as the All Blacks final score minus the opposition final score. Positive values indicate an All Black victory and negative values correspond to an All Black loss.

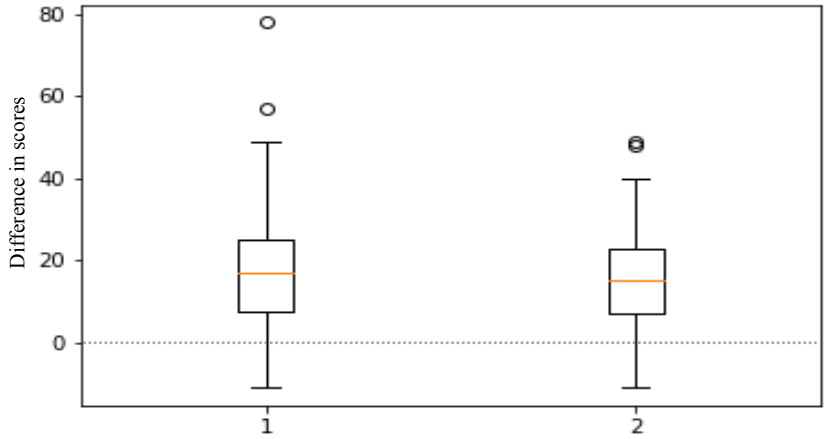


Figure 1: Box plot showing All Blacks margins and the impact of outliers

Figure 1 shows two outliers on the left box plot of the difference in scores. As the number of observations is small, we have removed these two outliers. Once the outliers had been removed, another box-plot was produced which displayed another two outliers as shown on the right of Figure 1, again these were removed.

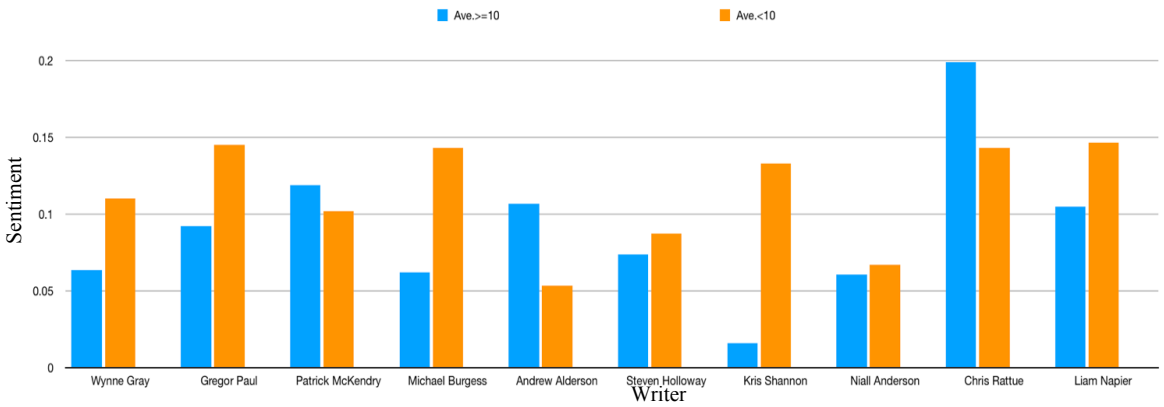


Figure 3. Shows the average sentiment for each author with orange representing games where the All Blacks win by less than 10 and blue by more than or equal to 10.

Figure 3 is a bar chart showing the average pre-match opinion of writers on occasions where the All Blacks win by more than 10 and when they don't, with sentiment on the y-axis and the x-axis represents a different game per point displayed. Figure 3 shows that 70% of the opinion writers have a more negative average sentiment for pre-match articles on the occasions where the All Blacks win by more than 10. Due to this finding, the correlation between average sentiment and difference was calculated and found to show a slightly negative relationship, r = -0.23. Gratch *et. al.* (2014) found that during the FIFA world cup, the number of negative tweets increased with the increasing margin of goals in the match. This shows evidence support the basic idea that there is a relationship between difference in score and general opinions whether it be live, post or pre-match analysis.

MODELING

To simplify the analysis, a model was fitted using the binned differences as the response variable rather than the raw difference. The scores where binned with values (-10 to 70) where -10 implies the All Blacks lost by (10 – 19) points, and 50 implies the All Blacks won by (50 – 59) points. The midpoints of these bins were used to characterize the magnitude of the win within the grouping. Using the binned differences as the response variable, and the average pre-match sentiment as the only response predictor, a model was formulated which showed no statistical significance at a 10% significance level (p-value = 0.11) and the R-Squared value was 0.04215. The margin of a match cannot be predicted using the average sentiment.

