MathSport International 2025

-- CONFERENCE PROCEEDINGS --

11th International Conference on Mathematics in Sport

Luxembourg (Luxembourg)

4 – 6 June 2025

Hosted by the University of Luxembourg, organized by team MIDAS

ISBN: 9789083581408

Preface

This volume presents a selection of papers from the 11th MathSport International Conference, held at the University of Luxembourg and organized by team MIDAS from 4–6 June 2025. The conference brought together researchers and practitioners from around the world to explore the intersection of mathematics and sport.

In addition to four keynote lectures—delivered by Prof. Zuccolotto, Prof. Spieksma, Prof. Seil, and Prof. Pawlowski—the conference featured 104 presentations covering a broad spectrum of topics. Of these, 26 contributions are included in this volume as short papers, reflecting the diversity of sports and methodological approaches represented at the event. This variety aligns with the vision of the MathSport committee to foster interdisciplinary dialogue and innovation in the field.

The papers are arranged in alphabetical order by the first author's surname. We hope this collection will serve as a valuable resource for researchers and practitioners, and that it will inspire further advances at the interface of mathematics and sport.

Sponsors

EWG OR IN SPORTS

UNIVERSITY OF LUXEMBOURG
Department of Mathematics

Content

Teaching probability theory through tennis

T. Barnett*, A. Bedford** and E. Mealy**

*Macquarie University + email address: tristan@strategicgames.com.au
** University of the Sunshine Coast + email address: {abedford,emealy}@usc.edu.au

Abstract

This article obtains distributional characteristics for the length of a tennis game, which aids in teaching students an application of key statistical and computing concepts. Although the mean and variance help to describe the distribution, it is demonstrated that these two characteristics are insufficient for measuring 'risk' and therefore other characteristics such as the coefficients of skewness and excess kurtosis are obtained. By setting up recursion formulas with the appropriate boundary conditions in spreadsheets, the first four moments of the total number of points played in a game conditional on the point score are obtained, which in turn are converted to distributional characteristics. Further, the distribution of the total number of points played is compared to the distribution of the number of points remaining, to show graphically that the variance and coefficients of skewness and excess kurtosis remain unchanged by adding a constant to all values of the variable. The above could form an interesting teaching exercise in using Excel and probability theory, and provide a live student-built solution of tennis matches whilst in-play.

1 Introduction

Suppose we wish to calculate the mean (average value) number of points remaining in a game. Using a standard formula for calculating the mean value of a discrete distribution, this can be calculated by $\mu = E(X)$. Similarly, the variance (standard deviation squared) of the number of points remaining in a game can be calculated by $\sigma 2 = E(X2) - E(X)2;$ which is recognized as a measure of the dispersion of a set of data from its mean. Barnett and Clarke (2002) apply backwards recurrence formulas to obtain the mean number of points remaining in a game from any point score within the game and show that when the server has a 54% chance of winning a point on serve, the mean number of points remaining from the outset is 6.7. Barnett et al (2006a) applies generating functions to calculate the mean and variance of the number of points remaining in a game from the outset and show that when the server has a 60% chance of winning a point on serve, the mean number of points remaining from the outset is 6.5 with a corresponding standard deviation of 2.6. Barnett (2013) applies backward recurrence formulas to obtain the mean and variance of the number of points remaining in a game from any point score within the game. For example, when the server has a 60% chance of winning a point on serve; from 30-0 the mean number of points remaining in the game is 5.6 with a corresponding standard deviation of 2.2.

Both the mean and variance contain important information to describe the shape of the distribution and these characteristics could be used to compare one distribution to another. For example, comparing the mean and variance of a tiebreak set to an advantage set to help identify why 'long' matches can occur (Barnett

and Clarke, 2005). However, if a distribution is not symmetric (as typically occurs in a game and an advantage set) the mean and variance do not 'adequately' describe the shape of the distribution. Two other characteristics that are used to describe the distribution and measure risk are skewness and kurtosis. Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the centre point. Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. Note that excess kurtosis will be used throughout the article such that $excess\ kurtosis = kurtosis - 3$, so the normal distribution has an excess kurtosis of 0, and therefore a kurtosis of 3.

This article uses techniques of recursion formulas and generating functions to obtain the mean, variance, and coefficients of skewness and excess kurtosis of the number of points remaining in a game conditional on the point score; by noting from above that the literature has only obtained the mean and variance. This article also obtains the distributions of the total number of points and the number of points remaining in a game, where the two distributions are compared to demonstrate graphically the invariance property of variance $V(X + c) = V(X)$. The methods can be readily set up in spreadsheets to obtain numerical results and could form an interesting teaching exercise in probability theory by allowing students to obtain distributional characteristics of a tennis game.

2 Method

2.1 Probability of winning a game

We have two players; player A and player B.

$$Let\ pA\ represent\ a\ constant\ probability\ of\ player\ A\ winning\ a\ point\ on\ serve \quad (1)$$

$$Let\ pB\ represent\ a\ constant\ probability\ of\ player\ B\ winning\ a\ point\ on\ serve \quad (2)$$

Thus, pA and pB become the two parameters for the model.

$$Let\ qA\ represent\ a\ constant\ probability\ of\ player\ A\ losing\ a\ point\ on\ serve \quad (3)$$

$$Let\ qB\ represent\ a\ constant\ probability\ of\ player\ B\ losing\ a\ point\ on\ serve \quad (4)$$

It follows that $qA = 1 - pA\ and\ qB = 1 - pB$

Barnett and Clarke (2002) use backwards recursion in an Excel spreadsheet to calculate the condition probabilities of winning a game. Barnett et al. (2006b) use forwards recursion in an Excel spreadsheet to calculate the chances of reaching scorelines. The latter calculations are used to calculate the distributions of the total number of points played and the number of points remaining in a game, and represented graphically in Section 2.4.

2.2 Moments of the number of points in a game

Let $XA(a,b)$ and $YA(a,b)$ be random variables of the total number of points played in a game and the number of points remaining in a game respectively at point score (a,b) for player A serving.

Let $E(XA(a,b))$ and $E(YA(a,b))$ represent the first moment (or expectation) of the total number of points played in a game and the number of points remaining in a game respectively at point score (a,b) for player A serving. It can be shown that

$$E(XA(a,b)) = pAE(XA(a+1,b)) + qAE(XA(a,b+1)) \quad (5)$$

$$E\big(YA(a,b)\big) = 1 + pAE\big(YA(a+1,b)\big) + qAE\big(YA(a,b+1)\big) \quad (6)$$

Let $XnA(a,b)$ represent the n[th] power of the random variable $XA(a,b)$ for each $n > 0$.

Then $E(XnA(a,b))$ represents the nth moment with the following important relation $XnA(a,b) = (a + b + YnA(a,b))$ which, when expanded involves various powers of $YA(a,b)$. Thus, calculation must proceed recursively, i.e. first moment, second moment, and so on. These higher moments can then be used to calculate other statistics such as variance, skewness and excess kurtosis. This is an excellent student activity, suited for Excel.

Taking expectations gives the following recurrence formula:

$$E\big(XnA(a,b)\big) = pAE\big(XnA(a+1,b)\big) + qAE\big(XnA(a,b+1)\big) \quad (7)$$

The boundary values for $XnA(a,b)$ are obtained as:

$$E\big(XnA(4,0)\big) and\ E\big(XnA(0,4)\big) = 4n, \quad (8)$$

$$E\big(XnA(4,1)\big) and\ E\big(XnA(1,4)\big) = 5n, \quad (9)$$

$$E\big(XnA(4,2)\big) and\ E\big(XnA(2,4)\big) = 6n \quad (10)$$

The boundary values at $E(XnA(3,3))$ are obtained as follows:

The moment generating function for the total number of points played in a game from (3,3) with player A serving is given by:

$$MXA(3,3)(t) = \frac{(pA2 + qA2)e8t}{1 - 2pAqAe2t} \quad (11)$$

Therefore:

$$E\big(XA(3,3)\big) = M(1)XA(3,3)(0) = \frac{4(3pAqA - 2)}{2pAqA - 1} \quad (12)$$

$$E\big(X2A(3,3)\big) = M(2)XA(3,3)(0) = \frac{8(18pA2qA2 - 23pAqA + 8)}{(2pAqA - 1)2} \quad (13)$$

$$E\big(X3A(3,3)\big) = M(3)XA(3,3)(0) = \frac{16(108pA3qA3 - 200pA2qA2 + 131pAqA - 32)}{(2pAqA - 1)3} \quad (14)$$

$$E\big(X4A(3,3)\big) = M(4)XA(3,3)(0) = \frac{32(648pA4qA4 - 1556pA3qA3 + 1462pA2qA2 - 655pAqA + 128)}{(2pAqA - 1)4} \quad (15)$$

Table 1 represents the first moment (equivalent to the mean) of the total number of points played in a game at various score lines for player A serving given p_A=0.6

Table 1 The first moment of the total number of points played in a game at various score lines for player A serving given pA=0.6

		B score				
		0	15	30	40	game
	0	6.5	7.0	6.8	5.8	4
A	15	6.2	7.0	7.5	7.0	5
score	30	5.6	6.7	7.8	8.3	6
	40	4.8	6.0	7.5	9.8	
	Game	4	5	6		

2.3 Parameters of distribution of the number of points in a game

Let $\mu(XA(a,b))$, $\sigma 2(XA(a,b))$, $\gamma 1(XA(a,b))$ and $\gamma 2(XA(a,b))$ represent the mean, variance, coefficient of skewness and coefficient of excess kurtosis of the total number of points played in a game at point score (a,b) for player A serving. It provides a challenge now for students to ascertain this.

