Proceedings

16th Australasian Conference on Mathematics and Computers in Sport (ANZIAM Mathsport 2022)

Victoria University, Melbourne

7-9 July 2022

Edited by Ray Stefani and Adrian Schembri

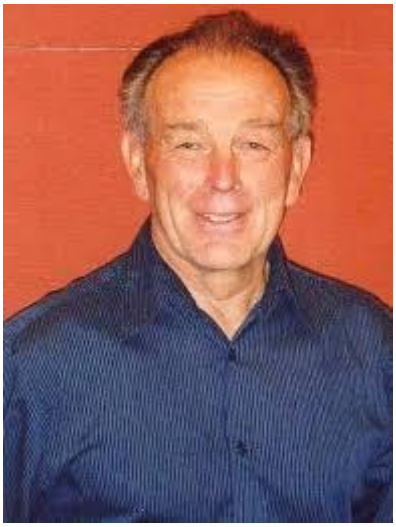
Published by ANZIAM Mathsport

All abstracts and papers have undergone a peer review.

In Memoriam

Neville de Mestre, 1938-2022

Founder of ANZIAM Mathsport



ISBN: 978-0-646-99402-4

Honoring Neville John de Mestre

Born June 15, 1938 in Wollongong, New South Wales, Australia. Died May 24, 2022

Family

Spouse Margaret Jean Fleming, December 15, 1961. Children: Nicole, Simone, Justine. Father: John Prosper Mestre. Mother: Edna May (Ramsey) Mestre.

Positions

Grantee Commonwealth of Australia, 1975, 84. Chairman board Campbell Primary School, Canberra, 1976-1979, Campbell High School, 1981-1983. Member Orienteering Federation Australia (national president 1988-1989). **Founder and first chairman of ANZIAM Mathsport, 1992.**

Education

Bachelor of Science, diploma in education, Sydney (Australia) U., 1959; Master of Science, U. Western Australia, Perth, 1966; Doctor of Philosophy, U. NSW, Canberra, 1974.

Career

Secondary school educator, NSW Department Education, 1960-1961; lecturer, senior lecturer, Royal Military College, Duntroon, Australia, 1962-1985; senior lecturer, Australian Defense Force Academy, Canberra, 1986-1989; associate professor mathematics, Bond U., Gold Coast, Australia, since 1990. Director Australian Capital Territory Mathematics Center, Canberra, 1975-1980.

Achievements

Neville John de Mestre has been listed as a noteworthy Mathematics educator by Marquis Who's Who.

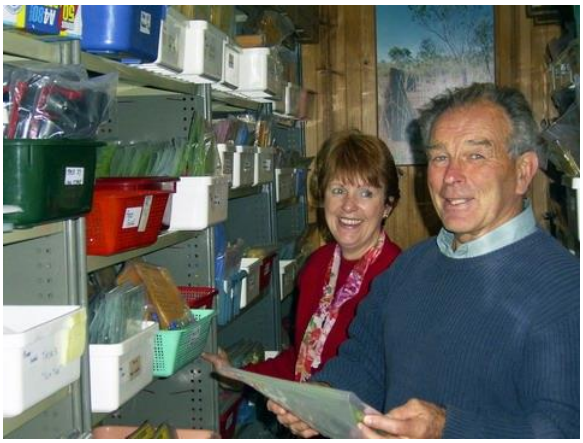
Works

Water polo; techniques and tactics Water polo; techniques and tactics
The Mathematics of Projectiles in Sport (Australian Mathematical Society Lecture Series)

Neville (left) 1995 BH Neumann Award Winner

Neville at his Passion

Neville at his Educational Materials Distribution Center



Neville at his Passion

J

Adrian Schembri, Keynote Speaker at Mathsport 2022

MENTAL HEALTH IN SPORTS

The psychological wellbeing and mental health of athletes presents unique challenges. Whilst athletes often report high life satisfaction, mental health difficulties remain prevalent given the high pressure and stressful environment that accompanies their pursuits. In this keynote talk, Dr Schembri will present on the prevalence of mental health issues among elite athletes. Wellbeing and mental health will be discussed from a lifespan perspective, considering challenges for young athletes as they emerge in their sport, unique challenges during an athlete's elite career, and risks to mental health following retirement. Provision of clinical care will be discussed. Dr Schembri will present perspectives of both prevention and intervention with regard to athlete mental health, including strategies for increasing mental health literacy, and identifying signs and triggers of mental health issues.

Dr Adrian Schembri is a Clinical Psychologist and Director of Welcome to Pod, a clinical psychology practice located in Richmond, Victoria. Adrian specialises in the treatment of adult mental health issues, including depression, anxiety, grief and adjustment. Adrian has a specific interest and experience working with men on their mental health, with a focus on relationship issues, communication at home and in the workplace, work life balance, executive coaching and supporting individuals with their navigation of complex and often overcrowded lifestyles.

Adrian completed a Doctorate in Clinical Psychology in 2010 and has since worked clinically and in academic and corporate settings. He has published peer reviewed articles and conference papers in the areas of clinical psychology, educational and developmental psychology, sports psychology and neuropsychology.

Whilst working as Director of Clinical Science at Cogstate, he supported the AFL and NRL with training club doctors on the administration and interpretation of cognitive tests used to guide return-to-play decisions following a concussion. Within his clinical work, Adrian frequently works with elite athletes during adolescence and early adulthood, and also supports retired players who are struggling with their mental health.

Adrian is a member of the Australian Psychological Society (APS) College of Clinical Psychologists and the Australian Clinical Psychology Association (ACPA).

Tim Neville, Keynote Speaker at Mathsport 2022

THE NUMBERS SUPPORTING OFFICIATING IN THE AUSTRALIAN FOOTBALL LEAGUE

Tim Nevile has extensive experience providing matchday coaching and feedback to AFL umpires. He has conducted initial assessment of performance to aid improvement of the national AFL umpiring group. In Queensland he has coached, assessed and mentored field umpires in the QAFL and QWAFL competitions. He as applied human factors theory and techniques to investigate team cognition in emergency management teams. He has conducted research in the use of procedures as a safety mechanism in the oil & gas industry.

Table of Contents

UNDERSTANDING PRESSURE METRICS AND THEIR IMPORTANCE IN AUSTRALIAN FOOTBALL

Steven Azzopardi [a,c], David Carey [a], Minh Huynh [a], Brent Manson [b]

[a] *La Trobe University, Melbourne*
[b] *Carlton Football Club, Melbourne*
[c] *Corresponding author: steven.azzopardi99@outlook.com*

Abstract

A team invasion sport involves a team in attack attempting to retain possession of an object leading to a score, with the defending team tasked with stopping them by applying pressure. The concept of pressure in team invasion sports is an important layer of detail to understanding the performance of teams, both in attack and defence. Quantification of the pressure applied to skill execution within matches has become increasingly common in invasion sports. This concept has been applied to professional Australian football in the form of 'pressure acts'. This, along with other derived statistics, are used by coaches, analysts and media to compare teams during a match, or to contrast players over the course of a season. However, published research into the relevancy of pressure metrics and its correlation with team success in Australian football is scarce. Due to the varying game plans and tactics utilised by professional Australian football teams, the way in which teams apply pressure may differ. Conversely, the performance of teams may change due to the way the opposition apply pressure. In this study, we assess the application of pressure by professional Australian football teams based on the pressure acts they apply when in defence, along with the pressure acts applied by their opposition. The pressure behaviour of teams based on ladder position is compared to assess variations between the most winningest sides and the rest of the competition. By calculating pressure points gained per minute of opposition possession, insight can be gained as to whether winning teams apply more pressure than their opponents. The first prototype of an expected pressure model – xPressure – is devised as a way to measure a team's defensive pressure against what is expected based on the actions of their opposition.

Keywords: AFL, Australian football, pressure acts, expected pressure

1. INTRODUCTION

Australian football has previously been dubbed the "most data rich sport" in the world (Watkins, 2016). This is due to the extraordinarily large number of statistics which can be collected from one match. The publication of basic statistic counts for Australian football can be traced back to the 1950s ("Second semi-final story in figures," 1953), and has continued to increase in prevalence and detail since. Statistics captured by the Australian Football League's official data provider (Champion Data) are extensively used by clubs, media and the general public. This includes the official AFL Player Ratings, fantasy ranking points such as Supercoach (Edwards, 2021; McIntosh et al., 2018), and basic counts of kicks and handballs in order to compare players and teams.

Physical pressure in team invasion sports is a tool used to assess the performance of the attacking team based on their reaction to the opposition's defensive actions (Leite et al., 2014). Teams that apply more physical pressure tend to perform better through limiting their opposition's ability to retain possession. Studies into the application of pressure in other sports, such as basketball (Leite et al., 2014) and association football (Çobanoğlu & Tereklİ, 2018) positively correlate defensive pressure with match success.

In response, Champion Data introduced 'pressure acts' (Watkins, 2016) – a metric designed to quantify the level of defensive pressure applied to the ball carrier. The inception of this metric provided an additional level of contextual detail to the data collected in Australian Football League games. For each disposal performed by a player in a professional Australian football match, one of six levels (set, none, corralling, chasing, closing and physical) is assigned to represent the amount of pressure applied to that disposal. Each level is pre-assigned a numeric value known as 'pressure points', ranging from 0.75 points for an instance of set pressure, up to 3.75 points for physical pressure (including a tackle). This allows for derived metrics such as 'pressure factor' to be calculated to represent the pressure applied by a team within a match.

Pressure acts and other pressure-related variables have been used in published research related to Australian football (Ireland et al., 2019; Sullivan et al., 2014; Vella et al., 2021). However, research into pressure metrics and their correlation with long-term team success has not been published. The objective of this paper is to analyse the ways in which different professional Australian football teams pressure their opposition, and whether differences can be observed based on single match or full season success. This provides insight as to

whether the most winningest teams pressure, or are pressured, differently than less successful teams.

An aim of this work was to create a metric known as expected pressure – xPressure. This was designed in a similar manner as expected goals metrics in association football and ice hockey (Hamilton, 2011; Macdonald, 2012; Rathke, 2017). The purpose of such a metric is to determine the level of pressure being applied by a team based on the actions of their opposition. The current metric used to compare the in-game pressure application of two teams, known as the pressure factor, computes how much pressure one team is applying based on the pressure points they accumulate during a quarter or a match. However, it does not consider the precise actions of their opposition. For example, if one team chooses to play an uncontested style of football with a focus on marking the ball and disposing from a set position, it does not allow much opportunity for the other team to apply high levels of pressure on the ball carrier. In this case, a relatively low pressure factor may still be indicative of strong defensive pressure in the situation. The adoption of an expected pressure metric would compare teams in-match with the historical pressure efforts of other teams in similar events to determine if their pressure application is above or below expectation.

2. METHODS

Match event data from all Australian Football League seasons from 2013 to 2019 inclusive, as collected by Champion Data, was used in this study. Finals and pre-season matches have been excluded, leaving twenty-two home-and-away season matches per team per season.

The data was filtered to only include events which are labelled with a pressure level. This includes kicks, handballs, tackles, dispossessions and pressure credits. Variables were added to clarify the team applying pressure and the team receiving pressure for each event.

The events were grouped twice based on team applying pressure, and team receiving pressure.

PROPORTION Z-TESTS – TOP 8 VS BOTTOM 10

The teams were separated based on their ladder positions at the end of the 2019 home and away season. In the AFL, the top eight teams qualify for the final series. The instances of each of the six pressure levels, both applied and received, were combined for the top eight and bottom ten teams. The proportions of each of the levels were compared between-groups using proportions z-tests to assess whether there were differences in the ways finals teams applied and received pressure compared with teams which didn't qualify.

PRESSURE PER MINUTE OF OPPOSITION POSSESSION

For each 2019 season match, time in possession for each team was calculated by marking times in a match which would imply taking possession (including hard and loose ball gets, intercept marks, free kicks, kick-ins and hitouts to advantage) and times which would imply the end of a possession (including scores, free kicks against, the ball going out of bounds and errors). Using the included time variable in seconds format, the time of each team possession could be deduced and added to calculate how long each team had possession of the ball over the course of the match.

The pressure points accumulated by each team were summed and used to calculate the pressure points applied per minute of opposition possession. For example, if Team A had possession of the ball for a total of thirty minutes, and Team B gained six hundred pressure points throughout the match, we'd say that Team B applied 20 pressure points per minute of opposition possession. From this, the individual team values from each game were assessed, along with team averages across the season.

For each match, the difference in pressure points per minute of opposition possession between the winning team and the losing team was calculated in search of a potential correlation between winning outcomes and pressure application.

YEAR-TO-YEAR PRESSURE POINTS PER MINUTE

The same process of calculating pressure points per minute of opposition possession was repeated for home-and-away season matches for each season between, and including, 2013 and 2018. Individual team values and differences between winning and losing sides were compared across seasons using ridgeline plots to explore the changes in mean and spread. For the differences between two sides, matches which resulted in a draw were excluded.

EXPECTED PRESSURE

The 2019 data, including finals, was filtered to include only kicks (not grounds kicks) and handballs. Several multi-level models were built, with the best chosen based on respective R2 and Chi-Square statistics. The models used pressure points as the target variable, while including other relevant metrics and their interactions as explanatory variables.

The chosen model was applied to each statistic in the 2019 data set. The expected pressure points for each team in each match were calculated to compare how teams applied pressure compared with the expectation. Comparisons were also made based on match result to assess whether winning teams were more likely to exceed expectation.

3.RESULTS

The proportions of instances of applied pressure which fell under each of the six pressure levels for the eighteen teams are displayed in Figure 1 for between-team comparison. Figure 2 shows the same data but representative of the pressure received by each team.

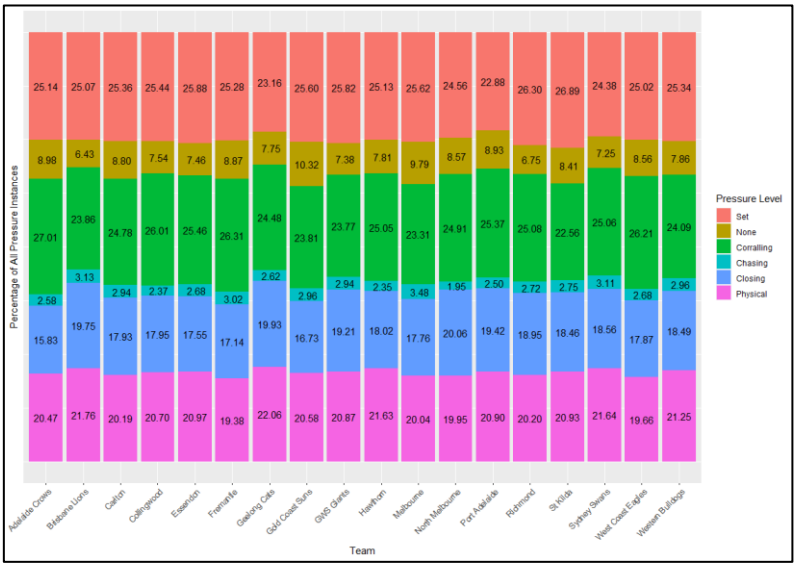


Figure 1: The proportion of each pressure level applied by AFL teams during the 2019 home-and-away season

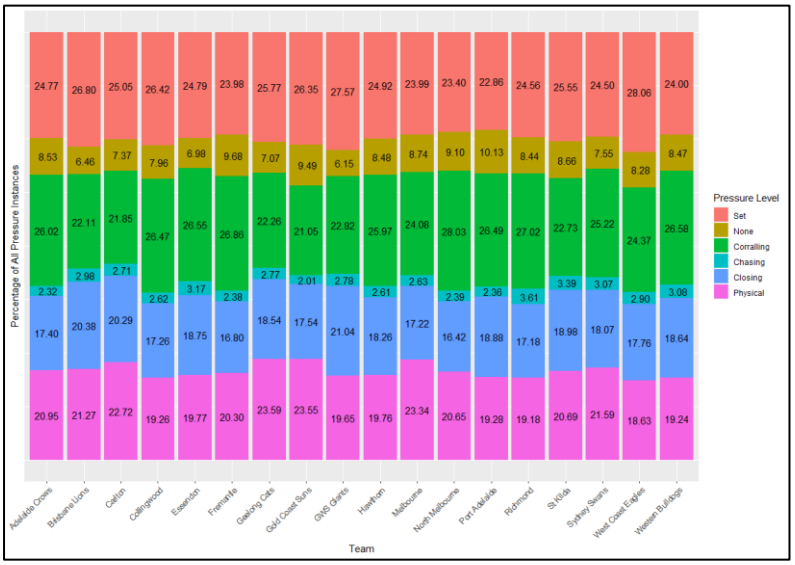


Figure 2: The proportion of each pressure level received by AFL teams during the 2019 home-and-away season

In 2019, St Kilda allowed their opposition the highest proportion of 'set' pressure (disposals taken from a mark, set shot, kick-in or free kick). Port Adelaide allowed the least. Geelong applied the greatest proportion of physical pressure (including tackles), with Fremantle applying the least.

West Coast were allowed the greatest proportion of set disposals, while Port Adelaide had the smallest. Geelong received the greatest proportion of physical pressure, while West Coast received the smallest.

Figures 3 and 4 show the z-scores and 95% confidence intervals for the proportions of each level, first applied, and then received, grouping the top 8 and bottom 10 teams respectively. For the pressure application data, the bottom 10 teams allowed a greater proportion of instances with no pressure compared with the top 8 teams. Top 8 teams applied a greater proportion of closing pressure than the bottom teams. No discernible difference

could be concluded for the other pressure levels.

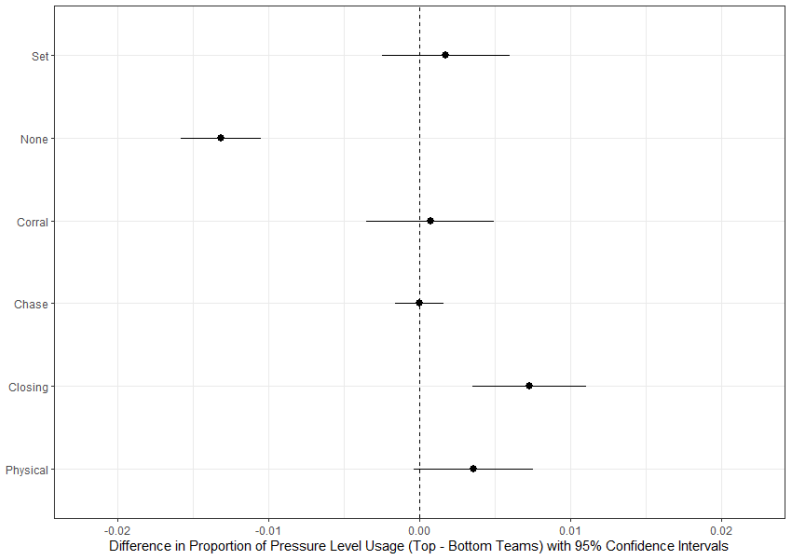


Figure 3: The z-scores (and 95% confidence intervals) from two proportion z-tests comparing the proportion of instances of applied pressure for each pressure level between teams in the top 8 and bottom 10 in 2019

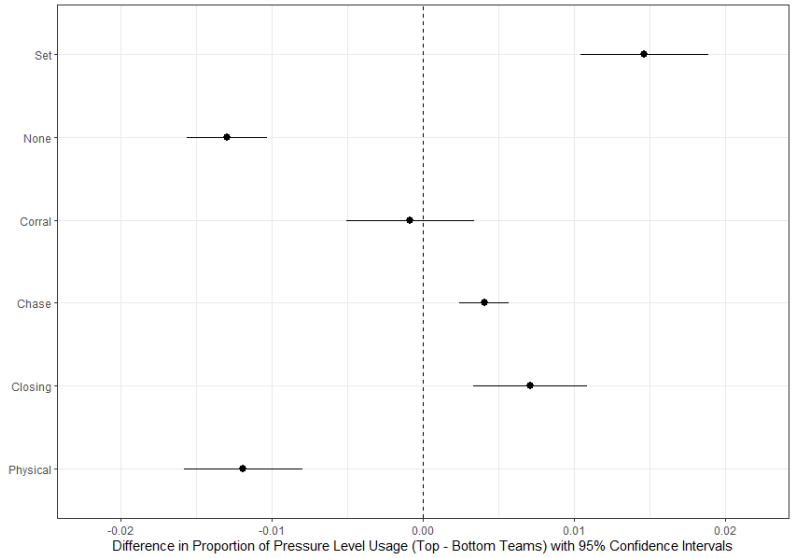


Figure 4: The z-scores (and 95% confidence intervals) from two proportion z-tests comparing the proportion of instances of received pressure for each pressure level between teams in the top 8 and bottom 10 in 2019

For the pressure received data, top 8 teams were allowed a greater proportion of play under set pressure. They also received a greater proportion of pressure being chased or closed. The bottom 10 teams were allowed a greater proportion of instances under no pressure, but also under physical pressure. No difference could be concluded for receiving corralling pressure.

Figure 5 highlights the spread of pressure point per minute of opposition values across the 2019 home-and-away season. The average value was approximately 21.25 pressure points. Figure 6 shows the difference in pressure points per minute of opposition possession between the winning and losing side in each match. On average, the winning teams recorded approximately 1.74 more pressure points than their opposition. Both metrics follow a normal distribution.

Figures 7 and 8 repeat the analysis of pressure points per minute of opposition possession and the difference in this metric across all seasons from 2013 to 2019. There appears to be no discernible difference in pressure points per minute across the seasons. For the difference between teams, the values appear to be closer to zero during the 2013 season, however the spread appears to be similar across-seasons.

Table 1 outlines the final multilevel built to represent expected pressure, along with estimates and random effects in Figures 9 and 10.

Figures 11 and 12 give examples of the output of the expected pressure model when controlling for

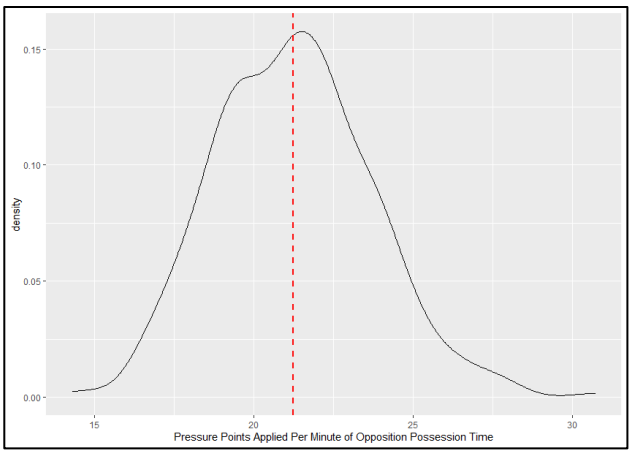


Figure 5: Density plot of all team match counts of pressure points applied per minute of opposition possession time during the 2019 AFL home-and-away season

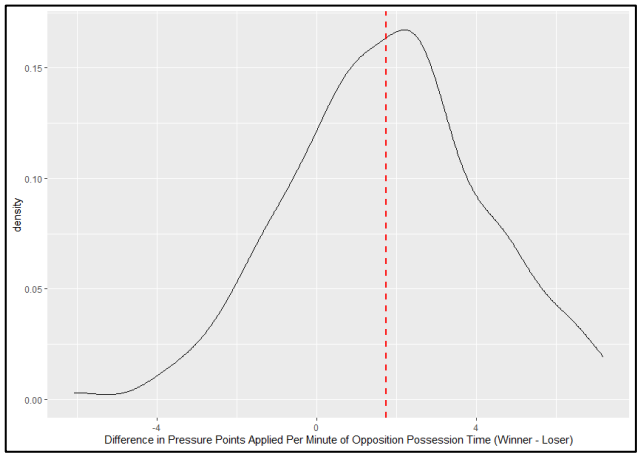


Figure 6: Density plot of the difference in pressure points applied per minute of opposition possession time (winner - loser) for each match during the 2019 AFL home-and-away season

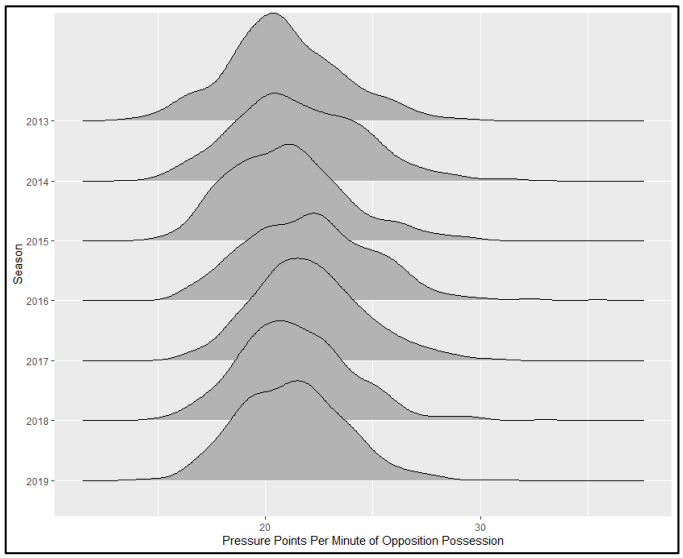


Figure 7: Density graphs of each count of pressure points per minute of opposition possession in AFL home-and-away matches, year-by-year from 2013 to 2019

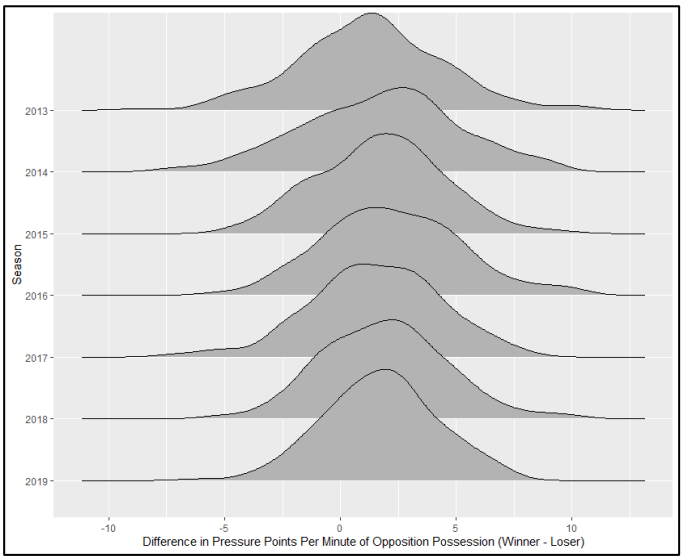


Figure 8: Density graphs of the difference in pressure points per minute of opposition possession (winner - loser) from each AFL home-and-away match (excluding drawn matches) between 2013 and 2019, faceted by season

certain variables, along with the slight difference in output based on the team that is applying the pressure. Figure 13 shows, for each team in each match in 2019, the difference between expected pressure and actual pressure points (for handballs and standard kicks only). As expected, the average falls extremely close to zero due to the model being built on this data.

For each team, the percentage of their actual pressure points when compared to expected pressure was calculated. For example, if a team registered 550 pressure points but were expected to gain only 500, they would register 110% of expected pressure points. Figure 14 shows the difference in this figure between the winning and losing team in each match, followed by the two parties separated in Figures 15 and 16. On average, winning teams performed 0.8% better than their opposition. Winning teams exceeded their expected pressure in 107 of 207 matches (51.7%), whereas losing teams did so in 96 matches (46.4%).

4. DISCUSSION

This work is the first to attempt an in-depth analysis of pressure metrics in Australian football and its correlation with long-term team success. This paper outlines an initial overview of the data and its potential applications.

	PRESSURE POINTS		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	0.96	0.88 – 1.05	**<0.001**
disposalType [Kick]	-0.18	-0.20 – -0.17	**<0.001**
clearanceTime [Pre-Clearance]	0.19	0.18 – 0.21	**<0.001**
fieldLocation [D50]	-0.03	-0.05 – -0.02	**<0.001**
fieldLocation [DM]	-0.04	-0.06 – -0.03	**<0.001**
fieldLocation [F50]	0.20	0.18 – 0.22	**<0.001**
possessionSource [Free Kick]	0.04	-0.04 – 0.12	0.335
possessionSource [Gather]	1.09	1.00 – 1.17	**<0.001**
possessionSource [Handball Receive]	0.89	0.81 – 0.97	**<0.001**
possessionSource [Hard Ball Get]	1.93	1.85 – 2.02	**<0.001**
possessionSource [Kick In]	0.03	-0.07 – 0.12	0.577
possessionSource [Kick In - Play On]	0.44	0.35 – 0.52	**<0.001**
possessionSource [Loose Ball Get]	0.93	0.85 – 1.01	**<0.001**
possessionSource [Mark]	0.07	-0.01 – 0.15	0.102
possessionSource [Out On The Full]	0.06	-0.03 – 0.15	0.196
possessionSource [Ruck Hard]	1.28	1.19 – 1.38	**<0.001**
startOfChain [CB]	-0.03	-0.05 – -0.01	**0.001**
startOfChain [KI]	-0.09	-0.11 – -0.06	**<0.001**
startOfChain [PG]	-0.07	-0.08 – -0.05	**<0.001**
startOfChain [TI]	-0.04	-0.05 – -0.02	**<0.001**
disposalType [Kick] * clearanceTime [Pre-Clearance]	-0.00	-0.02 – 0.02	0.997
disposalType [Kick] * fieldLocation [D50]	0.06	0.04 – 0.08	**<0.001**
disposalType [Kick] * fieldLocation [DM]	0.04	0.02 – 0.06	**<0.001**
disposalType [Kick] * fieldLocation [F50]	-0.14	-0.17 – -0.12	**<0.001**
Random Effects			
σ^2	0.55		
$\tau_{00\ pressureTeam}$	0.00		
$\tau_{11\ pressureTeam.disposalTypeKick}$	0.00		
$\rho_{01\ pressureTeam}$	-0.58		
ICC	0.00		
$N_{pressureTeam}$	18		
Observations	151954		
Marginal R^2 / Conditional R^2	0.426 / 0.427		

Table 1: The final multilevel model, predicting pressure points on a kick or handball using 2019 AFL data

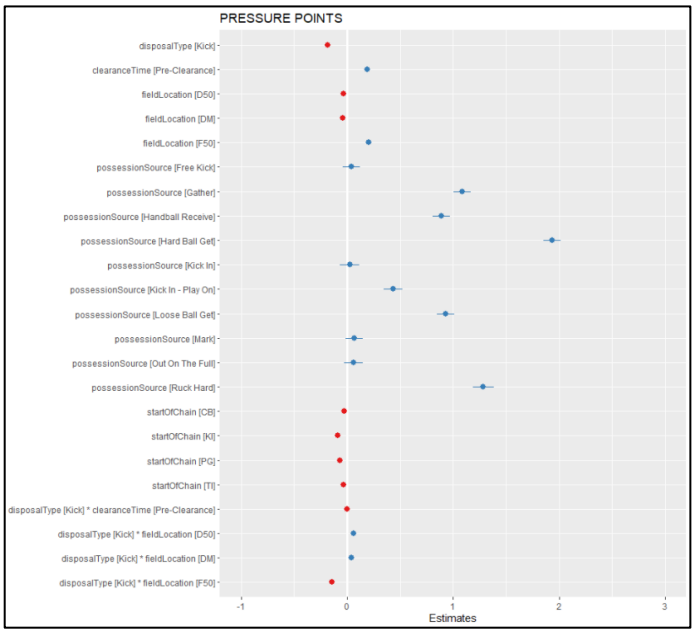


Figure 9: Forest plot of estimates from the chosen multilevel model

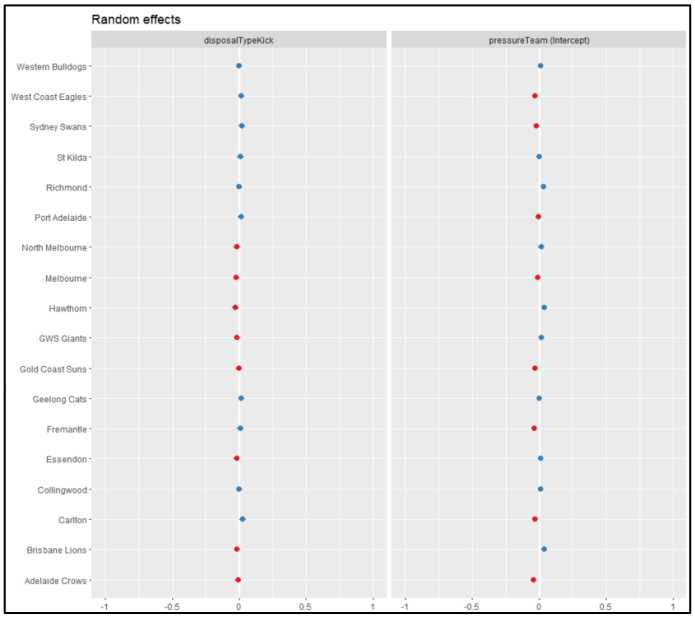


Figure 10: Random effects from the chosen multilevel model

On first glance, the comparison of the eighteen AFL teams based on the pressure they received and applied throughout the 2019 season appears to have minor between-team difference. Picking out individual teams and their usage of certain pressure levels infers greater differences. Brisbane and Richmond applied the lowest proportion of the 'none' pressure level (allowing the opposition to dispose of the ball in open play under no immediate pressure). The two teams finished the 2019 home-and-away season 2nd and 3rd, respectively. Conversely, Melbourne and Gold Coast allowed the greatest proportion of 'none' pressure and finished the season 17th and 18th, respectively.

Although this can be recognised for some pressure levels, it is not always the case. Despite finishing at opposite ends of the ladder, Geelong (1st) and Gold Coast received a higher proportion of physical pressure than any other teams. Differences in pressure level

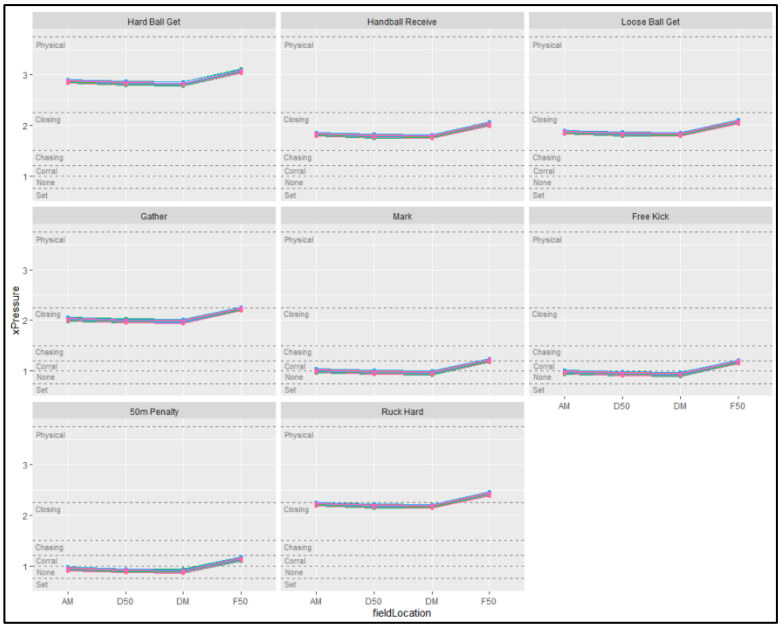


Figure 11: The expected pressure points based on field location, faceted by the source of possession, grouped by team, and controlling for disposal type (handball), time in relation to clearance (post-clearance) and how the chain started (centre bounce)

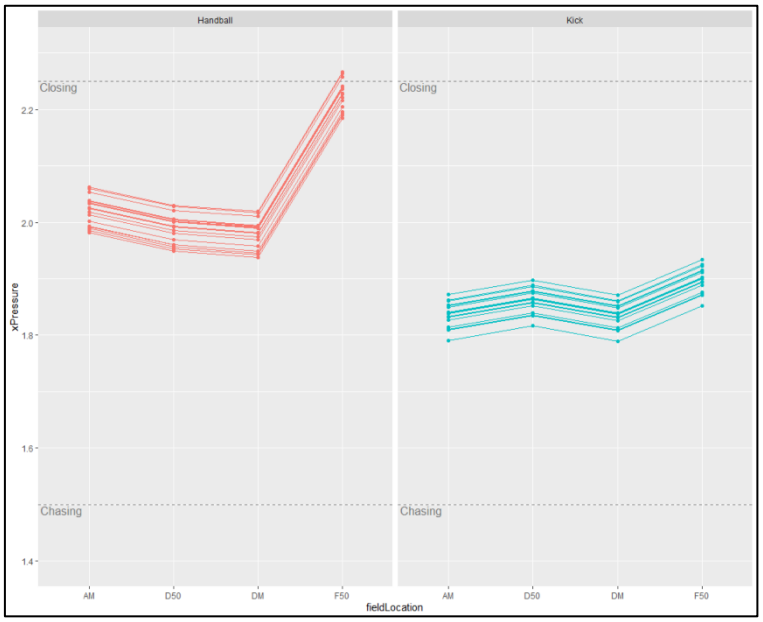


Figure 12: The difference in expected pressure between handballs and kicks based on field location, faceted by the source of possession, grouped by team, and controlling for source of possession (loose ball get), time in relation to clearance (post-clearance) and how the chain started (centre bounce).

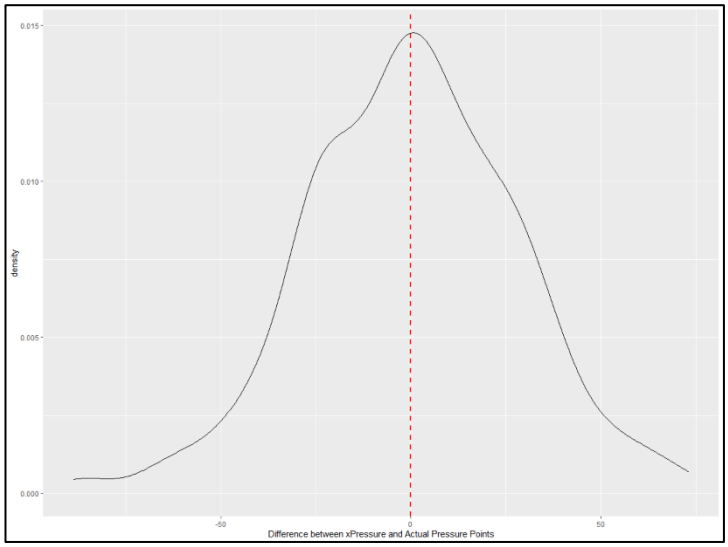


Figure 13: The difference between expected pressure and actual pressure points (on opposition kicks and handballs only) for all teams across all 2019 matches, including finals

proportions, both applied and received, may not be impacted solely by team strength but by game style. Some teams may invite physical pressure from their opposition to free up space for their teammates further up the field. Other teams may choose to play a kick-and-mark style, which would lead to an increase in 'set' pressure received. Another variable which would affect this may be the field of play, as Australian football grounds at the elite level are not all of the same dimensions. Geelong plays the majority of its home matches at Kardinia Park, which has the smallest width of any venues used for AFL games in 2019 (Staff writers from Fox Sports, 2018). Geelong received and applied a higher proportion of physical pressure than any other team in the 2019 season. Conversely, Fremantle and West Coast recorded the lowest proportions of physical pressure applied and were within the bottom four for physical pressure received. Both Fremantle and West Coast play their home matches at Perth Stadium, which has one of the largest playing areas of any regularly used venues. Although this remains one of several plausible factors, future work would endeavour to discern differences for teams who regularly play at a venue larger or smaller than normal to determine if playing surface makes a difference in pressure application.

The z-scores and their confidence intervals in comparing the difference in pressure applied and received, when grouping the top 8 and bottom 10, allowed for trends to be observed for teams that won the most games during the season. Teams in the bottom 10 allowed more instances of 'none' pressure than the top 8. This could be due to these teams focusing less on defending the ball carrier and choosing to stay close to other members of the opposition team. It may also be due to to their opposition having greater skill and, as such, being able to find more space while possessing the ball.

Top 8 teams applied more physical pressure, whilst bottom 10 teams received more physical pressure. This aligns with the notion that pressure is correlated with team success.

Bottom 10 teams happened to receive more 'none' pressure than top 8 teams as well as more physical pressure. This may be a sign of their opposition either choosing to physically pressure their opponent or, instead

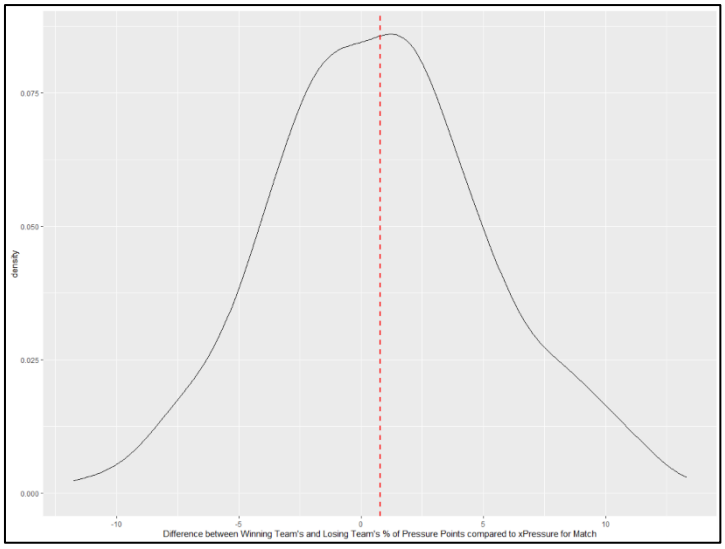


Figure 14: The spread of differences between the winning and losing team's percentage of pressure points (kicks and handballs) compared to expected pressure in a match

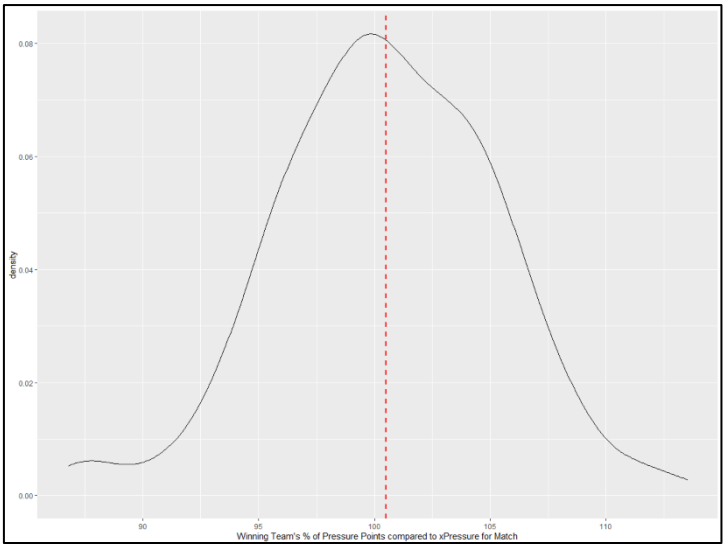


Figure 15: The spread of winning team's percentage of pressure points (kicks and handballs) compared to expected pressure in a match

of corralling, chasing or closing, opting to defend space further up the ground and allowing for the ball carrier to dispose of the ball under less duress. In theory, this could allow for the ball to arrive at a contested area further up the field.

By calculating the difference in pressure points per minute of opposition possession for each home-and-away match, it appeared to be most common for the winning side to record a higher frequency of pressure points. Due to the way pressure points are calculated in a discrete manner, stronger teams having more set shots at goal, which allow for approximately 30-45 seconds of possession time before disposal and the lowest assignment of pressure on the disposal, may be the cause of this outcome. Further work should search for ways for this unopposed possession time to be accounted for, whether by removing set pressure instances or calculating per a set number of disposals.

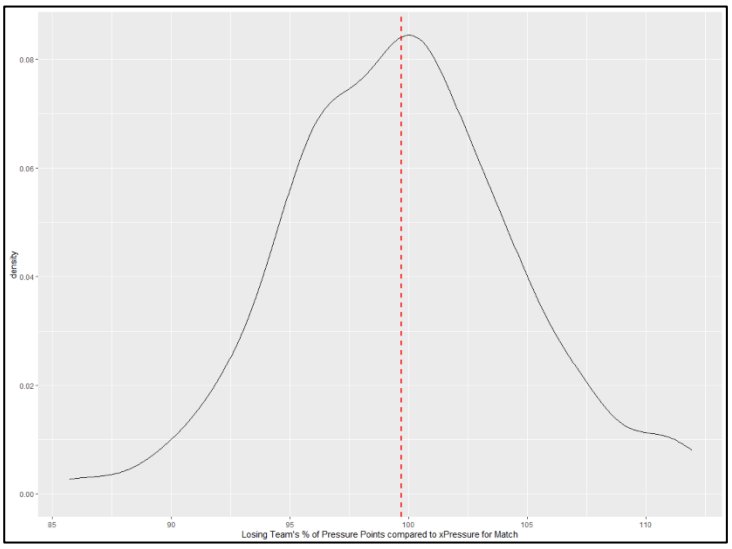


Figure 16: The spread of losing team's percentage of pressure points (kicks and handballs) compared to expected pressure in a match

The two ridgeline plots show minimal year-to-year difference in pressure points per minute of opposition possession, and little difference in this metric between winning and losing sides. The between-season similarity may be indicative of ongoing changes in rules and team game styles having minimal effect on the way teams apply defensive pressure. Data for these seasons may be combined for future research with the assumption that defensive pressure has changed minimally season-to-season.

The included model for calculating expected pressure is the first of its kind. The metric is able to compare a team's pressure performance to what should be expected in a match. A small positive correlation with winning was noted. A model of this kind would allow for coaches and analysts to assess their team's defensive performance against a more relevant benchmark than a comparison between the two team's pressure factors. An expected pressure above a team's true output would be indicative of the team overperforming defensively, and underperforming if their true pressure is below expectation.

5. LIMITATIONS

The nature of this research and its novelty has with it several limitations. Although moments of set play can result in differing outcomes, due to players only having a short amount of time to dispose of the ball from a free kick, mark or kick-in, the majority of these instances result in 'set' pressure. This could skew the data and impact the results of differences amongst other pressure levels. Future work should seek to separate set pressure from other instances, whether by removing marks, free kicks and kick-ins, or accounting for the time taken between possession and disposal.

The inclusion of pressure attained from tackles and other unique moments of play such as ground kicks could have drawbacks for calculating pressure points per minute of opposition possession, as many tackles would occur without the other team "gaining" possession of the ball. Removing these instances when calculating the metric may provide alternative results.

As the model was built using data from the 2019 season and tested on the same input, there is no proof that this model would remain relevant when applied on other matches. Further work will replicate the same process using a larger sample of matches before testing with more recent, unseen contests. The current model is also limited to kicks and handballs, with other instances of pressure such as disposal-less tackles and pressure credits excluded. Future models should aim to include these moments of pressure to give a true account of a team's defensive performance. Moments of set pressure, such as disposals from marks, kick-ins and 50m penalties need to also be accounted as higher pressure levels, though unlikely, are possible. Some disposals also yield an expected pressure less than 0.75, which is the lowest possible pressure points attainable. This should be controlled in future models. Once these changes are made, the final product will provide a new tool to analyse a team's defensive pressure during and after a game and in the long term.

REFERENCES

Çobanoğlu, H. O., & Tereklİ, M. S. (2018). Affects of defense unit on score (goals) in soccer. *International Journal of Sport Exercise and Training Sciences-IJSETS*, *4*(2), 57-63.

Edwards, S. J. (2021). Analyzing Fantasy Sport Competitions with Mixed Integer Programming. In *Data and Decision Sciences in Action 2* (pp. 167-182). Springer.

Hamilton, H. H. (2011). An Extension of the Pythagorean Expectation for Association Football. *Journal of quantitative analysis in sports*, *7*(2). https://doi.org/10.2202/1559-0410.1335

Ireland, D., Dawson, B., Peeling, P., Lester, L., Heasman, J., & Rogalski, B. (2019). Do we train how we play? Investigating skill patterns in Australian football. *Science and Medicine in Football*, *3*(4), 265-274.

Leite, N. M., Leser, R., Gonçalves, B., Calleja-Gonzalez, J., Baca, A., & Sampaio, J. (2014). Effect of Defensive Pressure on Movement Behaviour During an Under-18 Basketball Game. *Int J Sports Med*, *35*(9), 743-748. https://doi.org/10.1055/s-0033-1363237

Macdonald, B. (2012). An expected goals model for evaluating NHL teams and players. Proceedings of the 2012 MIT Sloan Sports Analytics Conference,

McIntosh, S., Kovalchik, S., & Robertson, S. (2018). Validation of the Australian football league player ratings. *International Journal of Sports Science & Coaching*, *13*(6), 1064-1071.

Rathke, A. (2017). An examination of expected goals and shot efficiency in soccer. *Journal of Human Sport and Exercise*, *12*(2), 514-529.

Second semi-final story in figures. (1953, 16 September 1953). *Sporting Globe*, 2. http://nla.gov.au/nla.news-article178191294

Staff writers from Fox Sports. (2018). Geelong's GMHBA Stadium playing surface to shrink as boundaries brought in for player safety. https://www.foxsports.com.au/afl/geelongs-gmhba-stadium-playing-surface-to-shrink-as-boundaries-brought-in-for-player-safety/news-story/03d334db55d584cd66e545a4bd7beff4

Sullivan, C., Bilsborough, J. C., Cianciosi, M., Hocking, J., Cordy, J. T., & Coutts, A. J. (2014). Factors affecting match performance in professional Australian football. *Int J Sports Physiol Perform*, *9*(3), 561-566. https://doi.org/10.1123/ijspp.2013-0183

Vella, A., Clarke, A. C., Kempton, T., Ryan, S., Holden, J., & Coutts, A. J. (2021). Technical involvements and pressure applied influence movement demands in elite Australian Football Match-play. *Science and Medicine in Football*, 1-6.

Watkins, A. (2016). Analytics in the AFL - The Most Data Rich Sport on Earth. In A. Watkins, *The Numbers Game*. VICE Sports.

DETERMINANTS OF SHOT AT GOAL ACCURACY IN AUSTRALIAN FOOTBALL AND OPTIMISING IT'S AFFECT ON MATCH OUTCOME

Dan B Dwyer [a,b] , Chris M Young [a]

[a] *Centre for Sport Research, Deakin University, Locked Bag 20001, Geelong 3220, Australia.*
[b] *Corresponding author: dan.dwyer@deakin.edu.au*

Abstract

Factors that affect shot at goal accuracy in AF have been explored previously, which include shot location, shot type and stadium design (Bedford & Schembri 2006). More recently, additional factors were identified such as playing position and weather, but not the experience of the player (Anderson et al., 2018). Furthermore, Browne et al. (2022) reinforced the importance of shot location and added the effect of pressure on the accuracy of shots in open play. The present study explored whether a single variable (arc angle) that represents the shot location, provides a more useful predictor of the likelihood of a goal. In addition, we aimed to explore whether it is more important to maximise the number of shots at goal, or the probability that shots will result in a goal. 32,694 shots at goal were analysed from 1260 team performances in three seasons (2017-2019) of the Australian Football League. The most important determinant of match outcome is the number of shots at goal, although there was a cohort (~14%) that won, without having more shots at goal. This cohort managed to win by having a higher shot at goal accuracy, which was achieved in large part by taking fewer shots at goal, from field locations that had a higher probability of scoring a goal. The probability of scoring a goal can be predicted using shot arc angle (61-65.8% classification accuracy), instead of the combination of shot angle and distance (60%). These results demonstrate that there is an interplay between the total number shots per match and the probability that those shots will result in a goal. All teams should consider this aspect of their performance as it relates to their match strategy and because points scored for and against them, may affect their final position on the league ladder.

Keywords: Technical performance, tactical performance, expected goal

Acknowledgements

We wish to thank Champion Data (Melbourne) for supplying the data for this analysis.

References

Anderson, D., Breed, R., Spittle, M., & Larkin, P. (2018). Factors affecting set shot goal-kicking performance in the Australian football league. *Perceptual and Motor Skills*, *125*(4), 817-833.

Bedford, A., & Schembri, A. (2006). An analysis of goal-kicking accuracy in Australian rules football. In *Proceedings of the Eighth Australasian Conference on Mathematics & Computers in Sport, Coolangatta, Australia*.

Browne, P. R., Sweeting, A. J., & Robertson, S. (2022). Modelling the Influence of Task Constraints on Goal Kicking Performance in Australian Rules Football. *Sports Medicine-Open*, *8*(1), 1-12.

FREQUENTLY RECURRING MOVEMENT PATTERNS IN THE AFL: A CLUSTERING ANALYSIS OF SPATIOTEMPORAL DATA

James Harrold [a,d] , Minh Huynh [a] , Matthew Varley [a]

[a] *Latrobe University*
[d] *Corresponding author: 20777618@students.latrobe.edu.au*

Abstract

AFL football is a physically vigorous sport where opposing players aim to outscore their opponent. Successful teams are often tactically and physically superior to their opponents, which makes it important for further analysis into how these athletes dynamically move during a game. With that, Sequential Movement Patterns can be identified and quantified using a clustering technique to explore how players predominantly move in play. Sequential Movement Patterns (SMP) are running metrics extracted from GPS (Global Positioning Systems) data to show the way in which players move during play. The advantage of utilising SMP to explore common movement patterns, is to associate these to specific player positions and demands. Usually, spatiotemporal data is used to identify threshold running zones which at times can be quite a subjective value. An alternative approach to this is to use a clustering technique on the collected spatiotemporal data (X, Y coordinates) to identify movements that are like each other in nature. This is a novel approach which has not yet been explored in the AFL. Therefore, the aim of this study us to identify and explore the differences in Sequential Movement Patterns among the distinct positions in the AFL. This can be beneficial for coaches and Sport Science staff to identify differing running patterns amongst positions within a team and adjust their running demands accordingly in training.

Keywords: *AFL, GPS, Movement Patterns*

Acknowledgements

We wish to thank the North Melbourne Football Club for their assistance in data retrieval.

ANGULAR DISPERSION AS A MEASURE OF SET SHOT GOAL-KICKING ACCURACY IN AUSTRALIAN FOOTBALL

Jasmine Farrugia [a] , Matthew Gloster [b] , Karl Jackson [b,c]

[a] *RMIT University – Melbourne, Australia*
[b] *Champion Data – Melbourne, Australia*
[c] *Corresponding author: karl.jackson@championdata.com.au*

Abstract

The traditional approach to modelling the accuracy of player actions across various sports has been to use a binomial (score/miss) outcome conditioned on various inputs related to the context of the action. Here we introduce a novel approach to assessing accuracy of set shots at goal in Australian Football, where the player can take the shot without intervention from defensive players. With official data from the 2014-2021 seasons nearly 35,000 set shots are included in the study, tagged with X,Y coordinates for the location of the shot and a seven-option multinomial output (goal, behind left/right, post left/right, out on the full left/right). For each shot, the location of the four goal posts is converted to the angular domain as degrees left or right of the centre of the goal line, and the angular dispersion is sampled from a uniform distribution in the possible outcome space given by the previously mentioned multinomial output. Distributions of angular dispersion are then derived for individual players to assess the reliability of their goal-kicking, and to establish expected results from set shots at goal as probabilistic outcomes. A further extension was applied to the assessment of decision-making related to snapped shots at goal to determine locations on the ground where this may be more accurate than a traditional drop-punt.

Keywords: Accuracy, Simulation, Australian Football, Expectation

Submitted to MATHSPORT 2022 *March 21, 2022*

EVALUATION OF AUSTRALIAN FOOTBALL STRATEGIES AS ADJUNCT STATES OF A MARKOV GAME MODEL

Darren M. O'Shaughnessy [a] & Chris McKay

St Kilda Football Club
[a] Corresponding author: darren.oshaughnessy@saints.com.au

Abstract

Markov Models are used in game theory to describe the evolution of the play through well-defined memoryless states. Scores and results can be encoded as absorbing states of the model, with Markov matrix operations enabling evaluation of all states' transition probabilities and their likelihood to end in a certain score. The on-field possession states have been modelled in several sports as a Markov or semi-Markov model, starting with Romer's (2002) dynamic programming approach in American Football.

More complex sports such as Australian Rules football (Meyer, Forbes & Clarke, 2006), ice hockey (Thomas, 2006) and soccer (Rudd, 2011) have employed these methods, based on notational analysis of events in the game. This paper uses (X,Y) event and descriptive data from Champion Data, combined with club-specific annotations that describe common scenarios in the game.

The Markov states in the model correspond to well-described events, such as shots at goal, stoppages, and turnovers. Additional state parameters include the position relative to goal, the relative position of the defensive team ("the bubble") and the pressure on the ball carrier. Adjunct states are built, corresponding to coaching scenarios such as "slow play coming out of defence". These are defined using coaches' descriptions, and they do not appear in the Markov matrix per se. Where there are multiple choices from the scenario (e.g. switch the play vs kick for distance), each choice is represented as a vector of outcome states, derived empirically.

When evaluating a game, the actual outcomes from these choices are compared with the historical average, in terms of metres gained, percentage turned over, and field equity (O'Shaughnessy, 2006) of the subsequent states.