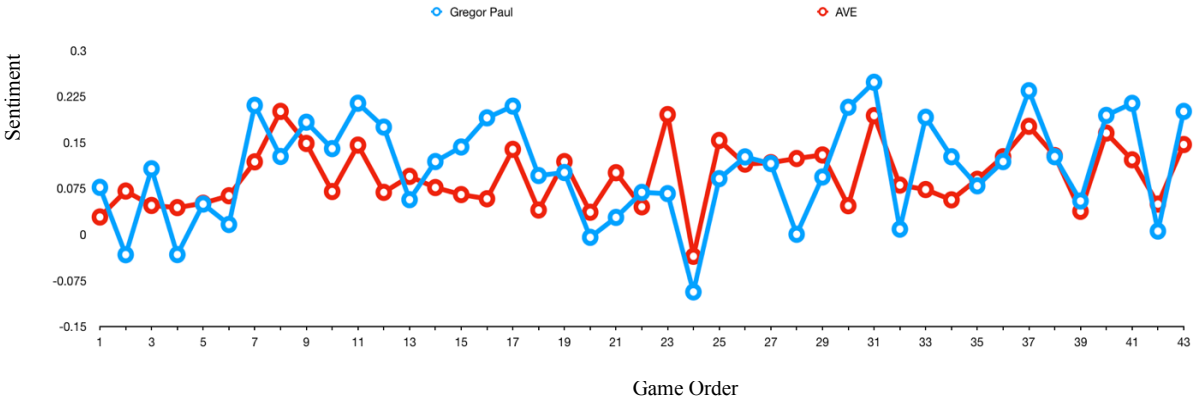


Figure 4. Pre-match sentiment for Gregor Paul articles compared with the average sentiment from nine other journalists across a sample of matches between 2011-17.

However, after comparing different writers, it was discovered that one writer stood out from the rest. Gregor Paul's opinion pieces occurred more in the data set than any other and his sentiment proved to be more significant than average. Figure 4 gives a visualization how his sentiment varied drastically from the rest on many occasions. Interestingly, his sentiment had a correlation of 0.58 with the average sentiment, showing only a moderate correlation (eq. 1). A second model was fit to using only the calculated sentiment from Paul's articles identifying a statistically significant relationship at the 1% level (p-value < 0.01, R-Sq = 0.26) (eq. 2). As shown in Figure 5, Gregor Paul's sentiment had a moderate negative correlation of r= -0.51 (2dp) with the binned difference in results.

$$Y = 12.941 - 68.325x \quad (1)$$

Using this model, an approximate margin can be calculated. Another model was then created with fitted Paul's sentiment against the raw differences. Paul's sentiment is a statistically significant predictor of score difference with an R-Sq if 15.83 (p-value = 0.0121). Figure 6 plots the raw difference in score values against Gregor Paul's pre-match article sentiment. Importantly, Gregor Paul's sentiment had a moderate negative correlation of r= -0.4 (1dp) with the difference in results, which is expected as this means more negative articles tend to be pre-cursor to larger victories.

$$Y = 16.721 - 53.341x \quad (2)$$

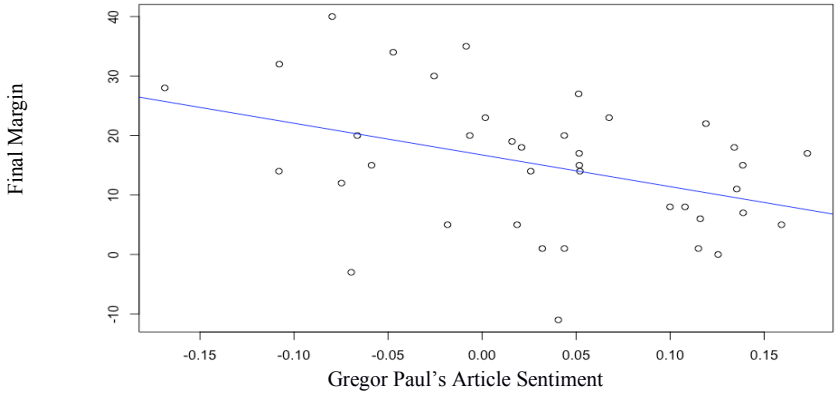


Figure 6. Gregor Paul's sentiment vs the raw difference in scores for All Blacks matches (2011-2017)

4. DISCUSSION

The intent of this analysis was to understand if sentiment is predictive of outcome, and if there is meaningful, collective consistency of sentiment across journalists. Not surprisingly, we found journalists often do not share the same sentiment in the lead up to games. This is evidenced with the difference in sentiment between Gregor Paul and the other writers ($r=0.56$).