The following standard results are used to obtain

$$\mu\big(XA(a,b)\big), \sigma 2\big(XA(a,b)\big), \gamma 1\big(XA(a,b)\big) and\ \gamma 2\big(XA(a,b)\big) \tag{16}$$

$$E\big(XA(a,b)\big) = \mu\big(XA(a,b)\big) \tag{17}$$

$$E\big(X2A(a,b)\big) = \sigma 2\big(XA(a,b)\big) + E\big(XA(a,b)\big)2 \tag{18}$$

$$E\big(X3A(a,b)\big) = \frac{\gamma 1\big(XA(a,b)\big)\sigma 2\big(XA(a,b)\big)3}{2} + 3E\big(X2A(a,b)\big)E\big(XA(a,b)\big) - 2E\big(XA(a,b)\big)3 \tag{19}$$

$$E\big(X4A(a,b)\big) = \gamma 2\big(XA(a,b)\big)\sigma 2\big(XA(a,b)\big)2 + 4E\big(X3A(a,b)\big)E\big(XA(a,b)\big) + 3E\big(X2A(a,b)\big)2 - 12E\big(X2A(a,b)\big)E\big(XA(a,b)\big)2 + 6E\big(XA(a,b)\big)4 \tag{20}$$

$$Let\ \mu\big(YA(a,b)\big), \sigma 2\big(YA(a,b)\big), \gamma 1\big(YA(a,b)\big) and\ \gamma 2\big(YA(a,b)\big) represent\ the\ mean, variance, coefficient\ of\ skewness\ and\ coefficient\ of\ excess\ kurtosis\ of\ the\ number\ of\ points\ remaining\ in\ a\ game\ at\ point\ score\ (a,b) for\ player\ A\ serving. \tag{21}$$

Finally, the following relations are used to obtain $\mu(YA(a,b))$, $\sigma 2(YA(a,b))$, $\gamma 1(YA(a,b))$ and $\gamma 2(YA(a,b))$

$$\mu\big(YA(a,b)\big) = \mu\big(XA(a,b)\big) - a - b \tag{22}$$

$$\sigma 2\big(YA(a,b)\big) = \sigma 2\big(XA(a,b)\big) \tag{23}$$

$$\gamma 1\big(YA(a,b)\big) = \gamma 1\big(XA(a,b)\big) \tag{24}$$

$$\gamma 2\big(YA(a,b)\big) = \gamma 2\big(XA(a,b)\big) \tag{25}$$

Table 2 represents the mean number of points remaining in a game at various score lines for player A serving given p_A=0.6.

Table 2 The mean number of points remaining at various score lines for player A serving given pA=0.6

		Bscore			
		0	15	30	40
Ascore	0	6.5	6.0	4.8	2.8
	15	5.2	5.0	4.5	3.0
	30	3.6	3.7	3.8	3.3
	40	1.8	2.0	2.5	3.8

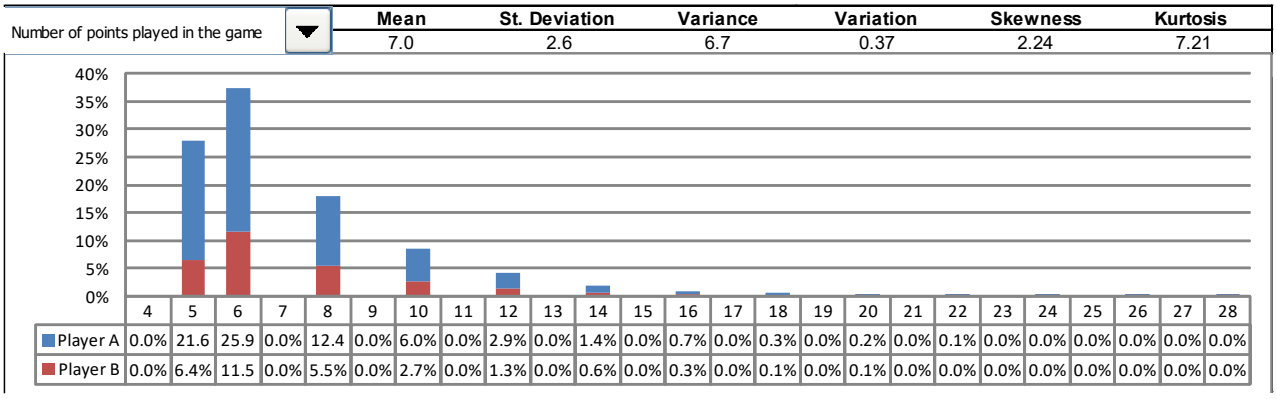
2.4 Distribution of the number of points in a game

Figure 1 represents the distribution of the total number of points played in a game from 15-15 (a=1, b=1) for player A serving given p_A=0.6 (Barnett, 2013). Notice how the blue colour is the chances of player A winning the game and the maroon colour is the chances of player B winning the game. For example, the chances of player A winning the game to 15 is given by the frequency distribution of blue for 5 total points played. This numerical value is 20.74%. Similarly, the chances of player B winning the game to 15 is given by the frequency distribution of maroon for 5 total points played. This numerical value is 6.14%. Therefore,

the game finishing with either player winning to 15 (or 5 total points played) is given by 20.74%+6.14%=26.9%. Figure 2 represents the distribution of the number of points remaining in a game from 15-15 for player A serving given p_A=0.6. Note that the shapes of both distributions from figures 1 and 2 are the same. In other words, the variance and coefficients of skewness and excess kurtosis remain unchanged by adding a constant (c) to all values of the variable. This is widely known as an invariant property in variance such that V(X+c)=V(X). The differences in both distributions are reflected only by shifting the horizontal scale by a constant; as reflected by the mean property M(X+c)=M(X)+c. As above, let $X_A(a,b)$ and $Y_A(a,b)$ be random variables of the total number of points played in a game and the number of points remaining in a game respectively at point score (a,b) for player A serving. By simple logic, $Y_A(a,b)+(a+b)=X_A(a,b)$, where a and b are represented by player's A's score and player's B's score respectively. It follows that $V(Y_A(a,b)+(a+b))=V(X_A(a,b))$. Using the invariant property of variance above $V(Y_A(a,b)+(a+b))=V(Y_A(a,b))$, since (a+b) is a constant. Therefore $V(X_A(a,b))= V(Y_A(a,b))$. Note that the coefficient of variation = standard deviation/mean.

3 Conclusions

It has been demonstrated in this article how setting up recursion formulas with the appropriate boundary conditions in spreadsheets can generate the first four moments of the total number of points played in a game conditional on the point score. Standard formulas are then used to obtain the distributional characteristics of the mean, variance, and coefficients of skewness and excess kurtosis of the total number of points played in a game and the number of points remaining in a game conditional on the point score. These two distributions are then compared and used to show graphically that the variance and coefficients of skewness and excess kurtosis remain unchanged by adding a constant to all values of the variable. The methods outlined could form an interesting teaching exercise in probability theory by allowing students to obtain distributional characteristics of a tennis game. This in turn allows students to build their own tennis calculator and become familiar with using spreadsheet software such as Excel. Similar methods can be obtained for the number of points remaining in a tiebreak game, number of games remaining in a tiebreak or advantage set, and the number of sets remaining in a best-of-3 or best-of-5 set match.