Keywords: Australian Rules Football, Strategic Evaluation, Coaching, Markov Model, Dynamic Programming, Markov Decision Process

Acknowledgements

The authors wish to thank the coaches at St Kilda Football Club for defining the scenarios used in this research.

References

Jackson, K. (2016). Assessing Player Performance in Australian Football using Spatial Data. *PhD Thesis, Swinburne University of Technology*.

Meyer, D., Forbes, D.G., Clarke, S.R. (2006). Statistical Analysis of Notational AFL Data using Continuous Time Markov Chains. *Proceedings of the 8h Australasian Conference on Mathematics and Computers in Sport, Gold Coast*, 81-90.

O'Shaughnessy, D.M. (2006). Possession Versus Position: Strategic Evaluation in AFL. *Proceedings of the 8h Australasian Conference on Mathematics and Computers in Sport, Gold Coast*, 226-238.

Romer, D. (2002). It's Fourth Down and what does the Bellman Equation say? A Dynamic-Programming Analysis of Football Strategy. *National Bureau of Economic Research, Working Paper 9024*.

Rudd, S. (2011). A Framework for Tactical Analysis and Individual Offensive Production Assessment in Soccer Using Markov Chains. *New England Symposium on Statistics in Sports, Harvard University*.

Thomas, A.C. (2006). The Impact of Puck Possession and Location on Ice Hockey Strategy. *Journal of Quantitative Analysis in Sports, 2, Article 6*.

ASSOCIATIONS BETWEEN THE TIMING OF PLAYER CONTRACT SIGNINGS AND PLAYER PERFORMANCE IN THE AUSTRALIAN FOOTBALL LEAGUE

Sam McIntosh[a,b] & Sam Robertson[a]

[a] *Institute for Health & Sport (IHES), Victoria University, Melbourne, Australia*

[b] *Dr Sam McIntosh: samuel.mcintosh@vu.edu.au*

Abstract

An integral incentive for an athlete in any team sport is the status of their contract. Despite multiple studies assessing the relationship of player performance on player value in the team sport notational literature, there is limited knowledge on the relationship between a player's contract status and individual player performance. This research analysed the extent to which player performance, defined as Australian Football League (AFL) Player Ratings, differs dependent on the status of a player's contract in the AFL. A measure of 'expected performance' was modelled allowing for an exploration into the differential with actual performance as a function of contract status. The results indicated that players who signed mid-season and were out-of-contract at the end of that season showed substantial differences between performance in the matches prior to and post signing. Furthermore, athletes who have more consistent performances (lower relative standard deviation) are less likely to see a reduction in performance post signing, as compared to more inconsistent performers. The findings and applications outlined in this research provide an explanation of the association between player performance with respect to the timing of player contract signings and could be used as an example of associations worth investigating to identify refined indicators of expected performance for matches post the signing of an AFL contract.

Keywords: decision support, performance analysis, data visualisation, player evaluation, team sport

References

Browne, P., Sweeting, A., Davids, K., & Robertson, S. (2019). Prevalence of interactions and influence of performance constraints on kick outcomes across Australian Football tiers: Implications for representative practice designs. *Human movement science*, *66*, pp. 621-630. doi:10.1016/j.humov.2019.06.013

Della Torre, E., Giangreco, A., Legeais, W., & Vakkayil, J. (2018). Do Italians really do it better? Evidence of migrant pay disparities in the top Italian football league. *European Management Review*, *15*(1), pp. 121-136. doi:10.1111/emre.12136

Frick, B. (2011). Performance, salaries, and contract length: Empirical evidence from German soccer. *International Journal of Sport Finance*, *6*(2), pp. 87-118.

Gómez, M., Lago, C., Gómez, M., & Furley, P. (2019). Analysis of elite soccer players' performance before and after signing a new contract. *PLOS ONE*, *14*(1). doi:10.1371/journal.pone.0211058

Jackson, K. (2016). Assessing player performance in Australian football using spatial data (Doctor of Philosophy). Swinburne University of Technology.

McIntosh, S., Jackson, K. B., & Robertson, S. (2021). Apples and oranges? Comparing player performances between the Australian Football League and second-tier leagues. *Journal of Sports Sciences*, *39*(18), pp. 2123-2132. doi:10.1080/02640414.2021.1921372

McIntosh, S., Kovalchik, S., & Robertson, S. (2018). Validation of the Australian Football League Player Ratings. *International Journal of Sports Science & Coaching*, *13*(6), pp. 1064-1071. doi:10.1177/1747954118758000.

McIntosh, S., Kovalchik, S., & Robertson, S. (2019). Multifactorial Benchmarking of Longitudinal Player Performance in the Australian Football League. *Frontiers in Psychology*, *10*(1283). doi:10.3389/fpsyg.2019.01283

Robertson, S., Spencer, B., Back, N., & Farrow, D. (2019). A rule induction framework for the determination of representative learning design in skilled performance. *Journal of Sports Sciences*, *37*(11), pp. 1280-1285. doi:10.1080/02640414.2018.1555905

Teune, B., Woods, C., Sweeting, A., Inness, M., & Robertson, S. (2021). The influence of environmental and task constraint interaction on skilled behaviour in Australian Football. *European Journal of Sport Science*, pp. 1-8. doi:10.1080/17461391.2021.1958011

BODYSURFING

Neville de Mestre

Bond University
Founder of ANZIAM Mathsport
margnev2@bigpond.com

Abstract

Australians have been bodysurfing for more than 120 years, and the Polynesians even longer. The early history of this sport is considered, followed by the mathematical and physical aspects of catching and riding ocean waves that break near the shore. Some comments on the difficulties of fluid dynamical research into bodysurfing involving drag, buoyancy, gravity and forward propulsion are considered. Finally, there is an awareness of the dangers and fun associated with this sport, which are covered in more detail in my book *Bodysurfing*, published in 2009, the only Australian book on the subject.

THE ANATOMY OF A RUN OUT IN CRICKET

Paul J. Bracewell [a,b,c] , Jason D. Wells [a,b]
[a] *Footmarx, New Zealand*
[b] *Play in the Grey, New Zealand*
[c] *Corresponding author: paul@playinthegrey.com*

Abstract

Advancements in machine vision have enabled this technology to become a viable alternative to GPS data for tracking player movement (McDonald et. al., 2020; Trowland et. al., 2020). Fixed camera footage is readily available on the internet, including those posted from national governing bodies, such as NZ Cricket (www.nzc.nz). The three-stage approach outlined by McDonald et. al. (2020) is used to convert raw match footage into a set of xy coordinates per player per frame. Here, we convert footage of two batters running between the wickets in a New Zealand First Class game played late in the 2020-21 season into machine readable data for further analysis.

We show the running paths of the two batters. In addition, we also show the acceleration and speed of the dismissed batter per frame, which can be used to infer timeliness of decision-making. In this real-world scenario, it took 78 Frames for the ball to come in from the field after leaving the bat (3.12 seconds). This required the dismissed batsmen to run at 20.4kph, the equivalent of running 100m in 17.65 seconds. Further assessment of the frame-by-frame data indicates the dismissed batter delayed his decision to run after first making a move to run by 9 frames (run out by 7 frames). In addition, this batter had a curved run, running an extra 15% to his running distance. This technology becomes potent for effective coaching of running between wickets.

Keywords: Machine Vision, Scouting, Acceleration

1. INTRODUCTION

Cricket continues to evolve as a game (Noorbhai et. al., (2015). Driven by spectator appeal, increased professionalism and athleticism, cricket has adapted to capture and retain both players and fans alike. Shortened formats of the game (T20 and The Hundred) create an approximately three-hour version of the game. The introduction of age and stage to retain junior cricketers also shows a willingness to adapt (Renshaw, 2017). Coaching and umpiring aids also continue to embrace emerging technology.

Shaheen (2021) outlined the technology used within game as part of the Decision Review System (DRS). This technology ranges from: Hawk-eye (ball tracking), Hotspot (detecting heat from contact between bat and ball), Snicko (profiling sound waves). These cover the use of ball tracking, imaging, and profiling sound. There are other implementations which are of interest. Chowdhury (2016) explored the use of machine vision to detect front foot no-balls. Bracewell et. al. (2020) showed how bowling speeds could be estimated from historical footage. Pai (2020) provided a tutorial for ball tracking using Python which is the extra detail required to implement what Bracewell et. al (2020) outlined.

Moodley et. al. (2022) state that there have been limited studies demonstrating the validation of batting techniques in cricket using machine learning. They showed how the batting back lift technique in cricket can be automatically recognised in video footage. Other approaches are outlined for cricket and other sports relating to movement and technique. As the cost of computing power decreases, the ability to process this type of footage becomes increasingly accessible. For example: Bracewell et. al. (2022) tracked the centre of play in rugby union; Bakhai (2020) outlined the use of machine vision in baseball; Faulkner et. al. (2015) investigated player detection for use in Australian Rules). New approaches and use cases will continue to emerge, driven by blogs and tutorials like Pai (2020) across multiple sports.

Here we take a different approach and assess the decision-making process of a batter. This is achieved by showing the running path. Importantly, we also show the change in speed which can be used to infer timeliness of decision-making.

2. METHODS

DATA

Using accessible footage available online we explore a run out to show how machine vision works within a cricket setting (https://www.nzc.nz/domestic/competition-centres/plunket-shield). Usefully, the footage from this site is a single fixed camera per game which makes the process of converting the human detections to a top-down view via a homographic transformation much simpler. This process is explained by McDonald et. al., (2020). Given cricket pitches are standardised, approximations of distance are readily obtained. To complete a run, batters move the running crease, also referred to as the non-strikers popping crease and the popping crease, which is 17.68m. This is shown in Figure 1.

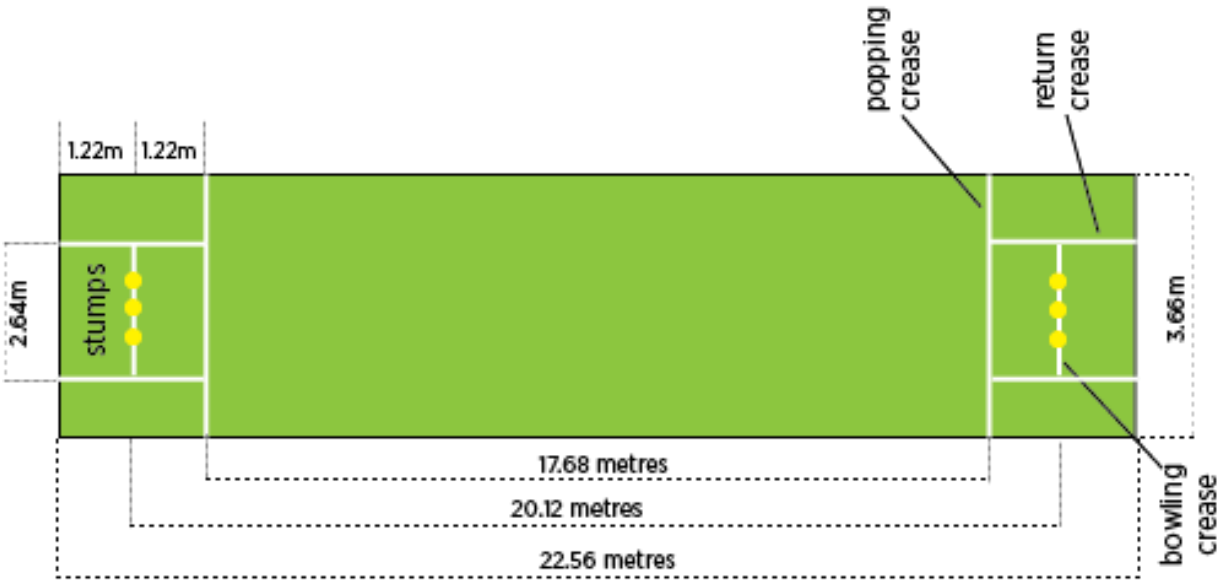


Figure 1: Stylised diagram showing the dimensions of a cricket pitch produced by the Government of Western Australia (2019).

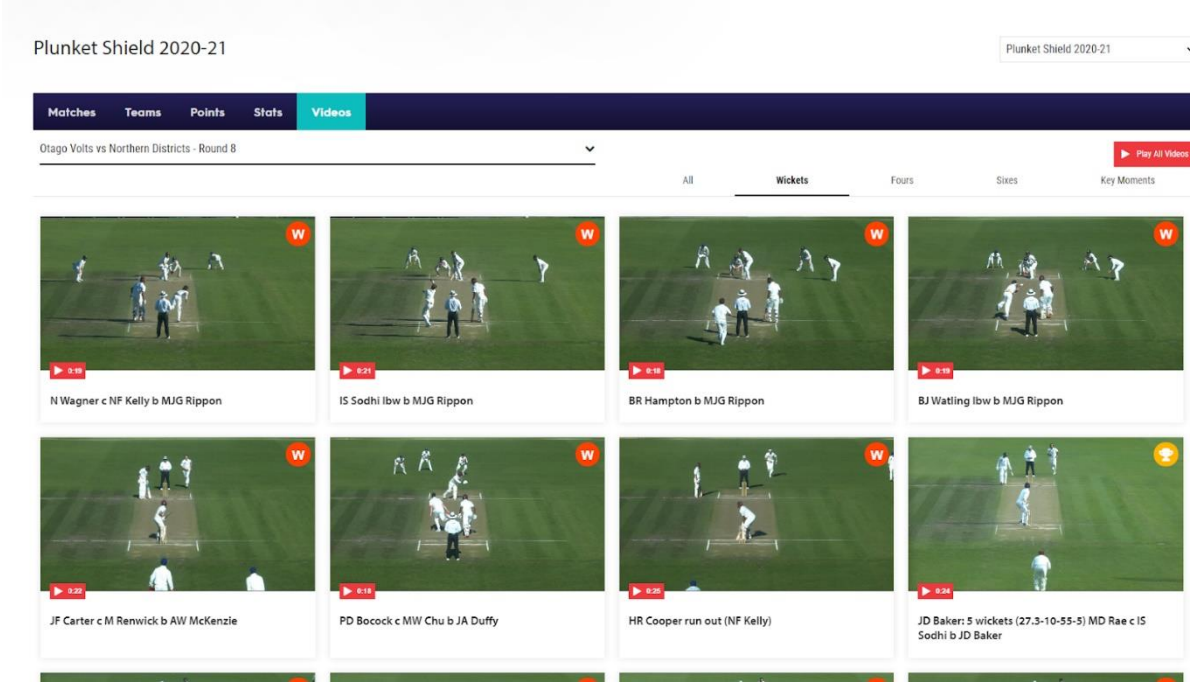


Figure 2: Accessing publicly available footage to assess Player Movement.

From the NZC website, we selected one run out for review. This is the clip third from left in the bottom row of Figure 2. This clip is titled “HR Cooper run out (NF Kelly)” and is from the Otago Volts vs Northern Districts Round 8 game of the 2020/21 Plunket Shield (New Zealand domestic first-class cricket).

From match footage, players are detected in each frame. Also, key features, such as crease, wickets and marking for danger zone are identified to help work out where a player is relative to the pitch. This aids in the homography and conversion to the top-down view.

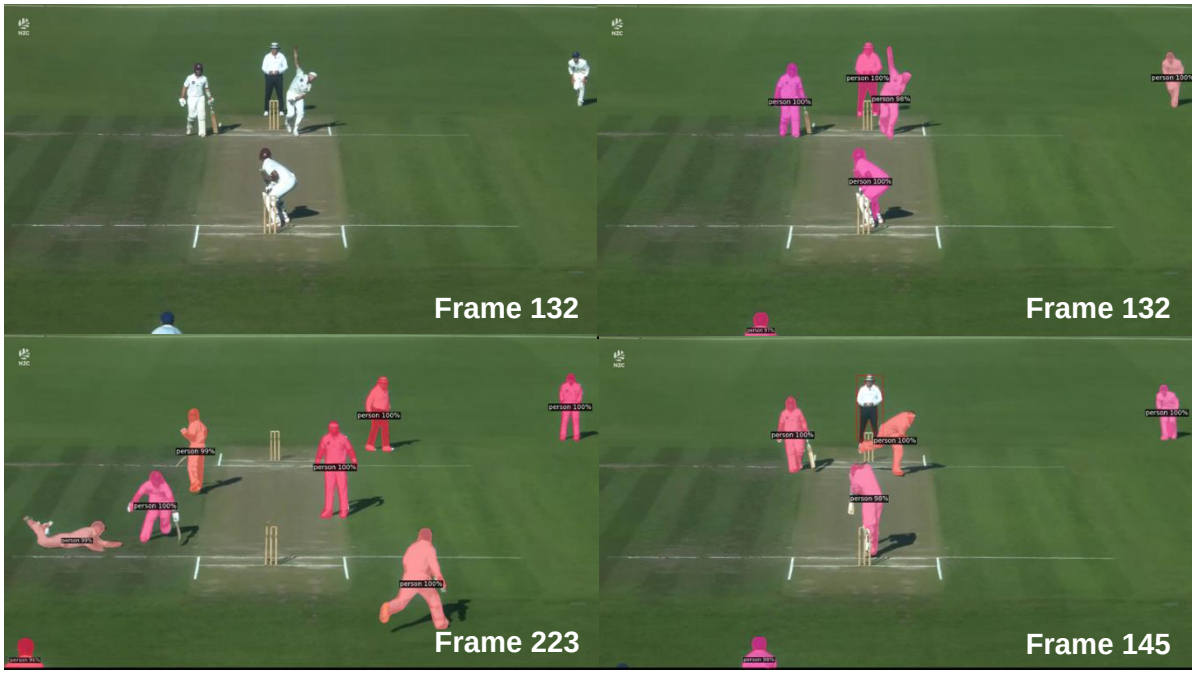


Figure 2: Mosaic showing key moments in HR Cooper's run out effected by NF Kelly in the Round 8 Plunket Shield match played between Otago and Northern Districts in the 2020/21 season.

In the top left, the image shows the point of release for the bowler at frame 132. The top right image is the same frame, but with human detections. Bottom right-hand image is frame 145 (revealing the ball was travelling at 122kph) which shows when the ball was stuck. The fourth and final image in the bottom left shows frame 223 which is when the bail first lifts after the wickets have been struck. In this final image, Cooper is short of his ground.

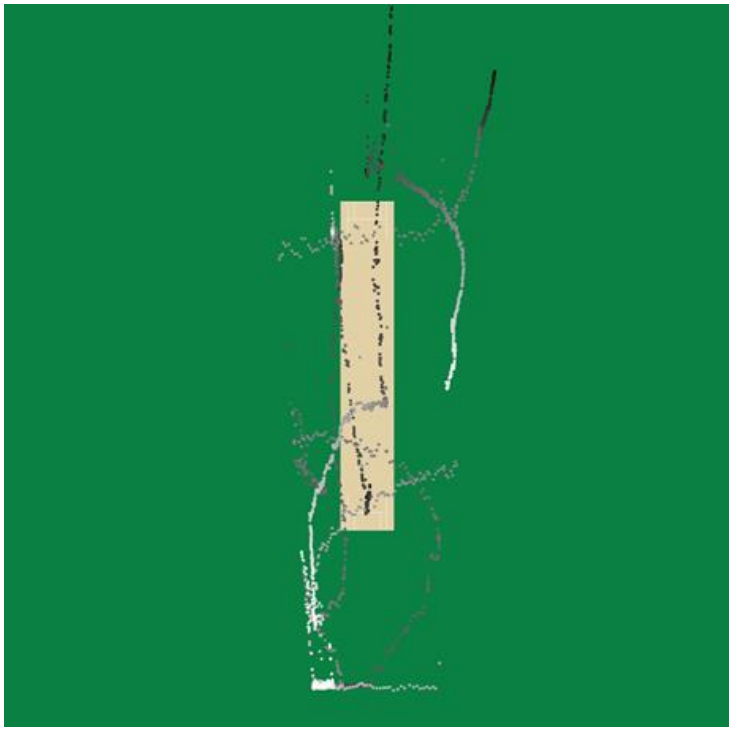


Figure 3. Paths of all detected humans from the clip for HR Cooper's dismissal

The player detections from Figure 2 are then converted into a top-down view as shown in Figure 3. The darker dots are the start of the detected human's movement, and the lighter dots are movements toward the end of the clip. The fielders and bowlers are seen which provides an additional perspective on fielding movement.

3. RESULTS

Understanding how people move allows insights to be extracted. It took 78 Frames for the ball to come in from the field (frame 223) after leaving the bat (frame 145) (3.12 seconds). This required Cooper to run at 20.4kph, the equivalent of running 100m in 17.65 seconds. For a first-class athlete, this is achievable.

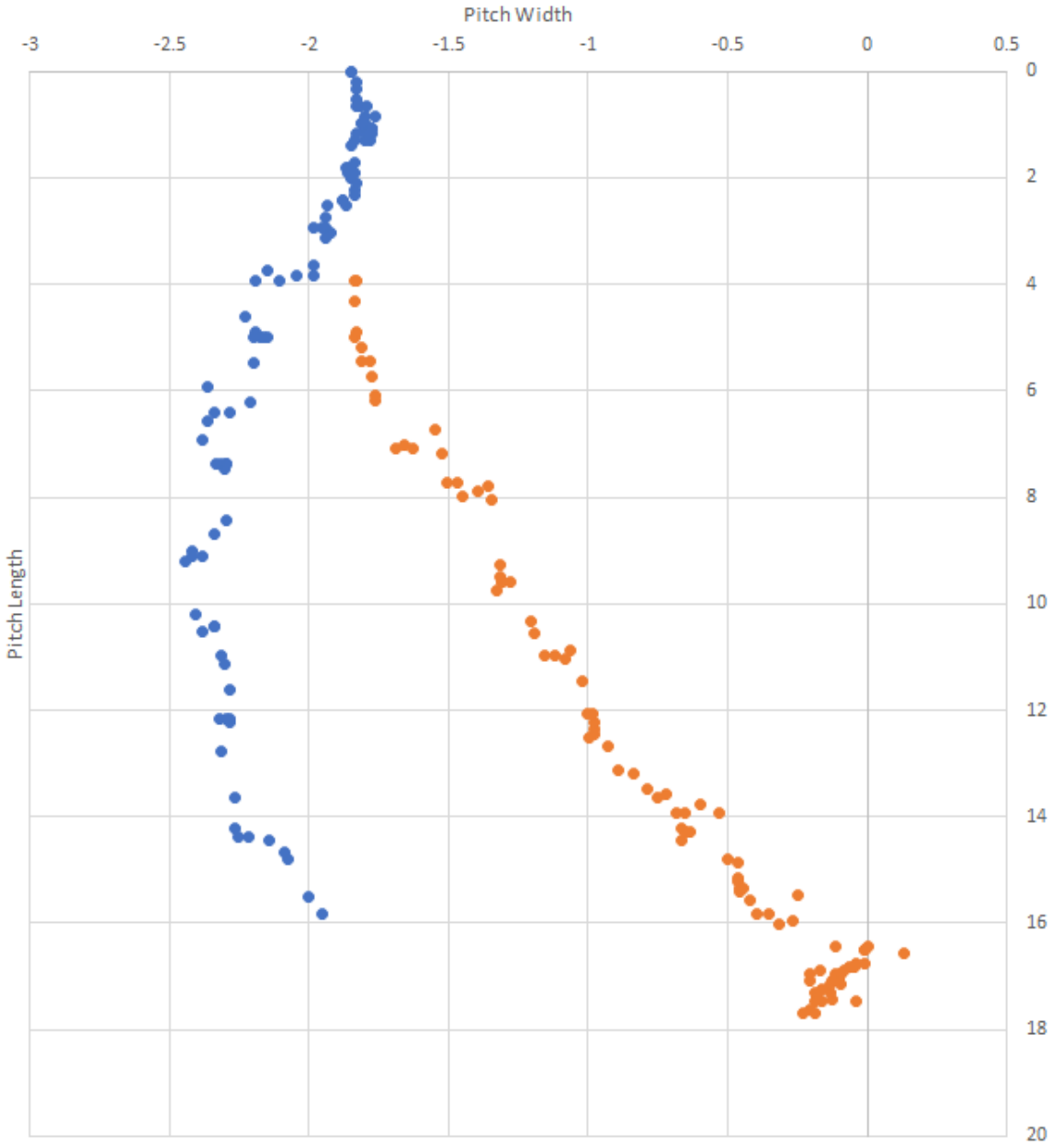


Figure 4: Running paths of Cooper (Blue dots, non-striker, starts at top of graph and top of image from frame 145 and Raval (Orange dots, striker, starts at bottom of graph and bottom of image from frame 145)

The distance travelled by each batter is calculated as the sum of the Euclidean distance between the x,y, coordinates for consecutive frames between frames 145 and 223. A feature of Cooper's path, shown as the blue dots in Figure 4, is the curve. The estimated total distance he ran was 20.32 metres, which is 15% more than necessary to cover the 17.68m between the popping and running creases. In further investigations we will explore the application of smoothing to remove some of the jitter.

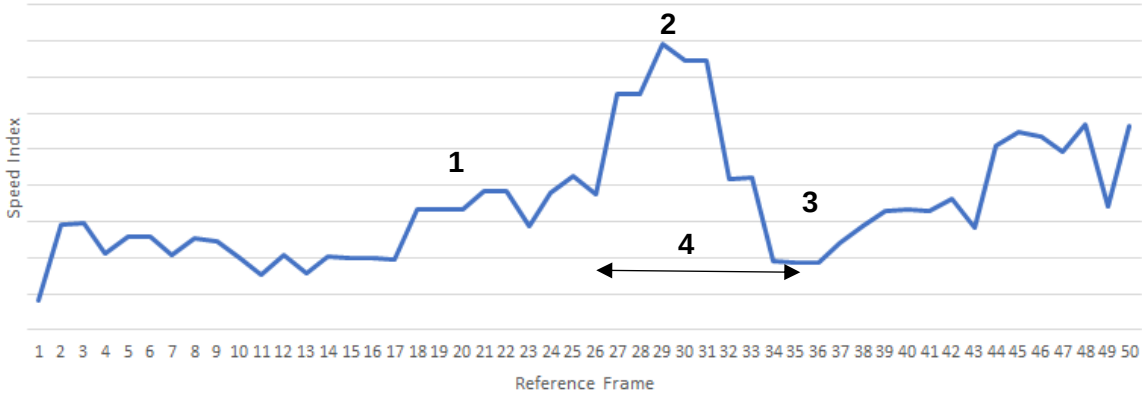


Figure 5: Speed Index per frame for the first 50 frames for Cooper's movement from when the ball is struck

Looking at the first two seconds of footage, we see Cooper does not react until after nearly 1 second (about 20 frames [1]). We can see the spike after approximately 30 frames [2] where he begins to run and hit top speed (after 27 frames), he then stalls and does not commence running again until 36th Frame [3]. Cooper was run out by 7 frames. Between the 27th and 36th frame Cooper "lost" 9 frames through indecision [4](0.36 of a second).

The profile in figure is based on the movement between pixels, so we refer to this as a speed index. The important elements to derive from the graph below are the stillness, up to frame 20 and then the relative start-stop-start movement that is evident from about the 20th frame.

3. DISCUSSION

The use of machine vision enables new data sources to be captured. Here, we have been able to assess how Northern Districts batter, HR Cooper moved while attempting a single run. Two features stand out. Firstly, the arc of Copper's running path added an extra 15% of running distance. Cooper delayed his decision to run after first making a move to run by 9 frames (run out by 7 frames).

The impact of those elements becomes important when considering the outcome. Had he run straight, this would have seen him complete the run in 13 fewer frames. This is estimated by assuming a constant speed across the 85 frames that it took Cooper to slide his bat across the line (frame 230). A 15% reduction is relative to 13 fewer frames (he was run out by 7 frames).

If Cooper had either run straight or committed early, he would have successfully completed the run successfully, despite the direct hit from NF Kelly.

4. CONCLUSION

Machine vision has increasing applications in cricket. These applications are beyond decisioning and entertainment. Here, we showed by tracking an individual, we were able to assess different elements of the run: namely the directionality of his path and the point at which he made a clear decision to attempt the run. Both features are useful for training and coaching purposes at all levels. Importantly, this is achieved from a fixed camera. This makes this type of approach accessible to coaches of all ages, particularly as Trowland et. al. (2022) showed that tracking could be undertaken from an iPhone. Extending beyond a single case to process many "runs" would help further insights into batter behaviour. Of particular interest is the reaction time and how quick a batter responds to a call or cue from their batting partner. Combing this type of processing to include fielding expands the capability to interrogate the viability of a run and evaluate the level of risk a batter is willing to assume, relative to the match context.

References

Bakhai, R. (2020, April 8). Using Computer Vision and A.I. to Enhance Coaching in Baseball. Medium.com Available at: https://medium.com/@ravibakhai/using-computer-vision-and-artificial-intelligence-to-enhance-coaching-in-baseball. Accessed: 3 December 2021.

Bracewell, P.J., Scott, F., & Bracewell, B.P. (2020). Historical evaluation of bowling speeds in cricket using machine vision. Proceedings of the 15th Australian Conference on Mathematics and Computers in Sports. Ray Stefani & Adrian Schembri eds. Wellington, New Zealand ANZIAM Mathsport. pp. 51-56.

Bracewell, P. J., Trowland, H.E., Wilson, J.K., & Tikhonov, A. (2020). Using machine vision to track the locus of play in rugby union. Proceedings of the 15th Australian Conference on Mathematics and Computers in Sports. Ray Stefani & Adrian Schembri eds. Wellington, New Zealand ANZIAM Mathsport. pp. 57-60.

Chowdhury, A. Z. M. E., Rahim, M. S. and Rahman, M. A. U (2016). Application of computer vision in Cricket: Foot overstep no-ball detection, 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 2016, pp. 1-5, doi: 10.1109/CEEICT.2016.7873086.

Faulkner H., & Dick, A. (2015). AFL Player Detection and Tracking. International Conference on Digital Image Computing: Techniques and Applications (DICTA), 2015, pp. 1-8, doi: 10.1109/DICTA.2015.7371226.

Government of Western Australia (July, 2019) https://www.dlgsc.wa.gov.au/sport-and-recreation/sports-dimensions-guide/cricket Accessed January 27th, 2020

McDonald, R.G., Bracewell, P. J., Trowland, H.E., & Tikhonov, A. (2020). Discounting camera movement in calculation of player paths using machine vision in rugby union. Proceedings of the 15th Australian Conference on Mathematics and Computers in Sports. Ray Stefani & Adrian Schembri eds. Wellington, New Zealand ANZIAM Mathsport. pp. 39-44.

Moodley, T., van der Haar, D. & Noorbhai, H. (2022). Automated recognition of the cricket batting backlift technique in video footage using deep learning architectures. Sci Rep 12, 1895 (2022). https://doi.org/10.1038/s41598-022-05966-6

Noorbhai, M. & Noakes, T. (2015) Advances in cricket in the 21st century: Science, performance and technology. Afri. J. Phys. Health Educ. Recreat. Dance 21(4.2), 1310–1320.

Pai, A. (2020, March 31). Howzatt! How to Build Your Own Ball Tracking System for Cricket. Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2020/03/ball-tracking-cricket-computer-vision/#h2_8. Accessed 1 December 2021.

Renshaw, I. (2017, September 18). Dr Ian Renshaw presents Junior Formats findings to Queensland Cricket coaches. Cricket Australia. Available at: https://www.community.cricket.com.au/coach/news/dr-ian-renshaw-presents-junior-formats-findings-to-queensland-cricket-coaches/2017-09-18). Accessed 1 December 2021.

Shaheen (2021, August 8). Hawkeye, Snicko, DRS and more. theSporting.Blog. Available at: https://thesporting.blog/blog/technology-in-cricket-the-rise-of-hawkeye-snicko-drs-and-more. Accessed 1 December 2021.

Trowland, H.E., Bracewell, P.J., McDonald, R.G., & Tikhonov, A. (2020). Validating player path tracking using machine vision. Proceedings of the 15th Australian Conference on Mathematics and Computers in Sports. Ray Stefani & Adrian Schembri eds. Wellington, New Zealand ANZIAM Mathsport. pp 45-50.

IDENTIFYING MULTIVARIATE CRICKET PERFORMANCE USING PARETO FRONTIERS

Tim Newans [a,b,c], Phillip Bellinger [a] , Clare Minahan [a]

[a] *Griffith University, Gold Coast, Australia*
[b] *Queensland Academy of Sport, Brisbane, Australia*
[c] *Corresponding author: t.newans@griffith.edu.au*

Abstract

In Twenty20 cricket, there is a trade-off relationship between batting average and strike rate as well as bowling strike rate, economy, and average. This study presents Pareto frontiers as a tool to identify athletes who possess an optimal ranking when considering multiple metrics simultaneously. 884 matches of Twenty20 cricket from the Indian Premier League were compiled to determine the best batting and bowling performances, both within a single innings and across each player's career. Pareto frontiers identified nine optimal batting innings and six batting careers. Pareto frontiers also identified three optimal bowling and five optimal bowling careers. Each frontier identified players that were not the highest ranked athlete in any metric when analysed univariately. Pareto frontiers can be used when assessing talent across multiple metrics, especially when these metrics may be conflicting or uncorrelated. Pareto frontiers can identify athletes that may not have the highest ranking on a given metric but have an optimal balance across multiple metrics that are associated with success in a given sport.

Keywords: cricket; visualisation; talent identification; optimal selection

1. INTRODUCTION

The need to identify attributes to quantify optimal performance is evident for every sport (1). With the exception of a few single-skill sports (2), most athletes require a number of attributes to perform in their given sport. These attributes can encompass physical (3), physiological (4), mental (5), or skill-based characteristics (6), that all can contribute to the performance of a player. Attributes such as speed, endurance, agility, strength, power, and accuracy are common across multiple sports (6), and each attribute can have multiple variables seeking to quantify that attribute. As such, coaches and support staff are consistently looking for new variables that could be used to either quantify new attributes of interest or develop more variables to better quantify already-identified attributes with the hope that these new variables can identify previously-hidden talent or interrogate subtle differences between different athletes. However, with the increase in the number of attributes of interest, the likelihood that an athlete excels in every attribute decreases. Consequently, methods are required that can analyse multiple attributes simultaneously, rather than viewing each attribute in isolation.

While traditional research statistical techniques focus around identifying the mean and standard deviation of a population (7), sports typically are not interested in the mean during talent identification processes, rather, they are looking for outliers. That is, coaches and support staff are looking for athletes that sit the furthest away from the mean in the direction that success is defined. Therefore, when multiple attributes are of interest, selection of athletes is by choosing athletes that sit the further away from the mean within each attribute. While this process can work when variables are positively correlated, this process can miss talent when variables are negatively correlated. For instance, at the elite level, there is a negative correlation between maximal sprint speed and endurance capacity (8). However, running-based team sports require athletes possess both speed and endurance to play at the elite level and, therefore, players necessarily need to trade off between having optimal speed and optimal endurance. In its simplicity, if both speed and endurance were equally required for success, selecting the top-*n* sprinters and the top-*n* endurance runners may not be the optimal athletes for that sport.

Consequently, both attributes need to be viewed in tandem. The process of optimising the balance of multiple attributes is termed 'multi-objective optimisation'. Mathematically, they aim to create the perfect balance of the attributes of interest. If a data point was defined as: $\vec{x}_1 \epsilon X$, it is, therefore, better than another data point defined by: $\vec{x}_2 \epsilon X$ if $f_i(\vec{x}_1) \leq f_i(\vec{x}_2)$ for all metrics $i \epsilon \{1,2,\dots,k\}$ and $f_j(\vec{x}_1) < f_j(\vec{x}_2)$ for at least one metric $j \epsilon \{1,2,\dots,k\}$. Once these conditions have been met, the remaining points are deemed Pareto-optimal and form what is called the Pareto frontier.

In Twenty20 cricket, there are multiple facets within both batting and bowling that can define success. Unlike Test cricket and, to an extent, One-Day cricket where scoring as many runs as possible regardless of how many

deliveries faced is of most importance, Twenty20 crickets requires batters to score faster (i.e., higher strike rate) and for bowlers to concede minimal runs which, in some cases, can come at the expense of preserving their wicket. Therefore, there is a trade-off relationship between batting average and strike rate as well as bowling economy, average, and strike rate within Twenty20 cricket. For example, early on in an innings the risk-return of attempting to hit six runs off a ball is significantly different than in the final over of an innings. Similarly, a bowler needs to balance taking wickets while also conceding as few runs as possible. For instance, when bowling four overs, it is again difficult to determine whether taking three wickets for 50 runs is of more worth than taking no wickets but only conceding eight runs as the three wickets may not have been worth conceding 50 runs. As both attributes within each domain are of interest, Pareto frontiers can be used to determine batters and bowlers that may not record the highest in either variable but display an optimal balance of the two attributes. Therefore, when assessing the quality of players, it is necessary to utilise tools that can analyse these datasets without favouring one metric over another. Therefore, the present study aimed to use Pareto frontiers to identify the best performing Twenty20 batters and bowlers.

2. METHODS

The present study comprised all 884 matches of the first 14 editions of the men's Indian Premier League (IPL), India's domestic T20 cricket competition. The dataset contained 566 batters and 467 bowlers. Collectively, there were 13,357 individual batting innings with observations ranging from 1-208 innings per batter, while there were 10,925 individual bowling innings with observations ranging from 1-180 innings per bowler.

To summarise the data, two summary statistics were generated for batting and three summary statistics were generated for bowling. The summary statistics were as follows:

- Batting Average: runs scored divided by frequency of dismissal
- Batting Strike Rate: runs scored divided by balls faced
- Bowling Average: wickets taken divided by runs conceded
- Bowling Strike Rate: wickets taken divided by balls bowled
- Bowling Economy: runs conceded divided by overs (i.e., 6 balls) bowled

To understand both the batting and bowling attributes within cricket, four Pareto frontiers for were established within the dataset:

i. *Pareto-optimal Batting Innings*
This analysis outlined the highest runs scored within an innings at the highest strike rate.

ii. *Pareto-optimal Batting Career*
This analysis outlined the highest batting average across a career at the highest strike rate. To provide a more accurate career report, batters required to have played a minimum of 20 innings which left 163 eligible batters.

iii. *Pareto-optimal Bowling Innings*
This analysis outlined the most wickets taken in an innings at the lowest economy.

iv. *Pareto-optimal Bowling Career*
This analysis outlined the lowest bowling average across a career at the lowest economy and lowest strike rate. To provide a more accurate career report, bowlers required to have bowled in more than 20 matches, which left 145 eligible bowlers.

The *rPref* package (9) was used in R v 4.1.0 (10) to determine the Pareto frontiers using the *psel* function with the 'top_level' argument set to 999 to ensure every athlete was assigned to a frontier.

3. RESULTS

Pareto-optimal Batting Innings

Nine Pareto-optimal innings were identified with extremities ranging from 6 runs off 1 ball (i.e., strike rate = 600) to 175 off 66 balls (i.e., strike rate = 265.15). Additionally, the solution of 6 runs off 1 ball has been attained eight times. The IPL batting innings Pareto frontier is displayed in Figure 1 and the batters are listed in Table 1.

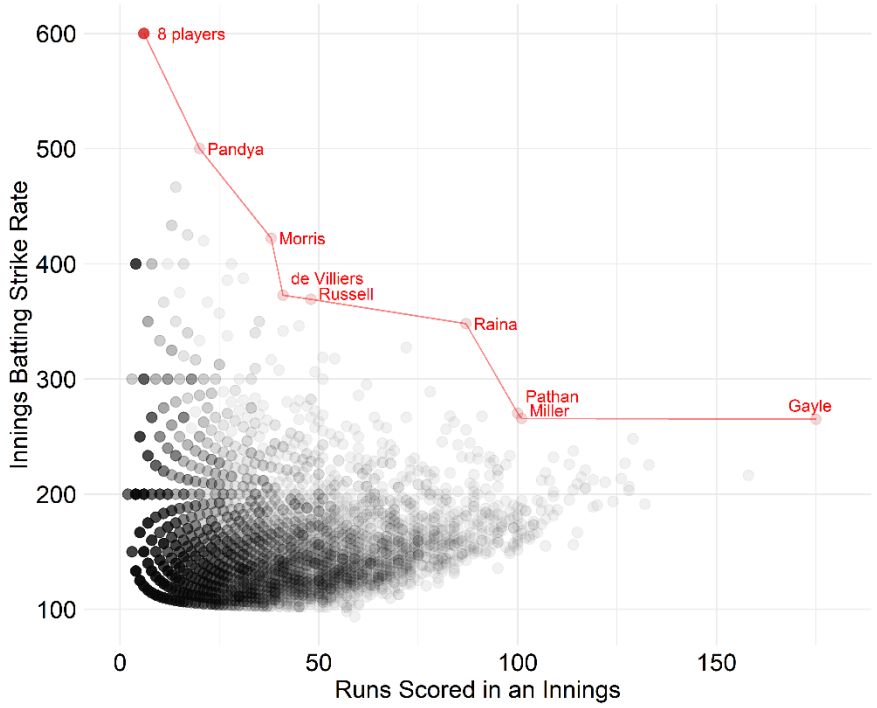


Figure 1. Pareto-optimal batting within an IPL innings. N.B. For illustrative purposes, points were filtered out i if both their runs scored was below 50 and their strike rate was below 100.

Table 1. List of all Pareto-optimal IPL batting innings

Batter	R (B)	Strike Rate	Match
Chris Gayle	175 (66)	265.15	IPL06 Match 31
David Miller	101 (38)	265.78	IPL06 Match 51
Yusuf Pathan	100 (37)	270.27	IPL03 Match 2
Suresh Raina	87 (25)	348.00	IPL07 Match 59
Andre Russell	48 (13)	369.23	IPL12 Match 17
AB de Villiers	41 (11)	372.72	IPL08 Match 16
Chris Morris	38 (9)	422.22	IPL10 Match 9
Krunal Pandya	20 (4)	500.00	IPL13 Match 17
Numerous	6 (1)	600.00	IPL04 Match 74 *1st occurrence*

Pareto-optimal Batting Career

Six Pareto-optimal batting careers innings were identified. Andre Russell recorded the highest career batting strike rate with 178.57 runs per 100 balls, while KL Rahul recorded the highest batting average with 47.43 runs per dismissal. The IPL batting career Pareto frontier is displayed in Figure 2 and the batters are listed in Table 2.

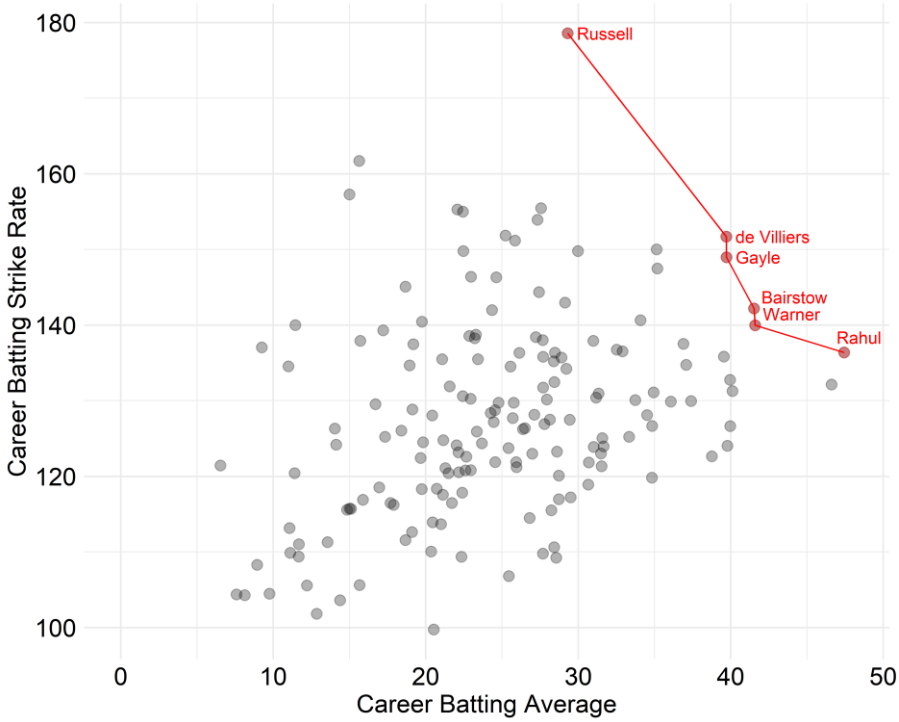


Figure 2. Pareto-optimal batting within across an IPL career. N.B. For illustrative purposes, points were filtered out if both their average was below 20 and their strike rate was below 100.

Table 2. List of all Pareto-optimal IPL batting careers

Batter	Innings	Average	Strike Rate
KL Rahul	85	47.43	136.38
David Warner	150	41.60	139.97
Jonny Bairstow	28	41.52	142.19
Chris Gayle	141	39.72	148.96
AB de Villiers	170	39.71	151.69
Andre Russell	70	29.31	178.57

Pareto-optimal Bowling Innings

Three Pareto-optimal bowling innings were identified: 2/0 by Suresh Raina, 5/5 by Anil Kumble, and 6/12 achieved by Alzarri Joseph. The IPL bowling innings Pareto frontier is displayed in Figure 3 and the bowlers are listed in Table 3.

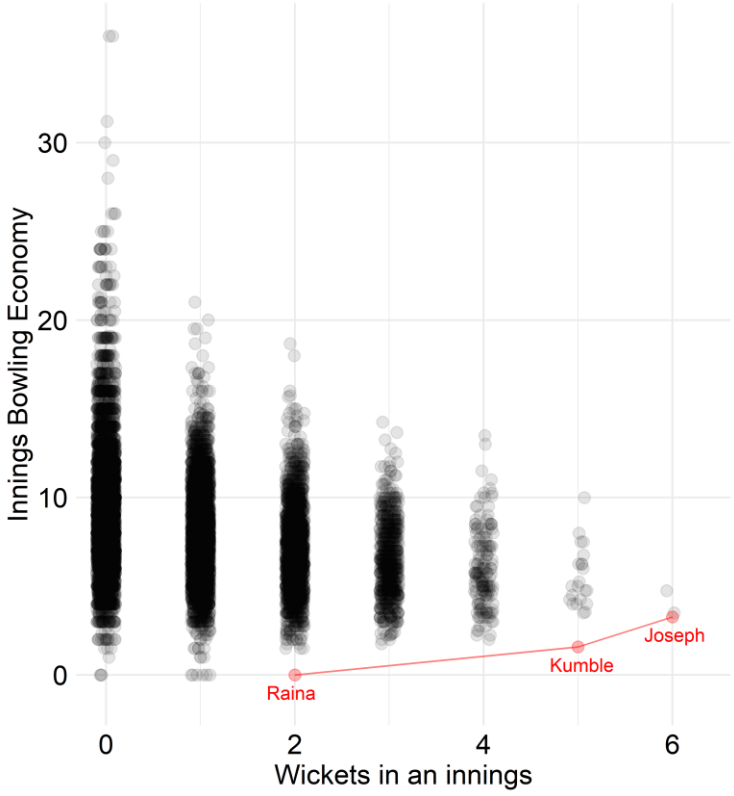


Figure 3. Pareto-optimal bowling within an IPL innings.

Table 3. List of all Pareto-optimal IPL bowling innings

Batsman	Overs	Wickets	Runs	Match
Suresh Raina	0.3	2	0	IPL04 Match 52
Anil Kumble	3.1	5	5	IPL02 Match 2
Alzarri Joseph	3.4	6	12	IPL12 Match 19

Pareto-optimal Bowling Career

Five Pareto-optimal bowling careers were identified, with Doug Bollinger achieving the lowest average, Rashid Khan achieving the lowest economy, while Kagiso Rabada recorded the lowest strike rate. The IPL bowling career Pareto frontier is displayed in Figure 4 and the bowlers are listed in Table 4.

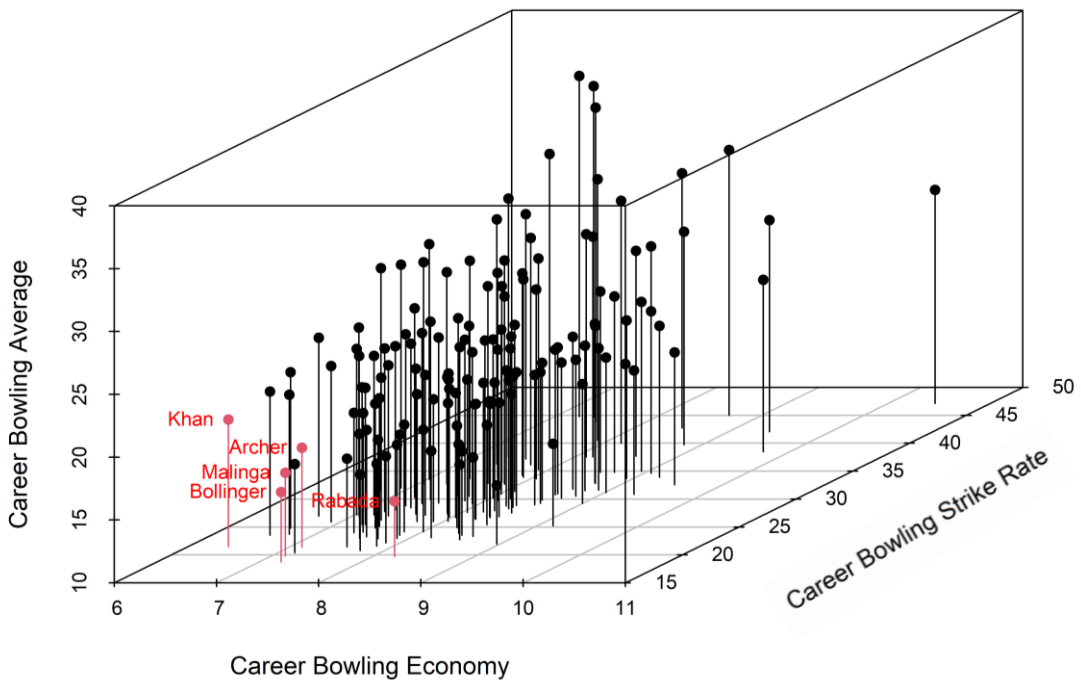


Figure 4. Pareto-optimal bowling across an IPL career.

Table 4. List of all Pareto-optimal IPL bowling careers

Batsman	Innings	Average	Economy	Strike Rate
Doug Bollinger	27	18.73	7.22	15.57
Kagiso Rabada	59	19.71	8.22	14.39
Lasith Malinga	122	19.79	7.14	16.63
Jofra Archer	35	21.33	7.13	17.93
Rashid Khan	86	21.46	6.40	20.12

4. DISCUSSION

This study sought to use Pareto frontiers to visualise optimal Twenty20 cricket batting and bowling performances, both within an innings as well as across a career. By analysing performance multivariately, rather than simply analysing multiple variables univariately, players can be deemed optimal despite not being objectively highest in a single variable. When conflicting attributes are of equal interest, Pareto frontiers can view these variables in tandem as the expectations of an individual to attain the highest level in both attributes univariately may be unfeasible. All four Pareto frontiers contained at least one athlete that was not the highest ranked athlete in any metric when analysed univariately, and yet was deemed Pareto-optimal due to their balance in the metrics of interest.

The main advantage of Pareto frontiers highlighted in the present study is identifying athletes who are optimal across multiple metrics even when they are not the highest ranked in any metric. This was most evident where Chris Gayle, when viewed univariately, has the 9th-highest career batting average (39.72), which is 6.71 runs per innings lower than the highest (Figure 2). Similarly, he has the 14th-highest strike rate, striking at 148.96 which is 29.61 runs per 100 balls lower than the highest. However, when considering both metrics simultaneously and visualising these metrics, he is one of the best batsmen across the 14 seasons of the IPL.

The present study also illustrated how Pareto frontiers can be used to visualise talent in more than 2 dimensions. For example, while Jofra Archer has the sixth-lowest bowling average, 14th-lowest economy, and the 19th-lowest strike rate (see Figure 4), he can be deemed a Pareto-optimal bowler as there are no other bowlers who supersede him across all three metrics. While there will be some correlations between the three bowling metrics (i.e., average, economy, and strike rate) as the metrics are related (e.g., wickets taken is the denominator of average and numerator of strike rate), visualising the third dimension is still necessary as the reader would still need to multiply the x and y values to understand where they would sit in the third dimension.

In the present study we chose to observe batting and bowling as purely independent roles within cricket; however, there are also avenues for Pareto frontiers to be established for all-rounders within cricket (i.e., players that are picked for both their batting and bowling ability). However, it should be noted that if an all-rounder Pareto frontier were to be established with both batting average and strike rate as well as bowling average, economy, and strike rate, the resulting five-dimensional outputs, while valid and executable, become increasingly difficult to interpret and visualise. To do such an analysis, a factor-reduction technique such as principal components analysis should be considered and the Pareto frontier could be built from the extracted components (e.g., batting and bowling).

While the present study is designed to be an introduction for sports scientists to the concept of Pareto frontiers, it should also be considered that there is some level of uncertainty surrounding each observation in the career Pareto frontiers due to the differing number of observations. For example, Jonny Bairstow is deemed Pareto-optimal as he is currently striking at 142.19 at an average of 41.52 after 28 innings; however, it is right to assume that it is more uncertain that he lies on the frontier than AB de Villiers who has 170 observations. Therefore, future research could consider providing confidence or credible intervals around the probability that an individual lies on the Pareto frontier. Consequently, it is then feasible that a probability that an individual sits on the first, second, or third frontier could be calculated.

While the present study used Twenty20 cricket to illustrate the power and usefulness of Pareto frontiers, the concept can be widely applied within sports science datasets, especially when the variables of interest are uncorrelated or negatively correlated. Pareto frontiers can still be established between two positively correlated metrics; however, it is likely that there will be less 'hidden' athletes on this frontier as naturally the athletes who are high in one metric will be high in the other metric. Future research should apply Pareto frontiers across different avenues within sports performance analysis which have multi-faceted determinants as there are many other possibilities within sports whereby Pareto frontiers can reveal athletes who possess the optimal balance of the metrics of interest.

5. CONCLUSIONS

With the proliferation of various physiological, mechanical, and skill-related attributes associated with performance, Pareto frontiers should be used within sports science to visualise multiple performance metrics. By analysing opposing data in tandem, more feasible expectations and benchmarks can be established to reveal talent that may have been missed when analysing multiple metrics univariately.

5. ACKNOWEDGEMENTS

The authors would like to acknowledge the support provided from the Queensland Academy of Sport's Sport Performance Innovation & Knowledge Excellence unit.

6. REFERENCES

1. Johnston K, Wattie N, Schorer J, Baker J. Talent identification in sport: A systematic review. Sports Med. 2018 Jan 1;48(1):97–109.
2. Rienhoff R, Hopwood M, Fischer L, et al. Transfer of motor and perceptual skills from basketball to darts. Front Psychol 2013; 4.
3. Kelly AL, Williams CA. Physical characteristics and the talent identification and development processes in male youth soccer: A narrative review. Strength Cond J 2020; 42(6):15–34..
4. Dodd KD, Newans TJ. Talent identification for soccer: Physiological aspects. J Sci Med Sport. 2018 Oct 1;21(10):1073–8.
5. Morris T. Psychological characteristics and talent identification in soccer. J Sports Sci 2000; 18(9):715–726.
6. Davids K, Lees A, Burwitz L. Understanding and measuring coordination and control in kicking skills in soccer: Implications for talent identification and skill acquisition. J Sports Sci 2000; 18(9):703–714.10.
7. Hopkins WG, Marshall SW, Batterham AM, et al. Progressive Statistics for Studies in Sports Medicine and Exercise Science. Med Sci Sports Exerc 2009; 41(1):3–12. Doi: 10.1249/MSS.0b013e31818cb278.
8. Sánchez-García M, Sánchez-Sánchez J, Rodríguez-Fernández A, Solano D, Castillo D. Relationships between sprint ability and endurance capacity in soccer referees. Sports. 2018 Jun;6(2):28.
9. Roocks P. Computing Pareto Frontiers and Database Preferences with the rPref Package. R J. 2016 Dec;8(2):393–40412.
10. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2019. Available from: https://www.R-project.org/

CRICKET'S NERVOUS NINETIES: FACT OR FANTASY?

Leo Roberts [ac], Ian Grundy [b], Matthew Spittal [a]

[a] *Centre for Mental Health, Melbourne School of Population and Global Health, University of Melbourne, Australia*
[b] *Mathematical Sciences Discipline, School of Science, RMIT University, Australia*
[c] *Corresponding author: leo.roberts@unimelb.edu.au*

Abstract

The *'nervous nineties'* is a famous colloquialism in cricket that conveys the mental challenge of batting within reach of 100 runs. The name itself (and its public usage) implies that batting in the nineties is more challenging than say batting in the eighties or once past 100, presumably due to internal pressure associated with the possibility of reaching (or failing to reach) the illustrious milestone. If it is true that the nineties are an especially difficult passage of an innings, historical data should show a change in the probability of dismissal, run rate or risk-taking at this time. Yet, despite the notoriety of this batting moment, our inspection of the relevant literature suggests that there are no formal attempts to verify the existence of the nervous nineties (either mathematically or phenomenologically). In fact, to our knowledge, only one peer-reviewed publication has explored the issue of batting performance near the century but did so with an interest in team organisational behaviour and focused on the arguably less prestigious one-day international game. To close the gap, we examined a combination of player-level, innings level and ball-by-ball level data using all available international test matches since 2004 – then modelled the regression discontinuity of performance indicators that might reflect nervousness around the 100 landmark. Preliminary analysis indicated no significant change in the probability of dismissal during the nineties but did suggest a tendency among batters to increase their run rate and score more boundaries. A separate multilevel logistic regression, used to specifically examine the predictors of getting out in the nineties, primarily revealed vulnerabilities in less skilled batters. Our analysis suggests that if players are nervous in the nineties, many have developed coping strategies to survive the period, possibly including playing more aggressively to rapidly get into three figures.