The average sentiment from the ten writers did not provide a strong predictor of the binned difference or the actual difference ($p=0.11$). However, the sentiment for Gregor Paul was statistically significant ($p<0.01$). This emphasises the most significant aspect of this paper: not all opinions carry equal weight in rugby analysis. Due to the negative relationship between Paul's sentiment and the margin of difference for both models, we can conclude that people should strive to read Paul's articles assuming the more negative his opinion is, the more likely the All Blacks are to win by a significantly large margin.

5. CONCLUSIONS

The relationship between the calculated sentiment from pre-match reports written by mainstream media was compared against the actual match outcome. Averaging the pre-match sentiment of ten journalists who wrote three articles or more between 2011 and 2017 revealed no statistically significant relationship with the eventual match outcome ($p=0.11$). Not surprisingly, the sentiment from these journalists differed.

On further investigation, Gregor Paul of the New Zealand Herald was identified as writing pre-match content where the sentiment was statistically significantly related to the final margin ($p<0.01$). Importantly, Gregor Paul's sentiment is negatively correlated with the margin. This means that when the sentiment of Paul's articles is negative, the margin tends to be wider in the All Blacks favour. Essentially, Gregor Paul is more positive in the lead up where the game is expected to be tough. He talks the team up. Conversely, when an easier win is anticipated, the article tends to be written more negatively. These results align with Gratch *et. al.* (2014) indicating that tweets became increasingly negative as the margin widened in football.

However, Ljajic *et. al*. (2015) stated that it was easier to classify sentiment based on a specific area. In this paper, the sentiment scores pre-existed as they were in a data base mapped to an article. Consequently, the sentiment scores were not calculated as sports comments/articles, but rather as general opinions. This may cause some error in sentiment analysis as the algorithm used will fail to detect certain sports jargon and slang words. In future the models could be trained using discussions regarding the Investec Super Rugby competition as it supplies more games.

Worth noting is the sentiment values were gathered from matches in which the All Black's played against high caliber opponents. Future research should be conducted were more matches are obtained against a larger range of teams. This issue steamed from technicalities when collecting the data which resulted in only being able to collect games that were played at night New Zealand Time. This means no away games in Europe or South America were used to train the model, but they were included in the initial summary statistics.

These insights are useful in providing a deeper understanding about how reporters communicate events. Therefore, the methodology outlined here can form the basis of classifying and rating sports journalists based on their pre-match summaries.

References

Shiladitya, S., Dyer, C., Gimpel, K & Smith, N (2013). *Predicting the NFL Using Twitter.* Presented at ECML/PKDD 2013 Workshop on Machine Learning and Data Mining for Sports Analytics.

Turney, P (2002). *Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews.* Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 417-424.

Li, N. & Dash Wu, D (2009). *Using text mining and sentiment analysis for online forums hotspot detection and forecast.*

Gratch, J., Lucas, G., Malandrakis, N., Szablowski, E., Fessler, E & Nichols, J (2014). *GOAL! Using Sentiment in the World Cup to Explore Theories of Emotion.* Published in Decision Support Systems, Volume 48 Issue 2, January 2010, pp. 354-368.

Zhai, J., Cohen, N & Atreya, A (2011). *Sentiment analysis for news articles for financial signal preparation.* S224N/Ling284 Final Projects 2010, 11.

Godbole, N., Srinivasaiah, M & Skiena, S (2007). *Large-Scale Sentiment Analysis for News and Blogs.* Presented at International Conference on Weblogs and Social Media, 2007.

Ljajic, A., Ljajic, E & Arsic, B (2015). *Sentiment analysis of textual comments in field of sport.* Presented at the 24nd International Electrotechnical and Computer Science Conference (ERK 2015), IEEE, Slovenia, September 21-23, 2015.

UNDERSTANDING OLYMPIC SPEEDSKATING: EFFECTS OF ICE, SKATES, GENDER DIFFERENCES, PERCENT IMPROVEMENTS

Ray Stefani [a]

[a] *California State University, Long Beach, USA*
[a] *Corresponding author:* Raystefani@aol.com

Abstract

Olympic speed skating champions at the 2018 Winter Olympics finally produced times as good as at Salt Lake in 2002, when ice conditions were exceptional. The causes of improvement in speed skating will be deduced and evaluated. Olympic speed skating began for men in 1924 and for women in 1960. An athlete produces power due to lean body mass and training. The efficiency of applying that power depends on coaching, technique and equipment (including the aerodynamics of the suit, skates and ice conditions). Technicians created outstanding ice conditions at Cortina in 1956, at Lake Placid in 1980 and at Salt Lake in 2002. Competition moved indoors at Calgary in 1988 where ice was not subject to the whims of weather, back outdoors in 1992 and back indoors (with better ice) for good in 1994. The clap skate was introduced in 1998. That skate allowed powerful extensor muscles to be used for the first time, as the skate remained in contact with the ice for a greater fraction of the power stroke. A linear regression analysis of Olympic winning times showed that the percent improvement per Olympiad (*%I/O*) was 0.8% due to general improvements in training and efficiency. An additional 3.3% was added for each of the five Games when ice improved, 4.2% was lost when competition moved back outdoors and an additional 2.1% was gained when the clap skate was introduced. A plot of the velocity ratio of the female/male Olympic champions shows an exponential increase from 87.5% in 1960, flattening out at 92% in 1980, as training and efficiency for women improved relative to men. The average *%I/O* has remained at 92% since 1980. Assuming equal training and efficiency, the 92% female/male velocity ratio should equal the female/male relative lean-to-weight ratio of elite speed skaters, which it does.