	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
Player A	0.0%	21.6	25.9	0.0%	12.4	0.0%	6.0%	0.0%	2.9%	0.0%	1.4%	0.0%	0.7%	0.0%	0.3%	0.0%	0.2%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Player B	0.0%	6.4%	11.5	0.0%	5.5%	0.0%	2.7%	0.0%	1.3%	0.0%	0.6%	0.0%	0.3%	0.0%	0.1%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Figure 1 Distribution of the total number of points played in a game from 15-15 for player A serving given pA=0.6

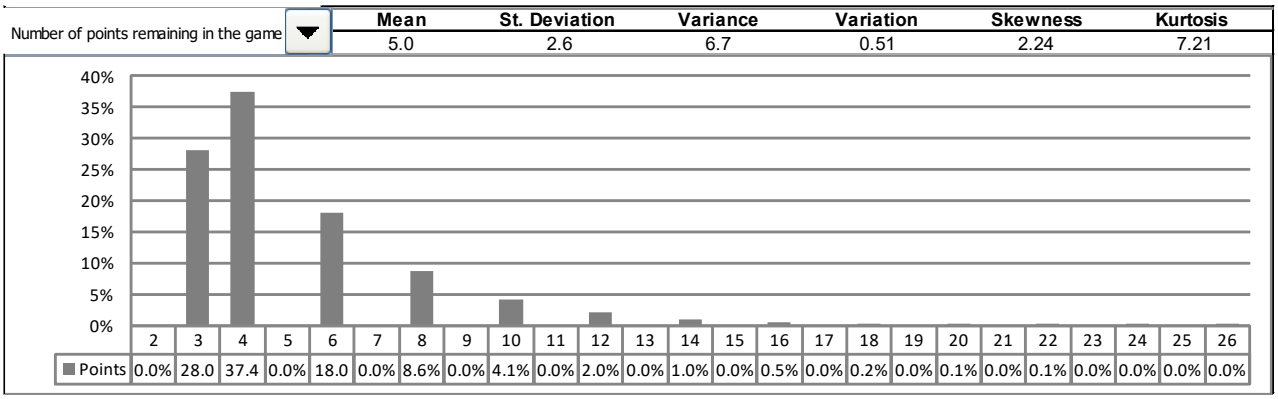


Figure 2 Distribution of the number of points remaining in a game from 15-15 for player A serving given pA=0.6

References

[1] Barnett T and Clarke SR (2002). Using Microsoft Excel to model a tennis match. Proceedings of the 6th Australian Conference of Mathematics and Computers in Sport.

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

[3] Barnett T, Brown A and Pollard G (2006a). Reducing the likelihood of long tennis matches. Journal of Sports Science & Medicine. 5(4), 567-574.

[4] Barnett T, Brown A and Clarke SR (2006b). Developing a model that reflects outcomes of tennis matches. Proceedings of the 8th Australian Conference on Mathematics and Computers in Sport.

[5] Barnett T (2013). Developing a tennis calculator to teach probability and statistics. Journal of Medicine and Science in Tennis. 18(1), 30-34.

Alternate Scoring Systems to a Test Cricket Series

T. Barnett*, G. Pollard**, A. Bedford*** and E. Mealy***

*Macquarie University + email address: tristan.barnett@students.mq.edu.au

** University of Canberra + email address: grahamhpollard@gmail.com

*** University of the Sunshine Coast, Sippy Downs, Queensland, Australia + email address: {abedford,emealy}@unisc.edu.au

Abstract

The relatively high draw probability in test cricket has fluctuated over the years from around 25% in 2003 to around 15% in 2022. These statistics indicate that players are playing more aggressively to score runs to increase their chances of winning the match due to the limited number of overs available to bowl the opposing side out twice to reduce the draw probability, and this strategy inadvertently increases the chances of the opposing side winning since by scoring runs faster there may be an increased chance of losing wickets. The draw probability can be reduced in test cricket by increasing the number of allowable overs, where the current system has a maximum of about 450 overs (90 overs over 5 days). Given that One Day International (ODI) cricket plays a maximum of 100 overs in a day, it could then appear 'practical' to extend the number of overs in test cricket from 90 to 100 overs per day. Also, an additional 6th day could also appear to be a 'practical' strategy to reduce the draw probability. Another method to reduce the draw probability in test cricket is by playing only one innings for each side (compared to the standard two innings). Thus, this presentation will discuss alternate scoring systems to a 3-test and 5-test series based on the discussion above using the following key objectives:

a) reduce the draw probability each match
b) increase the chances of the stronger team winning each match
c) reduce the draw probability of the series
d) increase the chances of the stronger team winning the series
e) reduce the length of the series

1 Introduction

The percentage of test cricket matches resulting in a draw has notably declined over time, with approximately 25% matches resulting in a draw in 2003 to only 15% in 2022 (Barnett and Pollard 2024). This reduction may be attributed to evolving team strategies aimed to force a match result within the limited number of available overs. Currently, test cricket allows for a maximum of 450 overs (90 overs across five days), which can limit match outcomes in slow-paced tests.

Another common scenario that contributes to drawn matches occurs when the batting side scores fewer runs in their first inning and lacks sufficient overs in the remaining match to realistically pursue a win. In such situations, the team often adopts a defensive strategy to avoid losing, rather than attempting to win. Again, it would seem of interest to extend the length of allowable overs in test cricket to increase competitiveness and spectator engagement.

An alternative method to reduce the likelihood of drawn matches in test cricket is to limit each side to one innings, rather than the traditional two. This method is permitted under Law 13.1.1 of the MCC Laws of Cricket, which states "A match shall be one or two innings for each side according to agreement reached before the match" (MCC Laws).

While single matches can benefit from these structural changes, it is important to note that test cricket is played over a series of matches. A well-known test cricket series is The Ashes, consisting of five test matches played over five days for each match, thus allocating 25 days of scheduled play. However, research by Pollard and Barnett (2024) suggests that shorter series formats, such as 3-test or 4-test series may offer a more efficient and balanced structure when combined with modifications to match durations

or inning limits. The Guinness World Records (2013) states the multi-format adopted for the Women's Ashes international test series was a world first. Utilising data from Cricket.com.au. (2025), a reduction in the percentage of drawn games and therefore drawn series can be seen, with Women's ashes from 1931 to 2011 resulting in a drawn series almost 39% of the time, while the multi-format series has experienced only a 25% draw rate.

Australian Summer weather presents another variable that affects the probability of drawn test matches, and accordingly drawn international test series. For instance, when analysing data from ESPN Cricinfo, of the 12 weather affected international test matches played at the Gabba since 1960, 10 of those games went on to finish in a draw.

This project will analyse alternate series structures with series composed of 3-test and 5-test series based on the discussion above using the following key objectives:

a) Reduce the draw probability in individual matches
b) Increase the likelihood of the stronger team winning each match
c) Reduce the probability of a drawn series
d) Increase the likelihood of the stronger team winning the series
e) Reduce the total duration of the series

2 Methods

This paper presents an alternative match and series structures for both 3-test and 5-test cricket series to reduce the likelihood of drawn results. These alternate systems will incorporate variations in match duration and format, specifically using one inning matches, six-day matches and a combination of both.

To ensure a systematic comparison of test series structures, the following assumptions are made based on the current test match format of a maximum of 90 overs per day across 5 days:

- All matches are played with a maximum of 90 overs/day
- Two-innings matches played over 6 days
- One-innings matches played over 3 days
- One-innings matches played over 4 days
- The total number of days used in a series must not exceed 25 days for a 5-test series and 15 days for a 3-test series

The current scoring systems will serve as a reference for evaluating the impact of proposed alternative scoring systems. Specifically, the standard structures are as follows:

Scoring system (1) 3 test series, two-innings, played over max 5 days, 90 overs per day (max days in series 15)

Scoring system (2) 5 test series, two-innings, played over max 5 days, 90 overs per day (max days in series 25)

Based on these assumptions and existing structures, a range of alternate scoring systems were developed for both 3-test and 5-test series. These combinations vary the number of one innings and two innings matches, as well as the number of days allocated to each match, while ensuring the total series length remains within the current limits. The proposed configurations for a 3-test and 5-test series are summarised below in Table 1 and Table 2 respectively.

To assist in evaluation, we use a Monte Carlo simulation incorporating standard cricket rules and probabilistic events. The simulation is based on a ball-by-ball approach, where each batsman's probability of scoring runs or being dismissed is modelled on their position in the batting order. The probabilities for scoring 0, 1, 2, 3, 4, 5, and 6 runs, chances of no balls, wides, and wickets are drawn from historical data gathered from 170 world test series names from 2020 to 2025. Furthermore, the probabilities of a wicket falling on a particular ball varies by batting position based on the historical data. The innings are simulated within a upper over limit upon the system, with differing cases. For example, for Scoring System (3) Team 1 being restricted to a maximum of 180 overs (i.e. if still batting they declare), and the total combined

overs for both teams capped at 270. These constraints ensure the simulation adheres to a more realistic test match format.