Keywords: cricket, batting, performance anxiety, the nervous nineties

WAIT PATTERNS FOR EVERY MAHJONG READY HAND

Alec Stephenson [a,d]

[a] *CSIRO, Melbourne, Australia.*
[d] *Corresponding author: alec.stephenson@csiro.au*

Abstract

Mahjong is a class of imperfect information games typically played with four players using a set of mahjong tiles. Players compete to form winning hands which are worth a certain number of points. It is a zero-sum game, where the winner obtains points from another player or players according to the game situation. Unlike bridge, there are no teams. The game originated in China and contrary to popular belief is relatively modern, with its origins dating from around 1880. The game was exported overseas and was extremely popular in America in the 1920s. There are many variants both internationally and within different provinces of China: Hong Kong, Sichuan, Shanghai, Taiwanese, Riichi (Japanese) are just a few examples.

In most forms of mahjong a standard winning hand consists of 14 tiles (17 in Taiwanese), and these 14 tiles consist of four sets of three (five sets of three in Taiwanese) and one pair. A pair is two identical tiles, whereas as set of three is either three identical tiles or a run of the same suit such as 1-2-3 or 7-8-9. A ready hand is a 13-tile hand which is one tile away from winning. This document for the first time lists every possible type of ready hand and every possible winning tile for each hand. This was achieved via brute-force, firstly by enumerating every case, and then using an axiomatic approach to define when two cases are of the same type. Previous combinatorial results on ready hands can now be derived by simply examining the table.

Keywords: Combinatorics, mahjong, mind games, ready hands

1. INTRODUCTION

A ready hand, or waiting hand, is a hand in mahjong that is one tile away from winning. The tables below give a wait pattern classification system for mahjong, covering all possible cases. There are 828 wait patterns, identified by integers from #1 to #828, although in the vast majority (99.67% in Riichi mahjong) of cases the wait classification will be a number between #1 and #60.

There are many mahjong variants (Lo, A., 2001); the system here corresponds to the regular hands in most 14-tile mahjong variants, where a regular hand is one that is completed by four sets and one pair. For consistency the terminology we use here comes from Riichi mahjong (Chiba, D., 2016, Miller, S. D., 2015). Our tables also provide another way of confirming previous combinatorial results in the literature such as Cheng, Y. et al. (2017). For recent research in mahjong AI, see Li, J. et al. (2020).

The wait classifications were derived by using a computer to generate all possible patterns and successively removing those that reduced to a simpler form. The most important thing for any wait classification system is mathematical consistency, but there is no one correct method, since different axiomatic choices on what patterns belong to the same wait category will lead to different outcomes. In order to avoid having too many categories, we categorize edge cases (where a theoretical zero-tile or ten-tile would complete the hand) according to the base form, so a wait such as 12345 is classified as #10 (Sanmenchan). One slight ambiguity is that in very rare cases a tile pattern will simplify to two different existing categories of wait. We use the obvious solution and allocate to the simplest category available.

The first table also gives the percentage frequency in Riichi mahjong of each wait classification. These are derived from classifying all games played in 2018 in the Houou room of Tenhou. Tenhou is the primary website for online Riichi mahjong, and the Houou room is for the top-level players. The algorithm used to derive the wait classification for each hand is given in the next section. There were 209,720 games played with 1,735,985 agaris (hands that did not end in a draw).

The E in the required tiles represents the East tile: in this case it is assumed that the tile pattern also contains a pair of East tiles. In general, it can be any pair unrelated to the main pattern. Double pattern waits can also occur, where the second pattern is not a simple pair but is of the form 22234 or 23334. For double pattern waits we take the computationally convenient approach of classifying the wait according to the pattern that is completed

by the winning tile, for example a Double Entotsu wait, where the hand contains the pattern 22234 in two different suits, is always classified as Entotsu (#11).

Patterns are ordered by number of tiles in the wait pattern, then by number of winning tiles (N-way), then by number of outs. If all of these are the same, they are listed by numerical order of the wait pattern. If a pattern that begins with the 2 tile and ends with the 9 tile can by completed the "10 tile", then the example case is given as starting at 1 and ending in 8. This maximizes the number of practical winning tiles and keeps the ordering consistent. Reversed patterns are not listed.

2. METHODS

The algorithm to identify the wait pattern of a ready hand uses a look-up table that contains every non-separable mahjong wait pattern and its corresponding wait category. A non-separable pattern is one that does not contain any gaps of two or more tiles, since if there exists a gap of this size the pattern can be split into separate pieces. For example, 3335777 is a non-separable pattern but 1234789 is a separable pattern (which separates into 1234 and 789). There are 19,273 non-separable patterns so the table has 19,273 rows and two columns: the pattern and the wait category. The table contains reversed patterns to ensure that only one look-up is needed. The look-up table means that we only need to construct an algorithm to extract the relevant non-separable pattern from the waiting hand, which is far simpler than directly calculating the wait category. The construction of the look-up table was via brute force enumeration.

The basic method is that we first cover non-regular hands (this may vary based on the variant; here we use Riichi mahjong), then a winning honour tile, then a winning suit tile. If the winning honour tile forms a pair, we have #1 (Tanki), if not we search for the relevant non-separable pattern in the rest of hand; if it exists then we look up the wait in the table, if not then the wait is #3 (Shanpon). For a winning suit tile, we identify the non-separable pattern closest to the winning tile. If it is anything
other than a simple pair we look up the wait in the table, otherwise we search for the relevant non-separable pattern in the rest of hand, looking it up if it exists, or returning #3 (Shanpon) if it does not.

There are two concepts to define first:
Split: This subroutine separates a set of tiles into its non-separable pieces. There will either be a single piece (no separation), two pieces or three pieces. For example, 111456999 separates into three pieces 111 456 999 as there are two gaps of two tiles.
EqualToZ: This subroutine tests whether the number of tiles in a set is equal to any of 2, 5, 8 or 11. This is important for identifying relevant patterns. At the points where we test if any of the suits/pieces are EqualToZ, there are only two possible outcomes: none are EqualToZ, or exactly one is EqualToZ.

The Algorithm
1. Check for kokushi (a non-regular hand type). If so, Return #0.
2. Check for chiitoi (a non-regular hand type). If so, Return #1 (Tanki).
3. If winning tile is an honour tile and forms a pair, Return #1 (Tanki).
4. If winning tile is an honour tile and forms a pung, then test if any suits are EqualToZ. If there are none, Return #3 (Shanpon). If there is one, apply Split to this suit and retain only the piece that is EqualToZ. Look-up this piece and Return result.
5. If winning tile is a suit tile, apply Split to the suit and determine the piece closest to the winning tile. If this piece is NOT a pair, look-up this piece and Return result. Otherwise, go to next step.
6. Determine if any of the other two suits or if any of the other piece(s) of this suit (if any exist) are EqualToZ. If there are none, Return #3 (Shanpon). If there is one and it is another piece of the same suit, look-up this piece and Return result. If there is one and it is one of the other suits, apply Split to this suit and retain only
the piece that is EqualToZ. Look-up this piece and Return result.

3. RESULT TABLES

Each table gives the wait pattern, the tile(s) required to win, and the number of outs, which is defined as the number of tiles that are available to win, considering that there are four identical tiles of each type (i.e., the number of required tiles multiplied by four, subtracting any that are already in the hand).

BASIC TABLE

	Tiles	Require	Outs	%
1	2	2	3	5.73
2	23	14	8	43.7
3	22	E2	4	13.17
4	24	3	4	14.15
5	2223	134	11	1.08
6	2224	34	7	0.6
7	2234	25	6	2.11
8	2345	25	6	1.56
9	2334	3	2	0.07
10	23456	147	11	4.68
11	22234	E25	7	0.93
12	22334	14	7	3.52
13	23345	14	7	4.48
14	23334	E3	3	0.13
15	22344	3	3	0.68
16	2223456	13467	17	0.142
17	2223444	12345	13	0.011
18	2333456	1247	14	0.128
19	2223345	1346	13	0.074
20	2233334	1245	13	0.007
21	2333345	1245	13	0.011
22	2223334	2345	9	0.016
23	2223344	2345	9	0.041
24	2224666	345	11	0.004
25	2224456	347	10	0.046
26	2224567	347	10	0.082
27	2222334	134	9	0.008
28	2222344	134	9	0.004
29	2223445	346	9	0.068
30	2234567	258	9	0.327
31	2344456	147	9	0.046
32	2344567	147	9	0.29
33	2345678	258	9	0.231
34	2233344	234	5	0.012
35	2223457	67	7	0.037
36	2233445	25	5	0.083
37	2234456	25	5	0.133
38	2333445	36	5	0.024
39	2334456	36	5	0.078
40	2223455	25	3	0.01
41	22234567	E258	10	0.063
42	12345678	369	10	0.237
43	22334456	147	10	0.098
44	23344556	147	10	0.041
45	23344567	258	10	0.21
46	23444456	E17	10	0.002
47	23445678	369	10	0.251
48	22233344	E14	8	0.006
49	22233445	E25	6	0.02
50	22234456	E25	6	0.038

51	22334455	E25	6	0.026
52	22234555	E25	4	0.003
53	22223344	E5	6	0.002
54	22333445	14	6	0.017
55	22334445	36	6	0.018
56	23333445	E6	6	0.002
57	23334445	25	6	0.006
58	23344456	25	6	0.046
59	23445567	36	6	0.079
60	22333344	E	2	0.000

Table 1: Basic sixty wait patterns with percentage frequencies.

TEN/ELEVEN TILE WAITS

	Tiles	Require	Outs				
61	2223456777	12345678	22	97	2223333456	1467	14
62	1112345678	235689	19	98	2224567999	3478	14
63	2223456678	134679	19	99	2333344445	1256	14
64	2333344567	124578	19	100	1222345678	1369	13
65	2344445678	235689	19	101	2222344666	1345	13
66	2223444456	123567	18	102	1113345678	2369	13
67	2223444567	123458	16	103	1113456678	2369	13
68	2223334567	234578	15	104	1113456789	2369	13
69	2224566667	34578	17	105	2233445556	1467	13
70	1234445678	35689	16	106	2333455667	1247	13
71	2223345678	13469	16	107	2333456678	1247	13
72	2233334567	12458	16	108	2222344456	1347	12
73	2333345567	12458	16	109	2222344567	1347	12
74	2333345678	12458	16	110	2223345567	1346	12
75	2344555678	13469	16	111	2223445678	3469	12
76	2222334456	13467	15	112	2233444456	1237	12
77	2223344556	13467	15	113	2234444567	2358	12
78	2223445566	13467	15	114	2333444456	1237	12
79	2223455667	13467	15	115	2334444567	2358	12
80	2233334456	12457	15	116	2222333445	1346	11
81	2333344456	12457	15	117	2222334445	1346	11
82	2333344556	12457	15	118	2223444556	3467	11
83	2344445567	23568	15	119	2223445567	3467	11
84	2223344445	12356	14	120	2223456677	5678	11
85	2223334455	13456	13	121	2233344445	1236	11
86	2224445566	34567	13	122	2222333444	1345	10
87	2223344567	23458	12	123	2223333444	1245	10
88	2223456667	25678	12	124	2223333445	2456	10
89	2233444567	12347	12	125	2223345666	1346	10
90	2223334456	23457	11	126	2223444455	2356	10
91	2223344456	23457	11	127	2223466678	2569	10
92	2233344456	23467	11	128	2344455567	1458	10
93	2223334445	23456	10	129	2344466678	1469	10
94	2223334555	23456	10	130	2223344455	3456	9
95	2223445666	23456	10	131	2233445566	2356	8
96	2222345666	1457	14	132	2223344555	2345	6
				133	2223457999	678	11

134	2222333344	145	10
135	2222333345	145	10
136	2222334444	135	10
137	2222344445	356	10
138	1112346678	569	10
139	1112346789	569	10
140	2223456778	679	9
141	2224455667	347	9
142	2224456678	347	9
143	2224556677	347	9
144	2224566778	347	9
145	2233344566	145	9
146	1123456789	147	8
147	1233345678	369	8
148	1233456678	369	8
149	1233456789	369	8
150	1234456789	147	8
151	2223455677	467	8
152	2233334455	256	8
153	2233445567	258	8
154	2233445677	147	8
155	2233445678	258	8
156	2234455667	258	8
157	2234566778	258	8
158	2234567789	258	8
159	2333445678	369	8
160	2334445567	147	8
161	2334455567	258	8
162	2334455678	258	8
163	2334456678	369	8
164	2334456789	369	8
165	2344456678	147	8
166	2344556678	258	8
167	2344566789	369	8
168	2223455777	256	7
169	2233334445	246	7
170	2223333455	245	6
171	2223444566	456	6
172	2223455567	258	6
173	2223455678	258	6
174	2223456788	258	6
175	2223467778	257	6
176	2223467888	258	6
177	2333455567	358	6
178	2333466678	369	6
179	2224444666	35	8
180	2334444556	17	8
181	1112345679	89	7
182	2224444567	37	7
183	2233344457	67	7
184	2234555567	28	6
185	2345555678	28	6
186	2233444556	25	4
187	2233444566	36	4
188	2333444556	36	4
189	2333445567	36	4
190	2334455667	36	4
191	2223445556	25	2
192	2333445556	35	2
193	2333456667	36	2
194	2333467778	37	2
195	2222334455	5	2
196	2234455556	2	2
197	22233344456	E147	11
198	22333444567	E258	11
199	11123456789	E147	9
200	22233445567	E258	9
201	22234455667	E258	9
202	22234556677	E258	9
203	22234566778	E258	9
204	22234567789	E258	9
205	22334455567	E258	9
206	22234567888	E258	7
207	11223345678	369	9
208	12233445678	369	9
209	12233456789	147	9
210	12333345678	E69	9
211	12334455678	369	9
212	22223344567	E58	9
213	22333444556	147	9
214	22334445566	147	9
215	22334445678	369	9
216	22334455667	147	9
217	22334456678	147	9
218	23333445678	E69	9
219	23334445556	147	9
220	23334445567	258	9
221	23344455667	258	9
222	23344455678	369	9
223	23344555567	E28	9
224	23344556678	147	9
225	23344566778	258	9
226	23344567789	258	9
227	23444456678	E17	9
228	23444556678	369	9
229	12334567789	258	9
230	22233334445	E14	7
231	22234555567	E28	7
232	22333344445	E25	7
233	22333444456	E25	7
234	22233344455	E25	5
235	22233444556	E25	5
236	22234445566	E25	5
237	22334445556	E25	5
238	22223334445	E5	5
239	22223344456	E5	5
240	22333344455	14	5
241	22334444556	36	5

242	22334445567	36	5	246	23344455567	36	5
243	23333444556	E6	5	247	22223344555	E5	3
244	23333445567	E6	5	248	22234455556	E2	3
245	23334444556	25	5				

Table 2: Ten/eleven tile wait patterns.

FOURTEEN TILE WAITS

	Tiles	Require	Outs				
249	1112345678999	123456789	23	292	2222334456678	134679	17
250	1112345666678	12345789	23	293	2223344445678	123569	17
251	2223456677778	12345689	23	294	2223344556678	134679	17
252	2223334567888	23456789	19	295	2223445566678	134679	17
253	2223344556777	12345678	19	296	2223455666778	134679	17
254	2333345677778	1245689	23	297	2233334455667	124578	17
255	2344445666678	1235789	23	298	2233445566667	124578	17
256	1112333345678	1245689	20	299	2233445666678	124578	17
257	2223444456678	1235679	20	300	2333344455667	124578	17
258	2222334456777	1345678	19	301	2333344555667	124578	17
259	2223333456777	1245678	19	302	2333344555678	124569	17
260	2223444455667	1235678	19	303	2333344556678	124578	17
261	2223444456777	1235678	19	304	2333344566778	124578	17
262	2223456667788	1346789	19	305	2334455556678	134679	17
263	2233444566667	1234578	19	306	2334455666678	124578	17
264	1112345666789	1234567	17	307	2334456777789	235689	17
265	1112223456789	1234679	17	308	2344455556678	134679	17
266	1112223456678	1234679	17	309	2223334444567	123578	16
267	2223444556677	1234578	17	310	2223334455678	134569	16
268	1122233345678	1235689	17	311	2223344445567	123568	16
269	2344455566678	1245689	17	312	2223444455678	235689	16
270	2223334445678	2345689	16	313	1112344567888	134679	16
271	2223334445666	1234567	15	314	1113334455678	234569	16
272	2223344556677	2345678	15	315	1113456667788	236789	16
273	2223455566677	1234567	15	316	2233444555678	123469	16
274	2333344445678	125689	20	317	2223333445567	245678	15
275	1113455556789	234679	19	318	2223334455567	134568	15
276	2222344566667	134578	19	319	2223334555567	234678	15
277	1113455556678	234679	19	320	2223444566778	123458	15
278	2333344555567	124678	19	321	2223444567789	123458	15
279	1222233456789	134679	17	322	2223445566667	234578	15
280	1111223345678	235689	17	323	2224445566678	345679	15
281	1112233445678	235689	17	324	2233344455556	123467	15
282	1112334455678	235689	17	325	2233444555567	123467	15
283	1112344556678	235689	17	326	1112345556789	145679	14
284	1112345566778	235689	17	327	2223334445556	134567	14
285	1112345667788	235689	17	328	1112223345678	123469	14
286	1112345677889	235689	17	329	2223344455567	345678	14
287	1222233456678	134679	17	330	1112233345678	123469	14
288	1223344445678	235689	17	331	1112345556678	145679	14
289	1233334555678	124569	17	332	2224445556667	345678	14
290	1233334567789	124578	17	333	2344455666789	134569	14
291	1233444455678	235689	17	334	2223334455667	234578	13
				335	2223334555678	234569	13

336	2223334556677	234578	13	390	1223345666789	12457	14
337	2223334566778	234578	13	391	1233334445678	45689	14
338	2223344455667	234578	13	392	1233334456789	12457	14
339	2223444567888	123458	13	393	1233344445678	23569	14
340	2223445678999	234569	13	394	1233444456678	23569	14
341	2233344456678	234679	13	395	1233444456789	23569	14
342	2233445556667	235678	13	396	1233445555678	13469	14
343	2333444555678	123458	13	397	2222333445678	13469	14
344	2223334445567	234568	12	398	2222334445678	13469	14
345	2223334456777	123457	12	399	1112333445678	23569	14
346	2223334445566	234567	11	400	1112334456678	23569	14
347	1111234567888	34679	16	401	1112334456789	23569	14
348	1112222345678	35689	16	402	2233334456678	12457	14
349	1112345555678	23689	16	403	2233344445678	12369	14
350	1113456788889	23679	16	404	2333344455678	14569	14
351	2222333344567	14578	16	405	2333344456678	12457	14
352	2222344445678	35689	16	406	2333445555678	13469	14
353	2222344567999	13478	16	407	2334445555678	13469	14
354	2222345566667	14578	16	408	2344445556678	15679	14
355	2222345666678	14578	16	409	2344445566678	15679	14
356	2223333456678	14679	16	410	2344445567789	23568	14
357	1113345555678	23469	16	411	2222333444556	13467	13
358	2333344445567	12568	16	412	2222333444567	13458	13
359	2344445555678	23689	16	413	2222334445566	13467	13
360	1122334445678	35689	15	414	2222334445567	13467	13
361	1223345556789	14679	15	415	2222334455667	13467	13
362	2222344445566	13567	15	416	2223333444567	12458	13
363	2222344445666	13567	15	417	2223333445678	24569	13
364	2222344456666	13457	15	418	2223334455566	14567	13
365	2223333444456	12567	15	419	2223344455566	13467	13
366	2223345566778	13469	15	420	2223344555667	13467	13
367	2223345667788	13469	15	421	2223345678999	13469	13
368	2223345677889	13469	15	422	2223444555567	12348	13
369	2223444455556	12367	15	423	2223444556678	34679	13
370	2223445567999	34678	15	424	2223445556667	13467	13
371	2233334556677	12458	15	425	2223445566778	34679	13
372	2233334566778	12458	15	426	2223445666678	23459	13
373	2233334567789	12458	15	427	1112344555678	13469	13
374	2233444455556	12367	15	428	2223455677778	25689	13
375	2233445556678	14679	15	429	2223456666778	25789	13
376	2233445677778	25689	15	430	1112345566678	45679	13
377	2333344455556	12467	15	431	2223456777788	25689	13
378	2333345566778	12458	15	432	1112346667788	56789	13
379	2333345567789	12458	15	433	2233334445566	12457	13
380	2333345677889	12458	15	434	2233344445567	12367	13
381	1223345556678	14679	15	435	2233344456667	25678	13
382	1233455556789	34679	14	436	2333344455567	14568	13
383	1234455556789	13467	14	437	2333444455567	13578	13
384	1112345567789	23568	14	438	2333444555567	12348	13
385	1122223345678	13469	14	439	2222345666777	14567	12
386	1122334555678	13469	14	440	2223344455678	34569	12
387	1222233345678	13469	14	441	2223444455567	23568	12
388	1222233445678	13469	14	442	2223445566677	35678	12
389	1223344555678	13469	14	443	2223455566667	24578	12

444	2223455666677	24578	12	498	1222345677889	1369	12
445	2223455666678	24578	12	499	2222333344456	1457	12
446	2223456667778	56789	12	500	2222333344556	1457	12
447	1112333456789	12347	11	501	2222333444456	1357	12
448	1234445666789	34567	11	502	2222334455666	1457	12
449	1112233456789	12347	11	503	2222334456666	1347	12
450	1112345678889	14789	11	504	2222344445567	3568	12
451	1122333456678	12369	11	505	2223333445566	1467	12
452	1122333456789	12369	11	506	2223333455667	1467	12
453	1234445566789	14567	11	507	2223455677999	4678	12
454	2222334445666	13456	11	508	1113344556678	2369	12
455	2223333444555	12456	11	509	1113344556789	2369	12
456	2223333444556	24567	11	510	1113345566778	2369	12
457	2223334444556	23567	11	511	1113345667788	2369	12
458	2223334455556	23467	11	512	1113345677889	2369	12
459	2223344445566	23567	11	513	1113445566678	2369	12
460	2223344445666	12356	11	514	1113445566789	2369	12
461	2223344556667	25678	11	515	1113455666778	2369	12
462	2223344566778	23458	11	516	1113455667789	2369	12
463	2223344567789	23458	11	517	1113456677889	2369	12
464	2223444455566	23567	11	518	1113456778899	2369	12
465	2233444556677	12347	11	519	2233334455556	1247	12
466	2233444566778	12347	11	520	2233344556667	1458	12
467	2233445566678	23569	11	521	2233444567778	3689	12
468	2223334455666	13456	10	522	1123344455678	2369	12
469	2223334456678	23457	10	523	2333444556667	2578	12
470	2223344456678	23457	10	524	2333445566678	1247	12
471	2223344555777	23456	10	525	2333455666778	1247	12
472	2223344555678	23458	9	526	1122223456789	1347	11
473	2223344567888	23458	9	527	1222234456789	1347	11
474	2233344455667	23467	9	528	1222234567789	1347	11
475	2223344456777	23457	8	529	1111233345678	2369	11
476	2223334555666	23456	7	530	1111233456678	2369	11
477	2223344455666	23456	7	531	1111233456789	2369	11
478	2222333345777	1456	14	532	1112234567789	2358	11
479	2224444566667	3578	14	533	1122333345678	1269	11
480	1111234555678	3469	13	534	1123333456789	1247	11
481	1112346788889	5679	13	535	1222333345678	1269	11
482	1234444555678	1369	13	536	1222345666678	1239	11
483	2222333345567	1458	13	537	1223333456789	1247	11
484	2222333345678	1458	13	538	1233334455678	4569	11
485	2222334444556	1367	13	539	1233345555678	3469	11
486	2222334444567	1358	13	540	1233444567789	2358	11
487	2222344678999	1345	13	541	1233455666789	2457	11
488	2223345666678	1349	13	542	1234444566789	3569	11
489	2224455667999	3478	13	543	2222333344555	1456	11
490	2233334444567	1258	13	544	2222333444455	1356	11
491	2233334555567	1248	13	545	2222334455677	1467	11
492	2333345555678	1248	13	546	2222334456667	5678	11
493	2334444555567	1238	13	547	2222344455667	1347	11
494	1122334567778	3689	12	548	2222344456678	1347	11
495	1222334455678	1369	12	549	2222344556677	1347	11
496	1222344556678	1369	12	550	2222344566778	1347	11
497	1222345566778	1369	12	551	2223334466678	1469	11

552	2223334467888	1458	11	606	2223466677889	2569	9
553	2223345556677	1346	11	607	1112355567789	1458	9
554	2223445667788	3469	11	608	2223466778899	2569	9
555	2223445677889	3469	11	609	2233344445566	2367	9
556	1112334567789	2358	11	610	1122334466678	1469	9
557	1112344566678	3569	11	611	2334455566678	2569	9
558	1112344566789	3569	11	612	1223344466678	1469	9
559	1112345666778	5689	11	613	2344455567789	1458	9
560	1112345667789	5689	11	614	2222333445666	1346	8
561	2233334455678	2569	11	615	2222344455566	1456	8
562	2233344456677	5678	11	616	2223333445556	2457	8
563	2233344466678	2569	11	617	2223344455556	2347	8
564	2233444455567	1258	11	618	2223344455677	2567	8
565	2233444455667	1237	11	619	2223444556667	4568	8
566	2233444456678	1237	11	620	2223444556777	3467	8
567	2233445555667	1467	11	621	2222333444555	1345	7
568	2233445555677	1467	11	622	2223333445666	2456	7
569	2234444566778	2358	11	623	2223455566678	2569	7
570	2234444567789	2358	11	624	1112344466678	1469	7
571	2333344567778	6789	11	625	2223455578999	2569	7
572	2333444455667	1237	11	626	2233344556677	2347	7
573	2333444456678	1237	11	627	2233344455566	2356	6
574	2334444566778	2358	11	628	2222333344445	156	11
575	2334444567789	2358	11	629	2224444567999	378	11
576	1112345566789	4567	10	630	2224677788899	345	11
577	2222333445567	1346	10	631	1113333456678	269	10
578	2223333444566	1456	10	632	1113333456789	269	10
579	2223333455777	2456	10	633	1113345666678	239	10
580	2223344445556	3567	10	634	1113456666789	239	10
581	2223455667778	4679	10	635	1122233346678	569	10
582	1112344567999	1478	10	636	1122233346789	569	10
583	1112345677999	1478	10	637	2222334444566	156	9
584	1112355566678	1469	10	638	2222334466678	569	9
585	1112355578999	1469	10	639	2223345566667	134	9
586	2233334445678	2469	10	640	1112346677889	569	9
587	2233334455567	2568	10	641	1112346778899	569	9
588	2233344455677	1467	10	642	2223466667788	259	9
589	2222333456777	1347	9	643	1112356666778	149	9
590	2222344456777	1347	9	644	2233334444556	267	9
591	2223333455567	2458	9	645	2233344456778	679	9
592	2223333455678	2458	9	646	2233444466678	169	9
593	2223344555567	2348	9	647	2234444555567	238	9
594	2223344566678	2569	9	648	1222233466678	569	9
595	1112233466678	1469	9	649	2334444566678	169	9
596	2223344578999	2569	9	650	2222344556777	356	8
597	2223444555667	3467	9	651	1112345677899	689	8
598	2223444566678	4569	9	652	2224455566677	347	8
599	2223445556677	3467	9	653	2224455666778	347	8
600	1112334566678	1469	9	654	2224456667788	347	8
601	1112334578999	1469	9	655	2224556667778	347	8
602	2223455556667	2678	9	656	2233334455566	467	8
603	2223455556677	2678	9	657	2234455556667	347	8
604	2223456666788	2578	9	658	2234455566678	347	8
605	2223456677788	6789	9	659	1112345678899	789	7

660	1122334456789	147	7	714	2334445566678	369	7
661	1122334566678	369	7	715	2334445566778	147	7
662	1122334566789	369	7	716	2334445566789	369	7
663	1122334567789	147	7	717	2334455566778	258	7
664	1122334567899	369	7	718	2334455567789	258	7
665	1123344556789	147	7	719	2334455667789	369	7
666	1123445566789	147	7	720	2334455677889	258	7
667	1123455667789	147	7	721	1223345567789	258	7
668	1123456677889	147	7	722	1223345677889	258	7
669	1223334456678	369	7	723	2344555667789	258	7
670	1223334456789	369	7	724	1233455567789	258	7
671	1223344456789	147	7	725	1112355577789	147	6
672	1223344566678	369	7	726	1233344466678	369	6
673	1223344566789	369	7	727	2223456777899	789	6
674	1223344567789	147	7	728	2223466678889	258	6
675	1223345677789	147	7	729	1112355578889	148	6
676	1233344556678	369	7	730	1112356778889	148	6
677	1233344556789	369	7	731	2233334445556	256	6
678	1233444556789	147	7	732	2233334445567	246	6
679	1233445566789	369	7	733	1222334566678	269	6
680	1233445567789	147	7	734	1222344466678	269	6
681	2222334444555	135	7	735	1222355566678	269	6
682	2223333444455	256	7	736	1233344456789	347	5
683	2223334445677	147	7	737	1112334577789	147	5
684	2223334456667	146	7	738	1112344456789	147	5
685	2223334467778	147	7	739	1112344567789	147	5
686	2223344555677	467	7	740	1112345677789	147	5
687	1112344678999	145	7	741	1222344456789	247	5
688	2233334455677	267	7	742	1233355567789	358	5
689	2233344455678	258	7	743	2223344555566	236	5
690	2233344456788	258	7	744	2223344556788	258	5
691	2233344467778	257	7	745	2223344567778	257	5
692	2233444555667	258	7	746	2223344578889	258	5
693	2233444556678	258	7	747	2223445556678	258	5
694	2233444566678	369	7	748	2223445566788	258	5
695	2233444566789	369	7	749	2223445667778	257	5
696	2233444567899	369	7	750	2223445667888	258	5
697	2233445566778	258	7	751	2223445678889	258	5
698	2233445567789	258	7	752	1112334578889	148	5
699	2233445667778	147	7	753	2223455566778	258	5
700	2233445667788	258	7	754	2223455567789	258	5
701	2233445677889	258	7	755	2223455667788	258	5
702	2234445556667	258	7	756	2223455677889	258	5
703	2234455666778	258	7	757	2223456778889	258	5
704	2234455667789	258	7	758	2233445556788	258	5
705	2234556667778	258	7	759	2233445567778	257	5
706	2234556677789	258	7	760	2233445578889	258	5
707	2234566777889	258	7	761	2333455566778	358	5
708	2333444556678	369	7	762	2333455567789	358	5
709	2333444556789	369	7	763	2333466677889	369	5
710	2333445566778	369	7	764	1222355567789	258	5
711	2333445677889	369	7	765	2223444555566	246	4
712	1222334567789	258	7	766	2223455567778	257	3
713	2334445556678	258	7	767	2223455567888	258	3

768	2223455578889	258	3	799	2233444467778	17	5
769	2223457777999	68	8	800	2233445555678	28	5
770	1223333445678	69	7	801	2234555566778	28	5
771	1112346666789	59	7	802	2234555567789	28	5
772	1122233345679	89	7	803	2234567788889	25	5
773	2334444555667	17	7	804	2333344566678	69	5
774	2334444556678	17	7	805	2333344566789	69	5
775	2222334455777	56	6	806	2333344578889	68	5
776	2224444556677	37	6	807	2333445555667	38	5
777	2224444566778	37	6	808	2333445666678	39	5
778	2224456677778	34	6	809	2333456666778	39	5
779	1233334577789	67	5	810	1222356666778	29	5
780	1112356777789	14	5	811	2334456666789	39	5
781	1123444456789	17	5	812	2223455556788	28	3
782	1123456777789	14	5	813	2233444556667	36	3
783	1222356777789	24	5	814	2333444555667	36	3
784	1223333466678	69	5	815	1233345677789	37	2
785	1233334566678	69	5	816	1233355577789	37	2
786	1233334566789	69	5	817	1222334577789	27	2
787	1233456666789	39	5	818	1222345677789	27	2
788	2222334455567	58	5	819	1222355577789	27	2
789	2222334455678	58	5	820	2333445678889	38	2
790	2222334456788	58	5	821	1222334578889	28	2
791	2222334467778	57	5	822	2333455578889	38	2
792	2222334467888	58	5	823	1222345678889	28	2
793	2223444455677	67	5	824	1222355578889	28	2
794	2223445555667	28	5	825	2222334445556	5	1
795	2223466777788	25	5	826	2233334456667	6	1
796	2223467788889	25	5	827	2233334467778	7	1
797	2233334466678	69	5	828	2333344556667	6	1
798	2233444456667	16	5				

Table 3: Fourteen tile wait patterns.

Acknowledgements

We wish to thank members of The Australian Riichi Mahjong Association (ARMA) for their assistance.

References

Cheng, S. T., Chan, A. C., & Yu, E. C. (2006). *An exploratory study of the effect of mahjong on the cognitive functioning of persons with dementia*. International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences, 21(7), 611-617.

Cheng, Y., Li, C. K., & Li, S. H. (2017). Mathematical aspect of the combinatorial game "Mahjong". *arXiv:1707.07345.*

Chiba, D. (2016). *Riichi Book I. A mahjong strategy primer for European players.* https://dainachiba.github.io/RiichiBooks/

Li, J.,Koyamada, S., Ye, Q., Liu, G., Wang, C., Yang, R., Zhao, L., Qin, T., Liu, T. Y., & Hon, H. S. (2020) Suphx: Mastering Mahjong with Deep Reinforcement Learning. *arXiv:2003.13590*

Lo, A. (2001). *The complete book of mahjong. An illustrated guide to the Asian, American and International styles of play*. Tuttle.

Miller, S. D. (2015). *Riichi mahjong. The ultimate guide to the Japanese game taking the world by storm*. Psionic Press.

Mizukami N, Tsuruoka Y. *Building a computer mahjong player based on monte carlo simulation and opponent models*. In 2015 IEEE Conference on Computational Intelligence and Games (CIG) Aug 2015, 275-283.

DETECTING CHANGE IN TONE OF TWEETS AND THE POTENTIAL IMPLICATIONS FOR MONITORING MENTAL HEALTH OF ATHLETES

Paul J. Bracewell [a,b] , Jason D. Wells [a,b] , Andy Craig [a,c]
[a] *Talennial, Australia*
[b] *DOT loves data, New Zealand*
[c] *Corresponding author: andy@talennial.com*

Abstract

Twitter is an online news and social networking site where people communicate in short messages called tweets. This platform and style of communication provides an opportunity for people and organisations to connect with followers world-wide. Twitter provides an API to enable programmatic access to tweets, enabling techniques like natural language processing (NLP) to scalably process this massive collection of data. There is no shortage of tutorials and guides on how to use NLP for detecting sentiment, emotions, or topics from machine readable test, like tweets. Schweinberger (2021) provides a simple to use guide for users of the R package. Bracewell (2022) extended the framework outlined in Schweinberger's tutorial to include brand attributes, like "Being Kiwi". Here, we introduce a proprietary "Aussie Index", an Australian equivalent of the "Being Kiwi" metric. In addition, principal components are used on 10 years of mainstream media to collapse the eight core emotions: Joy, Surprise, Trust, Anticipation, Sadness, Disgust, Anger and Fear to create two further proprietary metrics which summarise the "Level of Emotion" and the contrast between "Light versus Dark" emotions.

The tweets of a high-profile Australian sportsperson are explored. The emotions, tone and volume of tweets are assessed over a ten-year period. Prior to their publicised break from competing due to mental health reasons, there is a change in tone of messaging. A drop in the "Aussie Index", a change in the "Level of Emotion" and increasing "Darkness" in emotions are observed.

These results do not imply that the detected tone and emotions in tweets predict mental health issues. Instead, this case study suggests that this is a topic worthy of further investigation. Further research is required to determine if it is possible to create a monitoring system, akin to Statistical Process Control, for tracking changes in communication styles which may suggest potential issues.

Keywords: Natural Language Processing, NLP, Sentiment, Emotion Detection

1. INTRODUCTION

Natural language processing is a prevalent technique for scalably processing massive collections of documents. This branch of computer science is concerned with creating abstractions of text that summarize collections of documents in the same way humans can. This form of standardization means these summaries can be used operationally in machine learning models to describe or predict behaviour in real or near real time, as required. Bracewell et. al. (2022) outlined several approaches where sentiment had been used by DOT loves data (DOT), a Wellington-based data science firm. He also explained limitations with their existing approach, particularly regarding the overly simplistic nature of using just sentiment and volume of articles to summarise topics.

Bracewell et. al. (2019) outlined an approach which explored the relationship between on-field performance and mainstream media perception of athletes. An athlete's playing reputation was derived from a string of on-field performances. This is essentially an estimation of their ability as described by Bracewell (2003). When matches are previewed, this playing reputation informs the number of articles featuring an athlete and the associated sentiment. That is, players perceived to have better ability are talked about more often and more positively. Performances within a match appear to influence the media post-match review. That is, athletes who performed well in a game are more likely to be mentioned and talked about favourably. This work illustrates the potential to combine reputation risk management with both sports ratings and natural language processing of mainstream media. Such an approach will enable delivery of a scalable solution for professional athletes and their associates to understand the impact of their on-field and off-field behaviour on their personal brand. This would aid strategic decisions around the type of content to develop, the best timeline to deploy certain content and a measurement tool to assess the impact of that content. Moreover, such a tool could provide the ability to identify mental health risks. For example, the resilience of a player to public scrutiny could help understand which players need more support.

The concluding remarks by Bracewell et. al. (2019) about exploring the possibility of identifying mental health risks are of great interest. This forms the basis of this research. However, this is not a topic to be taken lightly. We do not seek to predict mental health issues. Instead, this case study suggests that this is a topic worthy of further investigation. Further research is required to determine if it is possible to create a monitoring

system, akin to Statistical Process Control, for tracking changes in communication styles which may suggest potential issues.

2. DATA

Twitter is an online news and social networking site where people communicate in short messages called tweets (https://twitter.com/). This platform and style of communication provides an opportunity for people and organisations to connect with followers world-wide. The Twitter API enables programmatic access to Twitter in unique and advanced ways (https://developer.twitter.com/en/docs/twitter-api). Twitter go further to provide case studies and recommendations regarding how the different endpoints and features available on the Twitter API can be used (https://developer.twitter.com/en/docs/twitter-api/what-to-build). Topics outlined for consideration by Twitter include: moderate conversations for health and safety, enable creation and personal expression, measure and analyse "what's happening", improve community experiences, curate and recommend content and impact the greater good.

3. NATURAL LANGUAGE PROCESSING

Bracewell (2022) extended the framework outlined in Schweinberger's (2021) tutorial to include brand attributes, like "Being Kiwi", and demonstrated how this peaked during Olympic games and appeared to ebb and flow within changes in New Zealand's response between 2020 and 2022 to the COVID-19 pandemic.

Here, a proprietary "Aussie Index" is introduced which is an Australian equivalent of the "Being Kiwi" metric. The same approach outlined by Bracewell et. al. (2022) is used where the R scripts provided by Schweinberger (2021) are modified to customise the word lists for more specific applications. Simply, within that script is a carefully compiled dataset, "nrc" which appears in this line:

```
dplyr::inner_join(get_sentiments("nrc"))        (1)
```

This data set was replaced by a proprietary dataset suitable for use within the Australian context. Importantly, the proprietary dataset is highly configurable. To build out this data set, synonyms and slang used in distinctly Australian settings were researched, curated and collated for use. These were developed in conjunction with Australian-based data led brand and collaborator marketing company, Talennial (www.talennial.com).

In addition, principal components are used on 10 years of mainstream media (1st January 2009 to 31st December 2018) to collapse the eight core emotions: Joy, Surprise, Trust, Anticipation, Sadness, Disgust, Anger and Fear to create two further proprietary metrics which summarise the "Level of Emotion" and the contrast between "Light versus Dark" emotions. These two principal components explain over 60% of the variation. In addition, the both eigenvalues exceed 1, and are the only two dimensions to do so.

4. SUBJECT

Importantly, mental heath issues are a fraught topic, and this research is not undertaken without serious consideration of the moral and ethical implications. We are dealing with personal and private information. However, the subject of this research has spoken openly in the media about this topic. Furthermore, we believe we are undertaking this research for the greater good. We are only using data that is publicly available.

Australian, Glenn Maxwell, is a talented batting Allrounder who has represented his country at international level in all three formats of cricket (see: https://www.espncricinfo.com/player/glenn-maxwell-325026). Middleton (2019, October 31) announced Glenn Maxwell had withdrawn from Australia's T20 International squad and would take a "short" break from cricket to deal with mental health issues. Later, on March 25th, 2020 the Sydney Morning Herald (www.smh.com.au) published an article online titled: "*Glenn Maxwell details his mental health demons*". Below is the opening few lines of that article:

> "*A mentally exhausted Glenn Maxwell wanted his arm to be broken during Australia's World Cup campaign last year so he could have a break from international cricket. He didn't realise it at the time but the star all-rounder was battling mild depression and anxiety. Falling into a dark place would eventually lead to Maxwell stepping away from the game last October for more than a month.*"

These two articles provide sufficient information to begin compiling data. Maxwell is verified on Twitter as "gmaxi_32". This provides sufficient information to extract his tweets using the Twitter API.

4. METHOD

Glenn Maxwell tweets, from his account, g maxi_32, were extracted between March 2011 and September 2021. These were then pass through the NLP algorithm described previously which appended to every tweet: sentiment polarity; eight emotions: Joy, Surprise, Trust, Anticipation, Sadness, Disgust, Anger and Fear; "Level of Emotion", "Light versus Dark" and the "Aussie Index". These attributes were averaged per day. In addition, a 28 day rolling average was applied. These metrics were plotted and step changes were assessed. For the emotion and the Aussie Index, these are rated per 100 content words. Content words include nouns, verbs, adjectives and adverbs. The two principal component derived metrics "Level of Emotion" and "Light versus Dark" are from the mainstream media population of articles, with population has a mean of 0 and standard deviation of 1.

5. RESULTS

Annotated graphs displaying the metrics per day (dots) overlaid with the 28-day rolling average (solid line) follow. The vertical line indicates 31 October 2019 which is the day of the Cricket Australia announcement.

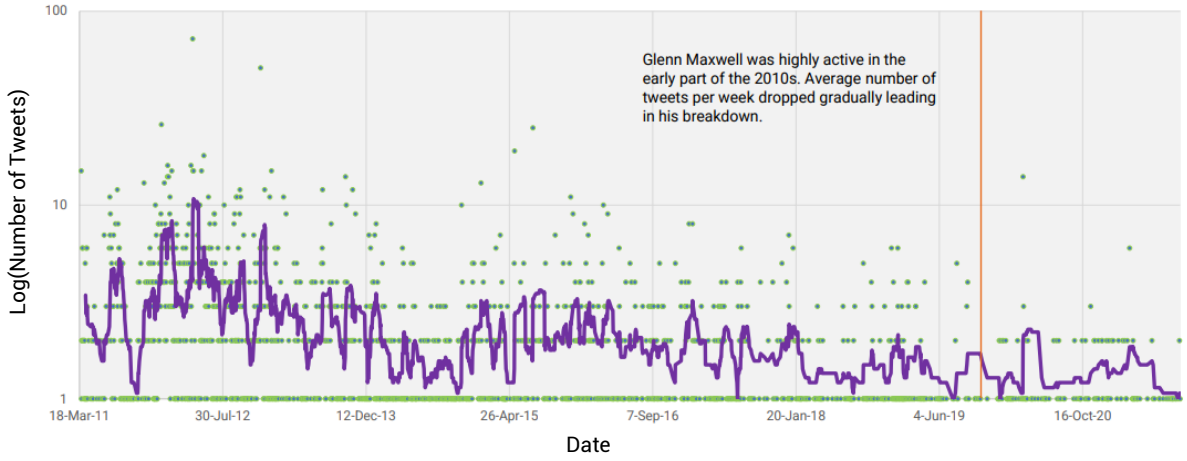


Figure 1: Log of the Number of Tweets per day between March 2011 and September 2021 by Glenn Maxwell

Figure 1 shows Glenn Maxwell was highly active in the early part of the 2010s. However, the average number of tweets per week dropped gradually leading to his break from the game in October 2020. This provides important context for reviewing the graphs that follow, particularly the continued activity.

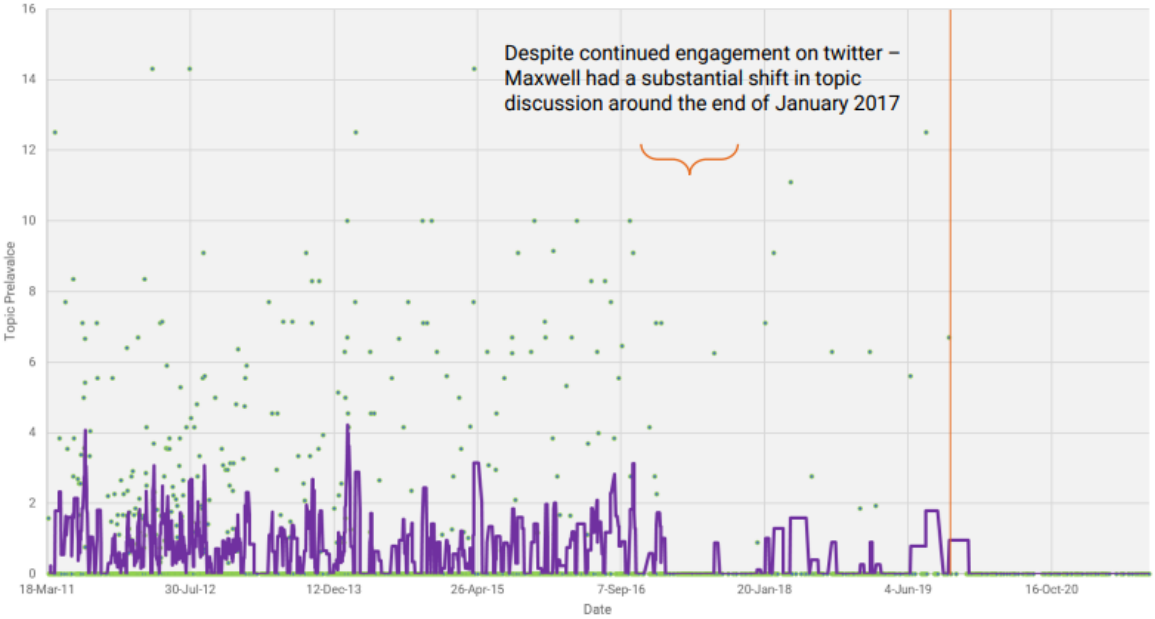


Figure 2: Time series plot showing the prevalence of the "Aussie Index" per 100 Content Words between March 2011 and September 2021 for Glenn Maxwell from his public tweets.

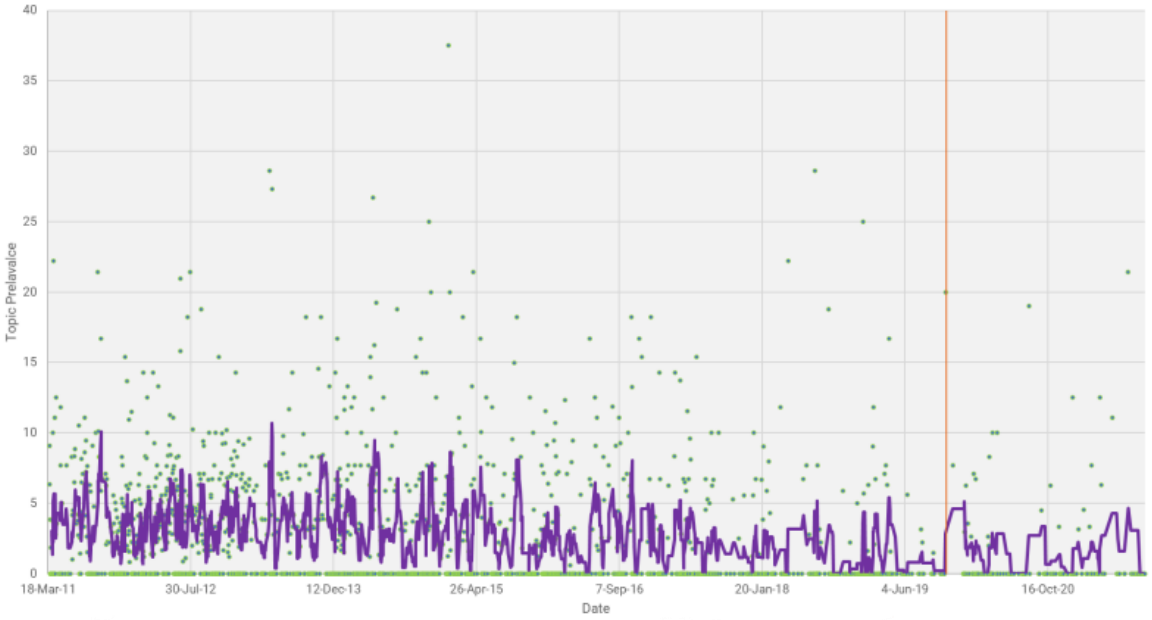


Figure 3: Time series plot showing the prevalence of the emotion, Trust, per 100 Content Words between March 2011 and September 2021 for Glenn Maxwell from his public tweets.

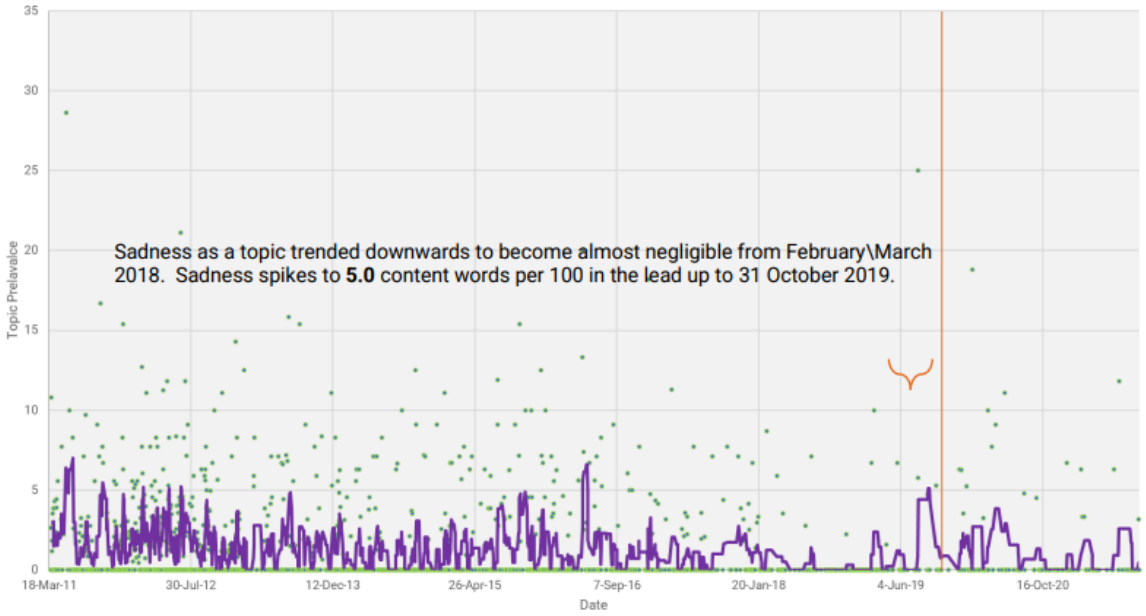


Figure 4: Time series plot showing the prevalence of the emotion, Sadness, per 100 Content Words between March 2011 and September 2021 for Glenn Maxwell from his public tweets.

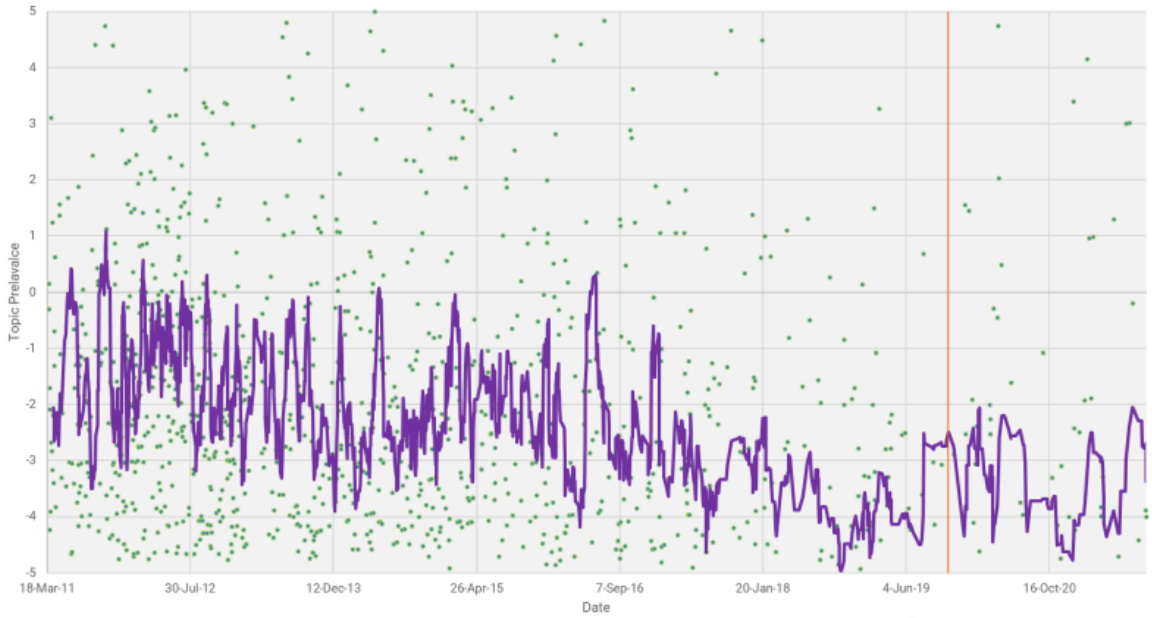


Figure 5: Time series plot showing the "Level of Emotion" in Glenn Maxwell's public tweets between March 2011 and September 2021

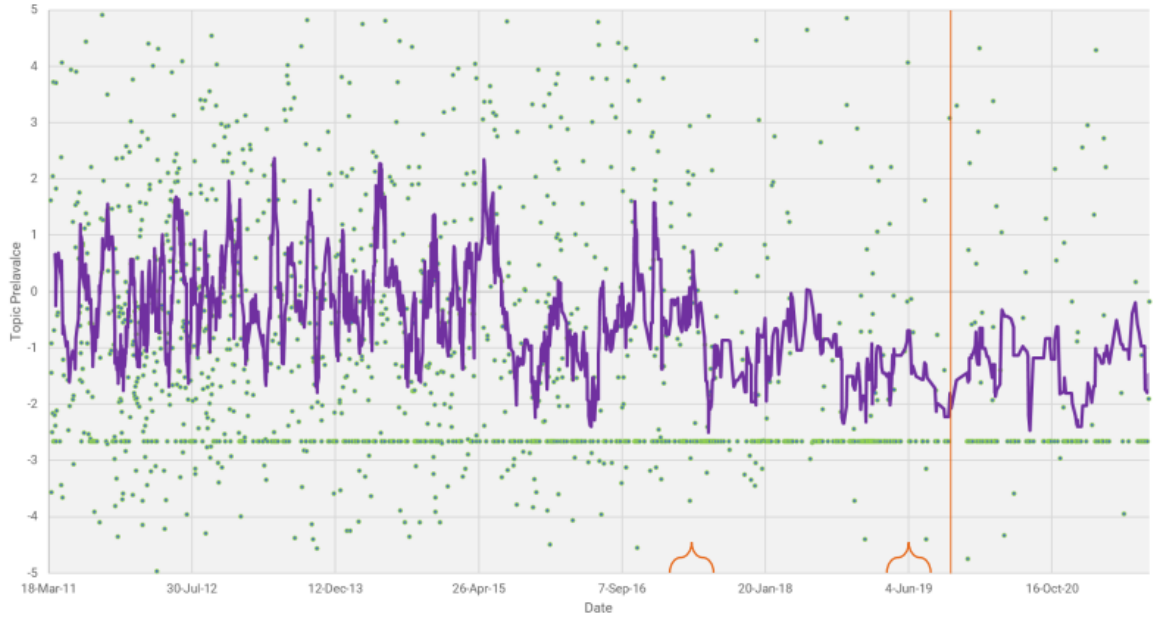


Figure 6: Time series plot showing the "Tone of Emotion" in Glenn Maxwell's public tweets between March 2011 and September 2021

5. DISCUSSION

The graphs (Figures 2-6) show a change in the topics, type of emotion and tone of tweets over a two year period of continued activity on Twitter by Glenn Maxwell (Figure 1) from his verified gmaxi_32 account. Importantly, this does not mean that the output can be used to predict looming mental health issues but serves as an indication this type of process could be worthy of further exploration.