Keywords: Speed skating, ice conditions, clap skates, training, efficiency, lean-to-weight ratio, improvement, velocity ratios, gender differences

1. INTRODUCTION

To be successful in an individual sport, an athlete has to efficiently apply force to some interface so as to move forward rapidly. The aerodynamic or hydrodynamic drag due to moving against some fluid has to be reduced. For example, in running, the interface is between the shoe and track; therefore, efforts at shoe design and artificial surfaces are important. The track suit has become more form fitting. In cycling, since the tire acts against the ground; tires are designed based on the terrain to be covered. A solid tire, carefully designed helmet and form-fitting suit reduce drag. In cross country skiing, the ski-snow interface is optimized by waxing the ski. Posture going downhill reduces drag.

In speed skating, the interface is between skate and ice. Both skate design and the clever engineering of the ice itself have been used to improve speed. Further, lightweight, form fitting suits now reduce drag. The vagaries of weather have been eliminated by moving indoors, when technology became available to create ice indoors.

The intention of this study is to evaluate the relative importance of the various activities of that sport aimed at improving the velocity of Olympic champions. Competition began for men in 1924 and for women in 1960. Further, the relative performance of female and male Olympic champions will be compared.

In Section 2, the kinesiology of speed skating is derived. Section 3 lists the milestones of speed skating history to identify those years when improvements were introduced. In Section 4, the percent improvement of each Olympic champion is found relative to the champion in the immediately preceding competition and then averaged across all events for each Olympics. A regression analysis will then evaluate the relative importance of the various milestones. In Section 5, the female/male velocity ratio is found for each event and then averaged for each Olympics. The relevance of the comparison is discussed. Section 6 presents the best three male and best three female speed skating Olympic champions. Conclusions follow in Section 7.

2. KINESIOLOGY OF SPEED SKATING

A serendipity emerges by drawing upon results from two fields: kinesiology (the science of human movement) and physics (generally dealing with the movement of inanimate objects). Athletic performance in general and speed skating performance in particular obey the following law.

$$\text{Power x Efficiency} = \text{Performance} \qquad (1)$$

From kinesiology, an athlete on a treadmill ergometer or on a cycling ergometer produces power depending on lean body mass times training, LBM Tr. That is, the power per unit of LBM is nearly a constant for equally trained athletes, Baker et al. (2001), Stefani (2006, 2007, 2014a, 2014b). Training includes physical conditioning, nutrition and psychology, the latter so the athlete will buy into the training regimen, making the regimen more effective. That produced power must be efficiently applied to speed skating where efficiency, E, depends on coaching, technique and equipment (such as the suit, skates and ice). The left side of (1) is therefore LBM Tr E.

To find the right side of (1), the laws of Newtonian mechanics can be used. Power applied equals the change in the skater's kinetic energy per unit of time (Performance). For each stroke, a skater pushes off to one side and also a bit upward. By assuming the body mass m acts at the skater's centre of gravity where the initial upward velocity is v_0, the kinetic energy is initially $m\ v_0^2/2$. The upward velocity under gravity is v_0 -gt at time t. At time t = vo/g, that kinetic energy become zero. We divide the initial kinetic energy by that time to get applied power. The right side of (1) is $m\ g\ v_0/2$. If we apply trigonometry to convert v_0 to forward velocity, v:

$$\text{LBM Tr E} = m\ g\ f(\text{angles})\ v / 2 \qquad (2)$$

The result is also rather intuitive. Power (besides meaning energy used per unit of time) is also defined as force times velocity. On the right side of (2), we see the product mg, which is the force acting on an athlete due to gravity. That force is then multiplied by vertical velocity (opposing the direction of gravity).

Assuming the angles are nearly the same for various skaters, the ratio of two velocities is:

$$v_2 / v_1 = (LTW_2 / LTW_1)\ (Tr_2 / Tr_1)\ (E_2 / E_1) \qquad (3)$$

where LBM / m is the lean-to-weight ratio, LTW. In Sections 4 and 6, the subscript "2" will refer to the velocity of an Olympic champion in a given Games while "1" will refer to the champion in the same event in the immediately preceding Games. In Section 5, "2" will refer to the woman winning an event in a given Games while "1" will refer to the man winning the same event in the same Games.