Table 1 Alternate Scoring Systems to a Standard 3-test Cricket Series

Scoring system	No. of matches	No. of one-innings matches	No. of two-innings matches	Max days one-innings matches	Max days in series
(3)	3	3	-	3	9
(4)	3	3	-	4	12
(5)	3	2	1	3	12
(6)	3	2	1	4	14
(7)	3	1	2	3	15
(8)	5	5	-	3	15

Table 2 Alternate Scoring Systems to a Standard 5-test Cricket Series

Scoring system	No. of matches	No. of one-innings matches	No. of two-innings matches	Max days one-innings matches	Max days in series
(8)	5	5	-	3	15
(9)	5	4	1	3	18
(10)	3	-	3	-	18
(11)	5	5	-	4	20
(12)	5	3	2	3	21
(13)	7	7	-	3	21
(14)	5	4	1	4	22
(15)	5	3	2	4	24
(16)	5	2	3	3	24

The simulation tracks match progression, recording individual batsman statistics (runs and balls faced) while continuously checking for stopping conditions. Key stopping criteria include exceeding the maximum overs, all-out conditions, or a chasing team's score surpassing that of the opposing team to win. The simulation stops under end of game conditions. We then evaluate a series results based upon these probabilities, eg. For (3) match series it is simple to calculate m draws in three matches as $P(D = m) = \binom{3}{m} P(D_1)^{\mathrm{m}}(1 - P(D_1)^{3-m})$ and so on for wins (W1). So series wins are easily derived $P(\text{Team 1 wins series}) = P(W_1)^3 + 3P(W_1)^2P(W_2) + 3P(W_1)^2P(D)$; $P(\text{Team 2 wins series}) = P(W_2)^3 + 3P(W_2)^2P(W_1) + 3P(W_2)^2P(D)$ and $P(\text{Drawn series}) = P(D)^3 + 3P(W_1)P(W_2)P(D)$. To win a 5-match series we have $P(\text{Team 1 wins series}) = \sum_{k=3}^{5}\binom{5}{k} P(W_1)^k \cdot P(W_2)^{5-k} \cdot P(D)^{5-k}$ in a non-mixed format series. In a mixed format, we need to adjust the winning and associated probabilities for the differing formats. This of course varies for series with team 1 winning under 1 innings 3 days and team 1 winning under 2 innings 6 days mixed in the same series.

3 Results

Tables 3 and 4 presents the converged probabilities of a simulation of 20000 games per match design utilising run, extras and wicket probabilities calculated from the analysis of 120 World Test Series matches from 2022-2025. Each table presents the probabilities of Team 1 winning ($P(W1)$), Draw(D), and Team 2 winning($P(W2)$). Table 3 presents the case in which no team declares in 1 innings games,

and any team declares if not all out by 180 overs in a 6 day, 2 innings match. Table 4 presents team 1 declares after 180 overs in 3 day games and 135 overs in either innings of 2 inning 6 day games.

Table 3 Converged probabilities for Team 1 win, Draw and Team 2 win

Match Design	Team 1 wins $P(W1)$	Draw D	Team 2 wins $P(W2)$
1 innings 3 days	0.5050;	0.0076	0.4874
1 innings 4 days	0.5118	0.0032	0.485
2 innings 6 days	0.4886;	0.0092	0.5022

Table 4 Converged probabilities for Team 1 win, draw and Team 2 win with declaration

Match Design	Declaration criteria	Team 1 wins $P(W1)$	Draw D	Team 2 wins $P(W2)$	$P(t1wins\|declared)$
1 innings 3 days	Team 1 decl @ 180 overs	0.4892	0.0088	0.5020	0.71
1 innings 4 days	Team 1 decl @ 180 overs	0.4960	0.0026	0.5014	1
2 innings 6 days	Team 1 decl @ 135 overs	0.4850	0.0046	0.5104	0.88

Table 5 3-test series

Scoring system	Match draw probability	Match draw probability with declaration	Series draw probability	Series draw probability with declaration	Maximum days in series
(3)	0.0076	0.0088	0.0128	0.0126	9
(4)	0.0032	0.0026	0.0050	0.0036	12
(5)	0.0076;0.0092	0.0088;0.0026	0.0132	0.0128	12
(6)	0.0032;0.0092	0.0026;0.0046	0.0070	0.0056	14
(7)	0.0076;0.0092	0.0088;0.0026	0.0134	0.0096	15
(8)	0.0092	0.0046	0.0168	0.0082	15

Table 6 5-test series

Scoring system	Match draw probability	Match draw probability with declaration	Series draw probability	Series draw probability with declaration	Maximum days in series
(8)	0.0092	0.0046	0.0168	0.0082	15
(9)	0.0776;0.0092	0.0088;0.0046	0.0134	0.0142	18
(10)	0.0092	0.0046	0.0122	0.0070	18
(11)	0.0032	0.0026	0.0054	0.0064	20
(12)	0.0076;0.0092	0.0088;0.0046	0.0174	0.0200	21
(13)	0.0076	0.0088	0.0132	0.0158	21
(14)	0.0032,0.0092	0.0026,0.0046	0.0072	0.0066	22
(15)	0.0032,0.0092	0.0026;0.0046	0.0130	0.0064	24
(16)	0.0076;0.0092	0.0088;0.0046	0.0172	0.0114	24

Tables 5 and 6 show we have reduced the draw prob for the series in all alternate formats (3-16) in comparison to the current 3-test and 5-test formats by playing two-innings over 5 days matches.

Furthermore, we show that the chance of Team 1 winning when they have declared is highest in 1 innings 4 days match design, but the declaration overall increases the probability that Team 1 wins.

Notably, the restriction of resources for Team 1 allows for a result, circumventing any adjustment of probabilities and variance in the simulations. Imposing a declaration allows the reduction in likelihood of a draw and the chances of a result favourable to Team 1. Any reduction in dismissal likelihood for Team 1 would lead naturally to an improved score (due to not shorting out on the declaration) and improvements therein are imagined.

4 Discussion

The rationale behind the study was to develop alternate series structures to reduce the draw probability in test cricket matches based on the following two observations from the current two-innings structure played over a maximum of 5 days:

1) players are playing more aggressively to score runs to increase their chances of winning the match due to the limited number of overs available to bowl the opposing side out twice to reduce the draw probability, and this strategy may inadvertently increase the chances of the opposing side winning since by scoring runs faster there may be an increased chance of losing wickets

2) a team batting second scored fewer runs compared to the other team in the 1st innings and is unable to win the test in the 2nd innings due to not having enough overs remaining and is thus playing defensively to play for a draw

An obvious and simple method to achieve the objective to reduce the draw probability is to allowing for an additional sixth day of play. By playing six days of cricket, a 3-test series a total of 18 days of play would need to be scheduled and in a 5-test series a total of 30 days would need to be scheduled. This increase in duration of playing time to reduce the draw probability in both 3-test and 5-test series may be unattractive to regulators, player conditioning teams and stadium and event management. Therefore alternate systems were devised in tables 1 and 2 by utilizing one-innings test matches. It is worth noting that an alternate system to the current 5-test series by playing all two-innings matches is given by system (10) where 6-day matches are played in a 3-test series for a maximum of 18 days. However, the only way to play a 5-test series (with two-innings matches played over a maximum of 6 days) and keep the maximum number of days to 25 as in the current format, is to play all one-innings matches or a combination of one-innings and two-innings matches. And thus, a total of 7 systems are given in table 2, where a one-innings match could be played over 3 or 4 days.

Thus, we are confronted with a situation on whether a one-innings match played over 3 or 4 days would reduce the draw probability in comparison to the current two-innings match played over 5 days. Using simple logic we can be confident that the draw probability of playing a one-innings match over 4 days will be less than the draw probability of playing a one-innings match over 3 days. Thus, one-innings matches over 3 or 4 days could be trialled in domestic matches to obtain estimates on draw probabilities in comparison to the current two-innings match played over a maximum of 5 days. Note also system (13) is a 7-test series playing all one-innings matches over 3 days for a maximum of 21 days in contrast to system (10) where a 3-test series playing all two-innings matches over 6 days for a maximum of 18 days. It is also worth noting that the proposed system (8) could be adopted as an alternate scoring system in both a 3-test and 5-test series by playing 5 one-innings matches over 3 days for a maximum of 15 days of play. It reasonable to assume that a two-innings match played over a maximum of 6 days will increase the chances of the stronger team winning in comparison to a two-innings match player over a maximum of 5 days, and a one-innings match played over 3 or 4 days. And thus, as is the case with many scoring systems, by devising a system to increase the chances of the stronger team winning will generally increase the length of the match/series

Consider systems (3) and (4) from table 3, where both systems play 3 one-innings matches over a maximum of 3 days and 4 days respectively. Hence system (3) has a maximum of 9 days in the series compared to system (4) a maximum of 12 days in the series. Thus, is it beneficial to schedule an

additional 3 days in the series to increase the probability of a match result, increase the probability of the stronger team winning the overall series and reduce the draw probability in the overall series.

This could be quite attractive to regulators in making decisions of cricket scoring systems. Note that systems (5), (6) and (7) also adopt a 3-test series, but they all utilise a combination of one-innings and two-innings matches within the series. But nevertheless, making a comparison of the key objectives with the current format of system (1) could be attractive to regulators. Also, system (8) adopts a 5-test series with all one-innings matches, and has the advantage that it could be adopted to replace both the current formats of system (1) and system (2). The equivalence of alternate systems to a 5-test series from table 4 is now given. All matches from system (8) and system (11) are one-innings and the number of matches played is 5. Systems (9), (12), (14), (15) and (16) also adopt a 5-test series, but they all utilise a combination of one-innings and two-innings matches within the series. System (10) adopts a 3-test series with all two-innings matches and system (13) adopts a 7-test series with all one-innings matches. Thus, system (10) has the property of keeping with the current two-innings structure for all matches.