Figure 2 shows a distinct change in the "Aussie Index". Did Glenn Maxwell lose his Authentic Self? Despite continued engagement on twitter, Maxwell had a substantial shift in topic discussion around the end of January 2017.

Trust as an emotion dropped considerably post this shift in topic discussion in late January 2017 from 3.45 content words per 100 to 1.79 content words per 100 prior to 31 October 2019, falling further to 1.60 from 31 October 2019 to September 2021 as shown in Figure 3.

Figure 4 shows that sadness as an emotion trended downwards to become almost negligible from February\March 2018. Sadness spikes to 5.0 content words per 100 in the lead up to 31 October 2019.

The concepts derived from the interpretation of the principal component analysis: "Level of Emotion" and emotional tone, "Light versus Dark" are shown in figures 5 and 6. These metrics are comparable to mainstream media where the population mean is 0 and standard deviation is 1. Figure 5 reveals that the level of emotion in Maxwell's tweets dropped substantially post January 2017, hitting a low in October 2018. However, there was a comparative jump in August 2019, but still low compared to his pre-2017 tweets. Increasingly dark emotions are shown in Figure 6. Min particular, the emotions expressed via Tweets become increasingly dark from August 2015, before climbing to being on the light side for the 2016/17 Cricket Season. However, post January 2017 his emotions continued to become darker with sustained low for over a month prior to his decision to withdraw from the Australian Cricket Team in October 2019.

6. CONCLUSION

The tweets of high-profile Australian cricketer, Glenn Maxwell were investigated using natural language processing. The emotions, tone and volume of tweets were assessed over a ten-year period. Prior to his publicised break from competing due to mental health reasons, there is a change in tone of messaging. A drop in the novel, proprietary "Aussie Index", a change in the "Level of Emotion" and increasing "Darkness" in emotions were observed.

These results do not imply that the detected tone, emotions and topics in tweets predict mental health issues. Instead, this case study suggests that this is a topic worthy of further investigation. Further research is required to determine if it is possible to create a monitoring system, akin to Statistical Process Control, for tracking changes in communication styles which may suggest potential issues.

Acknowledgements

We thank Australian Cricketer, Glenn Maxwell, for his public disclosure of his mental health struggles in 2020 and wish him well.

References

AAP (2020, March 25). Glenn Maxwell details his mental health demons. Sydney Morning Herald. (Available at: https://www.smh.com.au/sport/cricket/glenn-maxwell-details-his-mental-health-demons-20200325-p54dq9.html#:~:text=A%20mentally%20exhausted%20Glenn%20Maxwell,battling%20mild%20depression%20and%20anxiety. Accessed 28 September 2021)

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media.

Bracewell, P.J. (2003). Monitoring Meaningful Rugby Ratings. Journal of Sports Sciences, 21 (8): 611-620; 2003.

Bracewell, P.J., Hilder, T.A., & Birch, F. (2019). Player Ratings and Online Reputation in Super Rugby. Journal of Sport and Human Performance. Vol 7, No 2.

Bracewell, P.J., McNamara, T.S, & Moore, W.E. (2016). How Rugby Moved the Mood of New Zealand. Journal of Sport and Human Performance. 4(4). pp. 1-9.

Bracewell, P.J., Wells, J.D, & Craig, A. (2022). Detecting change in tone of tweets and the potential implications for monitoring mental health of athletes. Proceedings of the 16th Australian Conference on Mathematics and Computers in Sports. Ray Stefani & Adrian Schembri eds. Melbourne, Australia, ANZIAM Mathsport. (In Press).

Dissanayake, H., Ward, A.D., Hilder, T.S., Bracewell P.J. (2021). Family violence in the news: An analysis of media reporting of extreme family violence in New Zealand. Kōtuitui: New Zealand Journal of Social Sciences Online. Pp 1-18.

McIvor, J.T., Patel, A.K., Hilder, T., & Bracewell P.J. (2018). Commentary sentiment as a predictor of in-game events in T20 cricket. Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports. Ray Stefani & Anthony Bedford eds. Sunshine Coast, Australia: ANZIAM Mathsport, pp. 44-49.

Middleton, D. (2019, October 31). Maxwell withdraws from Aussie team. Cricket Australia. (Available at: https://www.cricket.com.au/news/glenn-maxwell-australia-t20-series-sri-lanka-break-from-cricket-mental-health/2019-10-31. Accessed 28 September 2021)

Pappafloratos, T.H., Hilder, T. A., & Bracewell, P.J. (2020). Societal bias and the reporting of top tennis players in New Zealand sport media. Proceedings of the 15th Australian Conference on Mathematics and Computers in Sports. Ray Stefani & Adrian Schembri eds. Wellington, New Zealand ANZIAM Mathsport. pp. 114-122.

Simmonds, P.P., McNamara, T.S., & Bracewell, P.J. (2018). Predicting win margins with sentiment analysis in international rugby. Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports. Ray Stefani & Anthony Bedford eds. Sunshine Coast, Australia: ANZIAM Mathsport, pp. 137-142.

Schweinberger, M. (2021). Sentiment Analysis in R. The University of Queensland. https://slcladal.github.io/sentiment.html

Soiferman, T.W. & Bracewell, P.J. (2022). Sign of the Times Sentiment Analysis on Historical Text and the Implications of Language Evolution. Advanced Practical Approaches to Web Mining Techniques and Application. Ahmed J. Obaid, Zdzislaw Polkowski & Bharat Bhushan eds. IGI Global (5). pp. 90-105

QUANTIFYING THE IMPACT OF OLYMPICS ON THE NEW ZEALAND NATIONAL IDENTITY IN MAINSTREAM MEDIA

Paul. J. Bracewell [a,b]
[a] *DOT loves data, New Zealand*
[b] *Corresponding author: paul@dotlovesdata.com*

Abstract

Natural language processing is a prevalent technique for scalably processing massive collections of documents. This branch of computer science is concerned with creating abstractions of text that summarize collections of documents in the same way humans can. Bracewell et. al. (2016) outlined a method for quantifying the collective mood of New Zealanders using mainstream online news content. Mood was quantified using a text mining pipeline built with the Natural Language Toolkit (Bird, 2009) in Python to measure the sentiment of articles and comments appearing in mainstream New Zealand media. A more refined approach is outlined by Schweinberger (2021) where distinct emotions are assessed. Here, that approach is extended further to include a proprietary attribute defined as "Being Kiwi", where over 200 words that are readily associated with New Zealand's national identity, like: "haka", "bbq", "bach", "kiwi", are tracked over time. Plotting the prevalence of "Being Kiwi" over the last decade reveals the highest points coincide with Olympic games. This has important implications for tracking brand values and attributes over time regarding attracting and aligning sponsors.

Keywords: Natural Language Processing, NLP, Sentiment, Emotion Detection

1. INTRODUCTION

Text Mining, Natural Language Processing (NLP) and Natural Language Generation (NLG) are well described in academic literature. Natural language processing is a prevalent technique for scalably processing massive collections of documents. This branch of computer science is concerned with creating abstractions of text that summarize collections of documents in the same way humans can. This form of standardization means these summaries can be used operationally in machine learning models to describe or predict behaviour in real or near real time as required. The predictive fill features within word processing tools like Google Docs and the Microsoft Outlook App show mainstream usage of NLP and NLG.

DOT has had success in applying NLP to a range of text sources, particularly looking at sentiment. Initial applications explored the concept of mood, derived from sentiment (Bracewell et. al. 2016). Sentiment was quantified using a text mining pipeline built with the Natural Language Toolkit (Bird, 2009) in Python to measure the sentiment of articles and comments appearing in mainstream New Zealand media. Commercially, DOT has explored the relationship between brand sentiment and churn rates. However, to protect client sensitivities, DOT ahs often reframed these problems to explore concepts relating to topical, current events to publish findings.

Different NLP applications have been previously published ranging from using commentary sentiment as a predictor of in-game events in T20 cricket (McIvor et. al., 2018), predicting win margins with sentiment analysis in international rugby (Simmonds et. al. 2018), exploring player ratings and online reputation in Super Rugby (Bracewell, et. al., 2019) and an analysis of media reporting of extreme family violence in New Zealand (Dissanayake et. al., 2021)

However, these approaches tended to lend themselves to one-off static reports and did not necessarily lead to ongoing engagements and systematic monitoring. The intent was to make use of one massive, publicly available source of data. Then, from that source, automatically deliver customisable and relevant outputs to many distinct clients from varied backgrounds: business, sport, and politics for example.

Dissanayake et. al. (2021) provides the most recent overview of DOT's collation of mainstream media articles, called Pressroom. To date, this archive contains approximately 10 million news articles published within publicly available main-stream media platforms dating back to 2005 and contains comprehensive collections of articles published on New Zealand websites. The Pressroom is used to report on current events and to track trends in both media reporting and social opinion.

We hypothesised that the difficulty in embedding the outputs within an organisation as an ongoing process stems from the lack of more granular insight, difficulties in objectively displaying the outputs and diminished transparency of outputs. The output of the initial processing, sentiment and number of articles, was too simplistic and lacked enough detail to help connect the user with their understanding of the problem domain. This inability to connect the user with the content creates a barrier to understanding and therefore diminishes trust in the output.

2. METHODS

DATA

The depth and breadth of data contained within the proprietary Pressroom provides a unique opportunity within NZ to configure unique tools for mass market use pertaining to reporting on current events. With millions of time stamped articles to explore, this rich source of data allows NLP tools to be trained within a New Zealand context. More importantly, given the volume of data, metrics generated can be scaled to create outputs that are relatable, comparable, and trackable. This helps bring context to the data by relating to the real world and helping build trust in the outputs. Furthermore, the temporal component can be used to create ongoing engagement, provided meaningful connection between the NLP outputs and client objectives can be achieved.

NLP ALGORITHM

To extract distinct emotions from the data, the more refined approach outlined by Schweinberger (2021) was applied. Exploration of the applicability of this tool formed the basis on an investigation of historical text and the implications of language evolution in a North American context (Soiferman et. al., 2022). Specifically, the change in emotion associated with key words can be aligned to major events. This research highlighted the need to evaluate the stability of characteristics, including features engineered based on word elements when deploying operational models. This is an important issue to ensure that machine learning models constructed to summarize documents are monitored to ensure latent bias, or misinterpretation of outputs, is minimized

As a further challenge, sentiment algorithms are based on text trained with overseas acquired data sets and do not necessarily reflect the New Zealand context. We have had repeated requests regarding the inclusion of Māori.

The tutorial prepared by Schweinberger (2021) provides a simple to use guide for users of the R package with code supplied. It is a straightforward task to work within that framework to customise the word lists for more specific applications. Simply, within that script is a carefully compiled dataset, "nrc" which appears in this line:

```
dplyr::inner_join(get_sentiments("nrc"))        (1)
```

This data set was replaced by a proprietary dataset suitable for use within the New Zealand context. Importantly, the proprietary dataset is highly configurable. As part of ensuring this data fits the intended use case, Māori terms are included, with our early work in modifying this data set covered in, national media outlet, Spinoff (Sowman-Lund, 2021).

Here, that approach is extended further to include a proprietary attribute defined as "Being Kiwi", where over 200 words that are readily associated with New Zealand's national identity, like: "haka", "bbq", "bach", "kiwi", are tracked over time. Bracewell (2022) extended the framework outlined in Schweinberger's tutorial to include additional brand attributes, like "Competitive" and a "Being Kiwi" equivalent, plainly named the "Aussie Index". These were developed in conjunction with Australian-based data led brand and collaborator marketing company, Talennial (www.talennial.com).

A core component of developing these techniques is validation against an external context. This is described in the next section.

3. RESULTS

The process for assigning sentiment, emotions and brand attributes strips out words that are not deemed important like: "the", "and", "then" and "a". This leaves content words, which include nouns, verbs, adjectives and adverbs. Content words are then matched with one of more emotions, themes or brand attributes. To create a prevalence score per 100 words. The first exploration of "Being Kiwi" is shown in Figure 1. The monthly average per month for the prevalence of "Being Kiwi" is approximately 0.244 for the period prior to January 2020. That is, for every 1000 content words, 2.44 of those words are related to "Being Kiwi. The period post January 2021 is explored in Figure 2. Over the last decade reveals the highest points coincide with Olympics games as shown in Figure 1. The 2012 Summer Olympics were held from 27 July to 12 August 2012 in London, England, United Kingdom. The 2016 Summer Olympics, known as Rio 2016, were an held from 5 to 21 August 2016 in Rio de Janeiro, Brazil. The 2020 Summer Olympics, known as Tokyo 2020, were held from 23 July to 8 August 2021 in Tokyo, Japan. To account for seasonality, and to account for any potential impact due to other annual events, we compare the month of the event with the same month the year prior. This reveals that "Being Kiwi" had a year-on-year increase of 19%, 17% and 28% for the 2012, 2016 and 2020 Olympics respectively. These are the major peaks as observed in Figure 1, which has important implications for tracking brand values and attributes over time regarding attracting and aligning sponsors.

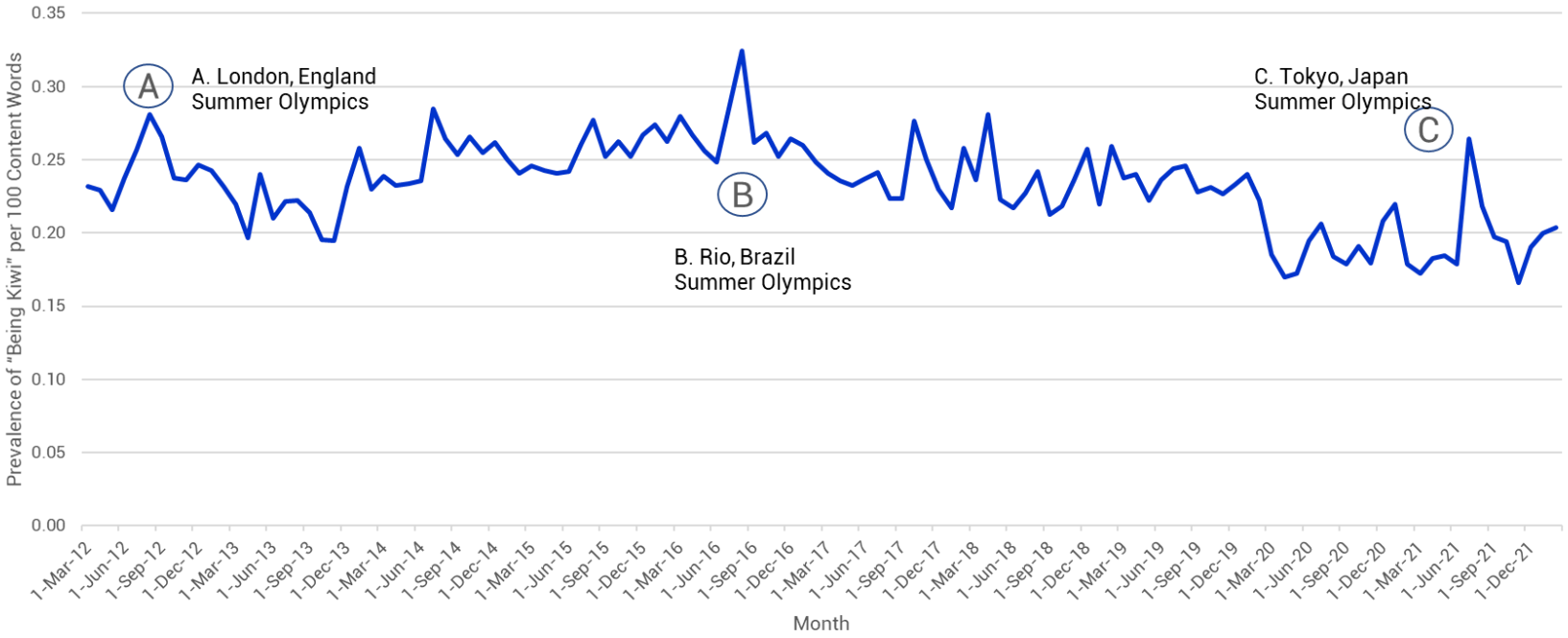


Figure 1: Time series plot showing monthly average of "Being Kiwi" words per 100 content words over a ten-year period (1 March 2012 to 28 February 2022)

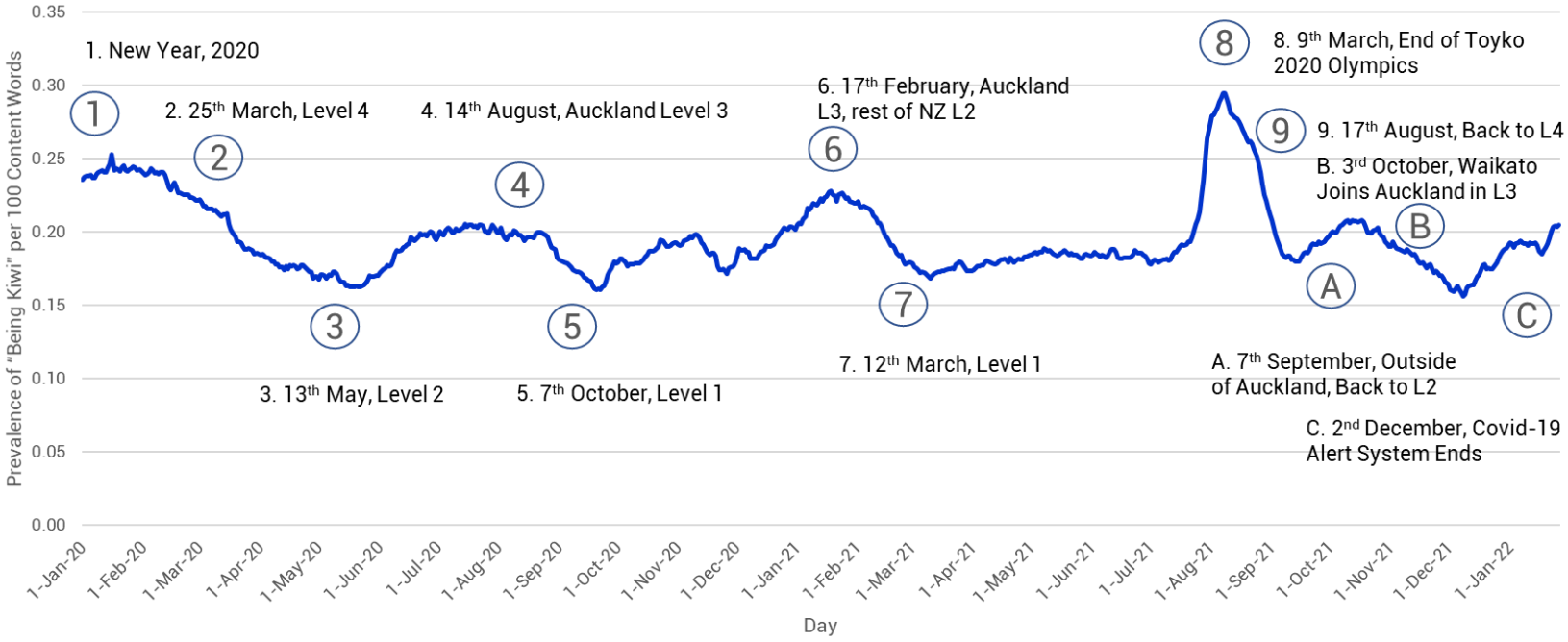


Figure 2: Time series plot showing 28 day rolling average of “Being Kiwi” words per 100 content words over an approximately two years (1 January 2020 to 28 February 2022)

Regarding Figure 1 and observing the spikes due to the Olympics raises questions about the role of rugby on New Zealand's national identity. This topic was explored by Bracewell et. al. (2016) where sentiment was found to but statistically significantly, and positively shifted by the All Blacks 2015 World Cup success. However, the Olympics appears to dominate on this novel index. This may be shaped by certain words used in creating the "Being Kiwi" metric. However, a more detailed investigation of an embedded period provides further support for this metric representing the New Zealand "way of life".

That is, an interesting feature of the decade long view of "Being Kiwi" is the apparent drop following January 2020. To investigate this further, a 28-day rolling average of the prevalence of "Being Kiwi" per 100 content words was explored and plotted over nearly two years. Several the peaks and troughs are explored.

Coinciding with this time frame is the COVID-19 pandemic in New Zealand. This is part of the ongoing pandemic of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The first case of the disease in New Zealand was reported on 28 February 2020. Key events are annotated on Figure 2, with more detailed information available from the NZ Government at the following website: https://covid19.govt.nz/about-our-covid-19-response/history-of-the-covid-19-alert-system/. That website provides specific details about how New Zealand handled the response to COVID. Interestingly, troughs appear to occur when parts of New Zealand went into Lockdown or faced restrictions. Level 4 was the harshest of the restrictions in New Zealand under the Covid Alert System and Level 1 the least restrictive. Late in 2021 this operating framework migrated to a Traffic Level System under the Covid Protection Framework, with the red setting the strictest and green the most open.

Of particular interest is the apparent increases in "Being Kiwi" when restrictions were eased. This suggests that the data may be identifying features of relevance in quantifying the New Zealand way of life.

4. DISCUSSION

A core component of developing these techniques is validation against an external context. In the previous section, two periods were examined. The first analysis showed that during periods of Olympic competition there were between 17% and 28% increases in the metric for the number of content words per 100 associated with "Being Kiwi", compared with the same month the year prior. Given the Olympics is an opportunity for New Zealand to present itself on the world stage, this suggests that this algorithm is detecting some useful features regarding how journalists write about New Zealand athletes during this period of competition.

Then, a more granular review of how this metric evolved daily whilst New Zealand was in the depths of a response to the COVID-19 global pandemic was undertaken. Here, it appeared that when New Zealand went into more restrictive states, such as lockdowns, journalists did not use as many words in articles that would typically be associated with "Being Kiwi". Furthermore, as restrictions eased, there appears to be a contemporary increase in the use of words associated with "Being Kiwi".

Given that changes in the proposed metric, "Being Kiwi" appears to change as events affecting New Zealanders unfolded, it is not unreasonable to make the observation that the "Being Kiwi" metric identifies phrases and words that are aligned with New Zealand's national identity.

5. CONCLUSIONS

As New Zealanders experienced a roller coaster of self-identification in the wake of various Covid-19 Alert System Levels, there was one shining light for Kiwis, the 2020 Tokyo Olympics, which ended on 9th August 2021. This was measured by expanding on the work by Schweinberger (2021). That approach was extended further to include a proprietary attribute defined as "Being Kiwi", where over 200 words that are readily associated with New Zealand's national identity, like: "haka", "bbq", "bach", "kiwi", are tracked over time. Plotting the prevalence of "Being Kiwi" over the last decade reveals the highest points coincide with Olympic Games. The implication is that the reporting in mainstream media around the Olympics uses key words that are associated with the New Zealand national identity. This has important implications for tracking brand values and attributes over time regarding attracting and aligning sponsors. The extension to explore daily movements relative to issues affecting all New Zealanders provides further validation of this metric.

References

Bracewell, P.J. (2022). Quantifying the impact of Olympics on the New Zealand national identity. Proceedings of the 16th Australian Conference on Mathematics and Computers in Sports. Ray Stefani & Adrian Schembri eds. Melbourne, Australia ANZIAM Mathsport. (Under Review).

Bracewell, P.J., McNamara, T.S, & Moore, W.E. (2016). How Rugby Moved the Mood of New Zealand. Journal of Sport and Human Performance. 4(4). pp. 1-9.

Schweinberger, M. (2021). Sentiment Analysis in R. The University of Queensland. https://slcladal.github.io/sentiment.html

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media.
Bracewell, P.J., Hilder, T.A., & Birch, F. (2019). Player Ratings and Online Reputation in Super Rugby. Journal of Sport and Human Performance. Vol 7, No 2.
Bracewell, P.J., McNamara, T.S, & Moore, W.E. (2016). How Rugby Moved the Mood of New Zealand. Journal of Sport and Human Performance. 4(4). pp. 1-9.
Bracewell, P.J., Wells, J.D, & Craig, A. (2022). Detecting change in tone of tweets and the potential implications for monitoring mental health of athletes. Proceedings of the 16th Australian Conference on Mathematics and Computers in Sports. Ray Stefani & Adrian Schembri eds. Melbourne, Australia, ANZIAM Mathsport. (In Press).
Dissanayake, H., Ward, A.D., Hilder, T.S., Bracewell P.J. (2021). Family violence in the news: An analysis of media reporting of extreme family violence in New Zealand. Kōtuitui: New Zealand Journal of Social Sciences Online. Pp 1-18.
McIvor, J.T., Patel, A.K., Hilder, T., & Bracewell P.J. (2018). Commentary sentiment as a predictor of in-game events in T20 cricket. Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports. Ray Stefani & Anthony Bedford eds. Sunshine Coast, Australia: ANZIAM Mathsport, pp. 44-49.
Pappafloratos, T.H., Hilder, T. A., & Bracewell, P.J. (2020). Societal bias and the reporting of top tennis players in New Zealand sport media. Proceedings of the 15th Australian Conference on Mathematics and Computers in Sports. Ray Stefani & Adrian Schembri eds. Wellington, New Zealand ANZIAM Mathsport. pp. 114-122.
Simmonds, P.P., McNamara, T.S., & Bracewell, P.J. (2018). Predicting win margins with sentiment analysis in international rugby. Proceedings of the 14th Australian Conference on Mathematics and Computers in Sports. Ray Stefani & Anthony Bedford eds. Sunshine Coast, Australia: ANZIAM Mathsport, pp. 137-142.
Schweinberger, M. (2021). Sentiment Analysis in R. The University of Queensland. https://slcladal.github.io/sentiment.html
Soiferman, T.W. & Bracewell, P.J. (2022). Sign of the Times Sentiment Analysis on Historical Text and the Implications of Language Evolution. Advanced Practical Approaches to Web Mining Techniques and Application. Ahmed J. Obaid, Zdzislaw Polkowski & Bharat Bhushan eds. IGI Global (5). pp. 90-105
Sowman-Lund, S. (2021, September 17). "11.30am: The Spinoff tops te reo usage in mainstream NZ media" . Spinoff. Available: https://thespinoff.co.nz/politics/17-09-2021/live-updates-september-17-trajectory-of-horror-what-national-mps-are-saying-about-their-own-party. Accessed at: 17 September 2021.

ANALYSING PLAYER SUBSTITUTION STRATEGIES IN THE NBA

Bart Spencer [a,b,c]
[a] *Victoria University, Australia*
[b] *San Antonio Spurs, USA*
[c] *Corresponding author: bartholomew.spencer@vu.edu.au*

Abstract

The timing of substitutions in the NBA is influenced by several in-match and pre-match factors. While some teams adhere to a static, pre-determined schedule, others adopt a more dynamic approach, rotating players based on the state of the game. The primary objective of this study was to define the range of substitution strategies used by NBA teams in the 2019-2021 seasons. Secondary objectives were to measure the similarity between teams' substitution strategies and to analyse their predictability. *K*-means analysis was used to cluster substitution strategies. A team's substitutions within a quarter were represented as an n-tuple, where each element denotes the time of a substitution (measured in seconds since the beginning of the quarter). The number of substitutions per quarter varies, hence tuples were transformed to achieve consistent dimensionality for clustering. Two methods were tested for this transformation. *K*-nearest-neighbor classification was used to predict the timing of a team's substitutions. Accuracy of these models were compared amongst teams to quantify the predictability (estimated via model error) for all teams in the NBA.

Keywords: NBA, basketball, clustering, interchanges, rotations, team sports, rotations

1. INTRODUCTION

Stint duration and substitution timing has been a commonly researched topic in team sports literature. In particular, there has been a focus on optimising player performance through substitution strategies. In AFL, for example, Corbett et al. (2017) found a weak relationship between physical performance and stint duration. Similar research in field hockey identified the fifth minute of play as being one where fresh players experience signs of fatigue (Linke & Lames, 2016). Linke and Lames (2016) additionally note, however, that player performance in the first minute of play following a substitution is significantly higher than the team average, noting the tactical advantage of substituting players. In an era where player management is of increasing importance (particularly to minimise injuries), further research into the trends and optimisations of substitution strategies is required.

In this study we present methods for analysing the timing of substitutions in the NBA. Games in the NBA contain an unequal number of substitutions due to limited rules regarding their timing, hence there is a great deal of diversity between teams in their approach to substitutions. The primary objective of this study was to define the variety of substitution strategies employed by teams. Unsupervised machine learning techniques were used to partition substitution time series into clusters. From this, we measure the similarity of teams in the NBA and produce a supervised machine learning model capable of predicting substitutions towards the end of a quarter based on the observed substation timings in the preceding minutes of play.

2. METHODS

DATA COLLECTION AND PRE-PROCESSING

Data was collected from all matches played during the 2019-2021 NBA seasons as of March 2022. Second Spectrum[1] player tracking data was used in this analysis. While player tracking data was not required for the analysis conducted in this study, these datasets can be used as a source for substitution times. At any point in time, the position of the five on-field players on each team is tracked in 25Hz using optical tracking systems. We obtain the exact frame that a substitution occurs by monitoring changes in the on-court player IDs.

An NBA match consists of four 12-minute (720 second) quarters in the case of no overtime play. In the event of a draw at the end of the fourth quarter, teams continue playing in five-minute overtime periods until a winner is declared. In this study, we analyse substitutions occurring in the four regular 12-minute quarters.

Player tracking data was incomplete or missing for a small number of matches in the analysed seasons. Incomplete matches were dropped from the analysis. All 30 NBA teams were included in this study, and a total of 3230 matches played between October 2019 and March 2022 were analysed.

[1] https://www.secondspectrum.com/index.html

SUBSTITUTION TIME SERIES SAMPLES

A team's substitutions within an individual quarter were represented as an *n*-tuple, with each element denoting the time (in seconds) where a substitution occurred. For example, the tuple (299, 377, 503, 503, 503) indicates a quarter where a single player was substituted at 299 seconds, another single player at 377 seconds, and finally three players at 503 seconds (4:59, 6:17, and 8:23 minutes respectively). For the purpose of clustering quarters, quarters where 0 substitutions occurred were dropped from the dataset. There is no strict rule requiring or restricting the number of substitutions in a quarter (or game), resulting in a varying number of substitutions in each sample. Across the analysed matches, an average of 5.4 ± 2.2 substitutions occurred per team each quarter. The total number ranged between 1 and 18 substitutions, and the median was 5.

The number of substitutions varies between samples; hence transformation was required before each sample could be clustered. Two methods for performing this transformation were trialled. The first approach involved partitioning each 12-minute quarter into one-minute bins (0-1 minutes, 1-2 minutes, etc.). Each sample was then represented as a 12-tuple, where each element (*n*) records the number of substitutions that occurred in the n^{th} minute. This process produces a dataset containing quarter substitution time series of equal size for all matches. The result is easy to interpret, however discretising data can produce misleading results in the case of samples falling close to the bounds of each bin.

The second transformation technique involved using kernel density estimation (KDE) to produce a continuous time series (at one-second intervals) of substitution density. KDE involves estimating the probability density function of a time series, producing a smooth density curve. The bandwidth of the kernel determines the amount of influence each data point (or substitution) has over the time around it. Scipy's implementation of KDE was used in this study (Virtanen et al., 2020). Bandwidth selection was set to Scott's factor (Scott, 2015) divided by four to minimise smoothing across the quarter (and produce a shape that retains the timing of substitutions). The advantage of this transformation technique is we produce a continuous time series; however, this may be harder for coaching staff to interpret.

Examples of each transformation technique are presented in Figure 1. In each plot, the continuous line is the KDE representation of a substitution time series, and the discrete bars are the binned representation of the same sample. The example on the left is of a quarter were 3 substitutions occurred in the third minute, following single substitutions in the 4^{th}, 5^{th}, 7^{th}, 8^{th}, 11^{th}, and 12^{th} minutes. The example on the right has a total of 11 substitutions, roughly grouped around the start, middle, and end of the quarter.

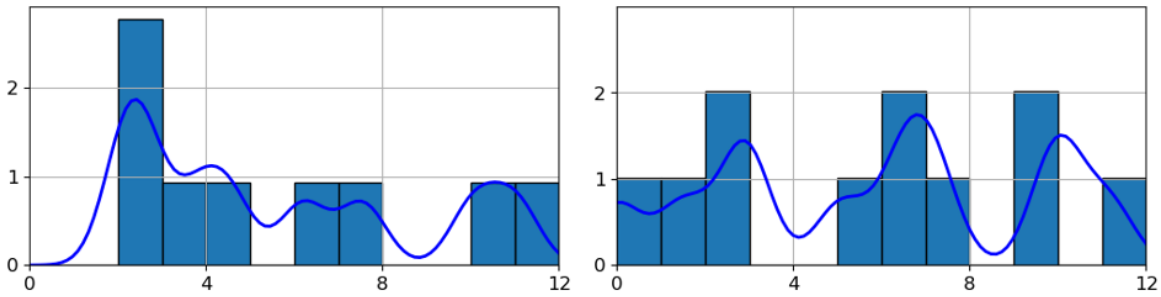


Figure 1. Histogram and KDE representations of two substation time series

MEASURING TEAM SIMILARITY

The similarity between each team in the NBA is measured based on the substitutions they performed across the analysed seasons. For each team, we produce a dataset of their transformed substitution samples. Euclidean bipartite matching between teams *A* and *B* is used to measure the similarity between the two teams. The Euclidean bipartite matching process involves pairing samples from dataset *A* and *B*, such that the total Euclidean distance between the matched pairs is minimised (Mézard & Parisi, 1988).

From this process, we produce a cost matrix measuring the distance between each team in the NBA. Teams are clustered via hierarchical clustering (using a 70% distance threshold for creating clusters). This process was performed for both methods of transformation detailed in the section above to compare results using either method.

CLUSTERING QUARTERS

K-means clustering was used to cluster substitution samples. *K*-means clustering is an unsupervised machine learning technique used to partition data samples into groups based on the distance (commonly the Euclidean distance) between samples and cluster centres. This technique has been applied to a variety of sports analytics problems such as the grouping of teams based on playing style (e.g., Spencer et al., 2016). The number of clusters (*k*) was chosen based on the elbow method, in which *k* is chosen as the point from which increasing *k* has a reduced effect on minimising the total error (in our case, the sum of squared distances between samples and cluster centres) (Kodinariva & Makwana, 2013).

PREDICTING SUBSTITUTIONS

A secondary objective of this study was to assess the predictability of substitution strategies within NBA teams. While this is a complex topic and grounds for future research, we propose the accuracy of predictive models as a proxy for quantifying teams' 'predictiveness'. To measure this, we use the *k*-nearest-neighbours (kNN) algorithm to predict a team's substitutions for the remainder of a quarter after time *t*, given their substitutions in the quarter up to time *t*. In this study, t = 360s (i.e., we observe behaviour in the first half of a quarter to predict behaviour in the second half). KNN classification involves classifying a sample's class based on the Euclidean distance between the sample and training samples (Peterson, 2009). Performance is optimised via adjusting the value for *k* which determines how many neighbours the prediction is based on.

Model inputs are the KDE representation of each quarter up to 360s, and the output (or response variable) is its cluster (from the *k*-means analysis detailed in the previous section). If we can accurately predict the cluster that a partial time series belongs to, we can output the likely shape for the remainder of the quarter. We use an 80/20 split for our training and testing datasets. Model performance was evaluated via raw predictive performance (i.e., the percentage of correctly labelled results), as well as via log-loss (Eq. 1) which considers the predicted probabilities when assessing model performance. A smaller log-loss value indicates a stronger model. To analyse the predictability of teams we output the log-loss of each team. Log-loss is calculated as follows,

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log\left(p_{ij}\right) \quad (1)$$

where,

N	No. of rows in test set
M	No. of classes
y_{ij}	1 if observation belongs to class j; else 0
p_{ij}	Predicted probability that observation belongs to j

3. RESULTS

TEAM SIMILARITY

Dendrograms produced from the distance matrices using discrete transformation via binning (Figure 2) and KDE transformation (Figure 3) are presented below. We focus on the positioning of the San Antonio Spurs (SAS). Using the discrete transformation method, SAS are grouped with Memphis Grizzlies (MEM), Milwaukee Bucks (MIL), and Utah Jazz (UTA). Using KDE transformation, SAS remain grouped with MEM and MIL, while being one group detached from UTA. Additionally, Charlotte Hornets (CHA) joins the group despite larger separation in the discrete transformation method.

QUARTER CLUSTERS

A total of 30 clusters were chosen for the *k*-means clustering of quarter time series. Clustering was performed on quarter samples transformed using the KDE transformation method. The centres of these clusters (0 through to 29) are presented in Figure 4. A variety of strategies are represented within these 30 clusters, such as quarters where substitutions occurred only at single points during a quarter (e.g., cluster 15, 16, 17) or quarters were substitutions occurred at the beginning of a quarter, followed by a second group of substitutions later in the quarter (e.g., at the beginning and middle of a quarter in cluster 0, or at the beginning and end of a quarter in cluster 6).

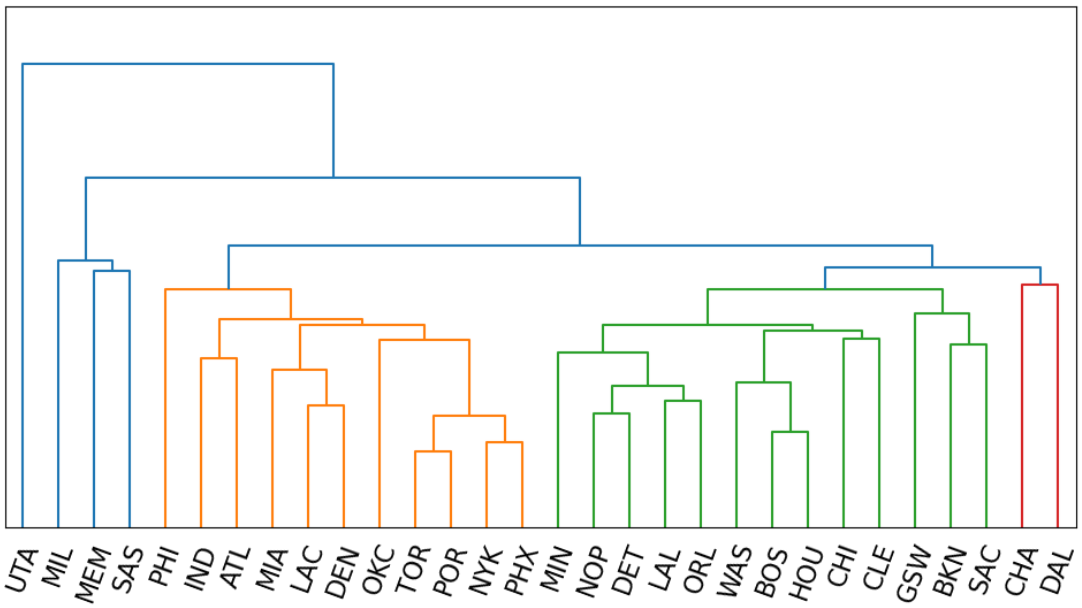


Figure 2. Hierarchical clustering of teams using the one-minute bins transformation method

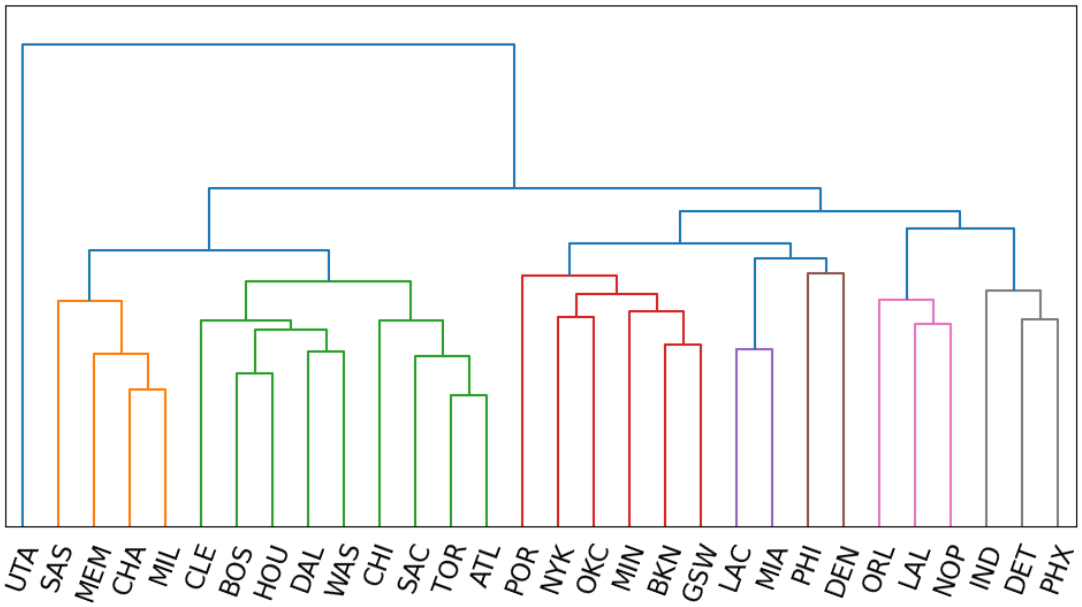


Figure 3. Hierarchical clustering of teams using the KDE transformation method

kNN PREDICTION RESULTS

A kNN model was trained using substitution time series for the first half of each quarter. A *k* value of 50 was used in this process. Values beyond 50 yielded minimal improvements to overall performance. Using a 360-second window (or half a quarter), the model was able to successfully predict 59% of quarter clusters.

Examples of the prediction process are displayed in Figure 5. As detailed above, we use a 360-second (half a quarter) training window to predict what teams will do in the remaining 360-seconds of a quarter. The blue line is the ground truth time series (transformed via KDE). The dotted lines represent the kNN predictions for the remainder of each quarter. The confidence of each prediction is denoted in the legend of each plot. In some

cases (e.g., Figure 7a) the algorithm has high certainty in a single cluster, while others (e.g., Figure 7b) the prediction can be spread across more clusters.

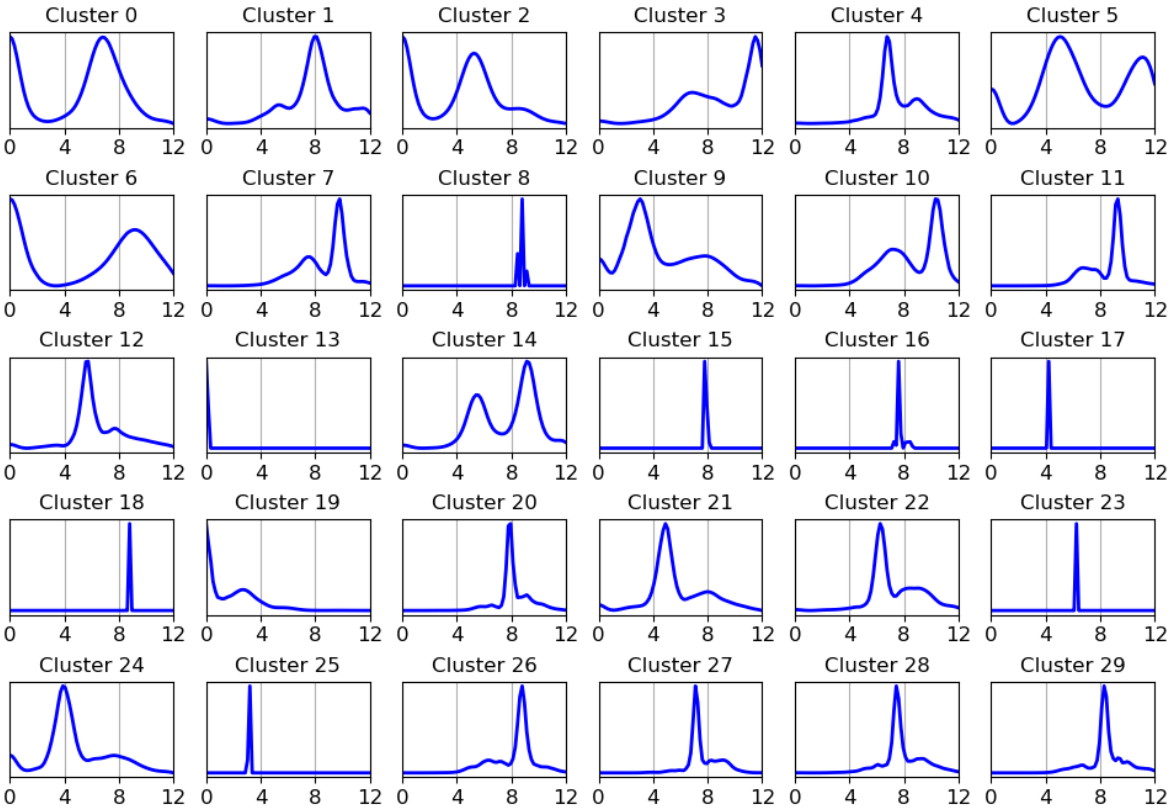


Figure 4. Time series plots of the 30 clusters produced by k-means clustering

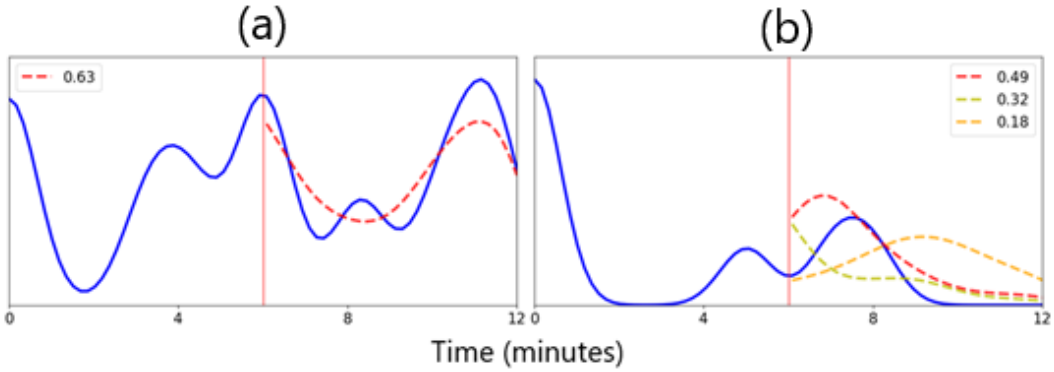


Figure 5. Example predictions after the 6th minute using k Nearest Neighbor (kNN) classification with labeled probabilities for predictions (dashed lines)

Finally, we measure the log-loss of predictions, grouped by team, and present these results in Figure 6, ordered by log-loss (where a smaller number indicates lower error or better predictability). After discussion with coaching and sport science staff, Phoenix Suns (PHX) and MIL were identified as two high performing teams across the analysed seasons (having topped the conferences in the 2021-22 season and made playoffs in the 2020-21 season). Based on log-loss, PHX and MIL were middle of the pack in terms of predictability.

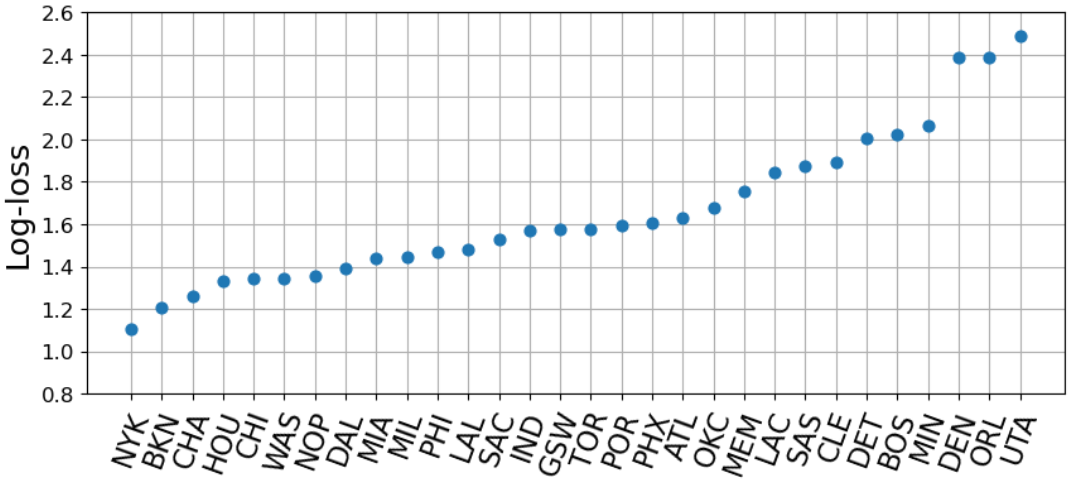


Figure 6. Loss-loss of kNN predictions for each team in the NBA

4. DISCUSSION

This study presented a method for defining substitution strategies in the NBA using *k*-means clustering on substitution time series for individual quarters. This output was used to measure the similarity between teams, and to predict future substitution behaviour based on substitutions up to the middle of the current quarter.

Two processes were trialled for transforming the substitution time series into samples of equal dimensions for clustering. The first involved partitioning the quarter into 12 one-minute bins and counting the number of substitutions that occurred in each bin. This process discretises the data. The second process retains the continuous nature of these substitutions by converting the time series into a smoothed time series measuring the density of substitutions across the 720 seconds of each quarter (via KDE). While this output produces a continuous dataset, its interpretability by coaching staff should be researched. In general, we might suggest that the discrete, binned dataset produces an output that is easier to interpret to a wider audience (Figure 1). The downside of the binning process, however, is that the use of arbitrary bounds can result in datasets having large Euclidean distances between them despite interchanges occurring at similar times. As an extreme example, substitutions occurring at 0:59 and 1:01 would be considered as far apart as substitutions occurring at 0:01 and 1:59. In the case of KDE transformation, the former pair would have minimal distance between them.

Predicting substitution behaviour in the second half of a quarter based on the timing of substitutions in the first half yielded strong results, with the correct cluster predicted in 59% of testing samples. Future research into how this performance changes based on a variety of match conditions should be conducted. For example, the addition of match information (e.g., score-line differential) may improve model prediction. In team sports literature it has been shown that the strategy and positioning of teams changes because of situational variables such as match location and opponent quality (Santos et al., 2017). It is likely that these factors would also have an impact on substitution timings. Behaviour in the fourth quarter, for example, is likely to be heavily influenced by the status of the game. These factors present an opportunity for future research to improve model prediction and quantify specific factors that cause NBA teams to make substitution decisions. An alternative approach to this topic could involve the use of association rules. Association rules can be used to determine the effects of match context on a fixed response variable (e.g., how kicking constraints affect kicking success in the AFL in the case of Robertson et al., 2019).

This study presented one method for measuring the predictability of teams via the log-loss of team-specific substitution predictions. In highlighting two top performing teams across the analysed seasons (PHX and MIL), we note that there is no clear trend as to whether being more or less predictable (as quantified via the presented methodology) improves a team's performance. Further research is required to analyse predictability relating to a variety of match factors, such as those touched on above.

5. CONCLUSIONS

This study presented a method for defining the range of substitution strategies employed by NBA teams in the 2019-2021 seasons. Using transformed substitution time series, team similarity was measured using Euclidean bipartite matching. Results revealed consistency in the similarity between SAS and a number of teams, however

there were differences in the number of clusters and positioning of some teams when taking a discrete (via binning) or continuous (via KDE) approach to the transformation of substitution time series samples. The predictability of teams was assessed via a kNN classification model to predict a team's likely substitution pattern in the second half of a quarter, given the timing of their substitutions in the first half of said quarter. We found that the predictability of teams varies greatly throughout the league. Finally, analysing PHX and MIL's performance (as recent conference winners) did not reveal clear trends in a specific substitution style relating to success in the NBA (or their ability to be more or less predictable).

Acknowledgements
We wish to thank Second Spectrum for providing the data used in this study.

References

Corbett, D. M., Sweeting, A. J., & Robertson, S. (2017). Weak relationships between stint duration, physical and skilled match performance in Australian Football. *Frontiers in physiology*, *8*, 820.

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal*, *1*(6), 90-95.

Linke, D., & Lames, M. (2016). Substitutions in elite male field hockey–a case study. *International Journal of Performance Analysis in Sport*, *16*(3), 924-934.

Mézard, M., & Parisi, G. (1988). The Euclidean matching problem. *Journal de Physique*, *49*(12), 2019-2025.

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, *4*(2), 1883.

Robertson, S., Spencer, B., Back, N., & Farrow, D. (2019). A rule induction framework for the determination of representative learning design in skilled performance. *Journal of Sports Sciences*, *37*(11), 1280-1285.

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Spencer, B., Morgan, S., Zeleznikow, J., & Robertson, S. (2016, July). Clustering team profiles in the Australian Football League using performance indicators. In *Proceedings of the 13th Australasian conference on mathematics and computers in sport, Melbourne* (pp. 11-13).

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & Van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, *17*(3), 261-272.

NETWORK ANALYSIS METHODS FOR LIVE PERFORMANCE OUTCOMES IN NETBALL

Anthony Bedford [a,b,d] , Kiera Bloomfield [a]

[a] *University of the Sunshine Coast*
[b] *New Zealand Silver Ferns*
[d] *Corresponding author: abedford@usc.edu.au*

Abstract

Live data collection is nothing new in sport. However, the ability to turn said data around into salient and actionable outcomes during a game remains a challenge. For the last six years, the authors have been analysing netball at both national and international levels. This includes, from team-based metrics to individual outcomes. As such, as performance analysts, they have been able to convey live data and live insights to coaches and players in their specific teams. However, being able to generate inferential outcomes in time-poor scenarios remains a challenge. The authors discuss existing techniques that have been used to convey outcomes during a match scenario, such as video, and some connections that are of influence. They then propose how to use network analysis from live possession data to ascertain pivotal players and then peel back the factors of influence in a game. Some examples of the models used are shown, and their relationship to outcomes.

Keywords: Coaching, network, netball, in-play

Acknowledgements

We thank Issy Packer for her assistance in data collection.

BROWNIAN MOTION MODELLING OF SCORING IN NETBALL TO DETERMINE THE MOST IMPORTANT MOMENTS

Anthony Bedford [a,b,d] , Kiera Bloomfield [a]

[a] *University of the Sunshine Coast*
[b] *New Zealand Silver Ferns*
[d] *Corresponding author: abedford@usc.edu.au*

Abstract

Using the scoring of goals, and their misses, we evaluate the importance of a shot in netball. We consider both national and international games which, in Australia, currently have very different scoring systems. The "Super Shot" in Suncorp Super Netball (SSN), the Australian domestic league at time of writing, affords two goals if successful, and is awarded from a range of three metres or more from the post while still inside the shooting circle. With its additional degree of difficulty to the shot comes a greater reward. Whilst Fox & Bruce (2020) published the success and expected value of this shot, we consider here the importance in the outcome of a goal made or missed from all four possibilities (in SSN), and two in conventional netball. Thus, through evaluation of two years of SSN data and International and Domestic games, we find that strategies require variation depending upon the margin, time of game, and mode of game. We utilise a Brownian motion variant to estimate the probability of a goal scored, and utilising modifications of Morris' (1977) model for Importance, this enables determination of strategy dependent upon outcome. Most notably, and obviously, the Super Shot yields big rewards and bigger consequences if missed, however the state of the game provides great insight into the desire to take risky shots.

Keywords: Brownian motion, netball, importance

Acknowledgements

We thank Issy Packer for her assistance in data collection.

References

Fox, AS. & Bruce, L. (2020). When does risk outweigh reward? Identifying potential scoring strategies with netball's new two-point rule. *PLoS ONE, 15(11)* : e0242716.

Morris C (1977). The most important points in tennis. In *Optimal Strategies in Sports*, S.P. Ladany and R.E. Machol eds., Amsterdam: North-Holland, 131-140.

THE TRUE MEANING OF THE OLYMPIC MOTTO IS NOT THAT RECORDS HAVE BEEN BROKEN BUT RATHER THE MINDSET BY WHICH RECORDS CONTINUE TO BE BROKEN AS WELL AS INSIGHT INTO MENTAL HEALTH ISSUES AFTER A SUCCESSFUL CAREER

Raymond Stefani
California State University, Long Beach (USA)
Raystefani@aol.com

Abstract

The Olympic motto, *Citius, Altius, Fortius*, was coined by Father Henri Didon in 1891, to teach athletes to do their best to run a bit faster, to jump a bit higher and to become a bit stronger than before. He didn't say anything about beating anyone, setting a record or being on a team. Following his meaning, when a young person enters sports, they accept the "new normal" of that day, using then-available nutrition, training and equipment to produce increasing amounts of power. They use then-available coaching and techniques to turn that power into increased performance. At the same time, others are working to improve those factors. The best of those improving athletes will compete on sports teams, the best of them will be on national teams, the best of them will win medals and the best of them will set records, due to the cumulative effect of self-improvement. For example, using Olympic championship performances from 1928 to 2020/21, the Percent Improvements per Olympiad (%I/O) led to cumulative improvements creating previously unimaginable performances. In running, the %I/O (with cumulative improvement in parentheses) for women were 0.7% (15%) and for men 0.5% (10%). For jumping, women achieved 1.4% (29%) and men achieved 1.0% (21%). In swimming women achieved 1.6% (33%) while men achieved 1.4% (28%). Michael Phelps' documentary shows that the top athletes experience mental health issues due to feeling meaningless when the next generation of athletes follow the same process and obtain similar performances compared to these top athletes, who also feel powerless when they leave competition and lose all means of financial support. Counsellors are needed to deal with the athletes' self-worth while in competition and with transitional planning for the future.