According to (3), changes in velocity depend on the athlete's physiology (LTW ratio), on training (physical conditioning, nutrition and psychology) and on efficiency (coaching, technique, skates, and ice conditions). That list can be narrowed by examining photos of Olympic champions over time. Such photos cannot be included herein due to copyright protection; however, anyone can search Google and search for "Olympic speed skating champions" followed with any year(s) of interest added to that string. Choose "Images". Over time, the champions show little if any changes in body composition (LTW) or in technique. That leaves training, coaching, skates and ice conditions as the primary causes of velocity changes over time.

3. MILESTONES

The important milestones of Olympic speed skating history depend heavily on whether the track was indoors or outdoors and on ice conditions. For the first five periods listed in Table 1, the track was on lakes or on dedicated outdoor tracks. Competition in 1956 was held at 1750 m on Lake Misurina at Cortina, Italy. The lake was wind sheltered and at altitude. The organizers chose an inlet shaped perfectly for a track. After the ice formed, every day, technicians cut the track away from the bank and continued a one-foot border completely around the rink, so ice never heaved or cracked. A special barrel with hot water was devised to resurface the ice regularly. That is the first milestone. That ice was the best skated on until then and perhaps the best ever, Stefani (2014c), IOC (1956).

Year(s)	Type of Track		Year(s)	Type of Track
1924-1932	Outdoor Track		1960-1984	Outdoor Track
1936	Lake		1988	Indoor Track (Calgary)
1948-1952	Outdoor Track		1992	Outdoor Track
1956	Lake (Cortina)		1994-2018	Indoor Track

Table 1 History of the Location of Olympic Speed Skating Competition: 1924-2018

Post 1956, competition continued outdoors through 1984. The track at Lake Placid in 1980 had been used for the 1932 Olympics and for many competitions between then and 1980. By 1980, the ice was exceptionally

well prepared, so 1980 is a milestone. The next three milestones are in 1988 when competition moved indoors for the first time, in 1992 when the move was in the opposite direction and in 1994 when the move was back indoors for good and when the timing of the Winter Olympics was change to alternate with the Summer Olympics.

A look at speed skaters prior to 1998 shows that the stride had to be shorted so as not to trip by stubbing the solid skate into the ice. Van Ingen Schenau reused an old patented hinged skate from before 1900 and was able to convince competitive skaters to use the devise and to do so at the 1998 Olympics (another milestone). The leg could be more fully extended, employing powerful extensor muscles which made skating more efficient as to power utilization, Van Ingen Schenau et al. (1985), De Koning et al. (1991).

In 2002, technicians at 1425 m Salt Lake City, used purified water, applied 12-16 layers to form the ice and used a state-of-the art climate control system to maintain exceptional ice. Thus, we have seven milestones related to the improvement of speed skating performances.

- 1956 Special Ice Preparation at Cortina (Type 1 Ice Improvement)
- 1980 Excellent Ice at Lake Placid (Type 1 Ice Improvement)
- 1988 First Indoor Track at Calgary (Type 2 Ice Improvement)
- 1992 Back Outdoors
- 1994 Indoors for good (Type 2 Ice Improvement)
- 1998 First Use of the Clap Skate
- 2002 Exceptional Technology used to create the Salt Lake Ice (Type 1 Ice Improvement)

4. ANALYSIS OF THE IMPROVEMENST OF SPEED SKATING OLYMPIC CHAMIONS

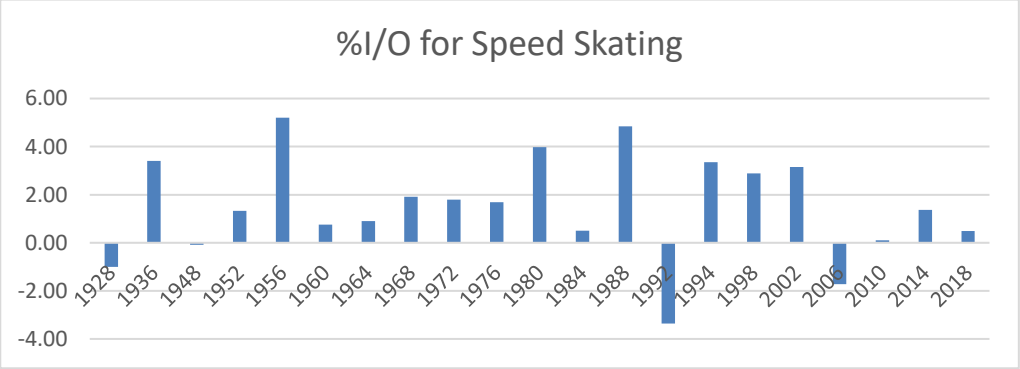


Figure 1 Percent Improvement Per Olympiad (%I/O) for Speed Skating, Averaged for each Games