6 Conclusions

This study showed our alternate series and match designs for Test cricket reduced draw probabilities. By simulating various combinations of match formats—particularly one-innings matches over 3 or 4 days and two-innings matches over 6 days—we demonstrated that structural adjustments can significantly influence match and series outcomes. Notably, one-innings matches over 4 days consistently reduced draw probabilities while maintaining competitive balance, and six-day two-innings matches further enhanced the chances of a stronger team securing victory.

The findings suggest that cricket regulators could consider trialling these alternate formats in domestic competitions to gather empirical data and assess feasibility. Systems such as (4) and (8) offer promising alternatives that preserve series length while improving result likelihood. Ultimately, adopting flexible match structures may enhance the strategic depth and spectator appeal of Test cricket, while aligning with modern scheduling and performance demands.

References

[1] Cricket.com.au. (2025). *Results | Women's Ashes Hub | cricket.com.au*. [online] Available at: https://www.cricket.com.au/womens-ashes/results [Accessed 1 Jun. 2025].

[2] ESPNcricinfo (2025). *AUS: Brisbane Cricket Ground, Woolloongabba, Brisbane Cricket Ground Test match team match results | ESPNcricinfo*. [online] ESPNcricinfo. Available at: https://www.espncricinfo.com/records/ground/team-match-results/aus-brisbane-cricket-ground-woolloongabba-brisbane-209/test-matches-1 [Accessed 1 Jun. 2025].

[3] Guinness World Records (2013). *First multi-format series in international cricket*. [online] Guinness World Records. Available at: https://www.guinnessworldrecords.com /world-records/112158-first-multi-format-series-in-international-cricket [Accessed 1 Jun. 2025].

[4] Nicholls, S., Pote, L., Thomson, E. and Theis, N. (2023). The Change in Test Cricket Performance Following the Introduction of T20 Cricket. Sports Innovation Journal, 4, pp.1–16. doi:https://doi.org/10.18060/26438.

[5] Pollard GH and Barnett T (2024). An analysis of a test cricket series. Proceedings of the 17th Australian Conference on Mathematics and Computers in Sport

[6] www.lords.org. (n.d.). Innings Law | MCC. [online] Available at: https://www.lords.org/mcc/the-laws-of-cricket/innings.

Revisiting Clutch Performance Among Elite Players in Tennis

Pascal Bauer* and Jan Bauer**

*Saarland University, Chair for Sports Analytics, pascal.bauer@uni-saarland.de
**Independent Researcher, Mannheim, Germany, jan.c.bauer@gmail.com

11th MathSport International Conference, Luxembourg, June 2025

Abstract

The triple-nested point structure (sets, matches, points) in tennis introduces some extraordinary effects: Players can win matches without winning more points than their opponent (*Quasi-Simpson paradoxon*). In addition, the ten best players in tennis history, on average, win 'only' 53.4% of the total points played, although they win 78.2% of their matches. Together, these insights have fueled the widely held belief that *clutch performance* is a major factor for success in professional tennis. We challenge this hypothesis, using purely match-statistic data from 93,884 matches spanning 23 years of professional tennis (1991–2024). Our findings indicate that overall point winning percentages, rather than over-performance on important points explain match outcome rates with an R^2-value of 93.1%. Additionally, we perform two simulations—each assuming a randomized dispersion of points won regardless of their importance: First, we simulate 100,000 tennis matches using artificial serve and return winning percentages. This reveals an s-shaped relationship between career points and career matches won. Second, simulating the careers of 500 players 1,000 times each using their actual match-level serve and return winning percentages yields an R^2 of 94.0% when predicting their real-life career match winning percentages.

1 Introduction

In a recent speech, Roger Federer mentioned that he won 81.7% of his 1,526 matches by winning only 54.1% of his points.[1] Federer phrased it as a message to never dwell with previous failures or successes to allow yourself to fully focus on the next point.[2] Patrick Mouratoglou a famous tennis coach, confirmed these statistics for Roger Federer (54.1%), Novak Djokovic (54.4%) and Rafael Nadal (54.5%),[3] inferring that these players won their matches at a few, very important points. Table 1 shows an overview of the most elite tennis players' career statistics compared to average using the data-set described in Section 2. Meffert et al. concluded that *"Big points exist in professional tennis"* as a result of a survey they conducted among experts ($n = 174$) agreeing distinctly on the subsistence of *big points* (97.3%), although they failed to find a clear definition for such [13].

[1] Full speech at Dartmouth `https://www.youtube.com/watch?v=pqWUuYTcG-o`, accessed 2024/20/12. The underlying data-set used is undefined; however, these statistics are roughly aligned with Table 1.

[2] *"In tennis perfection is impossible [...] In other words, even top-ranked tennis players win barely more than half of the points they played"*

[3] Link: `https://www.youtube.com/shorts/UrIihuSVtQ8`, accessed 2024/20/12

Several studies investigated the importance of potential key-points like break-points [13], match-points [7, 8, 23, 11, 14, 18, 19, 4] or tie-breaks [12, 2, 20]. More broadly, the influence of psychological factors like clutch performance, hot-hand [1, 6], choking under pressure [3] or back-to-the-wall-effect [6] has been researched under the umbrella of the i.i.d.-assumption[4] [7, 17, 16, 15]. The question of whether elite players can perform significantly better at important points has already been raised in 2004 [21]. Following Morris' definition of *important points*[5] [14], Pollard et al. concluded (on a basis of seven matches) that there is some evidence that top players in good form (investigated by the reference of A. Agassi) can perform above their career average at important points [21]. However, they motivate further research on a larger sample size. A more thorough study in 2012 analyzed $1,009$ matches from the US Open (1994–2006) and provided evidence that the top players perform better "*when it matters most*" [5]: When predicting both point outcomes and the ATP world ranking using basic player statistics, they showed that a proxy metric for clutch performance had a significant influence. Similarly, [9] weighted the points of 305 men's and 296 women's Grand Slam matches from 2011 by their influence on the match winning percentage according to [14][6], and showed that the weighted point winning percentage (*pw*) predicts match outcomes better than naive point aggregations. From then on, a series of researchers described this ability of the best players as self-evident [13, 23, 11], although none of the described approaches checked whether a raised clutch performance, i.e., a significantly improved *pw* at important points, discriminates between elite and average tour players on an appropriate sample size.

Thus, it is our objective to investigate an alternative hypothesis: a *pw* of 53.4%—with a random distribution—could naturally relate to a match winning percentage (*mwp*) of 78.2% due to the rules of the game itself. In the above example of Agassi [21], this would mean that an advantageous dispersion of his won points cause his 'good form' and not vice versa. Thus, we follow up on the research question raised in [21], i.e. whether an outstanding clutch performance discriminates among top players, by using only match-level (i.e. aggregated return and serve winning percentages) data on a significantly increased sample size ($n = 93,884$ matches, on average 330 matches per player).

2 Data

We use publicly available data compiled by Jeff Sackmann.[7] This dataset comprises match-level statistics for men's tour-level singles matches, focusing on major events, including Grand Slams, Masters, and ATP $250/500$ tournaments. Each match entry contains detailed information about tournament attributes as well as player- and match-specific details (including aces, double-faults, serve-/return-points, and break-point outcomes). Initially, the raw match data included $193,337$ matches spanning from December 1967, to May 2024. To focus on matches with comprehensive data, a first filter was applied to include only those matches with single-point statistics available, reducing the dataset to $96,442$ matches. Further filtering removed

[4] The assumption that points in tennis are independent and identically distributed.

[5] Assuming a fix win point winning percentage (*pw*) for servers (e.g. $p = 0.6$), they defined the importance of a point as the probability of winning the game under the condition of winning the respective point minus the probability of winning the match when losing it [14].

[6] They extended the definition of [14] from games to the whole match.

[7] `https://github.com/JeffSackmann/tennis_atp`, available under a Creative Commons BY-NC-SA 4.0 license `https://creativecommons.org/licenses/by-nc-sa/4.0/`

Player	**Matches** (*mwp*)	**Points All** (*pw*)	**Service** (*spw*)	**Return** (*rpw*)
Novak Djokovic	*84.2 (80.1)*	*54.5*	67.6	42.1
Rafael Nadal	83.0 (79.0)	*54.5*	67.4	*42.3*
Roger Federer	81.7 (79.1)	54.1	*69.5*	39.7
Pete Sampras	80.3 (74.0)	53.5	*69.5*	38.0
Carlos Alcaraz	79.1 (74.9)	53.2	65.6	41.7
Andre Agassi	76.9 (73.9)	53.3	65.9	41.6
Andy Roddick	75.3 (71.7)	53.0	71.1	35.9
Jannik Sinner	75.2 (70.1)	52.9	66.1	40.2
Stefan Edberg	73.3 (70.8)	52.8	64.9	41.2
Boris Becker	73.2 (68.8)	52.3	67.0	38.1
Average Top 10	**78.2**	**53.4**	**67.4**	**40.1**

Table 1: Top ten players according their match win percentages along with point-level career aggregates. The number in parentheses in the second column shows the outcome of Simulation (B) in Section 4. All values in %.

matches where players retired or won by walkover (2.65%), yielding a final set of 93,884 complete matches from January 1991 to May 2024.