Keywords: Olympic motto, performance improvement, breaking records, running, jumping, swimming, mental health, nutrition, training, equipment, coaching.

1. INTRODUCTION

Prior to the Tokyo Olympics, originally scheduled for 2020 but held in 2021, on 20 July 2021, the IOC modified the Olympic Motto to read "Faster, Higher, Stronger - Together", IOC (2021). By adding "Together", the IOC pledged that sport should advance by the unified efforts of all to conquer COVID and to further the goals of sport in general and the Olympic movement in particular. As the Tokyo 2020/21 Olympics progressed, as is usually the case with Olympic Games, world media implied that those who set records exemplified the "Faster, Higher, Stronger" manta, while similarly, individuals and nations that gained the most medals exemplified the purpose of the Olympics. None of those descriptions actually fit what the Olympic Motto truly means and what the Olympics themselves should evoke.

On 7 March 1891, at an Arcueil College sports assembly in France, Father Henri Didon, a Dominican Priest, taught that the students' goal in sport should be to do their best to improve little by little, that is, to run a bit faster, to jump a bit higher and to become a bit stronger than they had been before. He didn't say anything about beating anyone or setting a record. He gave them a Latin motto in sport as in life: *Citius, Altius, Fortius*, IOC (2002).

His friend, Baron de Coubertin, was present. When de Coubertin founded the modern Olympic movement in 1894, he chose *Citius, Altius, Fortius* as the Olympic Motto. Baron de Coubertin clarified the meaning of the Olympic Motto, by creating the Olympic Creed, IOC (2021), based on a talk given at the 1908 Olympics by Ethelbert Talbot, Bishop of Pennsylvania. Following Talbot, Coubertin's Olympic Creed reads "The most important thing in the Olympic Games is not to win but to take part, just as the most important thing in life is not the triumph but the struggle. The essential thing is not to have conquered but to have fought well."

The true essences of sport and the Olympic movement are therefore self-improvement and participation. When a young person becomes interested in physical activity, their mindset should therefore be to choose the methods of that day and age (that new normal), not thinking about the past or wondering if anyone in the future will do better. The vast majority are just having some fun with physical activity. There is a pyramid of methods that builds to performance. **For coverage of the physics and performance components of the sports in this paper, including the ratio of female/male winning velocities, see Stefani (2008) and Stefani (2014).**

Performance
Coaching Technique
Nutrition Training Equipment

The athlete employs the then-available methods of nutrition, training and equipment to create physical power. Next, then-available coaching and techniques covert that power to performance in the chosen sport. While the young athletes are doing their best to improve, using the methods of their era, sports scientists are working to improve nutrition, training, equipment, coaching and technique. Those technical improvements elevate what the athlete can achieve. The best of those improving athletes will compete on sports teams, the best of them will be on national teams, the best of them will win medals and the best of them will set records, due to the cumulative effect of self-improvement. *Citius, Altius Fortius* (Faster, Higher Stronger) explains **how** records are set, not **that** records have been set.

Recently, another important area of sports performance has come to the fore: mental health. Top athletes have spoken about mental health, such as tennis player Naomi Osaka, who was chosen by Japan to light the Olympic Torch, and Simone Biles, considered to be one of the greatest gymnasts ever. That topic will be covered herein.

The rest of this paper will begin by examining the remarkable improvements in Olympic winning swimming velocity, including a comparison of Johnny Weissmuller at 100m in 1924 and Grant Hackett at 1500m in 2004. The cumulative improvements for Olympic champion men and women in athletics (running and jumping) and in swimming, averaged over all events, will be evaluated, demonstrating application of the true meaning of the Olympic Motto. From changes to the rate of improvement, the effectiveness of some technological advances will be evaluated. The achievements at the Tokyo Olympics will exemplify how the "new normal" of COVID restrictions was accepted and conquered. Finally, mental health issues are covered, based on a documentary created by Michael Phelps.

2. IMPROVEMENTS IN OLYMPIC WINNING PERFORMANCES

LESSONS FROM THE MEN'S WINNING VELOCITIES IN SWIMMING AT 100M AND 1500M

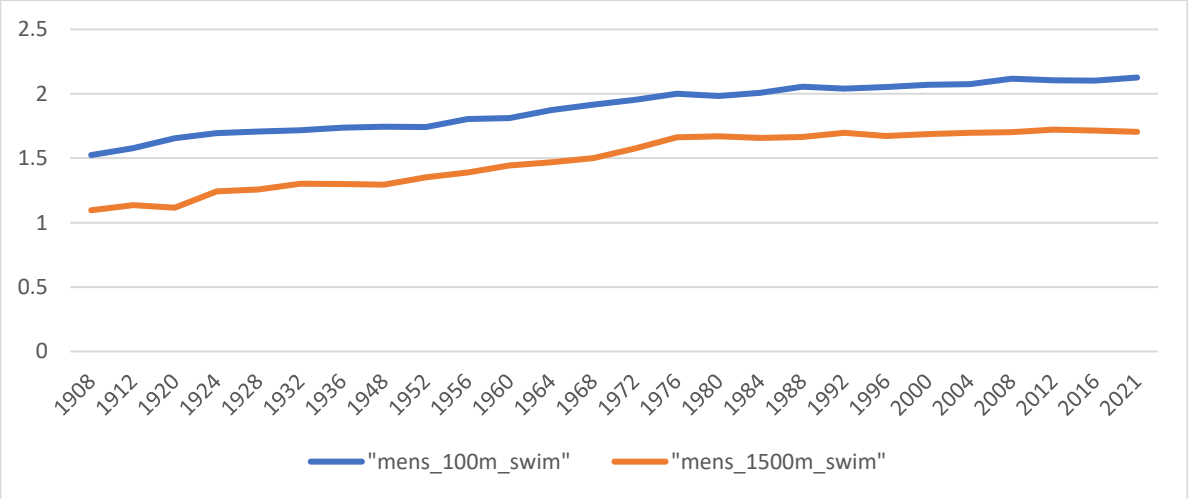


Figure 1: Velocity of Male 100m and 1500m Olympic Swimming Champions (1908-2021)

Figure 1 shows the winning velocities for male Olympic champions from 1908 through the 2020/21 Olympics at 100m (upper curve) and at 1500m (lower curve). The figure begins with 1908, the first year when swimming was held in a pool at metric distances. Prior to 1908, Olympic swimming was held in Piraeus Harbor at Athens in 1896, in the Seine River at Paris in 1900 and in a pool with distances in yards at St. Louis in 1904.

We see the consistent increase in winning velocities. Of particular interest is the competition in 1924. Johnny Weissmuller (USA) won at 100m in 59s, becoming the first to win in under 1 minute. That was such a world-wide phenomenon, it was one reason why Weissmuller was chosen to play Tarzan in the movies. Also in 1924, Andrew (Boy) Charlton (Australia) became famous for winning at 1500m in a time of 20:06,5, nearly 2 minutes better than the Olympic record of 22:00. Those two performances were so widely publicized and so spectacular, how long did it take competitors to accept the new normal of such improved times? As we look at Figure 1, that

new normal was accepted immediately. Velocities continued to improve immediately after 1924 at both distances until the post WW2 Olympics of 1948, followed by generally increased velocities.

Now, start with Weissmuller's 1924 velocity in the upper curve. Move horizontally for 80 years and you hit the 1500m curve in 2004. In 2004, Grant Hackett (Australia) won the 1500m with a velocity of 58.9s per 100m. That is, Hackett swam 15 times father than Weissmuller with a slightly faster velocity. Hackett's winning 1500m time of 14:43.40 was more than 5 minutes faster than Charlon in 1924. Hackett's performance would have been considered highly impossible in 1924, yet if you compare photos of Weissmuller, Charlton and Hackett, you see little difference in physicality. Of course, Hackett embodies improvements in nutrition, training, coaching and technique. Suits are similar. The human body has gracefully adapted to each new normal; hence there is strong likelihood of continued improvement in the future.

We now take a comprehensive look at accumulated improvement for men and women, averaged for all running, jumping and swimming events.

CUMULATIVE IMPROVEMENTS FOR OLYMPIC CHAMPIONS IN ATHLETICS AND SWIMMING

To calculate and plot the men's and women's cumulative improvements in Figure 2 (running), Figure 3 (jumping) and Figure 4 (swimming), the percent improvements over each Olympiad for each event in running, jumping and swimming were found and then averaged for each Olympics by gender. The average percent improvements per Olympiad (%I/O) were then accumulated and plotted. For each figure, over-all average %I/O and total cumulative percent improvement are shown by gender.

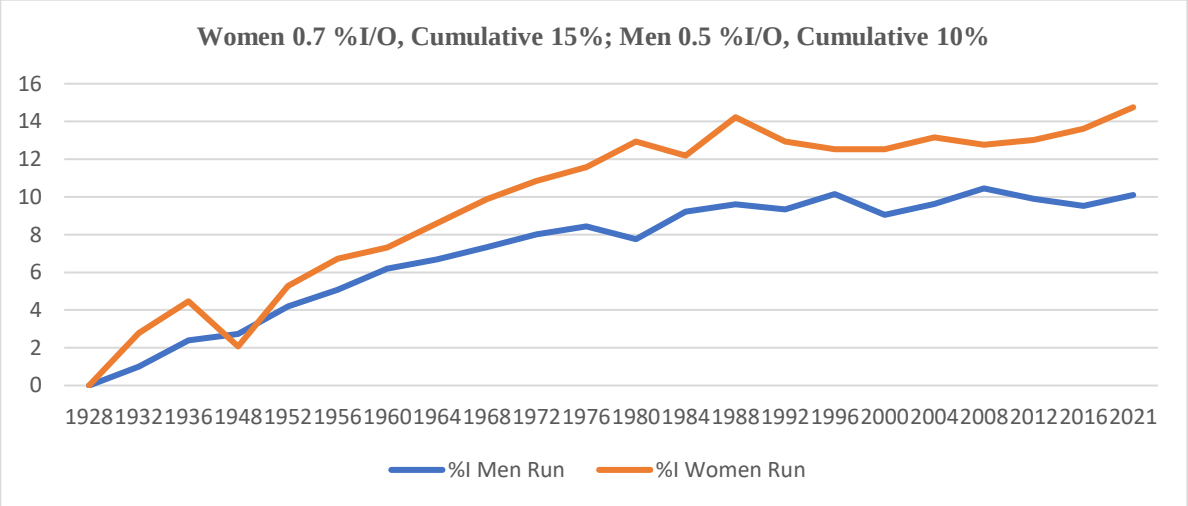


Figure 2: Cumulative % Improvement in Running for Male and Female Olympic Champions (1928-2021)

Figures 1-4 exhibit some changes in slope imposed on the athletes by various international happenings. Figures 2-4 begin in 1928 when women first competed in athletics. Although women began competing in swimming in 1912, starting with 1928 allows the comparison of men with women in running, jumping and swimming over a common time frame.

From 1928 to 1936, the build-up to WW2 featured rising nationalism resulting in a major focus on Olympics and high rate of improvement. The post WW2 Olympics of 1948 showed some reduced performances followed by a rebound in 1952. From 1956-1976 there was a Cold War emphasis on the Olympics. The Western Bloc boycotted the 1980 Olympics while the Eastern Bloc boycotted in 1984, with effects depending on which gender and bloc had been dominant in a sport. The Olympics of 1988 were fully attended, but with significant use of performance-enhancing drugs. From 1992 to the present, anti-drug efforts have been employed. Women generally improved more the men until the 1970s, as more women entered sports and nutrition, training, equipment, coaching and technique equalized compared to men until the rate of improvement of women and men equalized after the 1970s. For Figures 1-4, notice how athletes accepted those driving forces as each new normal, causing another upward movement of achieved performances due to self-improvement.

The improvements in running as shown in Figure 2 are less than those for jumping and swimming because running is dominated by working directly against gravity while jumping and swimming have more techniques to work with because the physics and kinesiology are more complex, Stefani (2008, 2014). In running, women

achieved an average percent improvement per Olympiad (%I/O) of 0.7%, creating a cumulative improvement of 15% while men achieved an average %I/O of 0.5% with a cumulative improvement of 10%. The 5% cumulative difference represents the gap women closed from 1928 through the 1970s.

If photographs are Googled for Elizabeth Robinson (USA) who won the women's 100m in 1928 and Shelly Ann Fraser-Price who was 12% faster while winning in 2008, 80 years later, one sees that the physicality is very similar as was true comparing Weissmuller with Hackett.

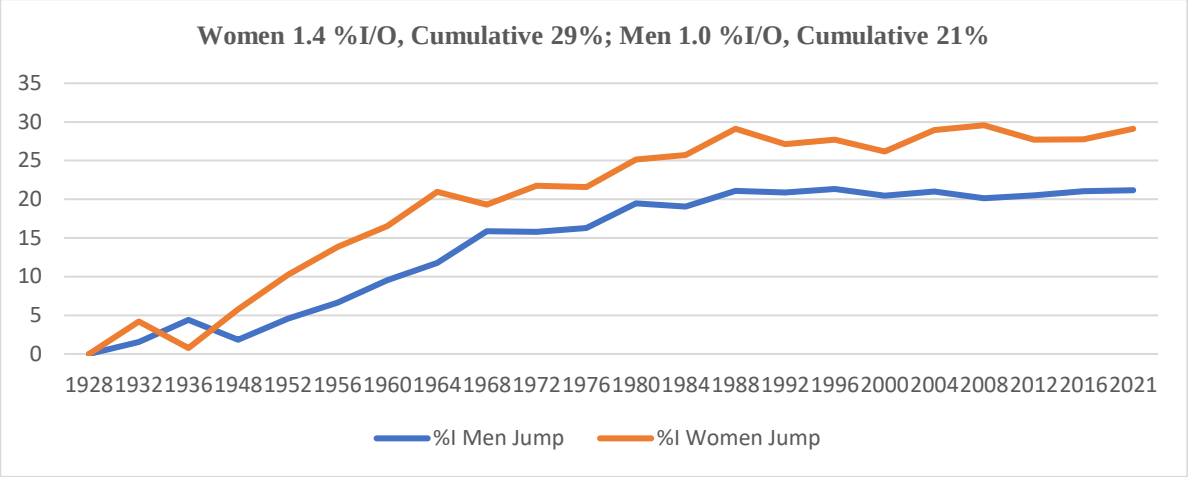


Figure 3: Cumulative % Improvement in Jumping for Male and Female Olympic Champions (1928-2021)

Figure 3 shows that in jumping, women achieved an average 1.4% %I/O with a cumulative improvement of 29%, while men achieved an average %I/O of 1.0% with a cumulative improvement of 21%. Those jumping values are essentially twice those from running. According to the laws of physics, in jumping, kinetic energy, depending on the square of the vertical component of velocity, is converted into potential energy, depending on the increase in the height of the centre of gravity. That increased height creates the vertical height or horizontal distance achieved by the jumper. If velocity is multiplied by (1 + i) for a small increase i, then the jump should increase by the square, $(1 + i)^2$, which is $(1 + 2i + i^2)$. Since i is small, i^2 can be ignored. The laws of physics therefore indicate that an increase of i in velocity should increase the jump by 2i, or twice as much, which is the relationship between the improvement parameters in Figures 2 and 3.

It is instructive to Google photos of Ethel Catherwood (Canada) the 1928 high jump winner and Tia Hellebaut (Belgium) who won in 2008 with a 29% higher jump. Catherwood vaulted the bar, while Hellbaut used the much more efficient Fosbury Flop. Both athletes have similar physicality. A jumper who vaults the bar must drive the centre of gravity well above the bar while with the Fosbury Flop, the centre of gravity only rises to the height of the bar of perhaps a bit lower, clearly a more efficient use of energy.

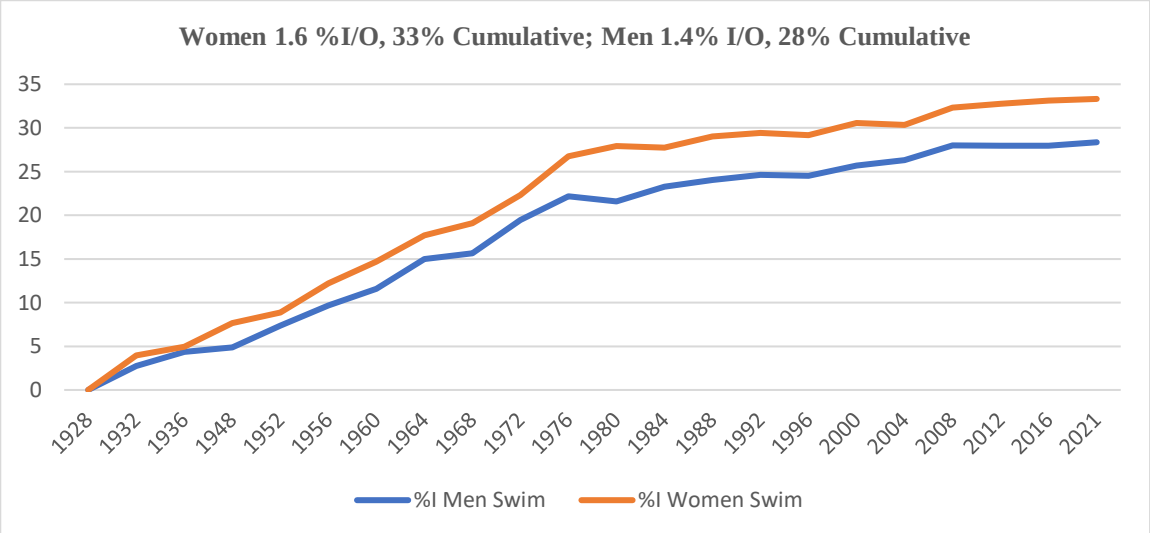


Figure 4: Cumulative % Improvement in Swimming for Male and Female Olympic Champions (1928-2021)

In Figure 5, female swimmers achieved an average 1.6 %I/O with a cumulative improvement of 33% while the figures for men were 1.4% and 28% respectively. The laws of hydrodynamics and kinesiology provide more techniques to exploit for improvement compared to jumping and running, Stefani (2008, 2014). If photographs are Googled for Ethel Lackie, the 100m champion in 1924 and Jodie Henry (Australia) who won in 2004, swimming 26% faster, again we see little difference in physicality Since there is no visible stress while performing so much better than years before, we can be confident of continued improvement.

THE EFFECTIVENESS OF TECHNOLIGIAL ADVANCES

%I/O can be used to evaluate the relative effectiveness of a technological breakthrough for an event or events in a given sport. The second panel of Table 1 shows the average %I/O before application of each of four breakthroughs, starting with 1956 when the effects of WW2 had ended. The third panel shows the %I/O for the Olympics when the breakthrough was first used. The fourth panel shows the percent increase in %I/O on first use.

The rowing ergometer, used for training, was by far the most effective, having increased %I/O by 508% on first use. International-competition exists just using the rowing ergometer. It is of interest to Google photographs of Jack Beresford (Great Britain), the 1924 winner at 2000m single sculls and Olaf Tufte (Norway) the 2004 winner who was 13% faster. Tufte could train 24/7 in hostile Norwegian winter weather and became Olympic champion in a country known for Winter Olympic excellence. Both athletes have similar physiques. The second most effective breakthrough is the fibreglass pole (419%) which allows more kinetic energy to become lifting potential energy, because the vaulter can run faster before bending the more flexible fibreglass pole. The Fosbury Flop (83%) and Clap Skate (58%) are also noteworthy.

Technological Breakthrough	Average %I/0 1956-Before	%I/O on First Use (Year)	% Change
Rowing Ergometer (Rowing Training)	1.22	7.42 (1980)	508
Fiberglass Pole (Pole Vault Equipment)	1.64	8.51 (1964)	419
Fosbury Flop (High Jump Technique)	2.25	4.12 (1968 M, 1972 W)	83
Clap Skate (Speed Skating Equipment)	1.82	2.88 (1998)	58

Table 1: Technological Breakthroughs and Their Effectiveness at Improving Performances

3. THE TOKYO OLYMPICS

One of the greatest challenges to the athlete mindset of accepting the new normal and simply trying to improve was the period before the Tokyo 2020/21 Olympics under COVID-induced restrictions. Training was interrupted. Competitions were cancelled. The Tokyo 2020 Olympics was delayed by one year to 2021. To further negatively affect the best intentions, there was a very real threat as the athletes left for Tokyo that the Games would be cancelled. On one hand, it would have been understandable if the winners at Tokyo performed worse that at Rio 2016. On the other hand, is the past history we have seen of the cumulative effect of self-improvement over a wide range of past international conditions (new normals) imposed on athletes.

	Tokyo 2020/21 vs Rio 2016	Rio 2016 vs London 2012	London 2012 vs Beijing 2008
Swimming	0.29%	0.18%	0.21%
Athletics	0.65%	0.31%	-0.26%

Table 2: Olympic Champion Percent I/O for Swimming and Athletics for Each of the Last Three Olympics

According to Table 2, the athletes met the challenge, Stefani (2021). Swimming winners performed an average of 0.29% better at Tokyo than at Rio 2016. Athletics winners, in athletics events held on the track, performed 0.65% better than at Rio 2016. Not only did the swimming and athletics winners do better than at Rio, both of those improvements were more than achieved under normal conditions by the Rio 2016 winners compared to London 2012 and more than achieved by the London winners in 2012 compared to Beijing 2008.

The Olympic motto for the winners at Tokyo 2020/21 should be Faster, Higher, Stronger-Truly Inspirational. I suggest the score at Tokyo was Athletes 1-COVID 0.

4. THE MENTAL HEALTH OF TOP ATHLETES

Recently, mental health degradation among top athletes has been widely discussed when tennis player Naomi Osaka withdrew from some competitions, citing mental health issues, while being so widely respected that she was chosen to light the Olympic Torch for Japan. The American gymnast Simone Biles withdrew from some events at the Tokyo 2020/21 Games, citing mental health issues.

Michael Phelps, having openly discussed his own mental health issues, produced an in-depth documentary called *The Weight of Gold*, Phelps (2020). Besides Phelps' participation, the following 11 highly successful athletes took part.

Jeremy Bloom, Alpine Skier	Steven Holcomb, Bobsled	Jeret Peterson, Alpine Skier
David Bodia, Diver	Lolo Jones, Hurdler and Bobsled	Katie Uhlaender, Skeleton Sled
Sasha Cohen, Figure Skater	Bode Miller, Alpine Skier	Shaun White, Snowboard
Gracie Gold, Figure Skater	Apolo Ohno, Short Track Skater	

Each one of them shared a common experience in the documentary. They each said that when newly-emerging athletes began to perform about as well as they had, each experienced loss of self-worth, leading to depression. Ironically, they did not conceptualize that these new athletes were just part of the same continuity of sport by which these superstars had themselves become established. For example, Michel Phelps has earned 28 Olympic medals, 23 of which are gold. The second most gold medal earner has won 9, yet Phelps said he never felt accomplished.

A number of them said that when their sports careers ended, so did the stipends by which they had lived while competing. They said they felt lost without having acquired another profession,

In addition to the nutritionists, trainers, equipment procurers, coaches, technical analysis and sports psychologists, all of whom work to help the athletes perform at their best, counsellors are needed to deal with the athletes' self-worth while in competition and with transitional planning for the future.

.

5. CONCLUSIONS

It is the mindset of young athletes living the real meaning of *Citius, Altius, Fortius* to try to improve, each using the methods of their day (their new normal) each aided by similarly-minded nutritionists, trainers, coaches, analysts and equipment makers. Winning performances keep getting better, simply due to the effects of that self-improvement, across the spectrum of competition from the lowest to very highest at the Olympics. As improvements accumulate over time, winning performances occur that would have been considered super-human in the past. Mental health issues are of concern. Counsellors are needed to deal with the athletes' self-worth while in competition and with transitional planning for the future.

References

IOC (2002). Olympism and Sport for All. Olympic Review, August-September 2002, 37-39.

IOC (2021). The Olympic Motto, 20 July 2021, https://olympics.com/ioc/olympic-motto .

Phelps, Michael (2020). The Weight of Gold. HBO Documentary, https://www.hbo.com/documentaries/the-weight-of-gold.

Stefani, R.T. (2008). The Physics and Evolution of Olympic Winning Performances. Chapter 3, Statistical Thinking in Sports, Chapman and Hall/CRC Press, 2008.

Stefani, R.T. (2014). Understanding the Velocity Ratio of Male and Female Olympic Champions in Running, Speed Skating, Rowing and Swimming. Proceedings of the 12th Australasian Conference on Mathematics and Computers in Sport, Darwin Australia, 25-27 June 2014, 106-111.

Stefani, R.T. (2021). The results are in from the Tokyo Olympics: Athletes 1, COVID 0. Significance Magazine Online, https://www.significancemagazine.com/sports/708-the-results-are-in-from-the-tokyo-olympics-athletes-1-covid-0, 7 September 2021.

What makes a good polo mallet? -vibrational characteristics using a cantilevered visco-elastic beam model

Ken Louie [a,c] , Paul Ewart [b]

[a] *Waikato Institute of Technology, Hamilton, New Zealand*
[b] *Waikato Institute of Technology, Hamilton, New Zealand*

[c] *Corresponding author: ken.louie@wintec.ac.nz*

Abstract

The shafts of most traditional polo mallets are made from the climbing stems of the rattan palm plant. This plant is in serious decline due to previous decades of neglect, habitat loss and exploitation for furniture. Some attempts have been made recently to replace the shaft of the traditional wooden mallet with human constructed materials. However, elite players comment that these modern mallets lack the "playability" and "feel" of the traditional mallets. Moreover, there is concern that use of modern composite materials could lead to more injury amongst the players' horses as they are sometimes struck on the stroke wind-up or follow-through. In this research, we make an initial enquiry into the vibrational characteristics of the traditional mallet shaft. This is done by modelling the shaft as a cantilevered visco-elastic beam and then representing this as a spring-damper-mass system. The oscillation frequency and decay parameters of this simplified system are compared with experimental data and this allows determination of the visco-elastic constant of the shaft without the mallet head. Good agreement with experimental data for oscillation frequency and decay is obtained when heads of different masses are then fixed to the same shaft. It is hoped that a sound theoretical understanding of the shaft's mechanical properties which determine its vibrational characteristics will lead to improved design of artificial shafts. This should also result in better player acceptance of these shafts and improved animal welfare.

Keywords: Polo mallet, vibrational analysis, visco-elastic beam model

1. INTRODUCTION

The gold standard for field polo mallets is fabricated from a combination of materials. The choice of material for: the head of the mallet is tipa, a south American hard wood; the shaft is manau cane, a genus of the rattan palm, and the handle, which is a shaped wood laminate, bound with a cloth backed rubber compound (Woods, n.d.).

Dwindling supply of 'quality' manau cane and the high variability of the root stock used for mallet shafts in general has seen the quality and availability of gold standard mallets decline. This had led makers to consider the use of other materials for the mallet shaft, including other species of cane which are considered inferior due to inconsistent performance required for high level match play (Woods, G. personal communication, July 21, 2021).

One approach to ensure that both the level of performance and the consistency of product is met is using engineered fibre composite materials. As with many sports, over the last 30 - 40 years, use of composite materials has proven to be well received, especially with the savings in weight corresponding to improvements in athlete performance (Easterling, 1993), (Jenkins, 2003). While this approach has worked for many sports there are aspects of engineered materials that may negatively affect their adoption. This includes player perception or "feel" of the equipment in use, which is a subjective consideration that is not easily quantified, as what constitutes good "feel" can differ from one player to the next (Steele, Jones, Leaney, 2007), (Curtis, Heller, Senior. 2021). Objective considerations include vibrational characteristics, strength, stiffness, impact response and general material properties (Jones, Betzler, Wallace, Otto, 2019). While the objective considerations can be readily assessed what we cannot currently determine is how these material properties combine to produce the response or feel desired by the player.

Another aspect which is not easily determined and unique to field polo relates to the complex interactions seen within the sport. For most sports activities where a ball (ball, puck, shuttle) is struck with a bat (bat, mallet, club, racket) the bat may contact the ball or the ground. With polo there is the added interaction of the horse and proximity of the opposing team and hence the bat may contact the ball, the ground, the horse, or an opposing player.

While it is theoretically possible that composite materials can be engineered to produce superior player-ball interactions, consideration also needs to be given for the horse and player interactions. This consideration, for

the other interactions, is necessary to ensure that any equipment developed does not produce unsafe conditions for the horse or an opposing player (Federation of international polo, 2018).
The gold standard polo mallets have been accepted for many years as they are universally considered to be safe for both horses and players under the current rules of the sport (Federation of international polo, 2018). Research is also underway to determine criteria that will ensure any future materials developed be as physically safe and perform comparably, if not better, than the gold standard during match play.
This paper presents an investigation that sees testing of the gold standard mallet to collect vibrational response data. These material properties will then be used to empirically validate the mathematical model produced to predict the dynamic response based on a cantilevered visco-elastic beam.

2. METHODS

Experimental set-up

A gold standard mallet (George Wood, Wood Mallets, Hawkes Bay, New Zealand) was clamped by the handle and then a load progressively applied, using the Lloyd LR30 universal tester, at a rate of 500 mm/min until the vertical displacement was 200 mm. The load was then released, and the dynamic response of the shaft measured by a three-axis accelerometer (AX3, 3-Axis Logging Accelerometer. Axivity Ltd, UK). The x-axis is taken along the shaft (positive direction away from the handle), the y-axis perpendicular to the shaft (in the horizontal plane) and the z-axis in the vertical direction (positive direction upwards). The sampling rate of the accelerometer was 100 Hz and the measurable range $\pm$ 16 g, where g is the gravitational acceleration constant 9.81 m/s^2. The test was repeated 5 times after each loading once the shaft had come to rest.

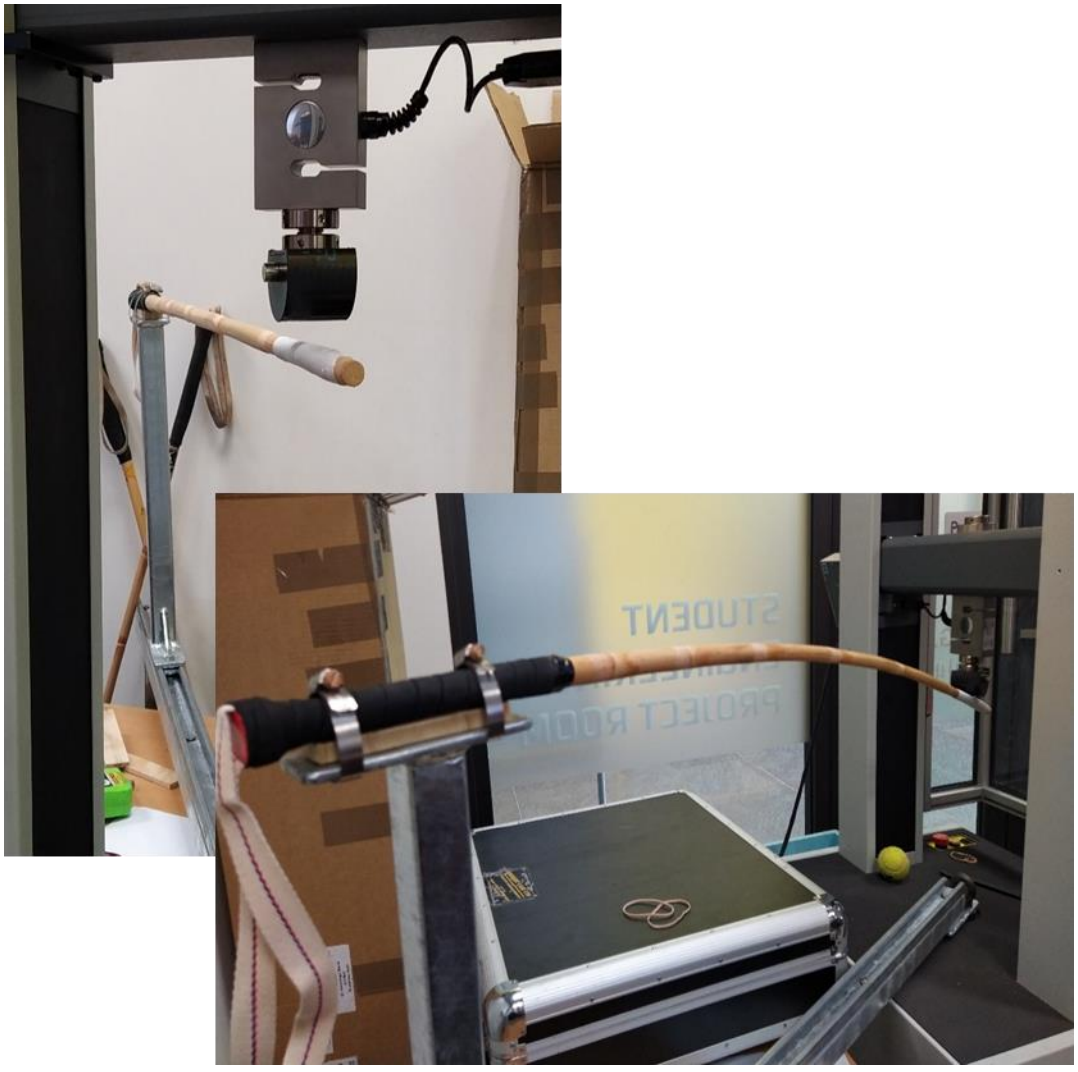


Figure 1: The mallet shaft before (top) and during loading (bottom)

The Fundamental Model

We modelled the mallet shaft as a cantilevered visco-elastic beam (Gürgöze, Doğruoğlu & Zeren, 2007) with the assumption that its visco-elastic properties fit the Kelvin-Voigt model (Perkins & Lach, 2011). The bending rigidity, length, mass per unit length and visco-elastic constant of the beam material are EI, L, m and α respectively.

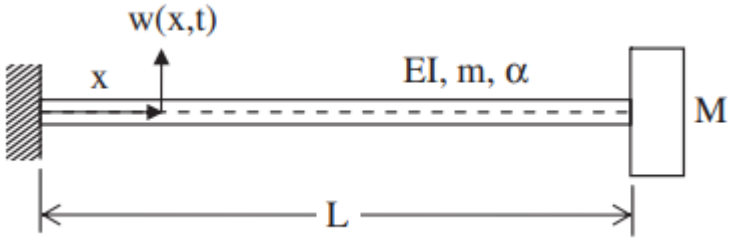


Figure 2: Fundamental model of the polo mallet (in the experimental set-up we removed the head so M = 0)

The equation of motion for this clamped visco-elastic beam can be obtained from the literature (Banks & Inman, 1991).

$$EIw^{\mathrm{IV}}(x,t) + \alpha I w^{\mathrm{IV}\cdot}(x,t) + m\ddot{w}(x,t) = 0$$

where I is the moment of inertia of the beam section, $w(x,t)$ represents the lateral displacement of the beam at the location x and time t, and primes and dots denote partial derivatives with respect to x and t.
Of the four parameters involved in the governing equation of motion, the visco-elastic constant α is most difficult to measure directly, so in this paper we use the data obtained from the experiment to "tune" this value. Then we can predict the dynamic response of the mallet when heads of different masses are added to compare with future experimental data. However, rather than numerically solving this complicated partial differential equation, an alternative approach which involves representing the fundamental model by an "equivalent" simplified model is now described.

The Simplified "Equivalent" Model

Gürgöze (2005) has shown that it is possible to represent the vibrational system in Fig. 2 by an "equivalent" simplified spring-damper-mass system as shown in Fig. 3. In the simplified system the value of the spring constant and damping coefficient must be taken as

$$k = \frac{3EI}{L^3} \text{ and } c = \frac{3\alpha I}{L^3} \qquad (1)$$

respectively. The parameter δ is the ratio of the beam mass (mL) to be added to the head mass (M) and is determined by "matching" the first eigenvalue of the system represented in Figure 2 with the eigenvalue obtained from the equivalent system in Figure 3. Further details on this "matching" procedure can be found in Gürgöze, Doğruoğlu & Zeren, 2007.

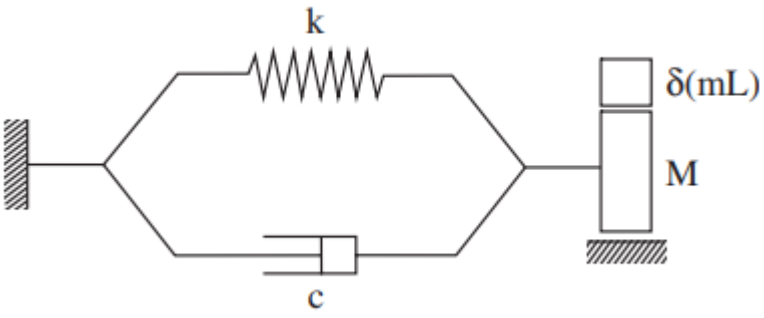


Figure 3: Equivalent spring-mass-damper system that represents the model in Figure 2

3. RESULTS

Experimental Data

In Figure 4 we plot the vertical (z-axis) acceleration in terms of g over one run. Note that the oscillation of the shaft continues for about 12 seconds and the peak values for acceleration exceed the range of the sensors ($\pm 16\ g$) at the start of the run. As expected, because of internal dampening in the shaft, the amplitude of the oscillation decays until a final value of $-g$ (representing downward acceleration due to gravity) is reached.

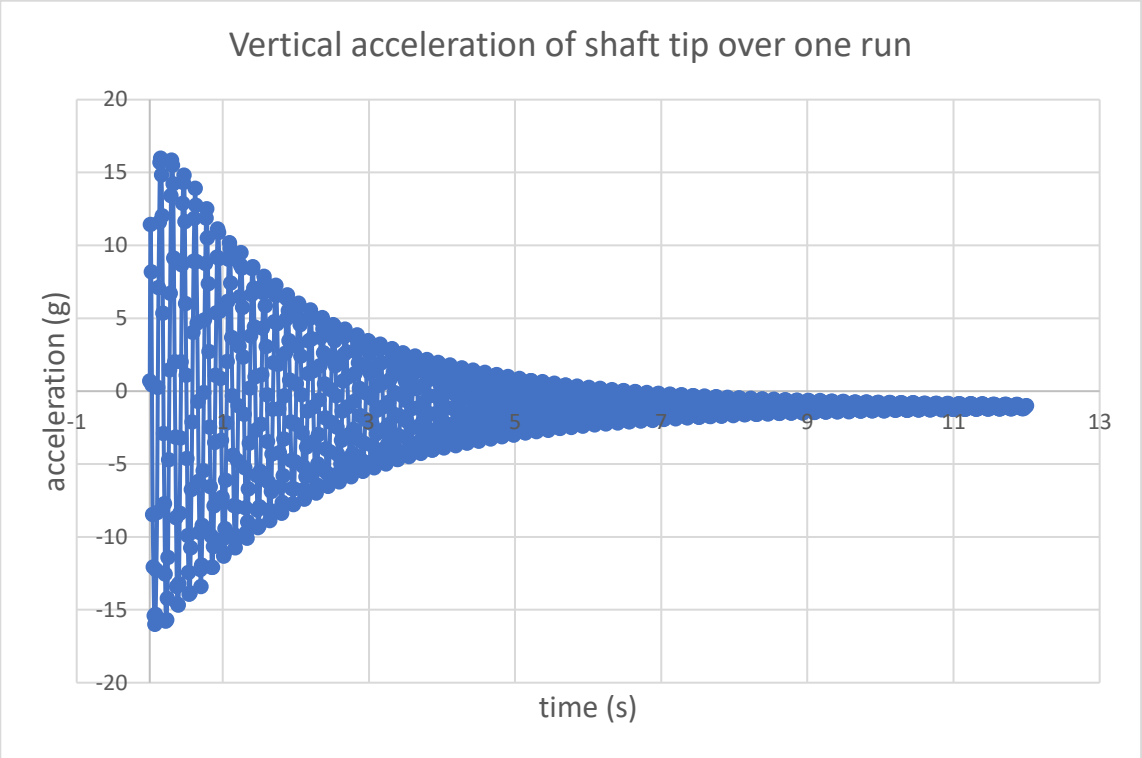


Figure 4: Vertical acceleration of the shaft tip after the load is removed

In Figure 5 we focus on the first second of the oscillations for the second run. These show some discrepancies in the acceleration values between sampling points in the initial period. For example, there is a decrease in acceleration from 15.8 g to 12.6 g between times 15.63 s and 15.64 s but this increases again to 15.85 g at the next sampling point of 15.65 s. This indicates that the sampling rate of 100 Hz, or the range of the sensor, may be insufficient to accurately calculate the acceleration. However, as time progresses these unexpected fluctuations appear to be resolved as seen from the smoother peaks and troughs from $t = 16.2$ s onwards.

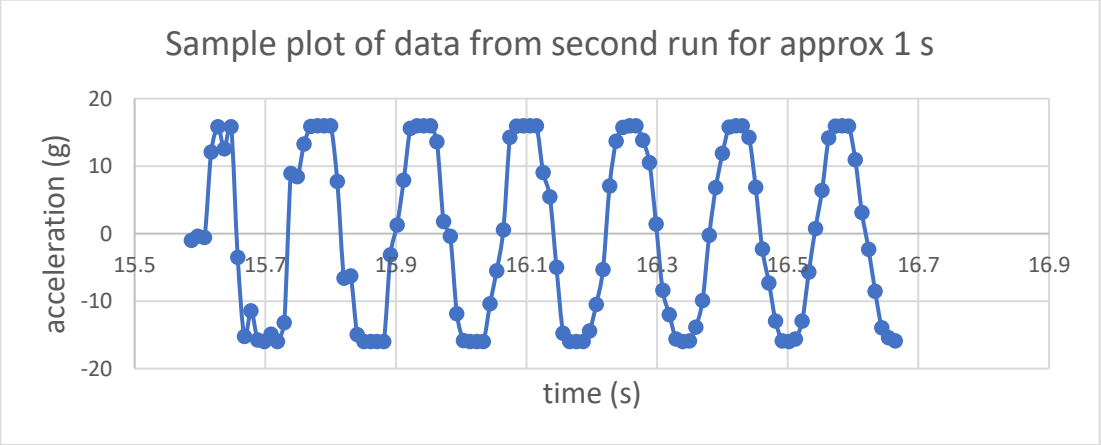


Figure 5: Vertical acceleration of the shaft tip in the second run after the load is removed

Analysis of simplified "equivalent" model

To proceed with analysis of the model in Figure 3 we need estimates of the parameters k, c and mL. As mentioned previously, the visco-elastic constant α is very difficult to measure directly, so in this analysis we tune the parameter c (which depends on α) so that results from this simplified model match the experimental data. We take values for the shaft of

$$L = 1.1 \text{ m}, I = 7.85 \times 10^{-9} \text{m}^4, E = 4.4 \times 10^9 \text{ Pa, and } mL \text{ (shaft mass)} = 0.2 \text{ kg.}$$

The value for moment of inertia I assumes the shaft is of uniform circular cross-section with radius $r = 0.01$ m and the Young's modulus value for E assumes it is made of rattan cane (Du & Wang, 2016). These values used in equation (1) give $k = 78 \text{ Nm}^{-1}$ and $c = 1.77 \times 10^{-8}\alpha \text{ Nsm}^{-1}$. The value for δ is obtained from equation (20) in Gürgöze, Doğruoğlu, & Zeren, (2007) which evaluates as 0.243. The simplified model to be solved is then

$$(0.243)(0.2)\ddot{w} + 1.77 \times 10^{-8}\alpha\, \dot{w} + 78w = 0 \tag{2}$$

or, on dividing through by the coefficient of the second derivative (0.0486),

$$\ddot{w} + 3.64 \times 10^{-7}\alpha\, \dot{w} + 1\,605w = 0 \tag{3}$$

with the value of the visco-elastic constant α to be chosen so that solutions match the experimental data. We are using w here to represent the lateral displacement of the shaft so the initial conditions for the simplified model should be $w(0) = -0.2$ m and $\dot{w}(0) = 0$. However, as noted in the experimental data, the early values for the acceleration (and hence displacement) should be treated with caution because of the insufficient sampling rate, so we focus mainly on matching the period and decay rate of the vibrations rather than exact time correspondence. Solutions of equation (3) will be damped sinusoidal curves with period $\frac{2\pi}{B}$ (Bronson, 1973) where

$$B = \frac{\sqrt{4(1\,605) - (3.64 \times 10^{-7}\alpha)^2}}{2} \tag{4}$$

and damping factor e^{-At} where $A = \frac{3.64 \times 10^{-7}\alpha}{2}$. From Figure 5 the period of the oscillations is approximately 0.16 s hence $B \approx \frac{2\pi}{0.16} = 39.3$. Solving equation (4) for the visco-elastic constant gives a value of $\alpha = 4.27 \times 10^7$ and a value of $A = 15.6$. Hence over each period of the oscillation, the model predicts that the peak value will decrease by a factor of $e^{-15.6(0.16)} \approx 0.08$. This decline in the peak values is clearly too rapid so this indicates that this estimate of α is incorrect. If instead we choose a value for $\alpha = 3 \times 10^6$ then the values for A and B are 0.56 and 40 respectively, leading to a period of 0.157 s and damping factor (over one period) of $e^{-0.56(0.157)} \approx 0.92$. This is clearly in much closer agreement with the data as shown in Figure 6, which shows the second derivative of w (acceleration, in units of g) with time.

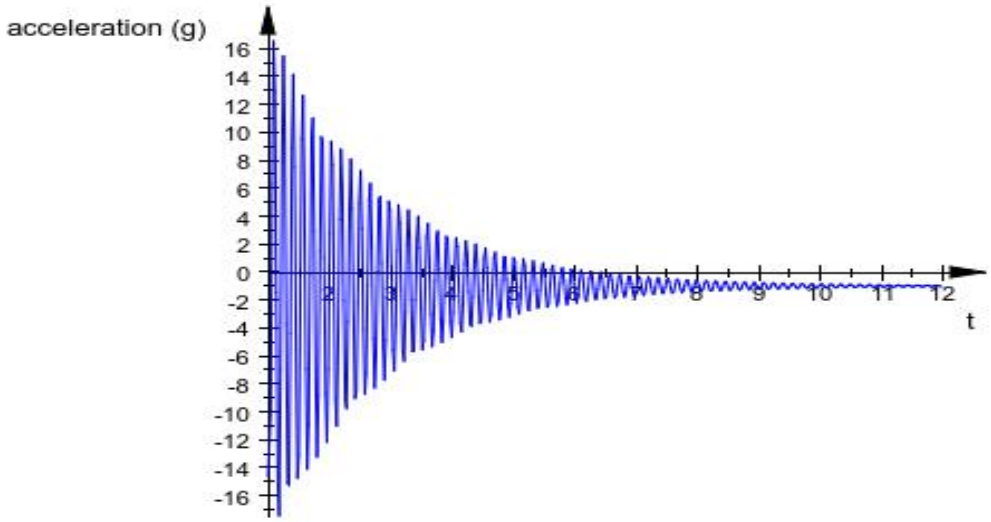


Figure 6: Results from analysis of the simplified "equivalent" model in equation (3)

For comparison purposes with the data, we also plot the model solution from $t = 3$ to 5 s. This clearly shows the period and damping rate with these parameter values is consistent with the data.

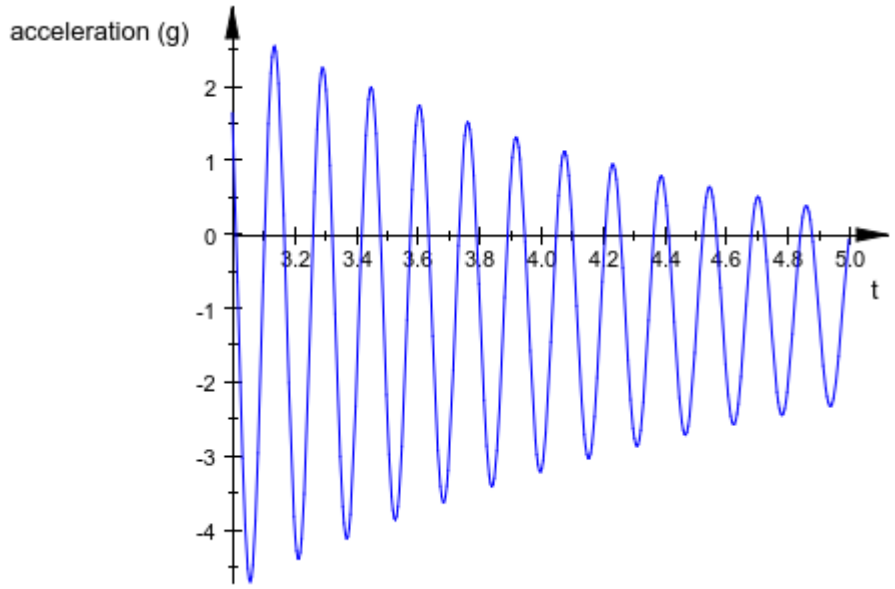


Figure 7: Close-up of simplified model solution from t=3 to 5 s showing good agreement with period and damping factor of the data.

4. DISCUSSION

Replacing the fundamental model of the mallet shaft with the simplified "equivalent" model allows more comprehensive understanding of the role that the material properties play in the resulting vibrations. Equation (4) shows that the impact of the visco-elastic constant is much greater on the damping factor than the period of vibrations. For example, a ten-fold decrease in α leads to only 1.5% decline in the period but an approximate ten-fold increase in the damping factor. The simplified model will also allow similar sensitivity analyses of the other material parameters E, I, m (mass per unit length) and α. This is useful as the values of these parameters would vary quite widely because of natural variability in the rattan cane used in the shaft. Knowing how the vibrational response of the shaft depends on the parameters will also lead to more efficient choice of new composite materials, whose parameter values can then be chosen to provide the appropriate behaviour.

As a further example of the use of the simplified model, we examine the effect of including the mallet head, which we take to be twice the mass of the shaft (so $M = 0.4$ kg in Fig. 3). Using Table 1 and Equation (20) from Gürgöze, Doğruoğlu, & Zeren, (2007) the new value of $\delta = 0.236$ and fixing all other parameter values as before, equation (2) now becomes

$$[(0.236)(0.2) + 0.4]\ddot{w} + 0.053\,\dot{w} + 78w = 0 \tag{5}$$

The acceleration from solving (5) subject to $w(0) = -0.2,\ \dot{w}(0) = 0$ is shown in Figure 8 from $t = 0$ to 10 s. In this case, the period has increased to 0.48 s and the damping factor over each period is reduced to $e^{-0.0593(0.48)} \approx 0.97$ which means the vibrations will persist for much longer.

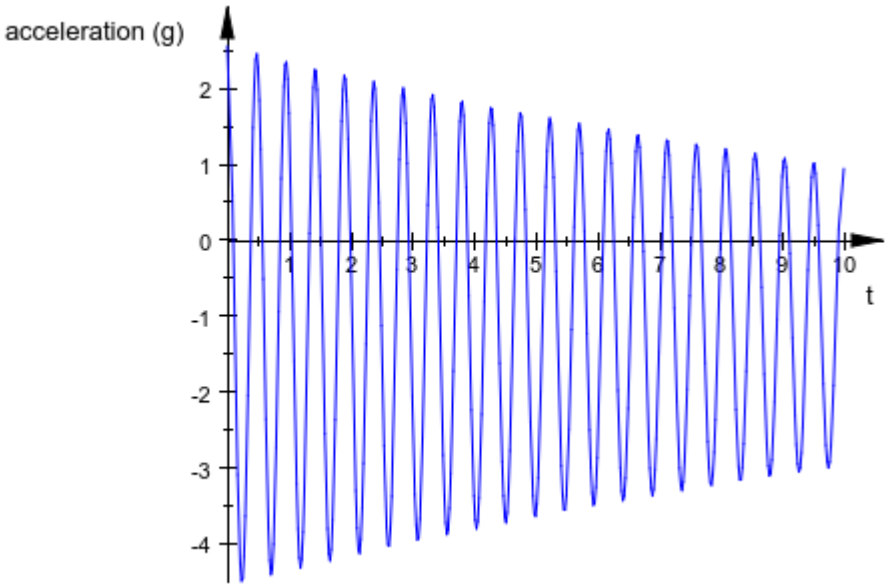


Figure 8: Effect of adding a head mass of $M = 0.4$ kg to the previous shaft

Whilst analysis of the vibrational characteristics of the polo mallet is important, it is only a first step in the classification and design of future mallets which might be made from composite materials. Also requiring study are the impact and reaction forces when striking the ball, which are very different from the quasi-static loading and clamped handle considered in this paper.

5. CONCLUSIONS

By using a simplified model of a visco-elastic beam available in the literature, we can determine the vibrational response of a polo mallet shaft with variable material parameters (Young's modulus, moment of inertia, mass per unit length and visco-elastic constant). We found that the visco-elastic constant has little effect on the period of vibration but greatly affects the damping factor. We can also easily predict the response when heads of different masses are attached to the shaft. Preliminary results show the period will increase and damping factor will decrease with increasing head mass. This study should lead to more informed consideration of alternative future materials for constructing polo mallets, given the increasing scarcity of traditional materials. Similar modelling approaches will hopefully also be useful when considering striking the ball with the mallet.

Acknowledgements

George Wood from Wood Mallets in the Hawkes Bay, NZ, for providing the shafts. Russ Best and the Polo Science team at Waikato Institute of Technology for sharing their player network.

References

Banks, H. T., & Inman, D. J. (1991). On damping mechanisms in beams. *Journal of Applied Mechanics, 58*, 716-723.

Bronson, R. (1973). Modern Introductory Differential Equations. Schaum's Outline Series. McGraw-Hill Incorporated.

Brashaw, B. K., Bucur, V., Divos, F., Goncalves, R., Lu, J., Meder, R., Pelle. Nondestructive testing and evaluation of wood: A worldwide research update. Forest products journal; Mar 2009; 59, 3; SciTech Premium Collection, pg. 7.

Curtis, D., Heller, B. & Senior, T. (2021) Methods for estimating moment of inertia of cricket bats. Sports Eng 24, 11.

Du D., & Wang, J. (2016). Research on mechanics properties of crop stalks: A review. Int J Agric & Biol Eng, Open Access at http://www.ijabe.org Vol. 9 No.6

Easterling, K.E. (Ed.). (1993). Advanced Materials for Sports Equipment. Chapman & Hall.

Federation of international polo (2018). The international rules for polo. (International rules of grass polo).

Gürgöze, M. (2005). On the representation of a cantilevered beam carrying a tip mass by an equivalent spring-mass system. *Journal of Sound and Vibration, 282*, 538-542.

Gürgöze, M., Doğruoğlu, A. N., & Zeren, S. (2007). On the eigencharacteristics of a cantilevered visco-elastic beam carrying a tip mass and its representation by a spring-damper-mass system. *Journal of Sound and Vibration, 301*, 420-426.

Jenkins, M. (Ed.). (2003). Materials in sports equipment. Woodhead Publishing Limited.

Jones, K.M., Betzler, N.F., Wallace, E.S., & Otto, S.R. (2019) Differences in shaft strain patterns during golf drives due to stiffness and swing effects. Sports Eng 22, 14.
Perkins, J. N., & Lach, T. M. (Eds.). (2011). Viscoelasticity: theories, types and models. Nova Science Publishers, Incorporated.
Steele, C., Jones, R. & Leaney, P.G. (2007). Internet based sensory analysis of tennis balls. Sports Eng 10, 33–47.
Wood, G. (n.d.). Croquet Mallets & Sets | Polo Mallets & Equipment - Wood Mallets. Retrieved 16 March, 2022, from https://www.woodmallets.com/

DEMOGRAPHIC DRIFT AND THE IMPACT ON JUNIOR RUGBY PARTICIPATION

Paul J. Bracewell [a,b], Oliver Hopkins [a], Emma C. Campbell [a], Tamsyn A. Hilder [a]
[a] *DOT loves data, New Zealand*
[b] *Corresponding author: paul@dotlovesdata.com*

Abstract

"The demography of New Zealand is changing, with major implications for rugby, especially for player participation (Spoonley, 2021)." To understand the implications, the challenge for sports administrators is to first quantify overarching demographic trends and then overlay and isolate the impacts this has on the participating population. Dissanayake et. al. (2020) found that socio-economic factors played a role in junior players leaving rugby in Auckland.

In this paper, this concept of demographic change in New Zealand and the subsequent impact on junior participation is extended further to utilise Explainable Artificial Intelligence (xAI). The impacts of socio-economic changes within the Wellington Region and the impact this has on junior rugby at a team level are explored.