In Figure 1, results for 1932 are omitted because mass starts were used rather than pairs, resulting in relatively slow times, Indeed, Figure 1 shows large %I/O values when there had been either outdoor or indoor competition with unusually good ice conditions (which will be called Type 1 ice improvement) in 1956, 1980 and 2002. There was a different cause of ice improvement visible when competition moved from outdoor ice to indoor ice (Type 2 ice improvement) in 1988 and 1994. Another improvement stands out when the clap skate was introduced in 1998. Of course, performances were degraded when competition went back outdoors in 1992. Notice that the average Olympic champion wasn't able to improve on the winning times of 2002 until 2018, when accumulated %I/O values finally showed a positive sum.

Linear regression was applied with the dependent variable being the %I/O values from Figure 1. Four dummy descriptive variables were chosen for the Type 1 ice improvement, for the Type 2 ice improvement, for the clap skate introduction and for the one movement back outdoors, (see Table 2).

Change	% Improvement Per Olympiad	P Value (%)	Significance (%)
Constant (Reference)	0.82 %	2.54%	97.46%
Good Type 1 Ice	3.29 %	0.07%	99.93%
Go Indoors for Good Type 2 Ice	3.28 %	0.30%	99.70%
Go Back Outdoors	-4.18 %	0.50%	99.05%
Introduce the Clap Skate	2.06 %	12.87%	87.13%

Table 2 Regression Coefficients for Percent Improvement Per Olympiad (%I/O)

It is surprising that the two types of ice condition improvements differed by only 0.01% and that each provided four times the general improvement of 0.82% per Olympiad. The introduction of the clap skate was 2.5 times as effective as the average improvement. Therefore, the considerable technical skill exhibited by ice makers far outpaced average %I/O due to better training and coaching. Ice makers also provided 50% more improvement than the clap skate.

5. VELOCITY RATIOS OF WOMEN/MEN FOR OLYMPIC CHAMPIONS

According to Equation (3), the velocity ratio of women/men for Olympic champions is

$$v_W / v_M = (LTW_W / LTW_M)\,(Tr_W / Tr_M)\,(E_W / E_M) \qquad (3)$$

A careful inspection of photos of female and male Olympic champions can be carried out by searching Google for "Olympic speed skating champions" followed by the gender and years of interest. Cleary, the physiology of the women is less muscular than the men, hence LTW_W / LTW_M would be less than one. Further, over time, the physiology of women appears to not change as is also true of men. That observation makes it likely that the LTW ratio was rather constant. After women began competing in Olympic speed skating in 1960, it is well known that societal changes moved women toward equality with men. We can therefore hypothesize that women became better trained, increasing the Tr ratio and that women became more efficient due to better coaching and equipment, increasing the E ratio. In summary, a working hypothesis is that the velocity ratio should generally have increased for women and then should have stabilize at the LTW ratio of women/men, when equal opportunity became the norm. Figure 2 sustains that hypothesis.

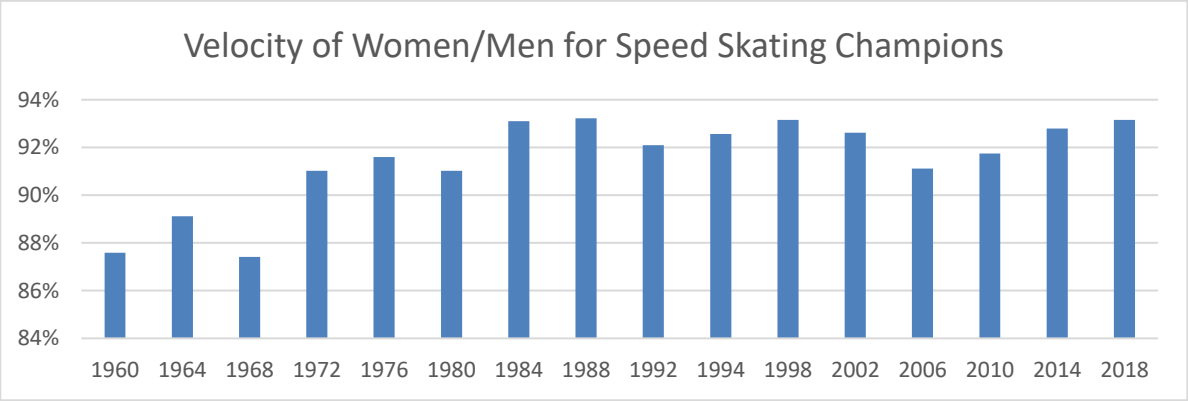


Figure 2 The average velocity ratio of women/men for each Olympics 1960-2018

Indeed, from 1960 to 1984, the average velocity ratio appears to follow an exponential curve starting at 87.4% and tapering off at 92% in 1984, remaining at an average of 92% thereafter. Is 92% the LTW ratio for women/men? Table 3 answers that question.