Using the filtered match-level data, we created a player-level dataset to extract relevant statistics for each individual player. To minimize statistical noise, we included only players with more than 100 recorded matches, reducing the dataset from 2,500 players to 500. The average number of matches (points) per player contained in this final dataset is 330 (53,000).

3 Career Level Regression Analysis

To predict the career *mwp* of a player, we considered several metrics as explanatory variables: *pw*, service point winning percentage (*spw*), return point winning percentage (*rpw*), and break-point ratio (*bpw*). Different combinations of these variables were used in separate regressions to assess their individual and joint contributions to the match winning percentage. For each combination, coefficients were estimated using ordinary least squares regression, minimizing the sum of squared residuals [10] on the full dataset. The statistical significance of each explanatory variable was evaluated to determine its impact on *mwp*.

The results are summarized in Table 2. All regression coefficients are statistically significant ($p < 0.1$). The first regression model (R1; also visualized in Figure 1a) uses only the overall point winning percentage as the explanatory variable. The high R^2 of 94.1% indicates that the point winning ability—without any information on their dispersion—is a strong predictor for *mwp*. If important points were to play a significant role, one would expect a much greater fluctuation in match outcomes for the same *pw*, which would result in a lower R^2. Interestingly, there is a large translation from an increase in *pw* to an increase in matches won: for each 1.0%-point increase in *pw*, *mwp* increases by 8.0%-points ($\beta_{pw} = 8.0$; R1)—at least within the observed sample range between 47% and 54% points won.

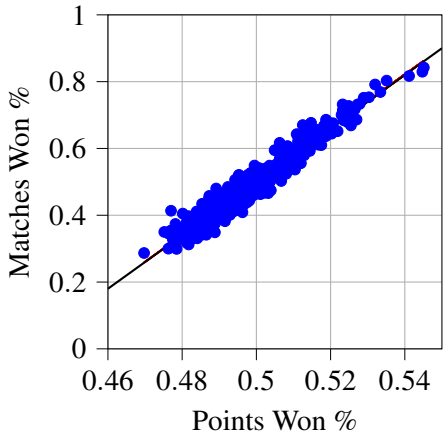
The second regression model (R2) incorporates *bpw*, defined as the ratio of the percentage of break-points successfully converted to the percentage of break-points faced and lost as an additional explanatory

variable. A higher bpw might reflect better performance in high-pressure situations during a match [14, 9]. Despite the perceived importance of break-points in tennis, including bpw in the model only marginally enhances the explanatory power ($\beta_{bpw} = 0.2$; R2 & R4).[8]

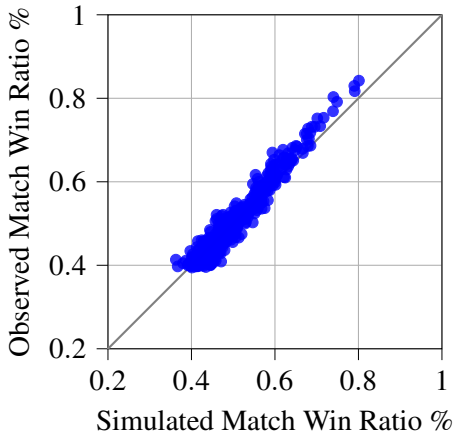
In the regressions that separately account for spw and rpw (R3, R4), the results remain consistent and do not alter the overall conclusions. Notably, the coefficients for service points ($\beta_{spw} = 0.37$) and return points ($\beta_{rpw} = 0.36$) are of similar magnitude, suggesting that improvements in both areas have a comparable impact on match performance.

Point winning percentages	**R1**	**R2**	**R3**	**R4**
Intercept	-3.5***	-3.2***	-3.1***	-2.8***
All points (pw)	8.0***	7.2***		
Service points (spw)			3.7***	3.1***
Return points (rpw)			3.6***	3.0***
break-points (bpw)		0.2***		0.2***
R^2	94.1	94.6	93.1	94.3

Table 2: Regression results for explaining a player's match winning percentage. Significance at the 0.1% level is denoted by ***. All numbers in %.



(a) Match winning percentage predicted by the overall point winning percentage. Each point represents a single player ($R^2 = .94$). Regression details in Table 2.



(b) Observed versus simulated match win ratios per player (R^2 =.94, RMSE = 2.4%) as the outcome of simulatinon (B) (Section **??**.

Figure 1: Relationship between point performance metrics and match outcomes.

[8]Note that we are introducing a bias by including the break-point ratio, since our regression does not separate between training and test data. Since break-points are rather at the end of games/matches/sets, they might cause an overfitted model. The results are in line with [5].

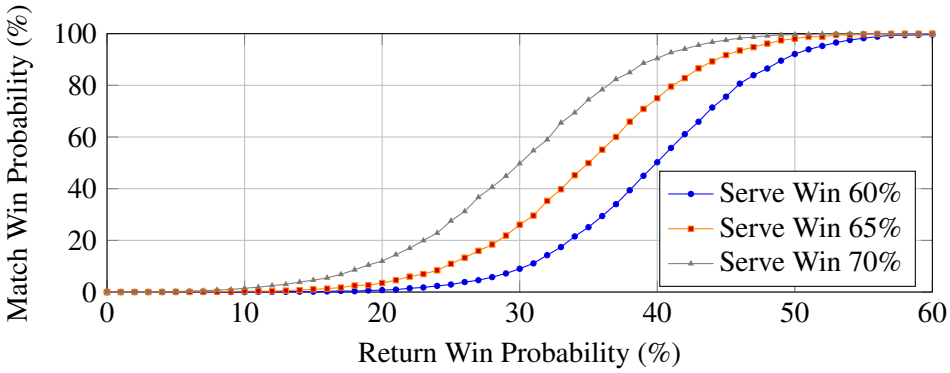


Figure 2: Simulation (A): For three different *spw*—60% (blue), 65% (orange), 70% (gray)—this Figure shows the relation between *rpw* and match-outcome. Each point in the diagram is based on 100,000 simulated matches.

4 Simulation Analysis

We developed a match simulation engine to analyze the relationship between *pw* and *mwp*. The model takes four inputs: *spw*, *rpw*, the total number of simulated matches, and the proportion of matches played as best-of-five-sets.

The model handles point winning probabilities as i.i.d. within a given state (serving or returning) and follows a Bernoulli distribution. Tie-breaks are assumed at 6:6 in every set.[9] Parametrization was based on the match-level dataset (Section 2). Using this engine, we run two different simulations: (A) First, a total of 100,000 matches were simulated using artificial combinations of *rpw* and *spw*.[10] (B) In a second step, we use actual match-level serve and return statistics to simulate each of the 500 player's careers 1,000 times respectively.

Figure 2 shows the relationship between *spw*, *rpw* and *pw* using purely artificial data. The i.i.d. distribution in our simulation suggests a situation where players constantly compete at their performance level without being influenced by psychological factors. Figure 2 also indicates that the alleged linear relationship between *pw* and *mwp* in Figure 1a only shows an excerpt of a rather s-shaped relationship: Given a static *spw* of 70%, increasing *rpw* from 25% to 35% increases the *mwp* by almost 50%, while the same increase from 0% to 10% won points on return improves the match winning percentage by only a few percent. Consequently, simulating top-players career *mwp* (as in Figure 2) using their aggregated *spw* and *rpw*, fails to handle the non-linear part of the relation.

To consider this, visualized in Figure 1b, we simulated all matches of each player's career with the respective match-level *spw* and *rpw* (1,000 career-simulations per player). Column two in Table 1 shows the predicted match winning percentage (in brackets) of our simulation. In general our naive simulations match the actual career statistics, however, the players listed in Table 1 consistently outperform their simulated results.

[9] Simplifying actual rules where, for example, Grand Slams may omit tie-breaks in the fifth set

[10] The share of best-of-five matches was set at 75%, and the ranges for *spw* and *rpw* were selected in line with observed data.

5 Discussion

Previous work on clutch performance [21, 9, 5] applied Morris' definition of important points [14] on point-by-point data. [9] showed that in-sample importance-weighted point winning percentages predict match outcomes better than naive point aggregations. However, similar to [21], the cause of this correlation could be overfitting. Compared to this [5] enriched their *spw* and *rpw* with up-to-date information at every point.[11] Using this dynamic definition of point importance, they detect a significant influence ($p < 0.05$ in all experiments) of a player's "*critical ability*" when predicting both the outcome of future points and when predicting out-of-sample ATP-rankings. In a second experiment, they showed that a players critical ability explains 13%[12] of his career average ATP-ranking. Overall, [5] found evidence for clutch performance being a separator between players, however, this finding is not consistent through all their experiments.