The basis of the xAI in this instance is to meaningfully group and adapt to the latent structures for those groupings based on socio-economic attributes. Constrained spectral clustering is used to group and explain similar small areas based on attributes such as deprivation, educational attainment, civic compliance and discretionary spend. Given the dynamic nature of the underlying data sources, this segmentation is updated regularly. This enables socio-economic change to be identified in a meaningfully relevant timeframe.

To explore the contemporaneous relationship with player participation, an inverse Huff model (Ward et. al., 2020) is used to distribute players in the vicinity of each club. Exploring the change in playing numbers over a two-year period, clubs in neighbourhoods which improved socio-economically also grew in playing numbers.

These findings have important implications for growing the game in New Zealand and can be used to help identify, not only new pools of junior players, but also sponsorship and advertising opportunities.

Keywords: Spectral Clustering, Churn

1. INTRODUCTION

Physical activity has many benefits for children, including improved academic performance, better cognition, elevated mood, and increased self-esteem (Rasmussen and Laumann, 2013). Junior sport plays a large part in creating an active and healthy lifestyle for children in New Zealand. According to Ministry of Health (2021), to maintain a healthy lifestyle, children should complete "1 hour of moderate or vigorous physical activity spread over each day". Participation in junior sport allows for this recommendation to be more routinely met.

In a survey done by Sport NZ (2019), it was found that only 58% of young people met the current recommended level of exercise. Of the young people who were under the recommended level of exercise, a large number did not lack the motivation to exercise, meaning there are other factors at play. The survey revealed that gender and deprivation impacted the likelihood of a young person being in this position.

Rugby has been a popular sport in New Zealand for many years, but in recent times, the number of junior players leaving the sport has increased (Dissanayake et al., 2020). Before making policy decisions in the hope of altering the current trend, it is important to understand what is driving this movement away from the sport.

Junior rugby in Wellington ranges from under 5 through to under 13 grades. It transitions from Rippa rugby in the younger grades to large team format tackle rugby in older grades. The players can enter 1 of 20 clubs across the Wellington region, giving options for a range of geographic locations.

Factors other than a child's interest in sport contributes to whether used to calthey can participate. There are many factors that can influence the participation of children in sport, including deprivation, support, accessibility and parental sports preference (Sport NZ, 2019; Taks and Scheerder, 2006). When specifically looking at rugby it was found that weight limits (grades with a maximum player weight), team size, rugby sentiment portrayed by the media, and deprivation all play a role is influencing participation (Dissanayake et al., 2020).

A study in Germany by Steinmayr et al. (2011) investigated the relationship between youth participation in sport, and their distance from sports clubs. It showed that up to a certain point, distance did not affect the rate of participation, and beyond that point, participation fell at a linear rate. Clubs that offered different types of facilities differed from each other, among clubs that had a sports ground, participation was constant at about 50% until 2Km, after which it fell linearly to about 40% at 6Km.

DOT's Dynamic Deprivation Index (DDI) and Unique Segmentation (Us) tool reflect changes in the communities throughout New Zealand (NZ). The DDI is an extension of the Socio-economic Deprivation

Indexes (NZDep) by Otago University, a detailed and resource intensive study into NZ deprivation (Ward et al., 2019). The DDI offers deprivation data at a much higher temporal granularity than NZDEP, but at the expense of the detail in the deprivation estimates. Us looks to split NZ into 16 unique segments, based on approximately 200 variables. These 200 variables for both clustering and descriptive purposes. The variables include attributes such as deprivation, urbanality, electricity consumption and spending habits.

Building on an implementation of the Huff model by Ward et al. (2019), the study aims to propose a method of estimating the spatial distribution of junior rugby players given the number of players at each rugby club. As part of this method, a distance score function is presented as an option to implement user defined distance-participation relationships. While the player estimations cannot be verified at this stage, it is thought that the framework can be calibrated in future studies.

Using the junior rugby player estimates, the study aims to compare the DDI and Us segments as predictors of participation. Participation is measured by modelling the expected number of players in an area after adjusting for the junior population (rate of participation). The study also aims to identify trends in junior rugby participation over space and time after adjusting for changes in other variables.

2. JUNIOR PARTICIPATION DATA

Data on the number of juniors participating at each club were collected from 2019 to 2021 via publicly available draws (see: https://www.wrfu.co.nz/junior/draws/). The address for each club is known. The study used SA1 and SA2 polygon data, originally downloaded from Land Information NZ (LINZ). This was downloaded in the New Zealand Transverse Mercator (NZTM) projection and contained the respective area code for each polygon. To find the distances between each SA2 and rugby club, network distances between SA1 pairs were used. When calculating these distances, the central point of reference for each SA1 was the average of the contained meshblock centroids. The average SA1 network distance within each SA2 was then calculated to fit with the project data.

Rugby club points were created by using DOT's proprietary geocoder. The given World Geodetic System (WGS84) latitude and longitude coordinates were used to create the rugby club points, which were then projected to NZTM. The points were intersected with the SA1 and SA2 polygons to tag each club with the underlying area code. The network distances could then be joined to the club points, giving the distance between a given club and SA2. When estimating the player count per club, it was important to have the counts split out by grade. Not only because the team size varies depending on grade, but also the distribution of number of teams per grade varies across different clubs. To find an estimate of player numbers, club team counts were multiplied by the team size for the given grades. Up to the end of the 2021 season, WRFU grades were based on the age of the participant as of 1st January. Under 7 (years of age) grades and below have 7 players. Grades up to Under 11 have 10 aside. Finally, the Under 12 & 13 grades play with 15 per team. A flat 3 substitute players were added to each of these numbers.

Annual population estimates for 0-14 year-olds at a SA2 level were obtained from Stats NZ. To convert these figures to a junior rugby age range (5 to 13), the counts were multiplied by 9/14 (0.64). The SA2 junior player count estimates were calculated by multiplying $P(Area = i|Club = j)$ by the N_j, the number of players at club j.

2. DATA PREPARATION WITH THE INVERSE HUFF MODEL

To explore the contemporaneous relationship with player participation, an inverse Huff model (Ward et. al., 2020) is used to distribute players in the vicinity of each club. The steps in construction of this model for junior rugby participation are outlined at follows.

ATTRACTIVENESS

The junior population at each SA2 was used as the attractiveness parameter, A_i. This specification ensures that players are distributed to SA2 proportional to their population.

DISTANCE

Steinmayr et. al. (2011) observed that distance is not a factor in participation rates up until a certain point (threshold distance), after which, participation falls linearly. A distance score function, D_{score} was created to emulate this relationship. There are 2 parts to the function: (1) convert each distance, d, to its equivalent participation rate (%) in Steinmayr et. al. (2011), (2) convert each participation rate to a distance score (so that a lower participation rate corresponds to a higher distance). D_{score} assumes a constant (flat) rate of participation until a threshold point, T_{dist}, whereafter, participation falls at a linear rate until a maximum distance M_{dist}. Based on the Steinmayr et. al. (2011) findings of clubs with sports grounds, the following assumptions were made: (1)

New Zealand and German youth exhibit similar behaviours around sport participation and distance from clubs, (2) the participation rate is a constant 50% until 2 Km, (3) after 2 Km, participation falls linearly to 40% at 6 Km.

$$D_{score}(d) = \begin{cases} 100 - T_{rate} & \text{if } d <= T_{dist} \\ 100 - \left(T_{rate} - (T_{rate} - M_{rate})\frac{d-T}{M-T}\right) & \text{if } T_{dist} < d < M_{dist} \\ 0 & \text{if } d > M_{dist} \end{cases} \quad (1)$$

Where:
d = distance to club
D_{score} = Distance score
T_{dist} = Threshold distance (2 Km)
T_{rate} = Estimated participation rate (%) up till threshold (50%)
M_{dist} = Maximum distance included (6 Km)
M_{rate} = Estimated participation rate (%) at maximum distance (40%)
While the parameters defined in D_{score} are a good starting point, they rely on several assumptions and are likely to differ in practice. To show how sensitive D_{score} is to changes in parameters, D_{score} was run over a grid of parameters, and Huff models were run for each iteration. Knowing that sport participation tends to vary with deprivation level (Dissanayake et. al., 2020; Sport NZ, 2019), the correlation between the raw deprivation score, used to derive the DDI, and participation rate (as predicted by the inverse Huff model) were calculated. M_{dist} has a significant impact on the correlation, $R2 \approx -0.22$ around 5-6 Km compared to $R2 \approx -0.05$ at 8 Km. T_{rate} and M_{rate} had a relatively small but noticeable effect, while T_{dist} had very little impact

WEIGHTING EXPONENTS
An attractiveness exponent $\alpha = 1$ was used under the assumption that the population of an area does not change the rate at which juniors participate in sport. To retain the relationship created using the distance score function, Dscore, the distance exponent was set to $\beta = 1$.

3. DIMENSION REDUCTION USING XAI

Explainable AI (xAI) is artificial intelligence (AI) in which the results of the solution can be understood by humans. To ensure humans can understand what the AI is trying to do, the outputs must be transparent, robust, meaningful and connect to the real world. This builds trust. Consequently, an approach is outlined that dynamically adapts to latent data changes yet remains interpretable by design.

DATA PREPARATION
Here, the approach used to develop the proprietary Unique Segmentation tool (Us) is outlined (https://dotlovesdata.com/products/us/).

The purpose of the segmentation in Us involves using a variety of datasets to describe different areas of New Zealand (at an SA1 level) and cluster them into several distinct groups that can be used to represent the diverse population of New Zealand in a way that is more targetable and actionable. This required the creation of a clustering model that would take all data and cluster in a dynamic way such that the resulting clusters would be truly representative of different communities, and their evolution, across the country.

The first major step involved collecting the enormous amount of data necessary for such a task. Many datasets were used, including deprivation metrics such as education, income, and housing; urbanality metrics such as proximity to urban areas and density; electoral metrics such as voter turnout, party votes, and referendum votes; spending metrics such as electricity and luxury goods; and a variety of data pulled from the New Zealand Census of Population and Dwellings (see: https://www.stats.govt.nz/topics/census).

Importantly, because each dataset is organised differently, they are adjusted to represent average values for individual SA1s. The importance of aggregating datasets to this level is so that the clustering is not based on individually identifiable data, but rather represents the small but sufficiently distinct general area. The SA1, or Statistical Area 1 scheme, was introduced by Statistics New Zealand as part of the Statistical Standard for Geographic Areas 2018 (SSGA18) (StatsNZ, 2017). SA1 is intended to allow the release of more detailed information about demographic characteristics than can be made available at the smaller meshblock level. Constructed from combinations of meshblocks, SA1s generally have a population range of around 100 to 200 people, and at most approximately 500 people.

Once the data has been aggregated to the SA1 level, the resulting dataset has around 200 columns. Each column represents a distinct feature, so it is then necessary to reduce the dimensionality of the dataset before

clustering. First, variables such as those related to cultural and ethnic groups must be omitted from the clustering dataset to avoid racial bias and potential racial targeting from the segments. If there exist differences in lifestyle and behaviour between different cultural groups, these will naturally show up throughout the other variables in the dataset, however it is important not to artificially introduce biases such as these into the model. After such variables have been omitted, the dataset is still very large, so further dimensionality reduction is applied using principal component analysis (PCA). With the size of the dataset in the initial version of Us, PCA can explain approximately 70% of the variation in the dataset using eight principal components, thereafter additional principal components explain less than 30% of the variance within the dataset. With the dataset transformed to these principal components, it is then ready to be clustered.

CLUSTERING DATA

A key challenge is constraining the number of clusters, as having two few or too many will result in clusters that are either too general or too specific, as well as constrain the size of the clusters such that each represents roughly equivalent proportions of the overall population. This attribute is often required for geo-targeting in market applications. Consequently, size constrained clustering methods are investigated.

Raykov, Boukouvalas, Baig, & Little (2016) discuss that the k-means algorithm is one of the most commonly used clustering algorithms in current use, due to its simplicity. However, this simplicity entails certain restrictive assumptions about the data, the negative consequences of which are not always immediately apparent. Zelnik-Manor & Perona (2004) also note the benefits and shortcomings of common clustering methods such as k-means, explaining that these methods typically estimate explicit models and return high quality results when the data is organised according to the assumed models. They caution that when data is arranged in more complex and unknown shapes, these models can fall short. Höppner & Klawonn (2008) additionally note that k-means generally tends towards clusters of approximately equal size, but only when the data density is uniform, again echoing that same sentiment in that it works well with the optimal dataset. But datasets can very often be less than optimal, and Höppner & Klawonn explain that clusters can become very imbalanced in their coverage of the data points, in these instances.

Zelnik-Manor & Perona (2004)'s recommendation in such instances is spectral clustering, which, rather than estimating explicit models of the data distribution, can perform a spectral analysis of point-to-point similarities within the data matrix. Li, Wang, Xu, & Yang (2018) also hold spectral clustering in high regard, regarding it as the most effective clustering algorithm due to its ability to deal with non-convex sample space distribution problems.

Figure 1. Results of clustering on dummy datasets. The first row shows the results of k-means clustering, while the second row shows the results of spectral clustering.

As demonstrated in Figure 1, spectral clustering is capable of clustering datasets that consist of more complex shapes, where traditional k-means falls short.

Li, Wang, Xu, & Yang (2018) discuss further how spectral clustering could be adapted into a constrained model. They explain that spectral clustering being an unsupervised learning method means that compared with the supervised learning it lacks the information of class labels, but that pairwise constraints can, in some cases, be obtained and encoded into the spectral clustering to get better results. With this, we can put together the Constrained Spectral Clustering algorithm:

Input. Dataset $X=\{x_1,...,x_n\}$, number of clusters C, pairwise constraints

Output. The c clusters of dataset X

1. Compute the local scale σ_i for each point $x_i \in X$ such that $\sigma_i = d(x_i, x_K)$, where x_K is the K^{th} neighbour of point x_i. NB: Zelnik-Manor & Perona explain that the selection of K is independent of scale and is a function of the data dimension of the embedding space, and suggest a value of $K = 7$, which gave good results even for high-dimensional data.
2. Form the locally scaled affinity matrix $A \in \mathbb{R}^{n \times n}$ such that $A_{ij} = exp(-\frac{d^2(x_i,x_j)}{\sigma_i \sigma_j})$ for $i \neq j$ and $A_{ii} = 0$.
3. Modify the affinity matrix A by the pairwise constraints.
4. Define D to be the diagonal matrix such that $D_{ii} = \sum_{j=1}^{n} A_{ij}$, and construct the symmetric normalised Laplacian matrix $L = D^{-\frac{1}{2}}(D - A)D^{-\frac{1}{2}}$.
5. Find the C largest eigenvectors of L and form the matrix $G = \{g_1,..., g_C\} \in \mathbb{R}^{n \times C}$, where C is the largest possible group number.
6. Recover the rotation R which best aligns G's columns with the canonical coordinate system using the incremental gradient descent scheme, and construct the matrix $Z \in \mathbb{R}^{n \times C}$ after rotating the eigenvector matrix G with R.
7. Take the alignment result Z of the top C eigenvectors and assign the original point x_i to cluster c if and only if $max_c(Z_{ic}^2) = Z_{ic}^2$. If the data is highly noisy, use this to initialise k-means clustering on the rows of Z.

CHOOSING K

Zelnik-Manor & Perona (2004) mention that when choosing the number of clusters, that the process is usually manual, and that there has been limited research as to how one might determine this automatically. They discuss analysing the eigenvalues of the affinity matrix as an intuitive solution to finding an optimal number of clusters but propose an alternative method: using the initial constrained spectral clustering algorithm with a defined maximum number of clusters C, and grading the cost of the alignment for each group number up to C, treating the largest group number with minimal alignment cost to be the optimal number of clusters. They define a cost function:

$$cost = \sum_{i=1}^{n} \sum_{j=1}^{C} \frac{Z_{ij}^2}{M_i^2} \text{ such that } M_i = max_j Z_{ij} \quad (2)$$

and Let $Z \in \mathbb{R}^{n \times C}$ is the matrix obtained after rotating the eigenvector matrix G.

Adding this into our algorithm above, starting from step 6:

6. Recover the rotation R which best aligns G's columns with the canonical coordinate system using the incremental gradient descent scheme, and construct the matrix $Z \in \mathbb{R}^{n \times C}$ after rotating the eigenvector matrix G with R.
7. Grade the cost of the alignment for each group number up to C according to the equation $cost = \sum_{i=1}^{n} \sum_{j=1}^{C} \frac{Z_{ij}^2}{M_i^2}$ such that $M_i = max_j Z_{ij}$, and set the final group number C_{best} to be the largest group number with minimal alignment cost.
8. Take the alignment result Z of the top C_{best} eigenvectors and assign the original point x_i to cluster c if and only if $max_j(Z_{ij}^2) = Z_{ic}^2$. If the data is highly noisy, use this to initialise k-means clustering on the rows of Z.

In testing, C=16 seemed to provide optimal results. This result was further checked against each model's distortion, as well as its Bayesian information criterion (BIC) (Schwarz, 1978), a criterion for model selection among a finite set of models. The BIC is formally defined as: $BIC=k(n)-2(L)$, such that $L=p(x|\theta,M)$ (the maximized value of the likelihood function of the model M), x is the observed data, n is the number of data points in x, and k is the number of parameters estimated by the model.

C=16 minimises BIC and distortion without being far beyond the point of diminishing returns. However, while this may well change in updates to the clustering as the input dataset evolves, the results indicate that this is a suitable number of clusters for our model in this version.

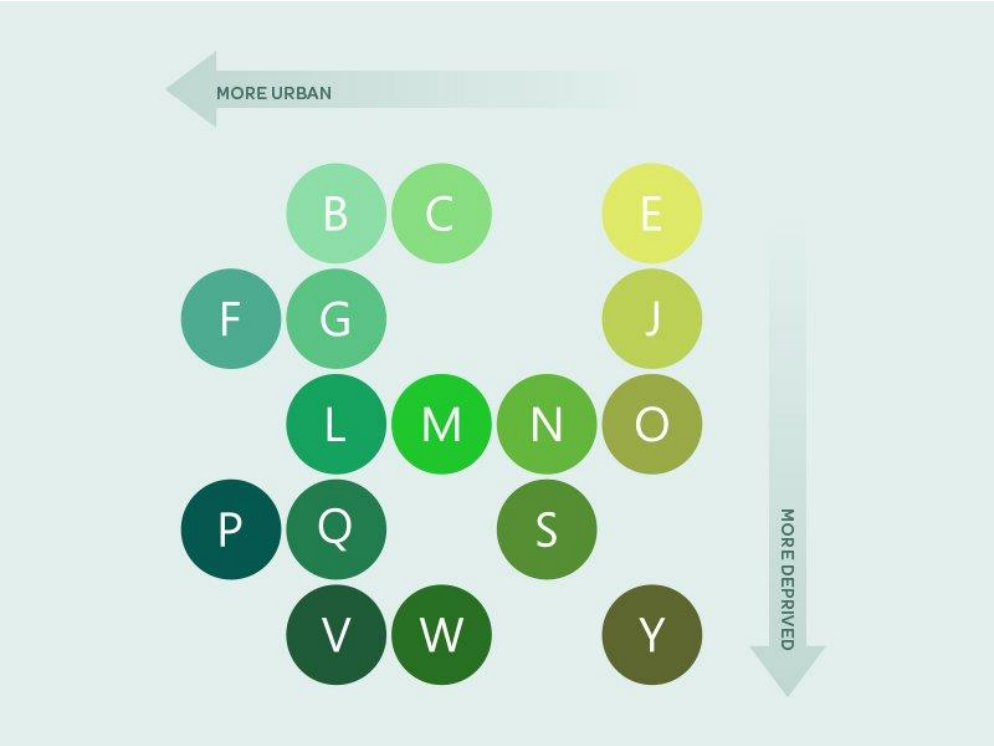


Figure 2: Overview of the DOT Us segments arranged by deprivation and urbanality.

4. METHOD

A Generalised Estimating Equation (GEE) with a Poisson link function was used to model the number of junior rugby players in each SA1. To account for varying SA1 populations, the estimated junior SA1 population was used as an offset term in the model. The SA1 code was used as the ID of each cluster in the model, while 'Year" identified the order of the repeated observations in each cluster.

Given that there are 16 clusters in the data set, it is important to be mindful of the number of parameters in each model. 3 groups of models were fitted, each corresponding to three core predictors: deprivation score (DS), DDI, or Us segment. Theses variables were not used in the same model due to the level of correlation they have with each other as DS is used to create DDI which in turn is a core component in Us.

Variables used for predicting the number of junior rugby players in each SA1 were:

DS: Deprivation score used to derive the DDI
DDI: Levels 1 to 10 from the DDI, with "10" used as reference level (most deprived)
Us: 16 Us letter codes segment "M" from 2021 used as reference level (most central segment given urbanality and deprivation)
Min Dist: The distance to the nearest club (Km)
TA: The Territorial Authority, Wellington City used as the reference level
Year: Year of the rugby season

Within each of the 3 groups, 4 models were fitted for a total of 12 models (13 including the null model). Modelling was carried out using the R package 'geepack' by Halekoh and Højsgaard (2006). Within the 'geeglm' function, 'SA1 Code' was used as the ID variable to identify repeated measures, while 'Year' was used to identify the order of the repeated measures. In each model, 'log(Junior pop)' is present as the offset term, accounting for the varying junior populations within each SA2. Models 1 to 4 were fitted for each core predictor

(DS, DDI, and Us). These models are later referred to as Model-0, Model-DS1, Model-DS2, ..., Model-US3 and Model-US4:

Model 0: log(N Players) = log(Junior Pop)
Model 1: log(N Players) = log(Junior Pop) + Core Predictor
Model 2: log(N Players) = log(Junior Pop) + Core Predictor + Min Dist
Model 3: log(N Players) = log(Junior Pop) + Core Predictor + Min Dist
Model 4: log(N Players) = log(Junior Pop) + Core Predictor + Min Dist + TA + Year

A first-order autoregressive, AR(1), correlation structure was specified, as consecutive observations were generally more similar than observations that were 2 years apart.

Using 'geepack', the quasi-likelihood under the independence model information criterion (QIC) was calculated to compare the performance of each model (Pan, 2001). For the sake of comparison between the core predictors, one DS/DDI model and one Us model were selected for evaluation.

When using GEEs, the standard observed versus predicted residuals can be used, but the variance is a function of the mean response (link function) (Fitzmaurice et al., 2011). To avoid this and make comparison easier, the Pearson residuals can be used.

5. RESULTS

Table 3 is a summary of Model-DI3 and Model-US3, showing estimates, rate ratios (RR) and the corresponding 95% confidence intervals (CI).

Variable	Parameter	Estimate	RR	RR 95% CI	Sig.
Model-DI3					
Intercept	Intercept	-3.667	0.026	(0.017, 0.038)	***
DDI	1	0.899	2.457	(1.771, 3.41)	***
DDI	2	0.839	2.313	(1.668, 3.209)	***
DDI	3	0.785	2.193	(1.565, 3.074)	***
DDI	4	0.842	2.321	(1.67, 3.225)	***
DDI	5	0.935	2.546	(1.832, 3.539)	***
DDI	6	0.848	2.335	(1.648, 3.31)	***
DDI	7	0.791	2.206	(1.524, 3.193)	***
DDI	8	0.688	1.991	(1.373, 2.886)	***
DDI	9	0.025	1.026	(0.818, 1.285)	
Min. Distance	Min. Distance	-0.099	0.905	(0.833, 0.984)	*
TA	Kapiti Coast District	0.413	1.512	(1.257, 1.818)	***
TA	Lower Hutt City	0.256	1.292	(1.043, 1.601)	*
TA	Porirua City	0.931	2.538	(1.8, 3.578)	***
TA	Upper Hutt City	0.418	1.519	(1.24, 1.861)	***
Model-US3					
Intercept	Intercept	-3.016	0.049	(0.034, 0.071)	***
US Segment	B	0.066	1.068	(0.752, 1.518)	
US Segment	C	0.481	1.617	(1.196, 2.186)	**
US Segment	F	0.059	1.060	(0.745, 1.509)	
US Segment	J	0.330	1.391	(0.962, 2.012)	.
US Segment	L	0.417	1.518	(1.074, 2.145)	*
US Segment	N	0.032	1.033	(0.772, 1.381)	
US Segment	P	0.423	1.527	(1.074, 2.171)	*
US Segment	Q	0.054	1.055	(0.663, 1.678)	
US Segment	S	0.464	1.591	(1.141, 2.219)	**
US Segment	V	-0.473	0.623	(0.428, 0.907)	*
US Segment	W	-0.809	0.445	(0.305, 0.651)	***
Min. Distance	Min. Distance	-0.070	0.932	(0.851, 1.022)	
TA	Kapiti Coast District	0.193	1.213	(0.856, 1.718)	
TA	Lower Hutt City	0.218	1.244	(0.913, 1.695)	
TA	Porirua City	0.851	2.343	(1.511, 3.631)	***
TA	Upper Hutt City	0.504	1.656	(1.168, 2.349)	**

Table 1: Model summaries showing parameter estimates and significance from Model-DI3 and Model-US3

Not all parameters were significant at a 95% significance level but were included due to the reduction QICC that they provided and assistance in interpretability. Model-DDI3 shows that DDI levels 1 to 8 are significantly different from levels 9 and 10 (which are not significantly different from each other). While the RR for each level from 1 to 8 are very similar (relative to their CI), a DDI of 5 is associated with the largest RR. This means SA1s with a DDI of 5 have 2.546 (1.832, 3.539) times the rate of junior rugby participation compared with SA1s with a DDI of 10. For each 1 Km a SA1 is from the nearest club, the SA1 has 0.905 (0.833, 0.984) times the rate of junior rugby participation. The RR for each of the TAs are significantly different from Wellington City, with Porirua City being the most different. Porirua City has 2.538 (1.8, 3.578) times the rate of participation compared with Wellington City. Looking at Model-US3, Us codes "C" and "W" tell a similar story to DDI results seen in Model-DDI3. Segments "C" and "W" are at opposing ends of the deprivation scale, while having similar urbanality (Figure 1). Compared to segment "M", segment "C" SA1s have 1.617 (1.196, 2.186) times the rate of participation, while segment "W" SA2s have 0.445 (0.305, 0.651). Interestingly, segment "S", has a RR of 1.591 (1.141, 2.219) despite being much closer to "W" than "C" in terms of deprivation. Porirua and Upper Hutt City have significantly higher rates of participation compared with Wellington City. Porirua City and has the largest difference, having 2.343 (1.511, 3.631) times the participation.

6. DISCUSSION

Model-DDI3 and Model-US3 indicate that SA2s with high levels of deprivation are associated with a lower rate of junior rugby participation. This aligns with the findings from Dissanayake et al. (2020) who found that deprivation had a negative effect of junior participation in sport. In both models it seems apparent that the relationship between deprivation and participation in nonlinear. In both cases, participation was only negatively affected at very high levels of deprivation and seemed to be similar at deprivation levels below the extreme.

In the case of the DDI and Us models, the QIC indicated that 'Year' was not useful, given that the models contained all other variables. This appears to contradict one of the motivations of the study, which was understanding drivers of the reduction in rugby participation. It may seem like a straightforward conclusion, but it would fail to account for the temporal variation considered by DS/DDI. Another consideration is that the study only contains 3 data points across time, making any evidence around this point weak. There was little variation in the Us segments at SA2 for this period

The 'TA' variable allowed for estimates to vary across space, revealing a trend that reflects negatively upon rugby participation within Wellington City. Model-DDI3 indicates that each of the 4 other TAs have a higher rate of junior rugby participation, while Model-US3 indicates the same but only in Porirua and Upper Hutt City. Demographic, socio-economic and cultural factors are known to effect participation in junior sport (Taks and Scheerder, 2006), future work may be needed to explain how this relates to the differing participation rates throughout the Wellington Region.

A key limitation of the study is that the number of players that are estimated to come from different SA2s cannot be verified. In its current state, the GEE is trying to predict the output of the IHM, as opposed a true player count per SA2. It is thought the IHM estimations can be calibrated in further studies to acquire more robust estimations.

IHM is not typically used for distributing individuals out from central points. This method is still in its infancy, having only been presented by Ward et al. (2018, 2019, 2020).

While the GEE accounts for the correlation within repeated measures of SA2s, spatial autocorrelation has not been accounted for. This adds to the caution needed when interpreting findings from this study, as parameter estimates could be inflated and standard error estimates may be optimistically small (Mets et al., 2017).

While not presented in detail here, in 2019 and 2020, the Pearson residuals have noticeable geographical pockets where the model does not fit well. This is particularly noticeable in 2021, around Khandallah, where there are areas of homogeneous negative residuals. Khandallah is a north-eastern suburb of Wellington, approximately 4km from the Central Business District. The two nearest junior clubs are Wests (6.3Km) and Johnsonville\Newlands (4.3Km). The centre of Khandallah is 550m from Nairnville Park, the home ground of Old Boys University Rugby Club, who folded their junior club in 2021. Consequently, these negative residuals indicate that this part of Wellington is systematically underserviced. This serves as a reminder to administrators about the importance of having access to clubs, with the distance parameter from the IHM helping to inform optimal distribution of clubs given population levels.

7. CONCLUSIONS

Rugby is a popular sport in NZ and a source of physical exercise in a time where youth struggle to meet physical exercise recommendations. During falling junior rugby participation, studies have identified relationships

between junior sport participation is affected by socio-economic, geographical and cultural variables. An IHM was used to estimate the number of junior rugby players in each SA2 across the Wellington region. The estimated player numbers were modelled by the DDI and Us, as well as spatiotemporal variables. High levels of deprivation were found to have a negative effect on participation rates while SA2s within Wellington City were associated with a lower rate of participation compared with the other TAs. A statistically significant trend across time was not found, but this may have been due to a lack of repeated measures. The study has some limitations around certainty of estimates but provides a start for future work.

References

Ballinger., G. A. (2004). Using generalized estimating equations for longitudinal data analysis.

Dissanayake, H., Bracewell, P.J., Trowland, H.E., & Campbell E.C. (2020). Latent drivers of player retention in junior rugby. Proceedings of the 15th Australian Conference on Mathematics and Computers in Sports. Ray Stefani & Adrian Schembri eds. Wellington, New Zealand ANZIAM Mathsport. pp. 92-97.

Fitzmaurice, G.M., Laird, N. M., & Ware. J. H. (2011). Generalized Estimating. Wiley.

Ghisletta, P., Spini, D., Riand, F. J., Vascotto, B., Cordonier, C. & Lalive., C. J. (2004). An introduction to generalized estimating equations and an application to assess selectivity effects in a longitudinal study on very old individuals. Journal of Educational and Behavioral Statistics Winter, 29:421–437, 2004.

Halekoh U., & Højsgaard, S. (2006). The R package geepack for generalized estimating equations. Journal of Statistical Software.

Huff. D. L. (1964). Defining and estimating a trading area. Journal of Marketing, 28:34–38.

Huff. D. L. (1963). A probabilistic analysis of shopping center trade areas. Land Economics, 39:81–90.

Liang K.-Y., & Zeger, S. L (2017). Longitudinal data analysis using generalized linear models. Biometrika, 73:13–22, 1986.

Mets, K. D., Armenteras, D. & Davalos, L. M. (2017) Spatial autocorrelation reduces model precision and predictive power in deforestation analyses. Ecosphere, 8, 5.

Ministry of Health. (2021) How much activity is recommended? URL https://www.health.govt.nz/your-health/healthy-living/food-activity-and-sleep/physical-activity/how-much-activity-recommended.

Pan. W. (2001). Akaike's information criterion in generalized estimating equations. Biometrics, 57:120–125.

Rasmussen, M., & Laumann, K. (2013). The academic and psychological benefits of exercise in healthy children and adolescents. European Journal of Psychology of Education, 28:945–962, 9.

Spoonley, P. (2021, May 14). Changing face of NZ poses big challenges for the future of rugby. Stuff.co.nz. Available at: https://www.stuff.co.nz/sport/rugby/opinion/125061179/changing-face-of-nz-poses-big-challenges-for-the-future-of-rugby (accessed 22 February 2022).

Sport NZ (2019). Determinants of physical activity in young people.

Taks M., & Scheerder. J. (2006). Youth sports participation styles and market segmentation profiles: Evidence and applications. European Sport Management Quarterly, 6:85–121.

Ward, A.D., Bracewell, P.J., Cui, Y. (2018). Tavern proximity, tavern density and socio-economic status as predictors of assault occurrence within New Zealand: a temporal comparison. Kōtuitui: New Zealand Journal of Social Sciences. 13(1) pp. 82-98.

Ward, A.D., McIvor, J.T., & Bracewell, P.J. (2020) The geographic distribution of gaming machine proceeds in New Zealand, Kōtuitui: New Zealand Journal of Social Sciences Online, 15(1), pp. 54-74.

Ward, A. D., Trowland, H. & Bracewell, P.J (2019). The Dynamic Deprivation Index: measuring relative socio-economic deprivation in NZ on a monthly basis. Kōtuitui: New Zealand Journal of Social Sciences, 14:157–176, 1.

Höppner, F., & Klawonn, F. (2008). Clustering with size constraints. In Computational Intelligence Paradigms (pp. 167-180). Springer, Berlin, Heidelberg.

Li, L., Wang, S., Xu, S., & Yang, Y. (2018). Constrained spectral clustering using Nyström method. Procedia Computer Science, 129, 9-15.

Raykov, Y. P., Boukouvalas, A., Baig, F., & Little, M. A. (2016). What to do when K-means clustering fails: a simple yet principled alternative algorithm. PloS one, 11(9), e0162259.

Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, 461-464.

Stats NZ (2017). *Statistical standard for geographic areas 2018*. Retrieved from www.stats.govt.nz.

Steinmayr, A., Felfe, C., & Lechner, M. (2011). The closer the sportier? children's sports activity and their distance to sports facilities. European Review of Aging and Physical Activity, 8:67–82, 10.

Wellington Rugby Football Union (2019). Junior rugby team handbook.

Wheeler., J. O. (2005). Geography. In K. Kempf-Leonard, editor, Encyclopedia of Social Measurement, pages 115–123. Elsevier, New York.

Zelnik-Manor, L., & Perona, P. (2004). Self-tuning spectral clustering. Proceedings of the 18th Annual Conference on Neural Information Processing Systems (NIPS'04).

Zorn. C. J. W. (2001). Generalized estimating equation models for correlated data: A review with applications.

MODELLING HUMAN GAIT USING A NONLINEAR DIFFERENTIAL EQUATION

Jelena Schmalz[a], David Paul[a], Kathleen Shorter[a], Xenia Schmalz[b], Matthew Cooper[a], Aron Murphy[c]

[a]University of New England, School of Science and Technology, Armidale, NSW, Australia

[b]Department of Child and Adolescent Psychiatry, Psychosomatics and Psychotherapy, University Hospital, LMU, Munich, Germany

[c]Faculty of Medicine, Nursing and Midwifery and Health Sciences, University of Notre Dame, Australia

Abstract We introduce an innovative method for the investigation of human gait, which is based on the visualisation of the vertical component of the movement of the centre of mass during walking or running, in the space of the coordinates position, velocity, and acceleration of the centre of mass. We collected data and numerically approximated the gait by the best-fitting curve for a non-linear model. The resulting equation for the best fitting plane or curve in this space is a differential equation of second order. The model that we suggest is a Duffing equation with coefficients that depend on the height of a walker or runner and on the angular frequency of the oscillation. We present statistical analyses of the distribution of the Duffing stiffness depending on the speed.

Keywords: dynamical systems, Duffing equations, non-linear differential equations, biomechanics, biodynamics, gait modelling

1. INTRODUCTION

Research on the mechanics of human gait is of interest to different disciplines, for example sport science, medicine, and robotics. In this paper, we introduce a model for the movement of the vertical coordinate of a person's centre of mass (COM) during walking and running.

Human locomotion is an inherently complicated process requiring the complex integration of neural control and musculoskeletal dynamics in response to both internal and external forces. In an attempt to strip away complexity and gain an understanding of the fundamental principles underpinning human locomotion, simple mechanical models have been developed [1, 2]. The mechanical simplification of locomotion allows the identification of just a few key parameters that can be manipulated to examine cause and effect relationships and identify which features most influence the system.

Blickhan suggested a linear spring-mass model for hopping in 1989, [3]. Other papers followed, for example [4–7]. The motion of the centre of mass is described by the equation $mz_{tt} + Kz = -mg$, where m is the body mass, z is the vertical deflection of the centre of mass with the origin on the treadmill surface and the direction chosen upwards. The constant K is the stiffness, and g is gravitational acceleration. By z_{tt} we denote the second derivative of z, i.e., the vertical acceleration of the centre of mass. There have been different approaches on how to calculate leg stiffness. Blickhan's approach uses the formula $K = m\omega_0^2$, where ω_0 is the stride's angular frequency of the oscillation, which, during gait, reflects the stride's angular frequency. [3, 6]. Another approach for calculation of the leg stiffness is to find the ratio of F_m, the maximum value of the vertical ground reaction force, and ΔL, the absolute value of the leg compression, i.e., $K = \frac{F_m}{\Delta L}$. This definition of leg stiffness is used in several publications [4, 5, 7–9; see 6 for an overview]. There is a third approach to leg stiffness calculation based on the measurements of loss of mechanical energy by walking/running, W. Leg stiffness K is derived from the formula $W = \frac{1}{2}K(\Delta r)^2$, where Δr is the shortening of a spring (e.g. [10]). In examining mechanical and metabolic determinants of the human walking gait, Kuo [11, 12] employed an anthropomorphic three-dimensional, passive-dynamic model, in which human legs were represented as rigid inverted pendulums with small point masses modelling each foot and a larger mass modelling the concentration of the COM at the pelvis. These studies drew on earlier models of a rigid swing leg during walking [13] and continued the view that walking and running were two distinct gaits that could not be

described using similar mechanical models. This view, however, was discounted by Geyer, Seyfarth, Blickhan (2006) [14] who demonstrated that a compliant-legged, spring-mass bipedal model consisting of two linear, equal and massless springs and a single COM point as an extension of Blickhan's one-dimensional model, reliably predicted ground reaction forces and COM behaviour in both human walking and running. Several subsequent studies further validated the efficacy of a bipedal spring-mass model of walking [2, 15–18]. While much of the twenty-first century research in the field has adopted the bipedal spring-mass model and focused on adapting or adding selected elements to improve prediction accuracy for both walking and running gait mechanics, Blickhan's spring-mass model remains largely valid and has been applied, with modifications to suit certain parameters, in recent studies [19, 20].

Our mathematical model is based on the analysis of three-dimensional movement of COM. We concentrate on the projection of the movement of COM on the vertical axis. We suggest a new approach for finding leg stiffness for the simple harmonic oscillation model, and then develop a more precise model assuming that the stiffness for a fixed speed is not a constant but depends on the displacement of the COM. The innovative idea of our method is to visualise the data for the vertical component of a motion of the COM, $z(t)$, as a curve in the three-dimensional space, (z, z_t, z_{tt}) (see Section 2). Here z_t and z_{tt} are the first and the second derivatives of the function $z(t)$, velocity and acceleration, correspondingly. The linear differential equation $z_{tt} + \frac{K}{m}(z - z_0) = 0$, which is a simple harmonic oscillation model of a gait, can be interpreted as an equation of the plane in the space (z, z_t, z_{tt}). Here K is the leg stiffness, m is the mass of a participant, and z_0 is the average z-coordinate of the COM by the movement. Finding the best fitting plane to the data curve gives us the slope of the plane $\frac{K}{m}$ and the value of z_0. This simple model suggests that the leg stiffness is a constant. But we can observe that the slope in general is not a constant (see for example Fig.2), but can be represented as a non-linear model with a cubic term. Thus, we use a non-linear differential equation model, approximating the curves by the Duffing equation. The best fitting curves have the form $z_{tt} + kz(z - z_0)(h - z) = 0$. Here h is the height of a participant in motion. We called the value km Duffing stiffness. The constant k is different for each participant and increases with the speed of walking/running. We first analyse how Duffing stiffness relates to speed using data for six participants collected by us, then verify the results using data publicly available for 42 walking participants and 29 running participants [21, 22].

Modelling of gait by Duffing equations emphasises the commonalities of stable walking and running. However, the curves also contain individual features for each person and velocity. We provide examples of the variety of these curves in Section 8. Studying outliers might be more interesting for sport science, because they are a sign of some anomalies and instability in gait, which might, for example, suggest an injury.

2. DATA RECORDING

We collected walking and running data from six participants (aged 18 to 55 years, 3 men and 3 women). The study followed ethical protocols as per ethics requirements (HE19- 239). We measured the vertical coordinates $z(t)$ of the COM for each participant walking or running on the treadmill. The markers were the Left and Right PSIS and ASIS; we then computed the average of all four. The data was collected for different integer velocities, at 100 frames per second, over 10 seconds, for each velocity, using an 8 camera, Qualisys Motion capture system with the COM reconstructed using a pelvic marker set within Visual3D.

Using MATLAB, we visualised the data as curves in the three-dimensional space (z, z_t, z_{tt}). We filtered out high frequency oscillations, such as noise and individual features, using the fast Fourier transform function in MATLAB. The direction of the motion along the curve can be found, for example, in the following way. Find the point with the greatest z-coordinate. This is the highest point of the centre of mass during the gait cycle. The velocity at this point is equal to zero. Hereafter, the movement goes down, i.e., the velocity z_t becomes negative. In both pictures of Figure 1, from those perspectives, the movement is clockwise.

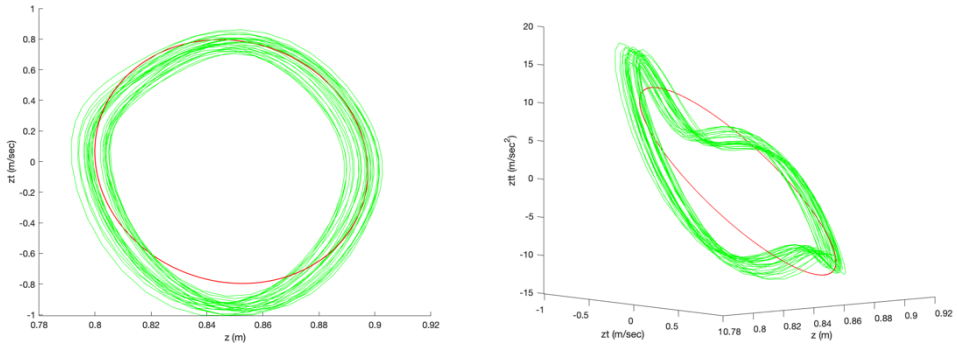


Fig 1. A MATLAB 3D figure shown from different perspectives. The green curve is a smoothed data curve with high-pass FFT threshold 0.03, the red curve is a smoothed data with FFT threshold 0.3.

3. DATA INTERPRETATION

The pictures in the space with the coordinates position, velocity, and acceleration, are rich in information. For example, Figures 2 and 3 show the data for walking (4 km/hr) and running (9 km/hr), respectively, of the same participant.

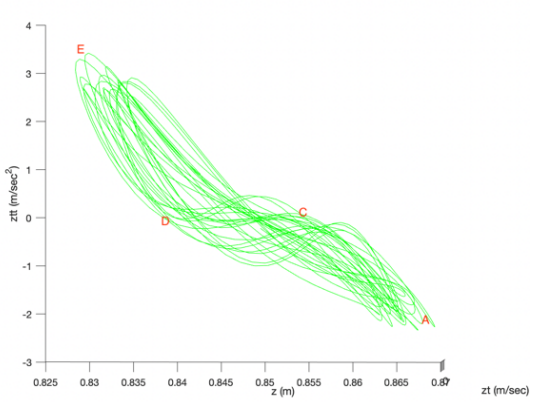


Fig 2

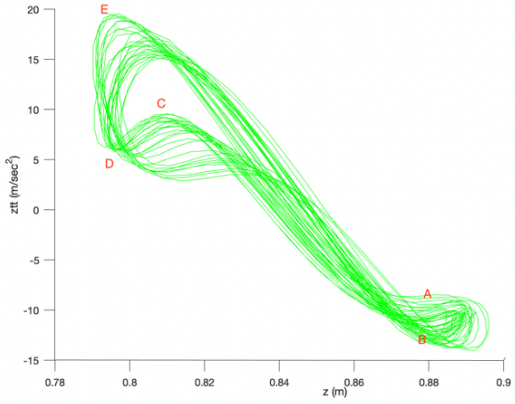


Fig 3

Figs 2 and 3 show data for walking, 4 km/hr and for running, 9 km/hr, correspondingly. The horizontal axis shows position and the vertical axis the acceleration of the COM. In both Figs, the part CD corresponds to the phase of the gait when a foot touches the surface, DE corresponds to the propulsion during toe-off, during EA the COM moves upwards and the acceleration diminishes. The arc AB appears only in Fig 3 and corresponds to the flight phase

Here we look into the projection to the plane z, z_{tt}, i.e., the horizontal axis shows the position of the COM (in meters) and the vertical axis shows the acceleration (in m/sec^2). The part CD of both curves corresponds to the phase of the gait when a foot touches the surface of the treadmill. In the case of walking, it is a flat line; vertical acceleration is close to zero. In the case of running, acceleration is diminishing because of braking during initial foot contact. The phase DE corresponds to the propulsion during toe-off, where acceleration (and consequently force) increases. On the segment EA acceleration diminishes, turns to zero when the COM reaches its average position, and is minimal at the point A. The minimum acceleration for the walking curve is -2 m/sec^2, the minimum acceleration for the running curve is about -10 m/sec^2, i.e., close to the gravitation constant g. The acceleration is less during walking than during running because the body is always in contact with the ground whereas during running there is a flight phase.

The arc AB on Figure 3 corresponds to the flight phase of running. This part of the curve is more complicated than just a flat constant $z_{tt} = -g$, because it is smoothly connected with the rest of the curve. Some information that we get from these curves is common for all participants and walking/running speeds, but some features are individual – for example not each participant has the "flight" component AB at running speeds, due to individualised transitions between walking and running gait patterns. In this paper we concentrate on their common properties, and suggest three models, based on differential equations. We build our three models based on data, purely numerically and mathematically.

4. MODELLING GAIT AS HARMONIC OSCILLATOR

By Hook's Law [23], a movement of a spring with stiffness K satisfies the equation

(1) $mz_{tt} + K(z - z_0) = 0,$

where m is the mass, K is stiffness and $z - z_0$ is the displacement of the COM. In three-dimensional space (z, z_t, z_{tt}), Eq (1) is the equation of a plane that passes through the point $(z_0, 0, 0)$. We find the coefficients in Eq. (1) numerically, by finding the best fitting plane for the smoothed data curve. The initial values specify the ellipse on this plane. The gravitation constant is included in Eq. (1) implicitly. We can rewrite the equation as

(2) $mz_{tt} + K(z - z_{st}) = -mg.$

The coordinate z_{st} is the average of the vertical coordinate of the centre of mass of a standing body but in a walking/running posture. It is not exactly the same as the coordinate of COM in a standing position. The relation between z_0 and z_{st} is calculated from Eqs. (1) and (2) and is $z_0 = z_{st} - \frac{mg}{K}$.

5. MODELLING BY A NON-LINEAR HOMOGENEOUS DIFFERENTIAL EQUATION

5.1. Best fitting curve, interpreted as a non-linear second order differential equation. Non-linear gait dynamics has been discussed, for example, in [24]. Stride-to-stride fluctuations, which are often considered to be noise, actually convey important information. To describe these fluctuations, we refine the method used for the harmonic oscillation model in section 4. We approximate the movement of the COM during walking or running by a Duffing equation, i.e., a homogeneous non-linear second order differential equation. We write

(3) $mz_{tt} + K(t)(z - z_0) = 0,$

where, unlike the harmonic oscillation model from Section 4, we consider stiffness to not be constant, but as dependent on time, $K(t)$, as stiffness and viscosity depend on a phase of a stride. For example, the slope on Figures 2 and 3 depends on $z(t)$, and the slope in the plane z, z_{tt} reflects stiffness: it is $\frac{K(t)}{m}$ at a time t. We divide both sides by m and consider the approximation of the function $k(t) = \frac{K(t)}{m}$ by polynomial,

(4) $z_{tt} + (k_1(z - z_0) + k_2(z - z_0)^2 + k_3(z - z_0)^3) = 0.$

As the value z_0 is not known beforehand, we look for the best fitting curves of the form:

(5) $z_{tt} + k_1 z + k_2 z^2 + k_3 z^3 + C = 0.$

The best fitting curve in the chosen coordinates is a differential equation of second order. Initial conditions are the values $z(0)$ and $z_t(0)$. We visualize the solution of this differential equation as a curve in the same coordinate system as the data curve. See, for example, Figure 4, where the red curve is the data curve, and the blue curve is the solution of the differential equation given by the best fitting curve. We computed a scaled

mean squared error between the Duffing equation output and the observed data. As the values z, z_t and z_{tt} have different units, we scaled each value by dividing by the difference between the minimum and maximum values on the corresponding axis.

5.2. Finding and analysing the fixed points. We are also interested in the fixed (equilibrium) points [25]. We first rewrite the second order differential Eq. (5) as a system of first order differential equations:

(6) $z_t = z_1,\ (z_1)_t = -k_1 z - k_2 z^2 - k_3 z^3 - C.$

We found the fixed points by solving the equilibrium equations (see for example [26])

(7) $z_1 = 0,\ -k_1 z - k_2 z^2 - k_3 z^3 - C = 0.$

The solutions to Eq. (5) are plotted in the same coordinate system as the data curves, see for example Figure 4. The solution curves are stable if we set one fixed point equal to zero, i.e., $C = 0$. Then the two other fixed points occur at z_0 (the centre of the closed curve, average coordinate of the centre of mass during walking/running), and at h for stable walking/running. The differential equation Eq. (5) becomes

(8) $z_{tt} + kz(z - z_0)(h - z) = 0.$

Numerical computations show that, for stable gait, the fixed point $z = z_0$ is a centre, while the fixed points $z = 0$ and $z = h$ are saddles.

5.3. Interpretation of the parameters in the model. Eq. (8) shows that we can model the movement of COM with a Duffing equation, up to a constant k, knowing only h and z_0. The values h and z_0 are close to the height of a person and to the average coordinate of the COM in motion, correspondingly. The gravitation constant g is involved implicitly in the differential equation, in a similar way as in the harmonic oscillation model in Section 4. The coefficient k does not have the meaning of the square of the angular frequency, $\omega^2 = \frac{stiffness}{mass}$, as in a linear case (Section 4), but k behaves in a similar way: it increases with an increase of the walking/running speed. We call this constant, multiplied by the mass, the **Duffing stiffness**. The meaning of the coefficient k is found from the following consideration. We rewrite Eq. (8) as $z_{tt} + k[-(z - z_0)^3 + (h - 2z_0)(z - z_0)^2 + z_0(h - z_0)(z - z_0)] = 0$. For z close to z_0 the linear approximation at z_0 is

(9) $z_{tt} + kz_0(h - z_0)(z - z_0) = 0.$

Comparing this equation with Hook's Law (Eq. (1)) we get an expression that relates the angular frequency ω with the coefficient k and the values z_0 and h: $\omega^2 \approx kz_0(h - z_0)$. Hence, the Duffing stiffness

(10) $mk \approx \frac{m\omega^2}{z_0(h-z_0)} = \frac{K}{z_0(h-z_0)}.$

We used that leg stiffness K is expressed as $K = m\omega^2$. Assuming the model when the centre z_0 is approximately in the middle between two saddle points, 0 and h, we get $h = 2z_0$, and Eq. (15) becomes symmetrical,

(11) $z_{tt} + kz_0^2(z - z_0) - k(z - z_0)^3 = 0$, where $k > 0$,

and $k \approx \frac{\omega^2}{z_0^2}$. Eq. (11) is the Duffing equation for a softening oscillator [26], i.e., the stiffness diminishes with the displacement.

5.5. Example based on the collected data. Figure 4 shows the observed smoothed data curve (red) and the curve corresponding to the solution of Eq. (8) (blue) for one of the participants, together with the fixed points.

The differential equation that describes the movement of the centre of mass of this participant running at the speed 8 km/hr is $z_{tt} + 279(z - 0.9)(1.8 - z) = 0$. We have rounded k to an integer, and z_0 and h to the first decimal place. From $k = 279$, $z = \frac{1}{2}h = 0.9$ we can compute $\omega \approx \sqrt{\frac{k}{z_0^2}} \approx 18.56$, i.e., the number of strides in a second is $\frac{\omega}{2\pi} \approx 2.95$. This participant made 24 strides in 10 secs, i.e., the real number of strides in a second is 2.4.

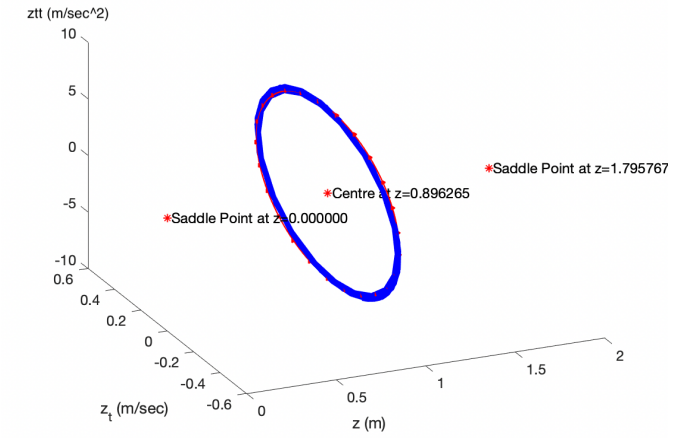


Fig 4. Curve (blue) for the solution to Eq. (8) for a participant (178cm tall, running at 8 km/h) compared with the data curve (red). The mean squared error is 0.015. The fixed points are saddles at $z = 0$ and $z \approx 1.8$ and the centre at $z_0 \approx 0.9$

6. STATISTICAL ANALYSIS

6.1. (University of New England (UNE) data. Fig.5 (left) shows that the Duffing stiffness depends on the speed of walking/running for each participant. By visually inspecting the graph, we see some outlier points. To further examine these, we calculated the standard deviation of Duffing stiffness for each participant, for the walking data. Two participants had standard deviations (SDs) > 400, while the rest of the participants had SDs < 100. Our observation of these two participants during the data recording showed that they were uncomfortable with some of the speeds, and they reported having no experience with treadmills. Therefore, we excluded these two participants from all further analyses. Fig.5 (right) shows the data for the remaining four participants.

The behaviour of the Duffing stiffness is different for walking (3 - 8 km/hr) and running (9 - 14 km/hr). Therefore, we separately fit data for walking and running speeds. We fitted a Linear Mixed Effect model in R for the walking data. We included speed as the fixed effect and Duffing stiffness as the dependent variable. We allowed both the slopes and the intercept to vary across participants. The model showed a slope estimate of $\hat{\beta}$=41.4, $t = 9.1$, $p = 0.004$. For the running data, we performed an equivalent analysis. Here, the slope associated with speed trended in a positive direction, $\hat{\beta} = 13.9$, $t = 3.0$, $p = 0.05$.

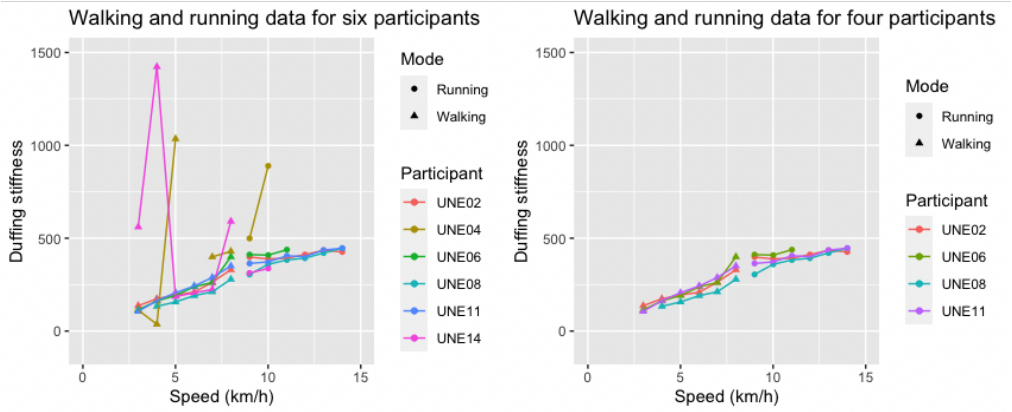


Fig 5. Duffing stiffness depending on speed with and without outliers. We consider the models for walking and for running separately. Each line represents a different participant

6.2. Comparing UNE dataset against public datasets. Next, we compare our data against public datasets for 42 walking participants [21], and 29 running participants [22]. As these sources were set up slightly differently than the UNE data, in order to match the data, we estimated the COM during walking or running, using the midpoint of the ASIS markers and subtracting the lowest value for the heel's marker.

Figure 6 shows the relationship between speed and Duffing stiffness.