	Running		**Swimming**		**Speed Skating**	
Period	LTW Ratio N=156	v Ratio N=103	Estimate N=1815	v Ratio N=181	LTW Ratio N=51	v Ratio N=46
1896-1924 (WW1 & Recovery)				83		
1928-1952 (WW2 and Recovery)		88		87		
1956-1976 (Cold War)		89		90		89
1980-1988 (Boycotts & Recovery)	92	91	91	91	92	92
1992-2018 (Anti-Drug)	91	90	90	90		92

Table 3 Velocity Ratios of Women/Men for Olympic Champions Estimated by Physiology and Actual

In Table 3, modern Olympic history from 1896-2018 is divided into five periods, based on major historical events affecting performance, including the recovery from the adverse effects of wars and boycotting. According to Fleck (1983), the typical elite female speed skater is 92% as lean, relative to her weight as her typical male counterpart. For speed skating, the velocity ratio indeed increased and then stabilized at the relative LTW value of 92%. Running and swimming data are included for comparison purposes, because those two sports offer the greatest number of Olympic medals. The data for running and swimming is taken from Stefani (2014a), which includes the physiological data upon which the expected values for running and swimming were calculated. As with speed skating, the velocity ratio of female/male Olympic running and swimming champions increased; however, after anti-drug activity ramped up in 1988, female Olympic champions lost 1%, relative to their male counterparts in running and swimming, a tendency that did not happen in speed skating. The physiological expected value, assuming equal training and efficiency, agreed with the velocity ratio for swimming but female runners were 1% slower (less efficient) that relative LTW would suggest. Stefani (2014a) posits that female runners overstride because a woman's hips are wider relative to their height than men, causing overstriding which, in turn, increases the likelihood of an ACL tear. Speed skaters do not fully extend their legs, which is probably why female speed skaters are as fast as physiology predicts while runners are one percent slower

6. THE BEST SPEED SKATERING CHAMPIONS

A data base containing the percent improvement of each Olympic champion, compared to the champion of the immediately preceding Games, provides a means for selecting the three best male and three best female champions. See Table 4.

Men				
%I	Year	Distance	Name	Time
8.40 6.94	1956 1956	1500 500	Yevgeny Grishin (SOV)	2:08.6 40.20
6.40	1988	5000	Tomas Gustafson (SWE)	6:44.63
5.96	1994	5000	Johan Olaf Koss (NOR)	6:34.96
Women				
6.01	1972	500	Anne Henning (USA)	43.33
6.00	1988	1000	Christa Rothenburger (GDR)	1:17.65
5.94	1968	3000	Johanna Schut (NED)	4:56.2

Table 4 Speed Skating Champions with the Best Percent Improvements

Of course, those that won under better-than-average ice conditions in 1956, 1988 and 1994 benefitted. However, the list contains skaters that are well known as being all time bests. Koss donated much of winnings to charity, which transcends sports and records.

7. CONCLUSIONS

Photos of past Olympic speed skating champions from 1924 to 2018 suggest that technique and physiology have changed little, leaving training and efficiency (due to coaching, skates and ice conditions) as causes of improved winning times. A regression analysis was applied to the average percent improvement per Olympiad of Olympic champions for each Olympics compared to the immediately preceding games. Training and coaching improved times by 0.82% per Olympiad over the entire period. The five cases of exceptional ice conditions (1956, 1980, 1988, 1994 and 2002) resulted in an average improvement of 3.3%, four times that due to training and coaching. The introduction of the clap skate (1996) improved winners by 2.01%, 2.5 times that due to training and coaching. The effect of good ice was further validated when competition moved from indoors to outdoors in 1992, causing 4% worse winning times.

Women entered speed skating competition in 1960. Photos of female and male Olympic champions suggest that there are clear physiology differences between male and female champions, although the physiology of each gender has remained rather constant. Technique has also remained unchanged in a period of improved opportunity for female athletes. That leads to a working hypothesis that the velocity ratio of women/men for Olympic champions should generally have increased for women and then should have stabilized at a constant,

the LTW ratio of women/men, when equal opportunity became the norm. In fact, that is what happened. The velocity ratio rose from 87.5% to 92% and then stabilized at an average of 92% which is the relative LTW ratio of elite female speed skaters relative to their male counterparts.