We revisit clutch performance from a less granular perspective using just aggregated career and match-level data on a significantly increased number of $93,884$ matches ([21]: $n = 7$, [9]: $n = 1,009$, [5]: $n = 305$). Our regression analysis in Section 3 (i.e. Figure 1a) shows that even a very naive model explains $\sim 94\%$ of player's match winning percentages. The remaining $\sim 6\%$, consequently, aggregates all other potential influence factors like clutch performance. This conclusion, narrowing the relevance of clutch performance down, is supported by our simulation study (B): match-level *spw* and *rpw* of players alone explain 94% of their match winning percentages.

Simulations, as applied in simulations (A) and (B), have been researched in the literature [17, 16, 15]. [17] used point conversion rates while serving in order to predict game, set, match, and tournament outcomes. Following up on this work, in [16], they investigated the robustness of Monte Carlo simulations for tennis matches (using *spw*) against disturbing effects like *the hot-hand-effect* the *back-to-the-wall-effect* and against deviating performances at *important points*.

Our study poses several limitations that should be overcome in future work: First, our regression in Section 3 should consider the non-linear relationship shown in Figure 2. Second, prediction accuracies should be considered on an isolated test data set. Future work on clutch performance should include player's career-aggregated performance per point importance (similar to [9]) in an out-of-sample prediction for players future success. Additionally, other sports showed that machine-learning based in-game-win-probability models can implicitly capture psychological effects [22]. Consequently, our exhaustive data-set should be used to compare Morris' rule-based in-game-win-probability [14] models against a purely data-driven model. Furthermore, we recommend rigor and granular statistical i.i.d. tests on a large point-by-point data set.

Overall, we contribute to existing literature by revisiting clutch performance, one relevant psychological component in tennis among others, using a large set of real-world data. Minor tendencies for psychological factors influencing the point distribution cannot be denied, however, our results help to put previous beliefs on a substantial influence of clutch performance into perspective and motivate future research.

[11] At the beginning of each match, they are estimated using past match winning percentages of both players. As the match is ongoing, the average values between these pre-match information and the past point winning percentages within that match are used.

[12] A correlation of 0.37 was reported

References

[1] Michael Bar-Eli, Simcha Avugos, and Markus Raab. "Twenty years of "hot hand" research: Review and critique". In: *Psychology of Sport and Exercise* 7.6 (2006), pp. 525–553.

[2] Danny Cohen-Zada, Alex Krumer, and Offer Moshe Shapir. "Testing the effect of serve order in tennis tiebreak". In: *Journal of Economic Behavior & Organization* 146 (2018), pp. 106–115. ISSN: 0167-2681. DOI: 10.1016/j.jebo.2017.12.012.

[3] Danny Cohen-Zada et al. "Choking under pressure and gender: Evidence from professional tennis". In: *Journal of Economic Psychology* 61 (2017), pp. 176–190. ISSN: 0167-4870. DOI: https://doi.org/10.1016/j.joep.2017.04.005. URL: https://www.sciencedirect.com/science/article/pii/S016748701630589X.

[4] Avinash Dixit and Susan Skeath. "The Most Important Situations in Tennis – and in R&D Competition". en. In: *Games of Strategy*. 1st ed. 1999.

[5] Julio González-Díaz, Olivier Gossner, and Brian W. Rogers. "Performing best when it matters most: Evidence from professional tennis". In: *Journal of Economic Behavior & Organization* 84.3 (2012), pp. 767–781. DOI: 10.1016/j.jebo.2012.09.021.

[6] David Jackson and Krzysztof Mosurski. "Heavy defeats in tennis: Psychological momentum or random effect?" In: *Chance* 10.2 (1997), pp. 27–34. DOI: 10.1080/09332480.1997.10542019.

[7] Franc J. G. M. Klaassen and Jan R. Magnus. "Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model". In: *Journal of the American Statistical Association* 96.454 (2001), pp. 500–509. DOI: 10.1198/016214501753168217.

[8] Franc J. G. M. Klaassen and Jan R. Magnus. "Testing some common tennis hypotheses: Four years at Wimbledon". In: (1996).

[9] Stephanie A. Kovalchik and Machar Reid. "Measuring clutch performance in professional tennis". In: *Statistica Applicata - Italian Journal of Applied Statistics* 2 (2018), pp. 255–268. DOI: 10.26398/IJAS.0030-011.

[10] Robert Ling. "Residuals and Influence in Regression (Review)." In: *Technometrics*. 1983.

[11] Dominik Meffert. "Big Points im Tennis? Zur spielsituativen Handlungsvermittlung für die Tennisausbildung: Erkenntnisse aus der Weltklasse". PhD Thesis. Cologne: German Sport University Cologne, 2021. URL: https://fis.dshs-koeln.de/en/publications/big-points-im-tennis-zur-spielsituativen-handlungsvermittlung-f%C3%BCr.

[12] Dominik Meffert et al. "Tennis at tiebreaks: Addressing elite players' performance for tomorrows' coaching". In: *German Journal of Exercise and Sport Research* 49.3 (2019), pp. 339–344. DOI: 10.1007/s12662-019-00611-3. (Visited on 05/30/2025).

[13] Dominik Meffert et al. "Tennis serve performances at break points: Approaching practice patterns for coaching". In: *European Journal of Sport Science* 18.8 (2018), pp. 1151–1157. DOI: 10.1080/17461391.2018.1490821. (Visited on 05/30/2025).

[14] Carl Morris. "The most important points in tennis". In: *Optimal Strategies in Sports*. 5th ed. North-Holland, 1977, pp. 131–140. ISBN: 0-7204-0528-9.

[15] Paul K. Newton and Kamran Aslam. "Monte Carlo tennis: A stochastic Markov chain model". In: *Journal of Quantitative Analysis in Sports* 5.3 (2009). DOI: 10.2202/1559-0410.1169. (Visited on 05/30/2025).

[16] Paul K. Newton and Kamran Aslam. "Monte Carlo tennis". In: *SIAM Review* 48.4 (2006), pp. 722–742. DOI: 10.1137/050640278. (Visited on 05/30/2025).

[17] Paul K. Newton and Joseph B. Keller. "Probability of winning at tennis I. Theory and data". In: *Studies in Applied Mathematics* 114.3 (2005), pp. 241–269. DOI: 10.1111/j.0022-2526.2005.01547.x.

[18] Peter G O'donoghue. "The most important points in grand slam singles tennis". In: *Research quarterly for exercise and sport* 72.2 (2001), pp. 125–131.

[19] Peter O'Donoghue. "Break points in Grand Slam men's singles tennis". In: *International Journal of Performance Analysis in Sport* 12.1 (2012), pp. 156–165. DOI: 10.1080/24748668.2012.11868591.

[20] G. H. Pollard. "An analysis of classical and tie-breaker tennis". en. In: *Australian Journal of Statistics* 25.3 (1983), pp. 496–505. DOI: 10.1111/j.1467-842X.1983.tb01222.x.

[21] Graham Pollard. "Can a tennis player increase the probability of winning a point when it is more important?" In: *Proceedings of the Seventh Australasian Conference on Mathematics and Computers in Sport* (2004). Ed. by R Hugh Morton and S Ganesalingam. Place: New Zealand Publisher: Massey University, pp. 253–256.

[22] Pieter Robberechts, Jan Van Haaren, and Jesse Davis. "A Bayesian Approach to In-Game Win Probability in Soccer". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. KDD '21. Virtual Event, Singapore: Association for Computing Machinery, 2021, 3512–3521. ISBN: 9781450383325. DOI: 10.1145/3447548.3467194. URL: https://doi.org/10.1145/3447548.3467194.

[23] Cédric Roure. "What are the key points to win in tennis ?" In: *ITF Coaching and Sport Science Review* 64 (2014), pp. 14–15. URL: https://www.researchgate.net/publication/267393027_What_are_the_key_points_to_win_in_tennis.

Models for Prediction and Analysis in Horse Racing

A. Bedford*, E. Mealy**, A. Koay*** and A. Velcich****

*University of the Sunshine Coast, abedford@usc.edu.au
** emealy@usc.edu.au ***akoay@usc.edu.au ****avelcich@usc.edu.au

Abstract

In our previous work we presented our computer vision (CV) platform that swiftly extracted horses from vision as analysable objects using segmentation modelling from semi-live footage. Building upon uses of CV and artificial intelligence (AI) through pre-race training, in-race tracking, and post-race adjudication and performance analysis, we propose two methods to obtain horse velocities which provides useful estimates for multiple needs: assessing if there are issues in a horses gait, cadence and stride in training and racing environments; provide real-time analysis for in-play betting; and provide pre-race analysis of runners for race prediction. The first method uses gate-based technology with global positional system (GPS) technology, the second a video-based transformation method using CV and AI.

We provide the framework for the system and demonstrate its utility in a few environments – training, race and post-race. We cover challenges and outcomes from the process and compare the velocities recording using speed gates to the CV models garnered from vision. We also outline the process of extracting baseline velocities and how this approach can also be utilised for in-play estimations of performance for setting prices and estimating outcomes.