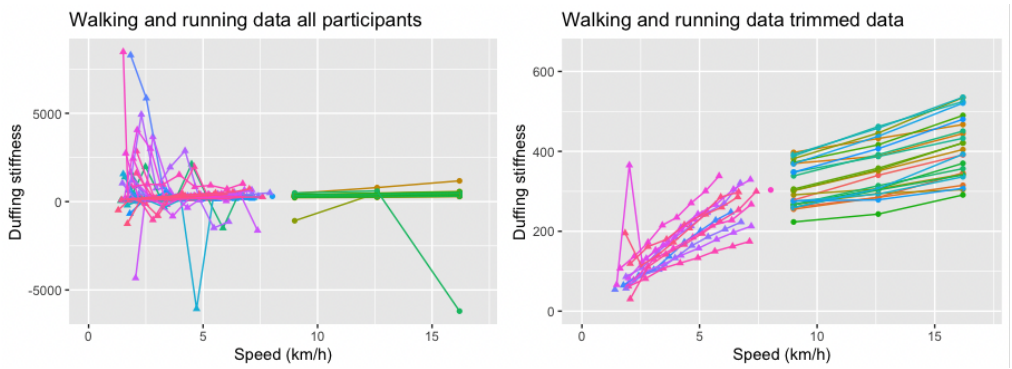


Fig 6. Duffing stiffness as a function of speed, with and without outliers. Note that the Duffing stiffness, k, depends also on the height of the COM of a participant, z_0, and the angular frequency of walking/running, ω: $k \approx \frac{\omega^2}{z_0^2}$.

First, we removed one participant with only one data point. Then, to identify outliers, we calculated the SDs in Duffing stiffness for each participant and excluded outliers for the walking and the running datasets, separately. For the walking data, [22], the SDs ranged from 39.2 to 3183.5. We chose to remove all participants with SD > 100. This left us with 12 participants for the walking data. We proceeded to fit the data with a Linear Mixed Effect model, akin to the UNE data. Again, we get a significant effect of speed on Duffing stiffness, $\hat{\beta}$=36.7, $t = 11.8$, $p < 0.0001$. For the running data, [21], we removed three out of 29 participants with SD > 100. The Linear Mixed Effect model showed a significant effect of speed, $\hat{\beta} = 13.5$, $t = 14.0$, $p < 0.0001$.

Finally, we compared the two datasets against each other. We created two models, one for running and one for walking, including both of the trimmed datasets. In the Linear Mixed Effect model, Duffing stiffness acted as the dependent variable, and the fixed effects were speed, dataset UNE versus [21, 22], with UNE acting as a baseline. We allowed the slope and the intercept to vary across participants. For walking, the effect of speed was, again, significant, $\hat{\beta} = 36.7$, $t = 12.4$, $p < 0.0001$. However, the effect of dataset was not significant, $\hat{\beta} = -43.1, t = -1.5, p = 0.2$, nor was the interaction between dataset and speed, $\hat{\beta} = 5.3,\ t = 0.9,\ p = 0.4$.

Similarly, in the running dataset, the effect of speed was $\hat{\beta} = 13.5,\ t = 13.2,\ p < 0.0001$. Neither the effect of dataset, nor the interaction of dataset and speed were significant, $\hat{\beta} = 27.5,\ t = 0.9,\ p = 0.4$ and $\hat{\beta} = 3.4,\ t = 1.1,\ p = 0.3$, respectively.

In summary, we found, overall, a significant positive relationship between speed and Duffing stiffness. For the running UNE data, the slope was not significant; however, when we combined the two datasets, the running slope was significant, and we found no interaction. Thus, the lack of significance in the UNE running data may be a result of low statistical power. We found no main effect of dataset, nor an interaction between dataset and speed. Thus, we find no evidence of a difference across datasets.

6.3. Identifying and examining outliers. From a practical perspective, an interesting aspect is participants whose data deviates from the fitted model. Here, we defined outliers based on SDs. This is the simplest method, which can easily be applied by a sport scientist without mathematical training. We drew a somewhat arbitrary threshold, where we treated all participants with a SD $>$ 100 as outliers. The reasons for high SD could vary across participants. For example, a typical problem for uncomfortable speeds less than 3 km/hr is an additional loop, as on the red line in Fig.7. The Duffing equation does not take into account the loop, as the blue approximation curve demonstrates. The other example of an outlier is illustrated in Fig. 8, where the curve breaks down into two parts, corresponding to the left or to the right leg. Red lines represent the data, blue lines are the solutions of the differential equation. A third reason for outliers that we noticed is an instability of walk, when each step varies in the amplitude and in the average height.

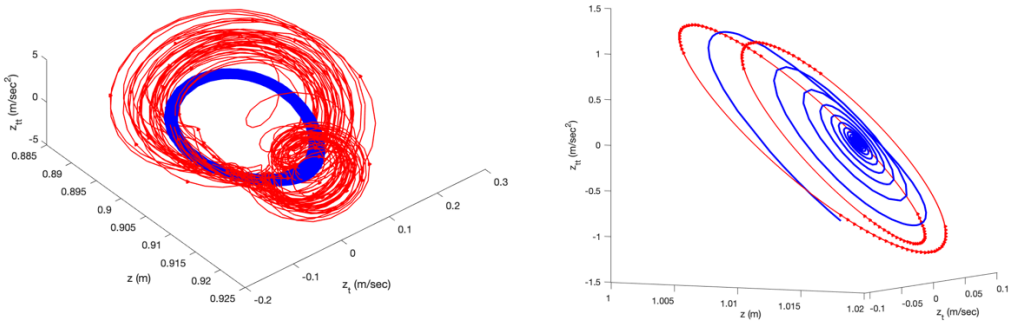


Fig 7. Slow walking speeds (< 3 km/hr, FTT=0.03) often contain an additional loop because of the compelled braking during each step

Fig 8. Asymmetry of the gait (4.7 km/hr) due to different strengths of the left and the right legs. The red curve brakes down into two parts, corresponding to the left or to the right leg

7. DUFFING EQUATION WITH VISCOSITY AND EXCITATION FORCE

7.1. Motivation. Approximation by a homogeneous equation with zero viscosity does not take into account the asymmetry caused by damping and excitation forces. For example, the red curve in Figure 9 shows a smoothed data curve for a running participant (14 km/hr, FFT threshold=0.3) in the phase plane (z, z_t). This curve is close to an ellipse, and the absolute value of the slope of the main axis of the ellipse, AC, is equal to $v = \frac{\nu}{m}$ if we assume that the viscosity is a constant. (If the viscosity $\nu = 0$, the slope vanishes, and the corresponding axis is horizontal.) The symmetry with respect to $z_t = 0$ is disturbed by the slope. Now we compare this slightly asymmetrical typical data curve with the symmetrical energy level curves, see Figure 9. Solutions of Eq. (8) with different initial values of $z(z_t = 0)$ give a set of energy level curves. The red curve is the data curve, the movement occurs in the direction ABCD. Energy is gained twice in each cycle (gait) in a phase of a "step", BC, and in a phase of a "fall" (in general not free fall), DA. The maximum of the energy occurs at

points A and C, the minimum occurs when at points B and D. There is a natural desire to find an approximation that considers the viscosity and the restoring force.

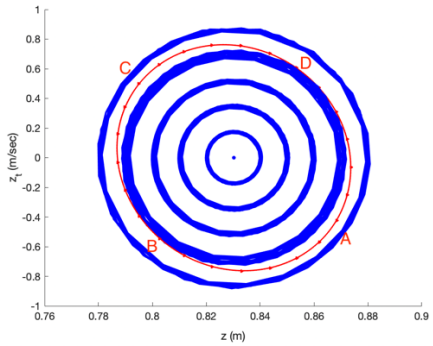


Fig 9. The energy level curves (blue) and the data curve (red) in the phase plane (z,z_t). The absolute value of the slope AC multiplied by mass m is the viscosity ν

7.2. A best fitting curve for a non homogeneous differential equation We identified the best fitting curve in the form

$$(12)\ z_{tt} + vz_t + kz_0^2(z - z_0) - k(z - z_0)^3 = fcos(\Omega t - \varphi),$$

where the term vz_t is the linear approximation of the damping force divided by mass. The constant $f = \frac{F}{m}$, where F is the amplitude of the excitation force, Ω is the angular frequency of the excitation force. The gravitation constant g is involved in the equation in a similar way as for the two already discussed models. As in the previous models, the equation for the best fitting curve was interpreted as a second order differential equation. The solution of the equation was computed and plotted in the same system of coordinates as the data curve. However, the balance between the viscosity and excitation force was too delicate, and most solution curves became stable spirals.

8. CONCLUSION

We introduced a new method for the investigation of human gait. This method is based on the visualisation of the vertical component of the movement of the COM during walking or running, in the space of the coordinates position, velocity, and acceleration of the centre of mass. We suggested a model by a non-linear homogeneous differential equation. We also had a partial success in approximation of the movement by a second order non-linear non-homogeneous differential equation. In this paper, we concentrated on features that are common for walking and for running. The individual features of the curves are of special interest for sport science, because they point to uncomfortableness in walking or running. We plan to investigate possible reasons for injuries by determining how stress is generated. One possible idea is to investigate why female runners have more frequent ACL (anterior cruciate ligament) tears than men [27].

ACKNOWLEGMENTS

The study followed ethical protocols as per ethics requirements (HE19- 239). Thanks to computer science students Ben Fisk, Danielle Galvin and Jarra McIntyre for developing a prototype of the software used for this paper and to sport science student Megan Bancks for the literature research. Thanks to Adam Harris and Gerd Schmalz for their critical comments and support.

References

1. Bullimore SR, Burn JF. Consequences of forward translation of the point of force application for the mechanics of running. Journal of Theoretical Biology. 2006;238:211–219.
2. Lipfert SW, Gunther M, Renjewski D, Grimmer S, Seyfarth A. A model-experiment comparison of system dynamics for human walking and running. Journal of Theoretical Biology. 2012;292:11–17.
3. Blickhan R. The spring-mass model for running and hopping. Journal of Biomechanics. 1989;22(11):1217–1227. doi:https://doi.org/10.1016/0021-9290(89)90224-8.
4. Blum Y, Lipfert SW, Seyfarth A. Effective leg stiffness in running. Journal of Biomechanics. 2009;42(14):2400–2405. doi:https://doi.org/10.1016/j.jbiomech.2009.06.040.
5. Ferris DP, Louie M, Farley CT. Running in the real world: adjusting leg stiffness for different surfaces. Proceedings of the Royal Society B: Biological Sciences. 1998;265(1400):989–994.
6. Nikooyan AA, Zadpoor AA. Mass–spring–damper modelling of the human body to study running and hopping – an overview. Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine. 2011;225(12):1121–1135. doi:10.1177/0954411911424210.
7. Silder A, Besier T, Delp SL. Running with a load increases leg stiffness. Journal of Biomechanics. 2015;48(6):1003–1008. doi:10.1016/j.jbiomech.2015.01.051.
8. Bullimore SR, Burn JF. Ability of the planar spring–mass model to predict mechanical parameters in running humans. Journal of Theoretical Biology. 2007;248(4):686–695. doi:https://doi.org/10.1016/j.jtbi.2007.06.004.
9. Farley CT, Gonza´lez O. Leg stiffness and stride frequency in human running. Journal of Biomechanics. 1996;29(2):181–186. doi:https://doi.org/10.1016/0021-9290(95)00029-1.
10. Dalleau G, Belli A, Bourdin M, Lacour JR. The spring-mass model and the energy cost of treadmill running. European Journal of Applied Physiology and Occupational Physiology. 1998;77(3):257–263. doi:10.1007/s004210050330.
11. . Kuo AD. Stabilization of Lateral Motion in Passive Dynamic Walking. The International Journal of Robotics Research. 1999;18((9)):917–930.
12. Donelan JM, Kram R, Kuo AD. Simultaneous positive and negative external and mechanical work in human walking. Journal of Biomechanics. 2002;35:117–124.
13. Mochon S, McMahon TA. Ballistic Walking. Journal of Biomechanics. 1980;13(1):49–57.
14. Geyer H, Seyfarth A, Blickhan R. Compliant leg behaviour explains basic dynamics of walking and running. Proc Biol Sci. 2006;273(1603):2861-2867. doi:10.1098/rspb.2006.3637. Proc Biol Sci. 2006;273(1603):2861–2867.
15. Hong H, Kim S, Kim C, Lee S, Park S. Spring-like gait mechanics observed during walking in both young and older adults. Journal of Biomechanics. 2013;46:77–82.
16. Jung CK, Park S. Compliant bipedal model with the centre of pressure excursion associated with the oscillatory behaviour of the centre of mass reproduces the human gait dynamics. Journal of Biomechanics. 2014;47(223-229).
17. Kim S, Park S. Leg stiffness increased with speed to modulate gait frequency and propulsion energy. Journal of Biomechanics. 2011;44:1253–1258.
18. Song H, Park H, Park S. A springy pendulum could describe the swing leg kinetics of human walking. Journal of Biomechanics. 2016;49:1504–1509.
19. Ludwig C, Grimmer S, Seyfarth A, Maus HM. Multiple-step model-experiment matching allows precise definition of dynamical leg parameters in human running. Journal of Biomechanics. 2012;45:2472–2475.
20. Maus HM, Revzen S, Guckenheimer J, Ludwig C, Reger J, Seyfarth A. Constructing predictive models of human running. J R Soc Interface. 2015;12.
21. Fukuchi CA, Fukuchi RK, Duarte M. A public dataset of running biomechanics and the effects of running speed on lower extremity kinematics and kinetics. PeerJ. 2017;5(e3298).
22. Fukuchi CA, Fukuchi RK, Duarte M. A public dataset of overground and treadmill walking kinematics and kinetics in healthy individuals. PeerJ. 2018;6(e4640).
23. Hook R. Micrographia: or, Some Physiological Descriptions of Minute Bodies Made by Magnifying Glasses, With Observations and Inquiries Thereupon. London: Printed by J. Martyn and J. Allestry; 1665.
24. Hausdorff JM. Gait dynamics, fractals and falls: finding meaning in the stride-to-stride fluctuations of human walking. Human Movement Science. 2007;26:555–589.
25. Jordan D, Smith P. Nonlinear Ordinary Differential Equations. fourth edition ed. Oxford University Press; 2007.
26. Brennan MJ, Kovacic I. The Duffing Equation: Nonlinear Oscillators and their Behaviour. John Wiley and Sons; 2011.
27. Stefani R. Kinesiology Analysis of Athletics at the Ancient Olympics and of Performance Differences Between Male and Female Olympic Champions at the Modern Games in Running, Swimming and Rowing. Athens Journal of Sports. 2017;4(2):123–138.

AN IMPROVEMENT TO THE TENNIS CHALLENGE SYSTEM

Tristan Barnett [a,c] , Vladimir Ejov [a] , Graham Pollard [b]
[a] *Flinders University*
[b] *University of Canberra*
[c] *Corresponding author: strategicgames@hotmail.com*

Abstract

Mathematics is applied to the tennis challenge system to derive a fairer challenging method. This method could be used in actual tournament play creating greater spectator interest. The concept of 'importance' is used such that a player has a free challenge if the importance of a point is above a certain threshold.

Keywords: Importance of points, Markov Chain model, player fairness

1. INTRODUCTION

The new challenge system for close line calls in tennis has been used on the ATP and WTA tour for Grand Slam events since the 2006 US Open, and was designed to increase fairness for players by obtaining accurate line calls and enhance spectator interest through video technology. In the current system, players have unlimited opportunity to challenge, but once three incorrect challenges are made in a set, they cannot challenge again until the next set. If the set goes to a tiebreak game, players are given additional opportunities to challenge (usually one extra). If the match is tied at six games all in an advantage set, the counter is reset with both players again having a limit of up to three incorrect challenges in the next 12 games, and this resetting process is repeated after every 12 games.

Strategies as to when players should challenge have recently appeared in the literature. Pollard et al. (2010) show that challenge decisions are based on the rate at which challenges occur, the expected number of points remaining in the set, the number of challenges remaining in the set, the probability of the challenge decision being successful and the importance of the point to winning the set. Clarke and Norman (2010) apply dynamic programming to the challenge system to investigate the optimal challenge strategy and obtain some general rules.

There appears to be problems with the current challenge system:

- Firstly, both of the above articles show that early in the set a player needs to decide whether to challenge or save challenges to later on in the set when the points are typically more important. Having to make such decisions is completely irrelevant to the game of tennis itself. The aim of the contest is to find the better player, and not to favour the player who is luckier within, or better at playing the challenge system. This is reflected by an article *Replay System Becomes Part of Players' Strategies* in The New York Times by Greg Bishop during the 2009 US Open.
  http://www.nytimes.com/2009/09/11/sports/tennis/11challenges.html

- Secondly, a player can run out of challenges because that particular set has a lot of balls that go close to the lines. This is perhaps particularly true in men's singles and men's doubles. The problem is exacerbated when each player does not have a similar number of challenges. A player who plays more balls near the lines is disadvantaged relatively. The player who, by chance has the need for more challenges, is disadvantaged.

- Thirdly, it would appear to be disappointing for the player and the spectators when that player runs out of challenges, the point is very important, and a challenge would have a clear likelihood of success. What is the chance that a grand slam final will be 'messed up' by an umpire making a wrong call and the player having run out of challenges, and subsequently losing the final when he might well have won it otherwise? This would be a very bad result for the player, the umpire and the game. Maybe this probability is not quite as small as some people might expect.

METHODS

a) Markov Chain model

We explain the method by first looking at a single game where we have two players, A and B, and player A has a constant probability p_A of winning a point on serve. We set up a Markov chain model of a game where the state of the game is the current game score in points (thus 40-30 is 3-2). With probability p_A the state changes from a, b to a + 1, b and with probability $q_A=1-p_A$ it changes from a, b to a, b + 1. Thus if $P_A(a,b)$ is the probability that player A wins the game when the score is (a,b), we have:

$$P_A(a,b)=p_AP_A(a+1,b)+q_AP_A(a,b+1)$$

The boundary values are:
$P_A(a,b) = 1$ if $a = 4$, $b \leq 2$, $P_A(a,b) = 0$ if $b = 4$, $a \leq 2$.

The boundary values and formula can be entered on a simple spreadsheet. The problem of deuce can be handled in two ways. Since deuce is logically equivalent to 30-30, a formula for this can be entered in the deuce cell. This creates a circular cell reference, but the iterative function of Excel can be turned on, and Excel will iterate to a solution. In preference, an explicit formula is obtained by recognizing that the chance of winning from deuce is in the form of a geometric series

$$P_A(3,3) = p_A^2 + p_A^2 2p_Aq_A + p_A^2 (2p_Aq_A)^2 + p_A^2 (2p_Aq_A)^3 + \ldots\ldots\ldots$$
where the first term is p_A^2 and the common ratio is $2p_Aq_A$

The sum is given by $p_A^2/(1-2p_Aq_A)$ provided that $-1<2p_Aq_A<1$. We know that $0<2p_Aq_A<1$, since $p_A>0$, $q_A>0$ and $1-2p_Aq_A=p_A^2+q_A^2>0$.

Therefore, the probability of winning from deuce is $p_A^2/(1-2p_Aq_A)$. Since $p_A+q_A=1$, this can be expressed as:

$$P_A(3,3) = p_A^2 /(p_A^2+q_A^2)$$

Excel spreadsheet code to obtain the conditional probabilities of player A winning a game on serve is as follows:
Enter **p_A** in cell D1
Enter **q_A** in cell D2
Enter **0.60** in cell E1
Enter **=1-E1** in cell E2
Enter **1** in cells C11, D11 and E11
Enter **0** in cells G7, G8 and G9
Enter **= E1^2/(E1^2+E2^2)** in cell F10
Enter **=$E$1*C8+$E$2*D7** in cell C7
Copy and Paste cell **C7** in cells D7, E7, F7, C8, D8, E8, F8, C9, D9, E9, F9, C10, D10 and E10

Notice the absolute and relative referencing used in the formula =$E$1*C8+$E$2*D7. By setting up an equation in this recursive format, the remaining conditional probabilities can easily and quickly be obtained by copying and pasting.

Similar recursion formulas with boundary conditions can be obtained for a tiebreak game conditional on the point score, set conditional on the game score and a match conditional on the set score. A predictions model is then applied to estimate the parameters of the probabilities of players winning a point on serve (Barnett et al, 2011).

b) Importance of points

Morris (1977) defines the importance of a point for winning a game (I_{PG}) as the probability that the server wins the game given he wins the next point minus the probability that the server wins the game given he loses the next point. The importance of a point to winning a game is thus:
$I_A(a,b)=P_A(a+1,b)-P_A(a,b+1)$.

Table 1 gives the importance of points to winning the game (I_{PG}) when the server has a 0.62 probability of winning a point on serve, and shows that 30-40 and Ad-Out are the most important points in the game. In a

similar way, we can define the importance of a game to winning a set and the importance of a set to winning a match. Table 2 gives the importance of games to winning a tiebreak set (I_{GS}) for player A serving. Player A and Player B were assigned point probabilities of 0.62 and 0.60 respectively to reflect overall averages in men's tennis. It is clear that every point is equally important for both players. Table 2 shows that the tiebreak game has the highest importance of 1.00, as the winner of this game wins the set. Similarly, table 3 gives the importance of sets to winning a best-of-5 set match (I_{SM}) and shows that the deciding set at 2 sets-all has the highest importance of 1.00, as the winner of this set win the match. Morris (1977) derived the following useful multiplicative result to obtain the importance of a point to winning the match (I_{PM}): For any point of any game of any set: $I_{PM} = I_{PG} * I_{GS} * I_{SM}$.

The definition of importance of a point in a match is a way of stating how much difference will result in the outcome of the match depending on whether a point is won or lost. In the context of a challenge system, importance of a point in a match can be viewed by how much percentage error will occur if a wrong decision is made. For example, suppose the score in a best-of-5 set match (all tiebreak sets) is 2-2 in sets, 5-5 in games and 30-30 in points and player A is currently serving. Suppose player A is winning 62% on serve and player B is winning 60% on serve. Using a Markov Chain model (Barnett and Clarke, 2005), player A has a 51.5% chance of winning the match from that position. If player A won the point, then his chance of winning the match would be 60.3%: whereas if player A lost the point then his chance of winning the match would be 37.3%. Therefore, the importance of the point in the match is given as 60.3%-37.3%=23.0%. If a wrong decision was made at that particular point in the match, then it would cost one of the players 23 percentage points in their chance of winning the match.

		Receiver's score				
Server's score		0	15	30	40	Ad
	0	0.25	0.34	0.38	0.28	
	15	0.19	0.31	0.45	0.45	
	30	0.11	0.23	0.45	0.73	
	40	0.04	0.10	0.27	0.45	0.73
	Ad				0.27	

Table 1: Importance of points to winning a game when the server has a 0.62 probability of winning a point on serve

		Player B's score						
Player A's score		0	1	2	3	4	5	6
	0	0.29	0.29	0.22	0.18	0.06	0.02	
	1	0.26	0.32	0.33	0.21	0.16	0.03	
	2	0.25	0.29	0.36	0.37	0.20	0.11	
	3	0.13	0.27	0.33	0.42	0.43	0.14	
	4	0.08	0.11	0.30	0.38	0.52	0.54	
	5	0.01	0.06	0.08	0.34	0.46	0.52	0.53
	6						0.47	1.00

Table 2: Importance of games to winning a tiebreak set when player A and player B have a 0.62 and 0.60 probability of winning a point on service respectively and player A is serving

		B's score		
A's score		0	1	2
	0	0.36	0.42	0.32
	1	0.32	0.49	0.57
	2	0.18	0.43	1.00

Table 3: Importance of sets to winning a best-of-5 set match when player A and player B have a 0.62 and 0.60 probability of winning a point on service respectively

3. RESULTS

a) Proposed new challenge system

It is proposed that the present challenge rule is modified in one way. Namely, that a player is allowed to challenge on points with sufficiently large importance, without risking that player's challenge point total.

Suppose the threshold value on when a player can always challenge a line call was given by the importance of the point in the match at 2 sets-all, 3 games-all, 0 points-all and player A serving. This is calculated as 1.00*0.42*0.25=0.104 when player A and player B have a 0.62 and 0.60 probability of winning a point on serve respectively. Then a player can always challenge at 2 sets-all and 3 games-all, only if the point score in the match has an importance of at least 0.104. This occurs at 2 sets-all and 3 games-all for 30-40 or Ad-Out (I_{PM}=0.305), 15-40 (I_{PM}=0.189), 15-30 (I_{PM}=0.188), 30-30 or deuce (I_{PM}=0.187), 0-30 (I_{PM}=0.161), 0-15 (I_{PM}=0.143), 15-15 (I_{PM}=0.132), 0-40 (I_{PM}=0.117), 40-30 or Ad-In (I_{PM}=0.115) and 0-0 (I_{PM}=0.104). This is represented in table 4 for a range of game scores in the deciding set, where an X indicates that a challenge is always allowable by both players. Note that a player can challenge at 2 sets-all and 6 games-all (tiebreak game), only if the point score has an importance of at least 0.231. This occurs for the majority of points in the tiebreak game, as expected.

	Score line at 2 sets-all (player A serving)					
Point score	0-0	1-1	2-2	3-3	4-4	5-5
30-40 or Ad-Out	X	X	X	X	X	X
15-40	X	X	X	X	X	X
15-30	X	X	X	X	X	X
30-30 or Deuce	X	X	X	X	X	X
0-30	X	X	X	X	X	X
0-15		X	X	X	X	X
15-15			X	X	X	X
0-40				X	X	X
40-30 or Ad-In				X	X	X
0-0				X	X	X
30-15					X	X
15-0, 30-0, 40-15 or 40-0						

Table 4: Indication as to whether a player can always challenge on a particular point in a match for a range of game scores in the deciding set given that the threshold value is given as 0.104

4. DISCUSSION

Being able to challenge 'free of charge' on some point scores later in the set, but not earlier, might cause confusion for some players in some situations. To get around this problem, you introduce a "challenge" screen visible to both players which gives a green light before the point is played if the point has a sufficient level of 'importance'. Otherwise the screen is empty (or a red light). Spectator's interest would also be lifted, quite possibly or naturally. It would give commentators an additional thing to talk about being the importance of points. Note that the free challenge light going on could be automated with the umpire's score card.

Instead of giving three incorrect challenges per set as proposed above, suppose players are given x challenges per set and have unlimited opportunity to challenge, but once x incorrect challenges are made in a set, they cannot challenge again until the next set. Further, players can always challenge when the point has a sufficient level of 'importance' = y without affecting their challenge point total, otherwise players cannot challenge if they have run out of their challenge point total.

Scenario 1)
When x=3 and the level of 'importance'=1, is equivalent to the current system.

Scenario 2)
When x=0 and the level of 'importance'=y, is "optimally" the best system in terms of minimizing time on player's challenging on "unimportant" points.

Scenario 3)
When $1 \leq x \leq 3$ and the level of 'importance'=y, is somewhere between Scenario 1) and Scenario 2)

At the very least Scenario 3) could be adopted such that players can always challenge when the point has a sufficient level of 'importance' = y without affecting their challenge point total. However, Scenario 2) could be obtained as an "optimal" system in terms of minimizing time on player's challenging on "unimportant" points.

However, whilst the fifth set is the most important set, it may be better to have the same procedure in each set. This is likely to be more easily accepted by the relevant people. An advantage of this is that the operation of the system would be identical for all sets. There is something nice about uniformity. Further, under the system described above, there would be points in earlier sets that are more important than some of the 'free challenge' points in the fifth set. This may present a problem. So just looking at set importances rather than match importances could be a preference.

Maybe every point in the tiebreak game should be a free challenge (with no 'additional' challenges given to the players at 6/6 because the set is 'long') and any point within the set at least as important as any point in the tiebreak game should also be free. This could be a useful selling point to the interested parties. If this was considered too generous, any point at least as important as say 2/2 within the tiebreak game could be a free challenge. The fact that players are given an extra challenge at 6/6 gives some merit to the ideas in this paper. The idea in this paper in fact parallels the present rules at 6/6. It just formalizes some present operational characteristics.

If the "challenge" screen was too much of a problem for players then you could have a system where a free challenge was given on every point in tiebreak games (representing a high level of importance) and at say set/match points.

5. CONCLUSIONS

Throughout this article it is demonstrated that a fairer method to the line call challenge system is such that a player should always be allowed to challenge at a score line with a certain level of importance without affecting their challenge point total.

References

Barnett T and Clarke SR (2005), Combining player statistics to predict outcomes of tennis matches. IMA Journal of Management Mathematics. 16 (2), 113-120.

Barnett T, O'Shaughnessy D and Bedford A (2011), Predicting a tennis match in progress for sports multimedia. OR Insight 24(3), 190-204.

Clarke S and Norman J (2010), An introductory analysis of challenges in tennis. Proceedings of the Tenth Australasian Conference on Mathematics and Computers in Sport. Edited by A Bedford and M Ovens. 43-48.

Morris C (1977), The most important points in tennis. In Optimal Strategies in Sports, Volume 5 in Studies in Management Science and Systems, Edited by S.P. Ladany and R.E. Machol, Amsterdam: North-Holland Publishing Company, 131-140.

Pollard GN, Pollard GH, Barnett T and Zeleznikow J (2010), Applying strategies to the tennis challenge system. Journal of Medicine and Science in Tennis 15(1), 12-15.

ASSESSING ON-COURT POSITIONING OF TENNIS PLAYERS WITH MACHINE VISION TO ESTIMATE THEIR CHANCES OF WINNING

Phillip Chung [b], Paul J. Bracewell [a,b,c] , Jordan K. Wilson [a], Yuying Xie [b]
[a] *Play in the Grey, New Zealand*
[b] *Victoria University of Wellington, New Zealand*
[c] *Corresponding author: paul@playinthegrey.com*

Abstract

The intent of this research is to investigate whether the players' location on the tennis court is related to their likelihood of success in a rally. Specifically, the players' distance from the centre net and centre line are key features that can be engineered. This type of data is obtained using a method for extracting the physical coordinates of the players on the court from broadcast footage. The physical coordinates of each player are extracted using a proprietary pipeline developed by Play in the Grey (McDonald et. al., 2020). This provides space-time coordinates for each frame in the footage which are used as input for the in-game prediction model. This research draws from previous research done on machine vision algorithms for Tennis footage (e.g. Chu et. al., 2010; Jiang et. al., 2009) and on various statistical methods for predicting Tennis match outcomes (e.g. Cornman, et. al., 2017; Kovalchik et. al., 2019; Kovalchik, 2020).

Initial results reveal that the closer a player is to the middle of the court towards the net, the greater the chance of winning the rally. The ability to process data in this way has implications for improving feedback to players.

Keywords: Feature Engineering, Logistic Regression

1. INTRODUCTION

The integration of new technology in sports is expanding the level of information and insight that can be extracted from video footage. Using automated computer vision systems could make in-depth insights more accessible to the benefit of athletes, coaches, broadcasters, as well as spectators.

Aside from the traditional statistics related to a sport, there has been a growing appetite for the creation of data-driven novel metrics that enhance the spectatorship of the sport. These metrics are not the usual statistics derived from historical performance; they are characteristically more viewer-friendly real-time analysis interwoven with graphic visualisations. An example of this type of new generation metric is the Premier League's development of 'momentum tracker' which calculates the likelihood of a goal within the next 10 seconds of gameplay.

The aim here is to apply a novel method of creating a real-time win-prediction model on Tennis rallies. We will derive physical metrics directly from broadcast tennis footage and use this data to create a model to determine which player is likely to win the current rally. The scope of this research is to accurately track the movement of the players relative to the centre net, in a singles tennis match rally.

The play of a single point in Tennis is by a rally. In simple terms, a rally in tennis starts with one player serving the ball from behind the baseline and ends when one player fails to return the ball within the opposition's court area. An average rally in a professional tennis match can last anywhere between 1 second to 10 seconds. The project will focus on a single rally event; footage used will be of an isolated rally which can vary in duration and a win-prediction model will be calculated on the variation of win probabilities within a rally.

The project requires two main components:

1. A reliable and accurate machine-vision algorithm to track player movement relative to the centre net within a given video footage.
2. An accurate predictive model to determine the players' likelihood of winning.

For the first component, we will be using a proprietary algorithm developed by the company Play in the Grey (www.playinthegrey.com) to calculate the movement of the player within a single rally (McDonald et. al. 2020; Trowland et. al., 2020). The footage fed into the model is solely from the court view of a rally. In televised tennis matches, this viewpoint is the standard way of showing a single rally and is captured from a height above and behind one of the players' showing the entire court from a down-angle view. For most rallies, the court view footage provides a relatively unobstructed and stable recording of the entire duration of a rally. The Play in the Grey model can take in this angled view of the court and players and project a 2-dimensional map of the match at each frame with X and Y coordinates of the player's location relative to the court. The model creates this 2D map from homography estimation which has been refined to minimise loss of

positional accuracy as outlined in Trowland et al. (2020). Whilst the court view footage tends to be relatively stable, there are slight camera movements during the rally which have the potential to affect the accuracy of the estimation. The model is also able to adjust for accuracy losses due to camera movements such as pan, zoom and tilt as explained by McDonald et al. (2020).

There is a substantive amount of literature on footage-derived object detection models developed specifically for tennis. Useful insights can be derived from literature for the purposes of this project, on the type of footage used (camera angle), separating useful elements from background noise and movement tracking.

FOOTAGE

Previous studies using computer vision on broadcast Tennis footage have opted for a similar approach by solely using the "court view" perspective footage as implemented by Archana and Geetha (2015), Jiang (2009) and Chu and Tsai (2010). To utilise raw broadcast footage, Chu and Tsai (2010) implemented a court view detection algorithm which recognised when the correct type of footage was being played and would begin the object detection algorithms. Detection of court view was carried out by identifying whether the dominant colours in the frame match that of the court view frame.

ALGORITHMS FOR PLAYER DETECTION IN TENNIS

A common issue with using the court view footage is that the player on the far court is represented by a smaller number of pixels and their movements are harder to track. The player on the far court is also more likely to be obstructed in view by the umpire, ground staff or advertisements. In dealing with this issue, both Archana and Geetha (2015) and Jiang (2009) opted for a partition of the two halves of the court with two search windows each focusing on one player.

Jiang (2009) implemented a further step in player detection by creating an 'Adaptive Search Window'. The court boundaries are established by the Initial Search Window for the upper and lower half of the court, then non-dominant colour detection is implemented to detect the player (dominant colour being the average colour of the full court). With the initial identification of the player, a smaller adaptive search window is created, which based on the possible speed of a human(2-7m/s) is only big enough to anticipate potential movements by the player but small enough that noise is not included in the search area. Only tested on 50 segments from 12 tennis games.

Archana and Geetha (2015) implemented the background subtraction method; develop a background model from a collection of background images then important elements of footage are derived by subtracting against the background model. Rate of success for player tracking – upper half player tracking 85.96% and lower half player tracking 91.23%.

PREDICTIVE ALGORITHMS

There is sufficient literature on applying building algorithms to predict tennis match outcomes, however, there is a limitation on the utility of these studies to this project's objective. These studies have only utilised statistics from determined and completed matches as well as external factors such as player rankings whereas the focus of this project is to make predictions based on real time in-game statistics and therefore more limited in the data available to make predictions.

Cornman et. al. (2017) is a good example of the current literature on Tennis match predictions. The study carried out match outcome predictions using several common machine learning algorithms including ANN, Random Forest, and others on historical match data. The features that were used for this study included but not limited to rankings, age, height, aces, double faults, and surface type. Whilst they yielded somewhat successful prediction rates, the algorithm is limited to historic data and is unable to account for real-time statistics from a match.

Literature on incorporating real-time statistics to create dynamic win-probabilities on sports such as football, NFL (US) are more established than tennis. Robberechts et al. (2019) outline that to calculate in-game win-probabilities, sport-specific features that are influential to scoring must be selected to make an effective dynamic predictor. Domain knowledge may be necessary to select the correct features to include. The study utilised a Bayesian model using eight features for each football team and predicted the future number of goals a team will score as a temporal stochastic process.

Kovalchik and Reid (2019) presents a novel approach to creating a dynamic model for tennis win predictions. The study has utilised a dynamic empirical Bayes updating rule to create a real-time prediction model which uses both in-game features as well as pre-match features. The study identifies serving as a key element in forecasting the probabilities of a players' success in a Tennis match. The model initialises a win-

expectation using the player's historical serving performance, then this win-expectation is adjusted throughout the match using the actual in-game serving performance of the player. The study found that the dynamic modelling provided a 28% reduction in error of in-match serve predictions as well as a 4% increase in overall win prediction.

Kovalchik (2020) also studied the optimization of the Elo rating system to predict the results of tennis matches. She carried out the research by combining MOV (Margin of Victory) and four different models: linear, joint additive, multiplicative and logistic. In that model, data is used to update the model based on historical data as well as data from matches, thus improving the accuracy of the new model's predictions. Still, the model is only 67% accurate, and that's mostly based on the underlying historical data.

2. METHODS

Here, the intent is to investigate whether the players' location on the court is related to their likelihood of success in any given rally. Two key dimensions are of interest: the players' distances from the centre net and centre line. We utilise computer vision for extracting the physical coordinates of the players on the court from broadcast footage (McDonald et. al., 2020).

DATA COLLECTION

To test the concept of whether player movement impacts their probability of winning a rally, we sourced data from the Wimbledon Tennis Tournament. Commonly known as The Championship, this is the oldest tennis tournament in the world and arguably the most prestigious.

Rally outcome and video footage are readily available online from https://www.wimbledon.com/. The data has two forms: already machine readable (information about the match and outcomes) and the footage. The footage requires substantial processing to first extract the tennis player's co-ordinates on court per frame and then feature engineering to extract metrics that may be representative of movement.

DATA PROCESSING

From the Wimbledon.com website, we collected a set of rallies and the corresponding footage. An important step in this process is ensuring the videos are suitable for the machine vision pipeline. This is dependent on factors such as the quality of the footage, height of the camera and stability of the camera.

Player on court location data was extracted using Play in the Grey's pipeline. As each frame is processed, time-based data of player movement is generated. From the initial player detection, shown in Figure 1, initial data representing the human perspective is captured. This is then transformed into a top-down view using homography as shown in Figures 2 and 3 and explained by McDonald et al. (2020). Once homography has been applied to transform the data we have coordinate tracking of player in a rally, where the two-dimensional data is represented as pixels. Pixel coordinates are translated to coordinates on the tennis court using a simple transformation as the dimensions of a tennis court are standardised. As explained on Olympics.com, "a competitive tennis court must be rectangular in shape, measuring 23.77 metres long. The width, however, differs for doubles (10.97 metres) and singles (8.23 metres)."

DATA DESCIPTION

To aid in construction a dataset for analytical purposes, the linkage file to connect the machine vision output data to the rally outcome data has the following features:

- *Gender*: 1=Men, 0=Women
- *Winner*: 1=the player is further away from the camera, 0= the player is closer to the camera
- *Server*: 1= Far court player, 0 = Close court player
- *TLx*, *TLy*, *TRx*, *TRy*, *BLx*, *BLy*, *BRx*, *BRy*: represent the pixel coordinates (x,y) for the four corners (Top\Bottom, Left\Right) of the Tennis court captured in one frame (can be from any part of the video, preferably middle of video). For footage where the camera moves slightly, use a frame which captures the most common coordinates of the Court in the video. This is used to convert the x-y data from pixels to metres.
- *xyFrameTime*: The time stamp for the frame used to determine the court pixel coordinates: *TLx*, *TLy*,…, *BRy* (An approximate time when the frame was taken is enough).
- *RallyLength*: the length of rally in seconds.
- *CamMvmt*: 0 = No camera movement in footage, 1 = Camera movement in footage.
- *MatchName*: Name of video used to extract footage

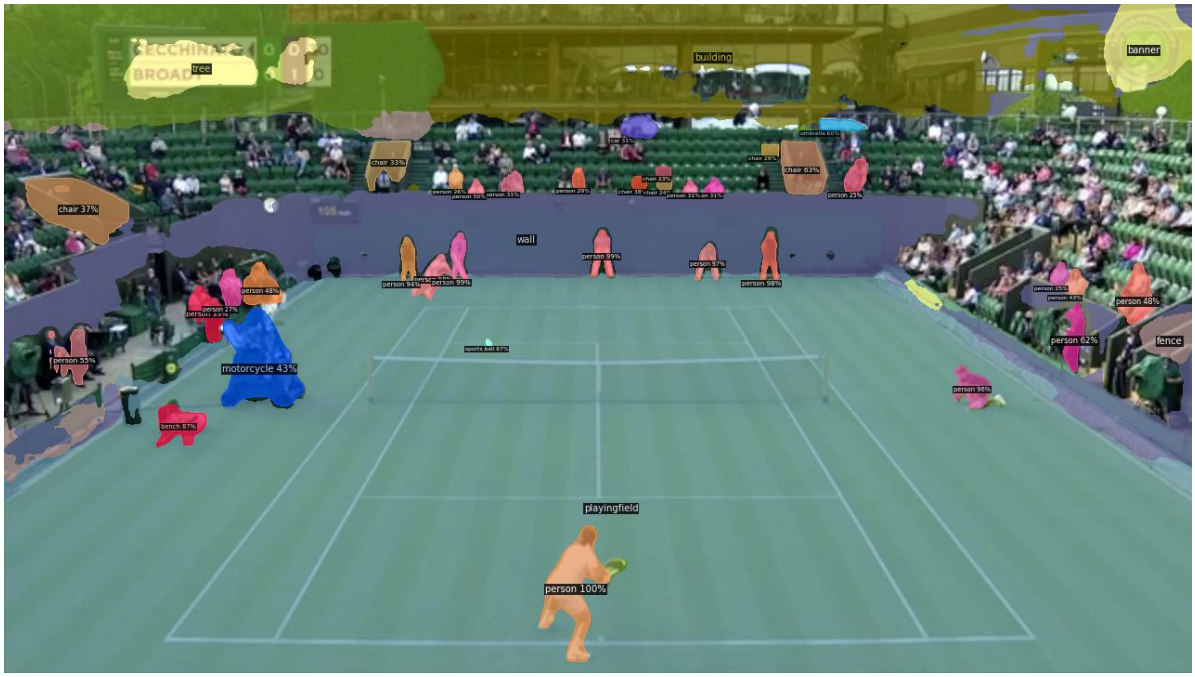


Figure 1: Example of humans detected in sample footage from Wimbledon 2020.

Figure 1 shows humans detected in a single frame. Other features are also detected, such as text and logos. As mentioned previously, to convert these detections into useful data, a homography transformation is applied. This can be simply understood as it is used to describe the object's positional mapping relationship between the world coordinate system and the pixel coordinate system.

DATA CLEANSING

Figures 2 and 3 show the output of the detection post homography. Importantly as shown in Figure 1, many humans are detected, some of which remain apparent within the court surrounds in Figure 2 (the left image). The following process was undertaken to improve the quality of data.

1. Remove non-player detections
2. Remove incorrect player coordinates
3. Impute missing player coordinates
4. Removal of due to camera movement issues.

In Figures 2 and 3, each dot represents a player's location on the court at each frame (darker dots early in rally, lighter dots late in rally). Comparing the two side-by-side images helps show the visible error values. This includes the umpires at the far end of the court and on the left-hand side. The two yellow dots at the top of the bottom court are post rally. To clean the data, bounding boxes are used to identify valid coordinates.

FEATURE ENGINEERING

Using the transformed data, we aggregate data of each player in a rally to obtain the maximum, minimum, mean, standard deviation for the following:

- Distance
- Speed
- Acceleration
- The distance between the player and the net
- The distance between the player and the centre line

In all instances, the data is converted to a half court view, so all data is converted. That is, the top half of the court undergoes a 180-degree rotational transformation through the centre of the court and the middle of the net. Table 1 lists the features constructed for player per rally along with the description. The prefix for most attributes is one of min (minimum), avg (arithmetic mean), max (maximum), std (standard deviation) and the refence point for comparison (net as a vertical reference) and centre as a (horizontal reference). Distance (dist) and acceleration are derived from the change in position from frame to frame.

Figures 2 & 3: Example of top-down view of detected humans over a series of frames, before (left) and after isolation of the tennis players (right). Positioning is represented by pixels. Darker colour indicates closer to start of rally and lighter colours closer to the end of the rally.

Variable	Description
min_fromNet	minimum distance from net
avg_fromNet	average distance from net
max_fromNet	maximum distance from net
std_fromNet	standard deviation of distance from net
min_fromCentre	minimum distance from centre
avg_fromCentre	average distance from centre
max_fromCentre	maximum distance from centre
std_fromCentre	standard deviation of distance from centre
avg_dist	average distance moved in rally
max_dist	maximum distance moved in rally
std_dist	standard deviation of distance moved
avg_acceleration	average acceleration in rally
max_acceleration	maximum acceleration in rally
std_acceleration	standard deviation of acceleration
frontHalf	proportion of time in front half of court
backHalf	proportion of time in back half of court
behindLine	proportion of time ibehind baseline
rallyLength	time taken to complete rally
Server	binary (server = 1, receiver = 0)

Table 1: Data features engineered from the data

3. RESULTS

PREDICTIVE MODEL

We constructed a simple logistic regression model using only the attributes listed in Table 1. The outcome of the rally (1=win, 0=loss) was the dependent variable. 152 rallies with data and corresponding footage were available for analysis, which was split in 121 training instances and 31 test cases.

Variable	Coefficient
min_fromNet:	-0.766
min_fromCentre	-0.629
avg_dist	-1.050

Table 2: Statistically significant variables from logistic regression to predict rally outcome based only on player movement within a rally.

In testing, the model had an accuracy of 71%. This suggests that there is sufficient explanatory power for the three statistically significant attributes listed in table 2 (at the 5% level of significant). Importantly, these attributes are also practically significant. Starting with avg_dist which has a negative coefficient. This indicates that the less a player has to move, frame to frame, the greater the chance they have of winning. As this measure is essentially speed (distance for a unit of time), this means that the longer a player must run at pace within a rally, the more likely they are to lose. This potential interpretation is given further weight by the first two rows: minimum distance from the net and minimum distance from the centre. As both coefficients are negative, this means that the closer a player gets to the net, or is from the centre, the more likely they are to win. Given the magnitude for the et attribute is larger, this implies that from the sample of 121 rallies used to construct the model, getting to the net was more important than being in the middle.

4. DISCUSSION

Whilst there was no literature available on utilising footage-derived data to predict win probabilities, there are still valuable insights we can draw from related literature to achieve our goal. Studies on models that solely utilised historical statistics are useful if the project were to incorporate such statistics into the model. Based on the findings from Kovalchik and Reid (2019), our model should consider the impact of serving and its significant influence on the likelihood of a player winning the rally. This could potentially be implemented through determining which player is serving, and initialising appropriate win-expectations based on the independent probability of a serving player's likelihood of winning the rally.

Following on the idea of utilising a combination of historic statistics and in-game statistics, we could extend this to the player movement information derived from the machine vision model and explore domain knowledge related to positional advantages in tennis. This would allow us to build a model that could determine whether a player's position on the court at any given time is statistically an advantageous position relative to the opposing player's position on their half of the court.

There is substantial potential to expand this exploratory analysis by including additional information and incorporating the findings from Kovalchik and Reid (2019). Specific future work includes: expand feature selection, use more data from more tournaments and use entire set footage instead of rallies.

To then expand upon a dynamic probability model, we need to expand feature selection: gender, age of player, types of surfaces for tennis courts. That is, we need to more thoroughly consider features that may affect the player's performance.

5. CONCLUSIONS

The progress made in machine vision algorithms now allows us to extract vast amounts of data from standard game footage which until recently was impracticable at an affordable cost. In creating new predictive models using this data, not only would we be able to enhance the experience for spectators, but there is also potential to generate an entirely new set of sports statistics using a mixture of domain knowledge and feature engineering. Athletes and coaches at various levels of the sport could access in-depth breakdowns and analysis that could improve performance and tactics.

In building the novel prediction algorithm, the study will explore the raw movement data extracted as well as using feature engineering to generate new insights that may aid in predicting the winner of a tennis rally. The data produced will be used to train and evaluate various machine learning algorithms to determine whether the in-game analytics could provide better accuracy over models that solely utilise traditional tennis statistics.

References

Archana, M., & Geetha, M. K. (2015). Object Detection and Tracking Based on Trajectory in Broadcast Tennis Video. Procedia Computer Science, 58, 225–232. https://doi.org/10.1016/j.procs.2015.08.060

Chu, W.-T., & Tsai, W.H. (2010). Modeling spatiotemporal relationships between moving objects for event tactics analysis in tennis videos. Multimedia Tools and Applications, 50(1), 149–171. https://doi.org/10.1007/s11042-009-0363-z

Cornman, A., Spellman, G., & Wright, D. (2017). Machine learning for professional tennis match prediction and betting. Stanford University.

Jiang YC., Lai KT., Hsieh CH., Lai MF. (2009) Player Detection and Tracking in Broadcast Tennis Video. In: Wada T., Huang F., Lin S. (eds) Advances in Image and Video Technology. PSIVT 2009. Lecture Notes in Computer Science, vol 5414. Springer, Berlin, Heidelberg. https://doi-org.helicon.vuw.ac.nz/10.1007/978-3-540-92957-4_66

Kovalchik, S., & Reid, M. (2019). A calibration method with dynamic updates for within-match forecasting of wins in tennis. International Journal of Forecasting, 35(2), 756–766. https://doi.org/10.1016/j.ijforecast.2017.11.008

Kovalchik, S.. (2020). Extension of the Elo rating system to margin of victory. International Journal of Forecasting, 36(4), 1329-1341.

McDonald, R.G., Bracewell, P. J., Trowland, H.E., & Tikhonov, A. (2020). Discounting camera movement in calculation of player paths using machine vision in rugby union. Proceedings of the 15th Australian Conference on Mathematics and Computers in Sports. Ray Stefani & Adrian Schembri eds. Wellington, New Zealand ANZIAM Mathsport. pp. 39-44.

McDonald, R. G., P. J. Bracewell, H. E. Trowland, and A. Tikhonov (2020). Discounting Camera Movement in Calculation of Player Paths using Machine Vision in Rugby Union. Proceedings of the 15th Australian Conference on Mathematics and Computers in Sports. Ray Stefani Adrian Schembri eds. Wellington, New Zealand: ANZIAM Mathsport.

Robberechts, P., Van Haaren, J., & Davis, J. (2019). Who Will Win It? An In-game Win Probability Model for Football.

Trowland, H. E., P. J. Bracewell, R. G. McDonald, and A. Tikhonov (2020). Validating Player Path Tracking Using Machine Vision. Proceedings of the 15th Australian Conference on Mathematics and Computers in Sports. Ray Stefani Adrian Schembri eds. Wellington, New Zealand: ANZIAM Mathsport.

NO-AD GAME SCORING WITHIN A BEST OF 5 SETS TENNIS MATCHES

Alan Brown [a], Tristan Barnett [b,e] , Graham Pollard [c] , Geoff Pollard [d] , Vladimir Ejov [b]
[a] *Affiliation not specified*
[b] *Flinders University*
[c] *University of Canberra*
[d] *Tennis Australia*
[e] *Corresponding author: strategicgames@hotmail.com*

Abstract

Tennis matches that take much longer than expected are a problem in several ways. They can delay the starting time of the following match, cause issues for broadcasters, lead to an increased number of injuries, and decrease the winner's chance of winning in the next round. In this paper several alternative game structures for possible use in reducing the length of best-of-5 set matches are studied. Also, criteria for comparing two or more tennis match scoring systems are outlined.

Keywords: Alternative game scoring systems for tennis, long matches in tennis, mean duration, variability of duration of tennis matches, efficiency of tennis scoring, changes to tennis scoring, parameter values in tennis

.

1. INTRODUCTION

The uncertain and highly variable length of games, sets and matches in tennis has been a concern for players, television, spectators, as well as tournament directors. It remains a concern.

Some matches have been observed to last more than 5 hours even though the 5th set was not a 'long' advantage set. For example, in the 2012 Australian Open Final, Djokovic beat Nadal, 5-7, 6-4, 6-2, 6-7, 7-5 in 5 hours and 53 minutes. Whilst long matches can be exciting as a stand-alone match (as in a final), it can be seen as unfair in the tournament setting as the winner can be too exhausted to do justice to his performance in the next round. This typically can occur in men's grand slam singles matches as they play best-of-5 set matches. Thus, it would be useful to have a scoring system that reduces the likelihood of such long matches, whilst keeping other match characteristics (such as the probability of the stronger player winning) much the same as they are at present.

Over the last several decades a considerable amount of research has been carried out on the match characteristics of various tennis scoring systems. When considering alternative scoring systems, it is not sufficient to consider just the mean and variance of the duration of a match. Issues related to the skewness of this distribution of duration are an important consideration. The probability that the better player wins also needs to be considered. Such measures or characteristics are available and are important in deciding whether a scoring system is acceptable or not. It is noted that there is typically a need for compromise when considering two or more scoring systems, as it is unlikely that one system is best on all such measures.

The major purpose of this research is to achieve a greater understanding of the characteristics of several match scoring systems using variations of the No-Ad game concept, which was described and studied by Pollard and Noble (2004) and is currently used in doubles (excluding grand slams where a standard deuce game is used). In the No-Ad game a player needs to win 4 points in order to win the game (and if the score line reaches 3 points-all in that game, the player who wins the next point wins that game). At most 7 points are played in the No-Ad game and this characteristic helps to reduce the skewness of the distribution of duration of a set and match of tennis. Some other game scoring systems are also considered in this paper.

The idea of "deuce" was introduced (at least as far back as 1490) for a simple reason … to ensure that the game could not be won by a one-point difference in the players' scores. Deuce was derived from the French "a deux du jeu"… two points away from game. It is reasonable to believe that there was no mathematics carried out on this deuce game back in 1490 concerning how it would affect the game of tennis into the future (Barnett, 2012).

Pollard and Noble (ibid) studied the efficiency and some duration characteristics of a best of 3 sets match using the No-Ad game. They also set up the 50-40 game in which server needs to win 4 points in a game whilst the receiver needs to win only 3 points to win the game. The concept behind such a game structure was

that the server has the advantage of serving within the game but has the disadvantage of needing to win one more point than the receiver in order to win. They noted, for example, the increased efficiency of best of 3 set tennis scoring when used for 'strong' servers as in men's doubles but did not study the best of 5 set matches, which are considered in this paper.

Recently, Pollard and Barnett (2018) reported on the 50-40 game and a few variations of it *within a single set of tennis*. One variation they studied was the 50-40, 40-0, 40-15 game which is a 50-40 game modified so that the server wins the game if the score reaches 40-0 or 40-15. The logic behind a scoring system such as this is that there is little point (in terms of efficiency and duration) in playing points that are relatively unimportant and unexciting. Further, in their discussion section they suggested a couple of further modifications of the 50-40 game that could be usefully studied. In their work they considered just a single set of tennis and whilst these single set results give some useful insights into the likely characteristics of a best of 3 or best of 5 sets match when using such games, the study of a complete match gives greater clarity regarding any preferred system. This is done in this paper.

2. METHODS

Alternative game structures to address the problem of 'long' best of 5 set matches

In this study we consider the **best of 5 tiebreak sets** using the two official game scoring systems of the Rules of tennis (advantage/deuce games and no ad games) as well as three possible alternatives.

(1) Advantage/Deuce games – a player needs to win 4 points but if the score line reaches 3 points-all, then a player must be 2 points ahead to win the game.

(2) No-Ad – a player needs to win 4 points in order to win the game. If the score line reaches 3 points-all, the player who wins the next point wins the game. At most 7 points are played in the No-Ad game.

(3) No-Ad* - a player needs to win 4 points but if the score line reaches deuce, then a player must win 2 more points to win the game. At most 9 points are played in the No-Ad* game.

(4) 50-40 (as defined by Pollard and Noble (ibid)) – the server needs to win 4 points whilst the receiver needs to win just 3 points in order to win the game. At most 6 points are played in this type of game.

(5) 50-40* - server needs to win 4 points and receiver needs to win just 3 points but if the score line reaches 3-2 (40-30) then the player who wins two more points wins the game. At most 8 points are played in this game.

Criteria for comparing tennis scoring systems

In this study of the best of 5 tiebreak sets matches for men, where every game is one of the above-mentioned game scoring systems, the match characteristics of interest are

(1) Probability that the stronger player wins, P
(2) Expected value of number of points played (duration) in the match, E(D)
(3) Standard deviation of the number of points played in the match, SD(D)
(4) Efficiency of the scoring system
(5) Coefficient of skewness of the number of points played in the match, $\gamma = E[(Z - \mu)^3]/\sigma^3$.
(6) The 95%, 99% and 99.5% points in the cumulative distribution of duration, denoted by CD95, CD99 and CD99.5.

Note that the efficiency of the tennis scoring system was devised in a very elegant paper by Roger Miles (Miles, 1984). The efficiency of a tennis scoring system with key characteristics P, the probability that the better player wins, and m, the mean duration (mean number of points played in the match), is equal to:

$$(2(P-Q)\ln(P/Q)) / (m(p_A-p_B)\ln(p_Aq_B/p_Bq_A))$$

where $Q = 1-P$, p_A is the probability player A wins a point on service, p_B is the probability player B wins a point on service, $q_A = 1-p_A$ and $q_B = 1-p_B$.