Based on the greatest rates of improvement over one Games, the best three male champions, in order, are Yevgeny Grishin (SOV), Tomas Gustafson (SWE) and Johan Olaf Koss (NOR). The best three female champions, in order, are Anne Henning (USA), Christa Rothenburger (GDR) and Johanna Schut (NED).

References

Baker, J.S., Bailey, D.M. and Davies, B. (2001) The Relationship Between Total Body Mass, Fat-Free Mass and Cycle Ergometry During 20 seconds of Maximal Exercise, *Journal of Science and Medicine in Sport*, 4(1), 1-9.

De Koning, J.J, De Groot, G., Van Ingen Schenau, G.J. (1991) Coordination of leg muscles during speed skating, *Journal of Biomechanics* 24(2), 137-146.

Erp-Baart, et al. (1989) Nationwide survey on nutritional habits in elite athletes, *Int. J. Sports Med*. 10, S3-S10

Fleck, S.J., (1983) Body Composition of Elite American Athletes, *American Journal of Sports Medicine*, 11(6), 398-403.

IOC, (1956) *Official Report of the VII Winter Olympic Games at Cortina d'Ampezzo,* downloaded from library.la84.org/6oic/OfficialReports/1956/orw1956.pdf.

Malina, R.M (2007) Body Composition of Athletes: Assessment and Estimated Fatness, *Clinics in Sports Medicine*, 26, 37-68.

Stefani, R.T. (2006) The Relative Power Output and the Relative Lean Body Mass of World and Olympic Male and Female Champions with Implications for Gender Equity, *Journal of Sports Sciences*, 24(12), 1329-1339.

Stefani, R.T. (2007) The Physics and Evolution of Olympic Winning Performances, Chapter 3, *Statistical Thinking in Sports*, Chapman and Hall/CRC Press.

Stefani, R.T. (2014a) Understanding the Velocity Ratio of Male and Female Olympic Champions in Running, Speed Skating, Rowing and Swimming", *Proceeding of the 12th Australasian Conference on Mathematics and Computers in Sport*, Darwin Australia, 25-27 June, 2014.

Stefani, R.T. (2014b) The Power-to weight relationships and efficiency improvements of Olympic champions in athletics, swimming and rowing, *International Journal of Sports Science and Coaching*, 9(2).

Stefani, R.T (2014c) Olympic speed skating and why all ice is not created equal, *Significance Online*, 27 March, 2014, https://www.significancemagazine.com/sports/49-olympic-speed-skating-and-why-all-ice-is-not-created-equal?highlight=WyJzdGVmYW5pIl0=.

Van Ingen Schenau, G.J., De Groot, G., De Boer, R.W. (1985) The control of speed in elite female speed skaters, *Journal of Biomechanics* 18(2), 91-96.

.

An Approach for Assessing Serve Predictability

Stephanie Kovalchik[a,b]

[a] *Victoria University*
[b] *Tennis Australia*
[d] *Corresponding author: skovalchik@tennis.com.au*

Abstract

The serve is one of the most important skills in tennis. The server has the strategic advantage in a point and can increase that advantage with a more effective service strategy. An important part of any service strategy is choosing a sequence of serves that is difficult for the opponent to predict. Despite the importance of this part of the service strategy, research on how to measure and assess serve predictability has been limited. In this paper, I present an approach for measuring the predictability of a serve using probabilistic suffix trees (PSTs). PSTs are an approach for sequence analysis using variable length Markov chains, a popular model for complex sequential data. In this paper, I apply the PST approach to characterize the predictability of the service location patterns of top male players at Grand Slams between 2015 and 2017.

Keywords: Tennis, Prediction, Sequence Analysis, Markov Model

References

Gabadinho, A., & Ritschard, G. (2016). Analyzing state sequences with probabilistic suffix trees: the PST R package. *Journal of Statistical Software*, *72*(3), 1-39.

PREDICTING THE OUTCOME OF TENNIS MATCHES USING GAUSSIAN PROCESSES

Martin Ingram[a,b]

[a] *University of Melbourne, Australia*

[b] *Corresponding author:* *ingramm@student.unimelb.edu.au*

Abstract

I present an approach to predict the outcome of professional tennis matches using Gaussian Processes. Rather than modelling each player's latent skill evolution over time as a random walk, as it is done in popular rating systems such as Elo and Glicko, I model the latent skill's evolution as a Gaussian Process. Elo and Glicko's random walk assumption corresponds to the choice of a particular kernel in the Gaussian Process, but other kernels may provide better predictive performance. I provide an evaluation of the model on tennis matches, comparing it to Elo and Glicko, and investigate whether its theoretical advantages translate to improvements in real-world performance.

Keywords: Tennis, Prediction, Paired Comparison