Given two scoring systems with the same mean duration, the one in which the better player A has a higher probability of winning has the greater efficiency. Correspondingly, given two scoring systems with the same likelihood of the better player winning, the one that has the smaller mean duration has the greater efficiency. Note that the efficiencies of tennis scoring systems are typically a lot less than 1 mainly because of the nested nature (points, games, sets) of tennis scoring using 'best of' structures. Very efficient scoring systems do not have 'best of' structures. They also have very large variances of duration, and this makes them quite inappropriate for scoring in tennis.

Note that the characteristics CD95, CD99, and CD99.5 should be sufficient for comparing the upper tails of the duration distributions, as (only) 127 five-set matches are played each year in each Grand Slam Men's Singles event.

Parameter values

The key input parameters in modelling a men's singles tennis match between player A and player B are p_A = probability that player A wins a point on his serve, and p_B = probability that player B wins a point on his serve.

Cross and Pollard (2011) noted that for men singles at the four Grand Slam events in 2008, the proportion of points won on service averaged 0.631, 0.621, 0.667 and 0.643 at Australian Open, French Open, Wimbledon and US Open respectively. They reported that these values 'had not changed much over the years [1999 to 2009]', except for the French Open ('associated with a considerable increase in first service speed' (Cross and Pollard (2009))). As these have an average value of 0.64, this value is used as the most appropriate average value for this study.

In the Cross and Pollard study the proportion of points won on service by the winner minus the proportion won on service by the loser was 0.11. This figure is biased in favour of the winner. For example, using simulation methods in a study of bias in sporting statistics, Pollard et al (2010) noted…"As the winner must have won the last point, last game and last set, the winner's service statistics can have an upwards bias, and the loser's service statistics a downwards bias.", and…"In the best of three tiebreak sets match between two *equal* players (with $p_A = p_B = 0.65$), the proportion of points won on service by the eventual winner is shown to be about 0.065 on average greater than the proportion of points won on service by the loser. For a best of five tiebreak sets match between these two equal players, this difference is shown to average about 0.049."

It is important that such biases are considered when working with reported statistics. We have done this in deciding the parameters to use in our study.

Taking 0.04 as a reasonable difference between the serving p-values in a 'moderately close' match, appropriate values for the parameters in a typical or average men's singles match are $p_A = 0.66$, $p_B = 0.62$.

It is noted here that, in an article using data from the 2016 Rio Olympics, Carl Bialik (2016) concluded that the service success rates for men's singles was 63%. It is noted that this percentage is quite similar to the figure above, and to the assumptions we make in our modelling.

It was anticipated that whilst no scoring system would be 'best' with respect to all characteristics, one scoring system might in some sense be best "overall'.

Method of Analysis

Most of these results are numerically exact and were developed using recursive formulas in an Excel spreadsheet (Barnett, 2016). The theory behind the recursive formulas is now outlined.

To analyse the progress of a match we denote by Z the total number of points played to date. Z is a discrete random variable with density f(z). The moment generating function of Z is
$M_Z(t) = E(e^{tZ}) = \sum_z e^{tz} f(z)$.

In a singles match there are two players, denoted by A and B, and the serve is rotated between player A and player B according to prescribed rules. Now Z = X + Y, where X is the number of points served by player A, and Y is the number of points served by player B. We make an important assumption that X and Y are independent random variables, i.e. the serving strength of player A does not influence the serving strength of player B, and vice versa. It follows from this that the joint distribution of these variables can be factorised.
$F(X,Y) = F_1(X) * F_2(Y)$

Using this independence of X and Y, we obtain $M_Z(t) = M_X(t)M_Y(t)$ as in Parzen(1960)
The formulas for extracting the moments of the sum of independent random variables is the following:

$m_{1Z} = m_{1Y} + m_{1X}$
$m_{2Z} = m_{2Y} + 2m_{1X}m_{1Y} + m_{2X}$
$m_{3Z} = m_{3Y} + 3m_{1X}m_{2Y} + 3m_{2X}m_{1Y} + m_{3X}$
$m_{4Z} = m_{4Y} + 4m_{1X}m_{3Y} + 6m_{2X}m_{2Y} + 4m_{3X}m_{1Y} + m_{4X}$

These expressions are obtained by successive differentiation the moment generating function with respect to t, and putting t to 0.

The score for a match in progress will be denoted by (a, b : c, d : e, f), where (a, b) is the score in points, (c, d) is the score in games, and (e, f) is the score in sets, for player A and player B respectively. We will use a truncated form of this notation whenever it is convenient so to do.

A tennis match consists of four levels - (points, games, sets, match). It becomes necessary to represent:
points in a point as pp,
points in a game as pg,
points in a tiebreak game as pg_T,
points in a tiebreak set as ps_T
points in a best of 5 all tiebreak set match as pm_{5T}.

Let s_A, s_B represent the condition that player A and player B, respectively served first at the beginning of a set. Let c_A, c_B represent the condition that player A and player B, respectively are currently serving in the set at the score (a, b : c, d). If (a, b) is not a boundary score for the current game then
$s_A = c_A$ and $s_B = c_B$, if (c + d) mod 2=0
$s_A = c_B$ and $s_B = c_A$, if (c + d) mod 2=1
except in the case of the tiebreak game of the tiebreak set, with c = 6, d = 6, when
$s_A = c_A$ and $s_B = c_B$, if (a + b) mod 4=0 or 3
$s_A = c_A$ and $s_B = c_B$, if (a + b) mod 4=1 or 2

Let $P^{psT}(a, b : c, d|s_A)$ represent the probability of player A winning a tiebreak set at this score, and player A serving first in the current set. Let $Y^{psT}(a, b : c, d|s_A)$ be the number of points remaining in the set at this score with player A serving first in the current set. This number is a random variable. Let $M_{YpsT(a,b:c,d|sA)}(t)$ be its moment generating function.

Similarly let $P^{psT}(a, b : c, d|w_A, s_A)$ represent the probability of player A winning a tiebreak set at this score, and player A serving first in the current set. Let $Y^{psT}(a, b : c, d|w_A, s_A)$ be a random variable of the number of points remaining in the set at this score conditional on player A both winning the set, and serving first in the current set. Let $M_{YpsT(a,b:c,d|wA,sA)}(t)$ be its moment generating function conditional on player A both winning the set, and serving first in the current set.

Many variants of this notation will be used. The representation of the score will be restricted whenever it is not essential to display the full score. Other symbols include B for player B, l for the condition of losing, and n for the condition of serving next.

The next step is to introduce weighted moment generating functions. Let X be a conditional random variable. Let C be the condition that X occurs with probability p_X.

Then
$W_{X|C}(t) = p_X M_X(t)$

This product of a probability and its associated moment generating function is defined as a weighted moment generating function. The weight is the probability measure such that the conditions applied to the random variable are true.

Denote by w_{nX} the weighted n^{th} moment of the random variable X. Then
$w_{nX} = p_X m_{nX}$ for n = 1, 2, 3, 4, ...

The more important situation for us arises when the score does change. Let X and Y be independent random variable with conditional probabilities p_X and p_Y, respectively, of occurring. Let Z denote the random variable for their sum, Z = X + Y when both X and Y occur. Then $p_Z = p_X p_Y$. It follows from the formula for moment generating functions that the weighted moment generating functions satisfy
$W_{Z|C1,C2}(t) = W_{X|C1}(t)W_{Y|C2}(t)$.

The formulas for extracting the weighted moments are the following:
$p_Z = p_X p_Y$
$w_{1Z} = p_X w_{1Y} + w_{1X} p_Y$
$w_{2Z} = p_X w_{2Y} + 2w_{1X} w_{1Y} + w_{2X} p_Y$
$w_{3Z} = p_X w_{3Y} + 3w_{1X} w_{2Y} + 3w_{2X} w_{1Y} + w_{3X} p_Y$
$w_{4Z} = p_X w_{4Y} + 4w_{1X} w_{3Y} + 6w_{2X} w_{2Y} + 4w_{3X} w_{1Y} + w_{4X} p_Y$

We now develop the algebra for weighted moment generating functions. We are able to add together two weighted moment generating functions whenever we encounter two mutually exclusive cases. Two simple examples where the score does not change are:
(a) Condition on initial server
$W_{Y psT(a,b:c,d)}(t) = W_{Y psT(a,b:c,d|sA)}(t) + W_{Y psT(a,b:c,d|sB)}(t)$

(b) Condition on winning or losing
$W_{Y psT(a,b:c,d|sA)}(t) = WY_{psT(a,b:c,d|wA,sA)}(t) + W_{Y psT(a,b:c,d|lA,sA)}(t)$

We now apply these ideas to the playing of a single point. In this case some of the notation appears to degenerate, so we must be careful. However, this analysis will be used whenever the score changes as a point is played in a game, a set, or a match.

Each point played is a single point, irrespective of the score. For player A serving, the probability of winning the point is denoted by p_A irrespective of the score and $q_A = 1 - p_A$. Let $P^{pp}(()|c_A, w_A)$ and $P^{pp}(()|c_A, l_A)$ represent the probabilities of player A winning and losing a point on serve respectively from score line () within the point. It follows that:
$P^{pp}(()|c_A, w_A) = p_A$
$P^{pp}(()|c_A, l_A) = q_A$

Let $Y^{pp}(()|c_A)$ represent the number of points remaining in the point from score line () with player A serving. Each point played is a single point, so $Y^{pp}(()|c_A) = 1$. Let $Y^{pp}(()|c_A,w_A)$ and $Y^{pp}(()|c_A,l_A)$ represent the number of points remaining in the point from score line () given player A won and lost the point respectively with player A serving.

Therefore:
$M_{Y pp(()|cA)}(t) = E(e^{Y pp(()|cA)t}) = E(e^t) = e^t$
$W_{Y\ pp(()|cA,wA)}(t) = P^{pp}(()|c_A,w_A) M_{Y\ pp(()|cA)}(t) = p_A e^t$

This is a fundamental brick in the model.

It is easy to check that
$w_n(Y^{pp}(()|c_A, w_A)) = p_A$ for n = 0, 1, 2, 3, 4, ...

Likewise $W_{Y pp(()|cA,lA)}(t) = P^{pp}(()|c_A, l_A) M_{Y pp(()|cA)}(t) = q_A E(e^t)$

and
$w_n(Y^{pp}(()|c_A, l_A)) = q_A$ for n = 0, 1, 2, 3, 4, ...

a) Number of points in a game

Let $W_{Y pg(a,b|cA,wA)}(t)$ and $W_{Y pg(a,b|cA,lA)}(t)$ represent the weighted moment generating functions of the number of points remaining in the game from score line (a, b) given player A is serving and player A won and lost the game respectively.

Theorem.
$W_{Y pg(a,b|cA,wA)}(t) = W_{Y pp(()|cA,wA)}(t) W_{Y pg(a+1,b|cA,wA)}(t) + W_{Y pp(()|cA,lA)}(t) W_{Y pg(a,b+1|cA,wA)}(t)$

Proof. $M_{Y pg(a,b|cA,wA)}(t) = E(e^{tY\ pg(a,b|cA,wA)})$ is an expectation that is calculated before the point at score (a, b) has been played. The point played is won and lost with probability p_A and q_A respectively, where $p_A + q_A = 1$ since there are only two possible outcomes. When we try to recalculate the original expectation after the point has been played, we obtain the weighted sum of two expressions

$M_{Y_{pg(a,b|cA,wA)}}(t)$
$= p_A E(e^{t(1+Y_{pg(a+1,b|cA,wA)})})P^{pg}(a+1, b|c_A, w_A)/P^{pg}(a, b|c_A, w_A)$
$+q_A E(e^{t(1+Y_{pg(a,b+1|cA,wA)})})P^{pg}(a, b+1|c_A, w_A)/P^{pg}(a, b|c_A, w_A)$

where the odds ratios
$P^{pg}(a+1, b|c_A, w_A)/P^{pg}(a, b|c_A, w_A)$ and $P^{pg}(a, b+1|c_A, w_A)/P^{pg}(a, b|c_A, w_A)$ reflect the changes in the chances of player A winning when the score is updated after winning or losing the point, respectively. The count of 1 for the point played is independent of the distribution of the remaining points after the point has been played, so, as for moment genrating functions, we can factorize the expectations to obtain
$E(e^{t(1+Y_{pg(a+1,b|cA,wA)})}) = E(e^t)E(e^{t(Y_{pg(a+1,b|cA,wA)})})$ and
$E(e^{t(1+Y_{pg(a,b+1|cA,wA)})}) = E(e^t)E(e^{t(Y_{pg(a,b+1|cA,wA)})})$.

After some rearrangement we find that
$P^{pg}(a, b|c_A, w_A)M_{Y_{pg(a,b|cA,wA)}}(t)$
$= p_A E(e^t)P^{pg}(a+1, b|c_A, w_A)E(e^{t(Y_{pg(a+1,b|cA,wA)})})$
$+q_A E(e^t)P^{pg}(a, b+1|c_A, w_A)E(e^{t(Y_{pg(a,b+1|cA,wA)})})$
The only step that is left is the identification of the various terms in this expression as weighted moment generating functions, to obtain
$W_{Y_{pg(a,b|cA,wA)}}(t)=W_{Y_{pp(()|cA,wA)}}(t)W_{Y_{pg(a+1,b|cA,wA)}}(t)+W_{Y_{pp(()|cA,lA)}}(t)W_{Y_{pg(a,b+1|cA,wA)}}(t)$

Note carefully in this result how first we are able to multiply the weighted moment generating functions on each path of this branching process which arises when scoring, because the steps on each branch are independent; and then add the results of this multiplication, because the paths are mutually exclusive.

It follows that $W_{Y_{pg(a,b|cA,wA)}}(t) = P^{pg}(a, b|c_A, w_A)M_{Y_{pg(a,b|cA,wA)}}(t)$
where $M_{Y_{pg(a,b|cA,wA)}}(t)$ is the moment generating function of the random variable $Y^{pg}(a, b|c_A, w_A)$.

By successive differentiation with respect to t from the Theorem, and setting t = 0 we obtain the following recurrence formulas.

$w_1(Y^{pg}(a, b|c_A, w_A)) = p_A w_1(Y^{pg}(a+1, b|c_A, w_A)) + q_A w_1(Y^{pg}(a, b+1|c_A, w_A)) + p_A P^{pg}(a+1, b|c_A, w_A) + q_A P^{pg}(a, b+1|c_A, w_A)$

$w_2(Y^{pg}(a, b|c_A, w_A)) = p_A w_2(Y^{pg}(a+1, b|c_A, w_A)) + q_A w_2(Y^{pg}(a, b+1|c_A, w_A)) + 2p_A w_1(Y^{pg}(a+1, b|c_A, w_A)) + 2q_A w_1(Y^{pg}(a, b+1|c_A, w_A)) + p_A P^{pg}(a+1, b|c_A, w_A) + q_A P^{pg}(a, b+1|c_A, w_A)$

$w_3(Y^{pg}(a, b|c_A, w_A)) = p_A w_3(Y^{pg}(a+1, b|c_A, w_A)) + q_A w_3(Y^{pg}(a, b+1|c_A, w_A)) + 3p_A w_2(Y^{pg}(a+1, b|c_A, w_A)) + 3q_A w_2(Y^{pg}(a, b+1|c_A, w_A)) + 3p_A w_1(Y^{pg}(a+1, b|c_A, w_A)) + 3q_A w_1(Y^{pg}(a, b+1|c_A, w_A)) + p_A P^{pg}(a+1, b|c_A, w_A) + q_A P^{pg}(a, b+1|c_A, w_A)$

$w_4(Y^{pg}(a, b|c_A, w_A)) = p_A w_4(Y^{pg}(a+1, b|c_A, w_A)) + q_A w_4(Y^{pg}(a, b+1|c_A, w_A)) + 4p_A w_3(Y^{pg}(a+1, b|c_A, w_A)) + 4q_A w_3(Y^{pg}(a, b+1|c_A, w_A)) + 6p_A w_2(Y^{pg}(a+1, b|c_A, w_A)) + 6q_A w_2(Y^{pg}(a, b+1|c_A, w_A)) + 4p_A w_1(Y^{pg}(a+1, b|c_A, w_A)) + 4q_A w_1(Y^{pg}(a, b+1|c_A, w_A)) + p_A P^{pg}(a+1, b|c_A, w_A) + q_A P^{pg}(a, b+1|c_A, w_A)$

Boundary Values:
$w_n(Y^{pg}(a, b|c_A, w_A)) = 0$, if $a = 4$ and $0 \le b \le 2$; $b = 4$ and $0 \le a \le 2$
$w_1(Y^{pg}(3, 3|c_A, w_A)) = 2p^2_A/(2p^2_A - 2p_A + 1)^2$
$w_2(Y^{pg}(3, 3|c_A, w_A)) = 4p^2_A(1 - 2p^2_A + 2p_A) / (2p^2_A - 2p_A + 1)^3$
$w_3(Y^{pg}(3, 3|c_A, w_A)) = 8p^2_A(4p^4_A - 8p^3_A - 4p^2_A + 8p_A + 1)/(2p^2_A - 2p_A + 1)^4$
$w_4(Y^{pg}(3, 3|c_A, w_A)) = 16p^2_A(1 - 2p^2_A + 2p_A)(4p^4_A - 8p^3_A - 16p^2_A + 20p_A + 1)/(2p^2_A - 2p_A + 1)^5$

Similar recursion formulas with boundary values can be obtained for $w_n(Y^{pg}(a, b|c_A, l_A))$.

Let $M_{Y_{pg(a,b|cA)}}(t)$ represent the moment generating function of the number of points remaining in a game at point score (a, b) for player A serving.

Using the rule for combining weighted moment generating functions with mutually exclusive conditions we obtain $M_{Ypg(a,b|cA)}(t) = W_{Ypg(a,b|cA,wA)}(t) + W_{Ypg(a,b|cA,lA)}(t)$ since the probability that a game will eventually end is 1.

Converting moments to parameters of distribution (mean, variance, coefficients of skewness and excess kurtosis) can readily be obtained.

Similar formulas and parameters of distribution can be obtained for when player B is serving such that $W_{Ypg(a,b|cB,wB)}(t)$ and $W_{Ypg(a,b|cB,lB)}(t)$ represent the weighted moment generating functions of the number of points remaining in the game from score line (a, b) given player B is serving and player B wins and loses the game respectively.

b) Number of points in a tiebreak game

The analysis of a tiebreak game is similar to that of a standard game except that it is necessary to allow for the rotation of service before each odd point in the tiebreak game.

c) Number of points in a tiebreak set

We study here the model for a tiebreak set. To account for the rotation of service in this type of set it is necessary to allow for the rotation of server at the beginning of each game. Using this convention, whenever a tiebreak game is required to resolve the winner of the set, this tiebreak game is marked to the server of the first point of the game, and hence to the server of the first point of the set when it comes to determining the first server of the next set. This rule applies irrespective of the outcome of the tiebreak game.

For player A serving in the first game of the set there are four cases to be dealt with separately. Consider the case where player A not only serves in the first game of the set, but wins the set, and serves in the first game of the next set.

Let $W_{YpsT\ (0,0:c,d|sA,wA,nA)}(t)$ represent the weighted moment generating function of the number of points remaining in a tiebreak set at point and game score (0, 0 : c, d) given player A served first, wins the set and is serving first in the next set to be played. Then by considering a complete game being played at that score we obtain, for $c + d < 12$

$$W_{YpsT(0,0:c,d|sA,wA,nA)}(t)=W_{Ypg(0,0|cA,wA)}(t)W_{YpsT(0,0:c+1,d|sA,wA,nA)}(t)+W_{Ypg(0,0|cA,lA)}(t)W_{YpsT(0,0:c,d+1|sA,wA,nA)}(t), \text{ for } (c + d) \bmod 2=0$$

$$W_{YpsT(0,0:c,d|sA,wA,nA)}(t)=W_{Ypg(0,0|cB,lB)}(t)W_{YpsT(0,0:c+1,d|sA,wA,nA)}(t)+W_{Ypg(0,0|cB,wB)}(t)W_{YpsT(0,0:c,d+1|sA,wA,nA)}(t), \text{ for } (c + d) \bmod 2=1$$

There is a special case for the tiebreak game, with $c = 6$, $d = 6$, where due to the rotation of serve player A cannot serve first in the next set, so
$W_{YpsT\ (0,0:6,6|sA,wA,nA)}(t) = 0$, which simplifies to $W_{YpgT\ (0,0|cA,wA)}(t)=0$

d) Number of points in a best of 5 all tiebreak set match

Because we have to take into account both the winner of the current set and the server at the start of the next set, the recurrence formulas have to allow for four-way branching rather than the two-way branching that we have previously met.

For player A winning the match and currently serving,

$$W_{Ypm5T(0,0:0,0:e,f|cA,wA)}(t)=W_{YpsT(0,0:0,0|sA,wA,nA)}(t)W_{Ypm5T(0,0:0,0:e+1,f|cA,wA)}(t)+W_{YpsT(0,0:0,0|sA,lA,nA)}(t)$$
$$W_{Ypm5T(0,0:0,0:e,f+1|cA,wA)}(t)+W_{YpsT(0,0:0,0|sA,wA,nB)}(t)W_{Ypm5T(0,0:0,0:e+1,f|cB,wA)}(t)+W_{YpsT(0,0:0,0|sA,lA,nB)}(t)$$
$$W_{Ypm5T(0,0:0,0:e,f+1|cB,wA)}(t)$$

The total number of points played in a tennis match has a discrete distribution. The moments of this distribution can be calculated using a lattice model with the Markov property and a few other modest assumptions. The Normal distribution has been widely studied, and tables of the probabilities for this distribution are readily available. The basic idea of the Normal Power approximation is to use these probability tables to estimate the tail probabilities of other distributions. This method uses the first four moments and produces a continuous approximation to the cumulative distribution.

The Normal Power approximation has a weakness in that it can fail when fitting distributions that have exponential tails. This weakness is exposed when attempting to fit the distribution of points in an advantage set where the number of games is not finite.

The Normal Power approximation has another weakness in that it can fail when fitting distributions that are multimodal. Therefore, special steps must be taken when estimating the distribution of points in a tennis match. The key observations are that the distribution of points in a tie-breaker set is unimodal, and the number of games is finite. The Normal Power approximation can be safely used to estimate this distribution. The quality of this estimate can be checked using simulation. The distribution of points in 3 set endings, 4 set endings and 5 set endings are each unimodal in a best of 5 set all tie-breaker match, as they inherit the properties of a single tie-breaker set. Each of these distributions can be estimated, and the distribution of points for the complete match can be obtained by weighed addition, where the weights to be used are the probabilities of each type of ending.

Let X be a random variable with a cumulative distribution F(x), so that $P(X \geq x) = F(x)$

Let μ, σ, γ_1, γ_2 be the mean, standard deviation, skewness and excess kurtosis of X. Let Z be a standardized random variable with mean 0 and standard deviation 1, with

$P(Z \geq z) = P(X \geq x)$

Denote the cumulative Normal distribution by $\varphi(.)$. Then the Normal Power approximation can be written as

$F(x) \approx \varphi(y)$

with
$z = (x-\mu)/\sigma$

and
$y = z - 1/6\ \gamma_1(z^2 - 1) - 1/24\ \gamma_2(z^3 - 3z) + 1/36\ \gamma_1^2\ (4z^3 - 7z)$

3. RESULTS

1. Firstly, we consider an 'average' or 'typical' men's singles match with $p_A = 0.66$ and $p_B = 0.62$. These (p_A, p_B) parameters represent an 'average' match in Grand Slam men's tennis and are particularly relevant for the US Open or Australian Open. The results for such a match are given in Table 1 for player A serving first in the match. Columns 2-6 in the tables are exact results (from the methodology). They were checked against the equivalent (exact) best of 3 tiebreak sets results in Pollard (1983).

B5 sets	P(A wins)	Mean	Efficiency	Stand Dev	Skew	CD95	CD99	CD99.5
Ad games	0.734	261.22	0.52	61.26	0.14	362	394	405
No-Ad	0.719	232.13	0.51	53.23	0.11	319	346	354
No-Ad*	0.728	247.58	0.52	57.15	0.12	341	370	379
50-40	0.718	198.50	0.59	46.27	0.13	274	299	307
50-40*	0.730	217.55	0.61	51.29	0.14	302	329	339

Table 1 Characteristics of a best of 5 tiebreak sets match when p_A=0.66 and p_B=0.62

The more relevant observations that can be made from Table 1 include

1. The 50-40 and 50-40* games, whilst producing more efficient match systems than the other game structures, reduce the mean duration by an amount that would appear to be excessive and undesirable *for Grand Slam tennis*.
2. The probability that player A wins the match is slightly reduced (relative to Ad games) when No-Ad* is used and reduced further under the No-Ad system.
3. The No-Ad and No-Ad* games produce similar efficiencies to the present matches using the Ad game. They reduced the means, the standard deviations, the skewness and the CDs.
4. The No-Ad* game produces a best of 5 set scoring system with characteristics roughly midway between those of the Ad game and the No-Ad game. Its mean duration is 13.6 points fewer than the present system, and its CD99.5 is 26 points smaller. It might be considered a useful solution to the issue being studied in this research.

2. Secondly, we consider a men's singles Grand Slam match in which the advantage of serving for both players are less than the Grand Slam average. These parameters could represent a typical match at the French Open between two players with weaker or less successful serves. Table 2 gives the relevant characteristics.

B5 sets	P(A wins)	Mean	Efficiency	Stand Dev	Skew	CD95	CD99	CD99.5
Ad games	0.741	260.68	0.58	61.82	0.15	363	396	407
No-Ad	0.721	229.69	0.55	52.92	0.11	316	343	352
No-Ad*	0.732	245.74	0.57	57.13	0.13	340	368	378
50-40	0.715	196.29	0.60	45.68	0.13	271	295	304
50-40*	0.728	215.15	0.63	50.71	0.14	299	326	335

Table 2 Characteristics of a best of 5 tiebreak sets match when p_A=0.62 and p_B=0.58

Whilst all of the observations made with respect to Table 1 apply also to Table 2, perhaps the most relevant comparison is the observation that P(A wins) decreases for both the 50-40 and the 50-40* games relative to Table 1 (whilst it increases for the other types of games). This is not a surprise as the advantage of serving is reduced with these parameter values.

3. Thirdly, we consider a men's singles Grand Slam match in which the advantage of serving is greater than average. These parameters could represent a typical match at Wimbledon between two players with stronger or more successful serves.

B5 sets	P(A wins)	Mean	Efficiency	Stand Dev	Skew	CD95	CD99	CD99.5
Ad games	0.725	263.30	0.45	60.98	0.12	364	394	404
No-Ad	0.717	236.11	0.46	53.86	0.11	324	350	359
No-Ad*	0.723	251.03	0.46	57.50	0.11	345	373	382
50-40	0.721	201.50	0.57	46.97	0.13	279	303	312
50-40*	0.732	220.49	0.57	51.87	0.14	306	334	343

Table 3 Characteristics of a best of 5 tiebreak sets match when p_A=0.70 and p_B=0.66

Whilst all of the observations made with respect to Table 1 apply also to Table 3, perhaps the most relevant comparison is the observation that P(A wins) increases for both the 50-40 and the 50-40* games relative to Table 1 (whilst it decreases for the other types of games). This is not a surprise as the advantage of serving is enhanced with these parameter values.

4, CONCLUSIONS

The statistical characteristics of five different best of 5 tiebreak sets scoring systems have been studied. The aim of the study was to see whether there was an alternative to the present system using advantage games that might lead to less occurrences of very long matches and thus might be of use in Grand Slam tennis. Several measures for comparing tennis scoring systems have been outlined.

The effect of five different types of games within the best of five sets structure has been analysed. The types of games included the Ad game and the No-Ad game as defined in the Rules of Tennis. The No-Ad* game was also considered. In this game the best of three points is played if deuce is reached. The 50-40 game in which the server needs to win 4 points whilst the receiver needs to win just 3 points in order to win the game, was also considered. The 50-40* game, a modification of the 50-40 game in which the best of 3 points is played if 40-30 is reached, was also considered.

Whilst the 50-40 and 50-40* games were shown to be typically quite efficient for many matches and very effective at reducing match length, they would appear to 'go too far' for consideration at the Grand Slam level. Further, they may appear problematic to some players due to their 'unbalanced' structure.

The No-Ad*, having characteristics somewhere between the Ad game and the No-Ad game resulted in a useful decrease in the number of very long best of 5 tiebreak sets matches. It would appear to be a useful addition to available tennis scoring systems.

References

Barnett T (2012). Analyzing tennis scoring systems: from the origins to today. Journal of Medicine and Science in Tennis 17(2), 68-77.

Barnett T (2016). A recursive approach to modelling the amount of time played in a tennis match. Journal of Medicine and Science in Tennis, 21(2).

Bialik C. (2016). Serving is a disadvantage in some Olympic Sports. https://fivethirtyeight.com/featurse/serving-is-a-disadvantage-in-some-olympic-sports/

Cross, R. and Pollard, G. H. (2009). Grand Slam men's singles tennis 1991-2009, Service speed and other related data. ITF Coaching and Sports Science Review, 16(49), 8-10.

Cross, R. and Pollard, G. H. (2011). Grand Slam men's singles tennis 1995-2009, Part 2: Points, games and sets. ITF Sports Coaching and Sports Science Review, 53(19), 3-6.

Miles R. E. (1984). Symmetric sequential analysis: the efficiencies of sports scoring systems (with particular reference to those of tennis). J. R. Statist. Soc. B, 46, 93-108.

Parzen, E. (1960). Morern probability theory and its applications. New York: Wiley.

Pollard G. H. (1983). An analysis of classical and tiebreaker tennis. Australian Journal of Statistics, 25(3), 496-505.

Pollard, G. H. and Barnett, T. (2018). Some new 'short games' within a set of tennis. International Journal of Computer Science in Sport. 17(1), 67-76.

Pollard. G. H. and Noble, K. (2004). The benefits of a new game scoring system in tennis, the 50-40 game. Proceedings of the seventh Australasian conference on mathematics and computers in sport,

Pollard. G. H., Pollard G. N., Lisle, I. and R. Cross. (2010). Bias in Sporting Match Statistics. Proceedings of the Tenth Australasian conference on mathematics and computing in sport, Darwin, edited by A, Bedford and M. Ovens. Mathsport (ANZIAM), 221-228, July, 2010.

“New balls, please”: Quantifying the effect of tennis ball fluffiness

Adrian Eassom [a, c], Sam Robertson [a], Machar Reid [b]

[a] *Victoria University*
[b] *Tennis Australia*
[c] *Corresponding author: adrian.eassom@gmail.com*

Abstract

In Grand Slam Main Draw tennis matches, tennis balls are replaced after the first seven games and thereafter every nine games (ITF, 2022). Over the duration of these seven/nine games, tennis balls degrade (Steele, 2006). A prominent feature of this degradation process is the change in ball “fluffiness”, which is attributed to the fibres on the surface of a tennis ball standing up after repeated ball impacts and in doing so changing the aerodynamic characteristics of the tennis ball (Mehta, 2001). Specifically, it has been demonstrated that the fluff on a tennis ball has a major influence on a tennis ball’s drag coefficient. The tennis ball drag coefficient is a dimensionless constant used to quantify the drag resistance through air of a moving ball, as defined within established tennis ball trajectory equations of motion (Cross, 2020). Using Hawk-Eye spatiotemporal ball motion tracking data collected from the 2022 Australian Open, this study presents methods to deconstruct ball trajectory time histories into equations of motion to estimate the variation in the drag coefficient throughout the duration of a tennis match. Using these methods and focusing on the serve, it was found that in general there was a gradual increase in the ball drag coefficient with usage, followed by a noticeable step change reduction in the drag coefficient at the changeover between old and new balls. The presented methods and findings can be used by a variety of tennis stakeholders, including tennis ball manufacturers/tournament organisers to monitor the degradation of tennis balls during tournament match play; and tennis players/coaches wishing to better understand the expected change in ball behaviour after the match umpire announces for “new balls, please”.

Keywords: Tennis ball, drag coefficient, fluffiness, ball trajectory, spatiotemporal data

1. INTRODUCTION

As per the official International Tennis Federation (ITF) rule book for Grand Slam tennis events, six tennis balls are provided for each Main Draw tennis match, which are then replaced with new balls after the first seven games (including warm-up) and thereafter every nine games (ITF, 2022). The changeover to new balls is announced by the chair umpire, prompting the opening of the vacuum sealed cans containing the new balls. The significance of this change to new balls is highlighted by the accepted tennis etiquette for the server of the next game to hold the ball aloft and make the receiver aware of the incoming new balls.

Over the duration of game play, new tennis balls will degrade due to a combination of factors (Steele, 2006). A major contributing factor to ball wear is repeated ball impacts with the racket and the court during play. Ball wear can be materialised through a decrease in ball stiffness, loss of ball mass and most notably the change in surface condition of the ball. The change in surface condition is commonly referred to as ball “fluffiness” and is attributed to the fibres on the surface of a tennis ball standing-up after repeated ball impacts and in doing so changing the aerodynamic characteristics of the tennis ball (Mehta, 2001).

One of these aerodynamic properties is the tennis ball’s drag coefficient; a dimensionless constant used to quantify the drag resistance through air of a moving ball, as defined within established tennis ball trajectory equations of motion (Cross, 2020). An increase in a tennis ball’s drag coefficient results in the ball slowing down through the air, causing the ball to land at a shallower distance into the court and extending the required reaction time of the receiver.

Measuring a ball’s drag coefficient can be accurately calculated under controlled wind tunnel test environments, such as the testing arrangements used in Mehta (2001) and Goodwill (2004). The estimated drag coefficient of new tennis balls from these studies were found to be in the range of 0.6 to 0.7. The testing by Mehta (2001) indicated that partially worn tennis balls resulted in an increase in drag coefficient, however both Mehta (2001) and Goodwill (2004) also found that heavily worn tennis balls resulted in a decrease in drag coefficient relative to new tennis balls. Flow visualisation studies by Mehta (2001) indicated that for tennis ball

speeds representative of match-play, the ball exhibited behaviour consistent with the transcritical flow regime (approaching independence of Reynolds Number) which is supported by the finding in both Mehta (2001) and Goodwill (2004) that the drag coefficient was generally independent of speed, with only a small decline with increasing ball velocity. This small decline in drag coefficient has been attributed to the flattening of the fibres on the surface of a tennis ball with increasing ball speed. Goodwill (2004) also investigated the effect of a spinning tennis ball, identifying an increase in drag coefficient associated with increasing spin rates.

Cross (2014) attempted an alternative approach to wind tunnel testing by firing new tennis balls from a ball launcher and measuring the speed of the balls at defined locations using video cameras, and then estimating the drag coefficient based on the trajectory equations of motion. The study found significant shot-to-shot variation in the measured drag coefficient, with measurements ranging between 0.45 to 0.57. However, the measurements remained lower than the drag coefficients derived from wind tunnel tests. The study also found that the drag coefficient was independent of the tennis ball speed and spin (in contrast to the wind tunnel studies).

A similar approach is required to measure a tennis ball's drag coefficient in tournament match play via the use of spatiotemporal ball motion tracking data. Choppin (2018) attempted to measure tennis ball drag coefficients using Hawk-Eye tracking data from Davis Cup and Fed Cup matches. Using the Hawk-Eye data, Choppin (2018) calculated the average horizontal deceleration of the tennis ball in the period from racket impact until contact with the court. The method made several simplifying assumptions, most notably not capturing the effect of spin. However, the findings of the study did estimate a 4% increase in the drag coefficient for used balls relative to new balls.

This study extends the methods used in Cross (2014) and Choppin (2018), by estimating the drag coefficient of a tennis ball from fitting trajectory equations of motions in 3 degrees of freedom to Hawk-Eye trajectory time histories. Importantly, these trajectory equations of motion also account for both topspin and sidespin, extending the application of the methods to any arc trajectory encountered in a tennis rally.

2. METHODS

The following presents a methodology to minimise the difference between the ball trajectory measured by spatiotemporal data and that predicted by equations of motion, and in doing so enable the prediction of ball motion properties including the ball drag coefficient from tournament match play.

SPATIOTEMPORAL DATA

For this study, Tennis Australia have provided access to Hawk-Eye spatiotemporal ball tracking data from the 2022 Australian Open (Men's and Women's Singles). The provided Hawk-Eye data enables the computation of the time varying motion of a tennis ball in 3 degrees of freedom, as per the global coordinate system presented in Figure 1. Considering a defined time-step, the location of the tennis ball in the X-Y-Z coordinate space as predicted by the Hawk-Eye system can be identified at discrete moments in time.

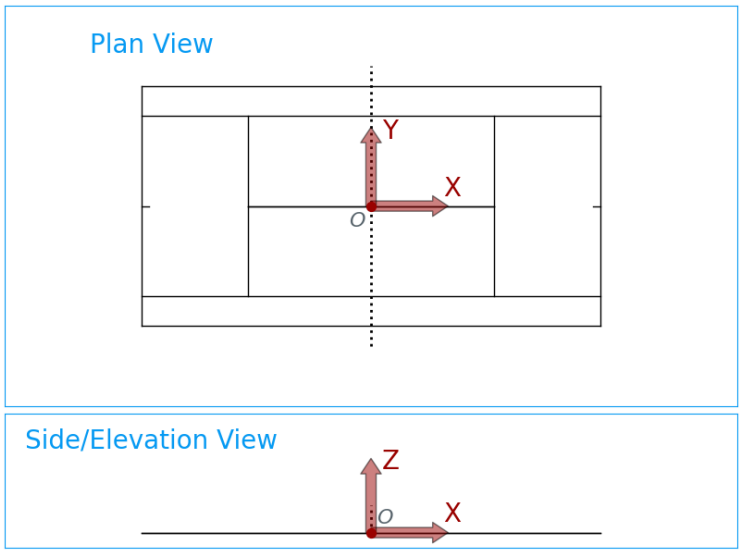


Figure 1: Global Coordinate System

EQUATIONS OF MOTION

The trajectory of a tennis ball can be calculated based on the gravitational and aerodynamic forces acting on a tennis ball (Cross, 2020) and assuming no other external forces, e.g. wind. From Cant (2020), this force balance in 3 degrees of freedom can result in the derivation of the following equations of motion for the trajectory of a ball:

$$dv_x/_{dt} = -kv\left[C_D v_x - \frac{C_L}{\omega}(\omega_y v_z - \omega_z v_y)\right] \quad (1)$$

$$dv_y/_{dt} = -kv\left[C_D v_y - \frac{C_L}{\omega}(\omega_z v_x - \omega_x v_z)\right] \quad (2)$$

$$dv_z/_{dt} = -g - kv\left[C_D v_z - \frac{C_L}{\omega}(\omega_x v_y - \omega_y v_x)\right] \quad (3)$$

where:

v = the absolute velocity of the tennis ball
v_x, v_y, v_z = the velocity of the tennis ball in the X, Y, Z axes respectively
ω = the absolute angular velocity of the tennis ball
$\omega_x, \omega_y, \omega_z$ = the angular velocity of the tennis ball in the X, Y, Z axes respectively
g = gravitational acceleration
k = ball constant
C_D = drag coefficient
C_L = lift coefficient

The ball constant k is defined as:

$$k = \frac{\rho \pi R^2}{2m} \quad (4)$$

where:

ρ = density of air
R = radius of tennis ball
m = mass of tennis ball

The drag coefficient, C_D, is assumed independent of speed and spin (Cross, 2014). However, there is potentially some variation with Reynolds Number and Spin Parameter (Goodwill 2004). The lift coefficient, C_L , has been found to be linearly proportional to the Spin Parameter, S (Cross, 2014):

$$C_L = C_L' S \quad (5)$$

where:

$$S = \frac{R\omega}{v} \quad (6)$$

The equations of motion, Eq. (1)-(3), can be solved numerically based on an assumed set of initial conditions:

- An initial ball position in X, Y, Z coordinate space, x_0, y_0, z_0.
- An initial ball speed, v_0.
- Ball topspin, θ, and sidespin, φ, assumed constant through the trajectory.
- An initial launch angle β_0 (relative to the horizontal plane) and launch heading, γ_0 (relative to the vertical plane) to translate the initial ball speed and spin into translational and rotational components in the X, Y, Z axes, $v_{x0}, v_{y0}, v_{z0}, \omega_{x0}, \omega_{y0}, \omega_{z0}$.

The translation of the ball topspin and ball sidespin into the rotational velocities required by Eq. (1)-(3), is achieved by the following transformation (Ivanov, 2021):

$$\omega_x = \varphi \sin(\beta) \quad (7)$$

$$\omega_y = \theta \cos(\gamma) - \varphi \cos(\beta) \sin(\gamma) \quad (8)$$

$$\omega_z = \theta \sin(\gamma) + \varphi \cos(\beta) \cos(\gamma) \quad (9)$$

OPTIMISATION

Assuming a fixed time-step, Eq. (1)-(3) can be solved numerically based on an assumed set of initial conditions for every arc trajectory of a tennis ball (e.g. from racket contact to impact with court, or from bounce off the court to racket contact). The initial ball position x_0, y_0, z_0, as well the initial launch angle β_0 and heading γ_0 can be derived from the Hawk-Eye ball trajectory data. However, the numerically solved trajectory can be varied based on the selection of the initial ball speed v_0, the topspin θ, the sidespin φ, the drag coefficient C_D, the lift coefficient C_L and the ball constant k. The ball constant k can be assumed based on the manufacturer specification for the ball properties (mass and diameter), whilst acknowledging there may be small variation from ball to ball. The remaining 5 parameters can be passed into an optimisation algorithm to minimise the mean-absolute-error between the Hawk-Eye trajectory and the numerically derived trajectory across all time-steps of the trajectory arc. This minimising optimisation was done using the SciPy python package, implementing the "Powell" method that is based on the methods from Powell (1964).

This optimisation can be done on every arc trajectory defined in the Hawk-Eye data. However, for the purposes of investigating the drag coefficient and specifically the new ball effect, this study investigated the predicted drag coefficient on serves greater than 50 m/s (180 km/h) in Men's Singles matches. The intention of this restriction is to isolate the data set to fast flat serves only, removing any potential dependency of speed and spin on the drag coefficient. Focusing the study on a specific shot type will assist in isolating variation in drag coefficient to ball degradation. Further investigation of the drag coefficient and lift coefficient predicted by these methods for all shot types is warranted as a future study. For instance, a kick serve with heavy spin would have an increase in lift coefficient that may make the ball more "livelier" in lateral directions.

3. RESULTS

Using the optimisations methods, Figure 2 presents an example of the data fitting between the Hawk-Eye trajectory and the numerically derived theoretical trajectory solution, with the very close match demonstrating the effectiveness of the optimisation algorithm in finding a solution to match the Hawk-Eye trajectory.

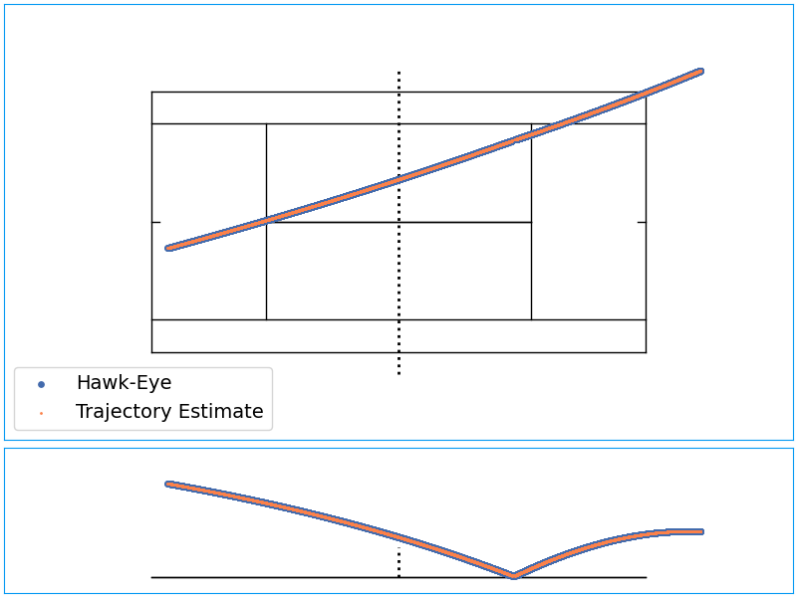


Figure 2: Example Trajectory Fitting

To examine the estimated drag coefficients in more detail, a representative Men's Singles match was selected that extended to 5 sets. Throughout the match, 7 different batches of tennis balls were used, with new balls called for by the umpire on 6 occasions. The drag coefficient estimation method was applied to every serve of the match of speeds greater than 50 m/s. Segregating each of these serves to their game number of use between 1 to 9, Figure 3 presents the drag coefficient estimate distributions. The figure indicates the trend of a progressive increase in drag coefficient with game number, with the highest rate of increase occurring in the earlier games. The variation in the drag coefficient within each game is indicative of the differing ball properties, surface condition and asymmetry of each ball played due to their varying accumulated use.

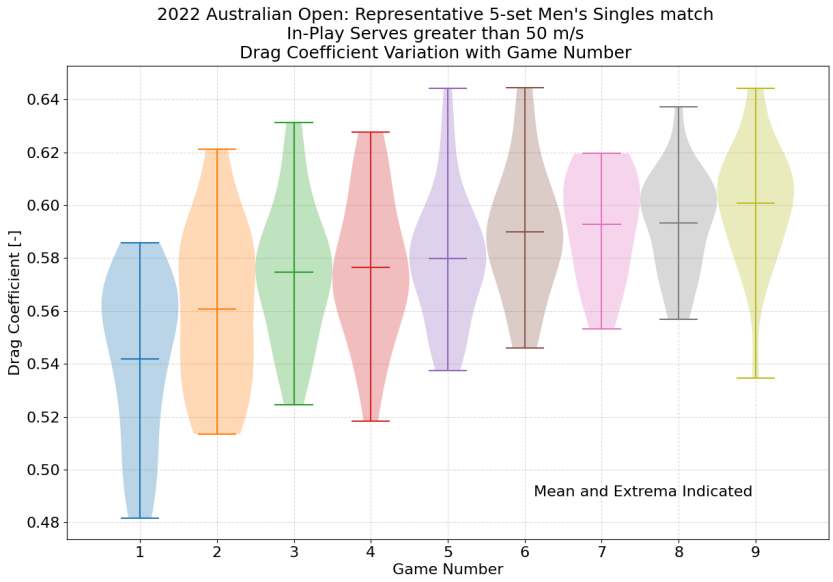


Figure 3: Drag Coefficient Variation with Game Number

Reviewing these serve drag coefficient estimates as a time history (using the match rally number as the time scale) enables highlighting the specific trends associated with the changeover to new balls. An example of this is shown in Figure 4, presenting the drag coefficient estimates from every serve with speed greater than 50 m/s from the third set of the match. New balls were called for by the umpire in the 5th game of the set, with the following two games characterised by serve drag coefficients between 0.475 and 0.55, on the lower end of the distribution relative to the rest of the match. With these lower drag coefficients, all other things being equal (e.g. no racket/string changes) the ball will tend to fly further and faster due to the lower drag resistance. Hence, if adjusted for correctly by a player, lower ball drag coefficients can reward aggressive play as opponents will have less time to react. Interestingly, there was rapid degradation of the tennis balls after these 2 games. Rapid degradation can be the result of extended rallies or due to players playing with high shot heaviness (combination of speed and spin).

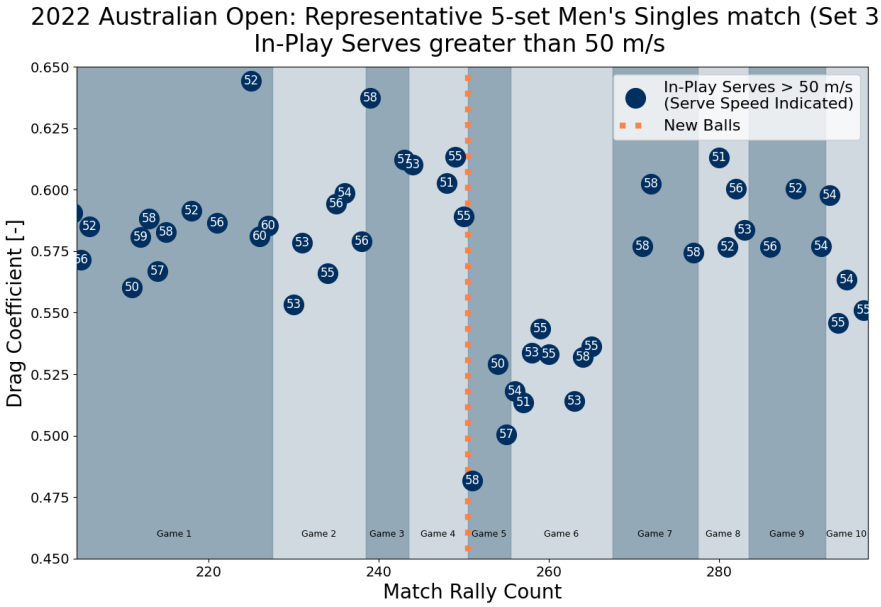


Figure 4: Drag Coefficient Variation during Set 3 of Representative 5-set Men's Singles Match

4. DISCUSSION

The drag coefficient results derived from the representative Men's Singles match demonstrated the step change reduction in drag coefficient with the changeover to new balls, as well as the gradual increase in drag coefficient with ball usage. This was a consistent trend across all matches that were assessed, albeit there was variance in the response that can be attributed to differences in ball properties and other variables such as the environment (temperature and humidity).

This information when viewed across the tournament can provide useful feedback for tennis ball manufacturers. Ball degradation is an accepted phenomenon of the sport (hence the official rule to change the balls after every 9 games), however, developing balls with a consistent specification and performance is assumed to be a targeting trait for manufacturers to gain approval from professional tennis players.

Professional tennis players can also use this information to adequately prepare for forthcoming matches. This may include developing strategies with respect to changing rackets with modified string tension to counteract the ball flying longer with new balls or at the very least awareness of this step-change in ball behaviour. From the perspective of the receiver, awareness of the reduction in reaction time may dictate court positioning when receiving to new balls.

Aside from the drag coefficient, deconstructing the ball trajectory time histories into equations of motion also enables the estimation of other trajectory parameters, including the lift coefficient and the ball topspin and sidespin. The lift coefficient quantifies the influence of the Magnus Effect (sideways force) of the spinning tennis ball, whilst differentiating between topspin and sidespin provides a more complete description of how spin influences the ball trajectory. As such, decomposing a ball trajectory into trajectory equations enables a higher fidelity interpretation of an executed shot. The effect of a small variation in any of the trajectory variables can then be investigated individually. The trajectory equations can also be extrapolated to answer hypothetical scenarios, for example, whether a volleyed ball would have landed out.

This paper has focused on the ball motion through the air, however, there is also the ball dynamics through impact with the racket and the court that could also be influenced by ball degradation (Steele, 2008). The ball motion into and out of the bounce as characterised by the spatiotemporal data could be used to calculate the ball's coefficient of restitution and coefficient of friction with the court using bounce equations such as those presented (Cross, 2020). Specifically, the ball's coefficient of restitution, the ratio of vertical velocity outbound and inbound of the bounce, could be investigated for evidence of association with ball degradation.

The methods presented in this paper are dependent on the accuracy of the spatiotemporal data from which the trajectory solution is being attempted to replicate. The methods presented are agnostic to the spatiotemporal data capture technology used. However, Hawk-Eye Technologies has an established presence in professional tennis through their electronic line-calling and broadcast applications. Hawk-Eye have previously indicated a mean error of 2.6 mm for their electronic line calling technology (Hawk-Eye Innovations, 2016), however, there has been no published results for the accuracy of the entire Hawk-Eye ball trajectory prediction.

Another potential source of error in the calculations is due to the assumed properties of the tennis balls. The manufacturer specification for the ball properties (mass and diameter) can provide a baseline assumption. However, there will be variation from ball to ball, and it is currently not possible to identify the exact ball used (and the associated properties) for every shot played in a tennis match.

5. CONCLUSIONS

This paper has presented a method to deconstruct ball trajectory time histories into equations of motion, with the focus of the study to estimate and discuss the estimated drag coefficient from match-play. However, this method can be equally used to extract and investigate other properties from the ball trajectory, including the lift coefficient and the different types of spin on the ball. Furthermore, the methods can be used for other applications such as trajectory extrapolation, e.g. enable prediction of whether a volleyed ball would have landed out.

Considering fast flat serves greater than 50 m/s (180 km/h) in Men's Singles matches from the 2022 Australian Open, it was found that in general there was a gradual increase in the ball drag coefficient with usage, followed by a noticeable step change reduction in the drag coefficient at the changeover between old and new balls. This information can be used by tennis ball manufacturers/tournament organisers to monitor the degradation and consistency of tennis balls during tournament match play. Whilst tennis players/coaches can also use this information to adequately prepare for forthcoming matches by ensuring they have developed strategic plans to adjust or counteract the effect of the step change reduction in drag coefficient associated with new balls.

Acknowledgements

We wish to thank Tennis Australia for their provision of Hawk-Eye data from the 2022 Australian Open.

References

Cant, O., Kovalchik, S., Cross, R., & Reid, M. (2020). Validation of ball spin estimates in tennis from multi-camera tracking data. *Journal of Sports Sciences, 38*(3), 296-303.

Choppin, S., Albrecht, S., Spurr, J., & Capel-Davies, J. (2018). The effect of ball wear on ball aerodynamics: An investigation using hawk-eye data. In *Multidisciplinary Digital Publishing Institute Proceedings* (Vol. 2, No. 6, p. 265).

Cross, R., & Lindsey, C. (2014). Measurements of drag and lift on tennis balls in flight. *Sports Engineering, 17*(2), 89-96.

Cross, R. (2020). Calculations of groundstroke trajectories in tennis. *Sports Engineering, 23*(1), 1-10.

Goodwill, S. R., Chin, S. B., & Haake, S. J. (2004). Aerodynamics of spinning and non-spinning tennis balls. *Journal of wind engineering and industrial aerodynamic, 92*(11), 935-958.

International Tennis Federation (2022). Grand Slam Tournament Regulations. *2022 Official Grand Slam Rule Book*, www.itftennis.com.

Hawk-Eye Innovations (2016). How it works. *Electronic Line Calling Technology, ELC_How_it_Works.pdf (pulselive.com)*.

Ivanov, A. I. (2021). Three-dimensional motion of tennis ball. *Technical Journal of Daukeyev University, 1*(1), 57-68.

Mehta, R. D., & Pallis, J. M. (2001). The aerodynamics of a tennis ball. *Sports Engineering, 4*(4), 177-189.

Powell, M. J. (1964). An efficient method for finding the minimum function of several variables without calculating. *The computer journal, 7*(2), 155-162.

Steele, C., Jones, R., & Leaney, P. (2006). Factors in tennis ball wear. In *The Engineering of Sport 6* (pp. 373-378). Springer, New York, NY.

Steele, C., Jones, R., & Leaney, P. (2008). Improved tennis ball design: incorporating mechanical and psychological influences. *Journal of Engineering Design, 19*(3), 269-284.

CHALLENGES AND CONSIDERATIONS IN DETERMINING THE VALIDITY OF PLAYER TRACKING SYSTEMS IN TEAM SPORTS

Sam Robertson [a]

[a] *Institute for Health & Sport, Victoria University, Australia*
[b] *Corresponding author: sam.robertson@vu.edu.au*

Abstract

Electronic performance & tracking systems (EPTS) are commonly used to track the location and velocity of athletes in many team sports. A range of associated applications using the derived data exist, such as assessment of athlete characteristics, informing training design, assisting match adjudication and providing fan insights for broadcast. Consequently the quality of such systems is of importance to a range of stakeholders. The influence of both systematic and methodological factors on this resulting quality is non-trivial. Highlighting these allows for the user to understand their strengths and limitations in various decision-making processes, as well as identify areas for research and development. In this paper, a number of challenges and considerations relating to the determination of EPTS validity for team sport are outlined and discussed. The aim of this paper is to draw attention of these factors to both researchers and practitioners looking to inform their decision-making in the EPTS area. Addressing some of the posited considerations in future work may represent best practice; others may require further investigation, have multiple potential solutions or currently be intractable.

Keywords: GPS, computer vision, performance analysis, coaching, validation

This Page is Left Intentionally Blank

Otherside – Using Expected Points to Evaluate Defensive Actions in Australian Rules Football

Liam Crowhurst [b], Robert Nguyen [a,]

[a] *University of New South Wales*

[b] *Corresponding author: liam.m.crowhurst@gmail.com*

Abstract

Expected points are a measure of the quality of an attempted shot on goal. Typically, expected points are calculated using covariates such as field location, shot context, type of shot among other things. Summary statistics of expected scores can provide additional context to results, insights into general shot quality and a team's overall effectiveness in front of goal.
Recently, European Football has looked at assigning a similar metric to keepers, known as expected saves, however an equivalent metric has not been developed for Australian Rules Football. Firstly, we have developed a reproducible and publicly available expected points model for Australian Rules using a 'play by play' dataset recently released online. Secondly, we use the 'play by play' dataset to extract defender actions and their effects to derive a new metric, defensive points saved. Lastly, we produced interactive visual maps that allow analysts, media and fans to explore the data, providing visual context to analytical insights,

Keywords: expected points, generalised additive models