1 Introduction

In this work, we aim to determine the speed of a horse and thereby impose modelling attributes for multi-uses, including betting, protests, prediction and horse welfare. Most horse racing is typically filmed with moving cameras and limited camera angles, which makes conventional speed estimation difficult, leading to inconsistent decisions, poor data and trying viewing. Existing global positioning systems (GPS) and gate systems are at times unreliable, and the detail required in a live environment with gaps in times is considered too slow. In the event of a protest, relying heavily on human decisions from video can potentially result in errors, discrepancies, variances, and unavoidable inconclusiveness.

Therefore, to improve the decision making, increase the transparency, accuracy, and reliability, we propose two solutions: a computer-vision (CV) prototype; and a modelling system utilising existing speed gate data. Data utilised for both methods is via publicly available data sources. Our CV prototype calculates real-time data of each horse, including speed, acceleration, and position, with a locally estimated scatterplot smoothing (LOESS) model to estimate prerace speeds from long form speed gate data.

2 Methodology

As we are undertaking two approaches, we shall outline them sequentially. Firstly, we shall cover the methods used for speed estimation utilising gate speed (sectional) times. The data we have covers most racecourses in Queensland and provides key information such as the time taken to run between track

sections – in Australian racing, this is typically 200 metres (approximately one furlong). While many models exist for horse racing [1, 2], we primarily focused on speed estimation due to its potential utility in supporting the CV model's ability to predict speed. The model overview was to utilise R to scrape data, optimise times, create a 'long form', simulate times based on regression, and estimate pricing, speeds and winning. A structure of this process is provided in the presentation. The model overview is as follows.

2.1 Statistical Sectional Estimation

Let $\hat{t}_{hi}$denote the sectional time (per furlong) for horse *h* in race *r* at section $i \in \{1,\ldots, S\}$, where *S* is the total number of sections per race. To estimate the expected sectional time, we fit a generalized additive model (LOESS regression with covariates) as seen in (1).

$$\hat{t}_{hi} = f_h(i) + \beta_1 \cdot \text{Distance}_r + \beta_2 \cdot \text{Weight}_{hr} + \beta_3 \cdot \text{Barrier}_{hr} + \cdots + \varepsilon_{hi} \quad (1)$$

where $f_h(i)$ is the LOESS smoothed by section *i*, β_i is set by overarching adjustments for track by each variable (by race *r*), and ε the standard residual term. Covariates are drawn from race-day variables allowing the model to account for varying race conditions. LOESS was used due to the non-parametric nature of data, and 'smooths' over poor race times and outliers, and has a nice recency. For each horse, we simulate *N* race outcomes using the model:

$$\tilde{t}_{h_1}^{(j)} = \hat{t}_{h_1} + \epsilon_{h_1}^{(j)}, \epsilon_{h_i}^{(j)} \sim \mathcal{N}(0, \sigma_h^2) \quad (2)$$

The total simulated time for horse *h* in simulation *j* is

$$T_h^{(j)} = \sum_{i=1}^{S} \tilde{t}_{h_1}^{(j)} \quad (3)$$

where we run a Montecarlo with a stopping rule using variance stability (<0.1 change) for all competitors. It is then simple to produce outputs of utility such as winning likelihoods, pricing models, and for our purposes, sectional speed estimates for horse *h*.

Winning probabilities are obtained simply ($\mathbb{I}$ as indicator function) for the quickest time:

$$P_h = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}\left\{T_h^{(j)} = \min_{h'} T_{h'}^{(j)}\right\} \quad (4)$$

and sectional leaders, *L*

$$L_{hi} = \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}\left\{\sum_{k=1}^{i} \tilde{t}_{hk}^{(j)} = \min_{h'} \sum_{k=1}^{i} \tilde{t}_{h'k}^{(j)}\right\} \quad (5)$$

The use of these statistical models becomes apparent once we integrate with the CV models. The process for the CV models is as follows.

2.2 Computer Vision Models to obtain Centroids and Posts

This research extends existing ComfyUI workflows [3] by integrating enhanced model architectures, notably YOLOv11, to address motion-induced tracking issues in equine performance analysis. It introduces a multi-model detection and segmentation pipeline—leveraging YOLOv8, YOLOv11, SAM2 and GroundingDino—capable of identifying, tracking, and assigning unique identifiers to

individual horses across video frames. By applying Kalman filters and Holography the system enables accurate trajectory mapping and performance metric extraction. The prototype is developed to analyse live footage and derive actionable insights, with applications in both real-time and post-race contexts.

2.3 Homography-Corrected Centroid Tracking with Sectional Time Smoothing

Using the centroids, post and rail points, we use Homography to estimate the horse's velocity, then merge this with (3) to provide reasonable estimates to 'fill in' any holes due to occlusions, jitter or camera changes from broadcast. This process was adapted from Zhang et al. [4] and modified in Figure 1, where we provided an overview of the process.

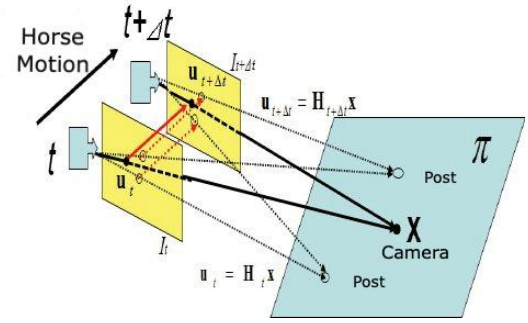


Figure 1 Homography modified from [4]

Let $\boldsymbol{c}_t = \mathbf{u}_t$ (as in Fig.1) be the image obtained centroid position, $c_{h,t} \in P^2$, for each horse in projective 2D space, and we repeat this at the shift of frame $c_{h,t+\Delta t}$. To correct for movements, we estimate H for the posts/rail junctions $c^*_{h,t+\Delta t} = H^{-1} \cdot c_{h,t+\Delta t}$. This projects the new centroid into the previous frame

$$\boldsymbol{c}^*_{\boldsymbol{h,t+\Delta t}} = \frac{1}{\omega}\begin{bmatrix} x \\ y \\ \omega \end{bmatrix} = \begin{bmatrix} x/\omega \\ y/\omega \\ 1 \end{bmatrix} \tag{6}$$

The centroid is smoothed, $\hat{c}_{h,t+\Delta t} = \alpha \cdot c^*_{h,t+\Delta t} + (1-\alpha) \cdot \hat{c}_{h,t}$, with α set to 0.25 in initial tests, and the first frame set at $\hat{c}_{h,t} = c_{h,t}$. We estimate velocity, $v_{h,t} = (\hat{\boldsymbol{c}}_{h,t+\Delta t} - \hat{\boldsymbol{c}}_{h,t})/\Delta t$ and thus speed $|v_{h,t}| = \left(\sqrt{(x_{t+\Delta t} - x_t)^2 + (y_{t+\Delta t} - y_t)^2}\right)/\Delta t$. Utilising the estimated times, let $T_{h,s} =$ Estimated time to complete section s, we get $\bar{v}_{h,s} = \frac{D_s}{T_{h,s}}$. Actual centroid movement is given by $\Delta d_{\text{obs}} = |\hat{c}_{h,t+\Delta t} - \hat{c}_{h,t}|$ and expected $\Delta d_{\text{exp}} = \bar{v}_{h,s} \cdot \Delta t$, so if $| \Delta d_{obs} - \Delta d_{exp} | > \delta$ then we flag for a second image sweep/post-race correction. Finally, a fusion between the model and observation is needed, such that $\hat{c}_{h,t} = \lambda \cdot p_{obs} + (1-\lambda) \cdot \boldsymbol{p}_{model}$, whereby λ is optimised.

3 Results

In reverse order of the methods, we will firstly discuss the CV model: YOLO11 object tracking provides accurate detections with the added functionality of remembering tracked objects rather than just detecting if an object exists in the frame. This aligns with our work using Homography, however horses are often 'lost' and reallocated a new id (due to occlusion, camera switching, etc.) An example of tracking is shown in Figure 2. YOLO11 object tracking models have been optimised for real-time use and can be employed to quickly and efficiently process data.

Figure 2: YOLOv11 (a) with centroids (b) Instance Segmentation & (c) speed estimation

The instance segmentation model (Fig 2.b) takes it a step further and applies a mask over individual instances of horses. This provides specific information about each tracked horse and will attempt to remember it over frames. Just like object tracking, this method can run into issues related to occlusion and requires exported data to be further analysed by stewards to make fully educated decisions. It should be noted that while instance segmentation is very effective at visualising the exact pixels detected as a horse, it is much more computationally intensive when processing videos. This can be mitigated by using less accurate versions of the segmentation models, or by using more powerful hardware. The speed estimation model provides a strict setup and provides a very rough speed reading on horses, as seen in Figure 2(c). At this point, it is preliminary and considered too unstable for use.

The SAM2 + GroundingDINO segmentation model (Figure 3(a)) is aimed at detecting any specified object within a frame and applying a mask over the detected objects [5]. The flexibility and ease of use are the standout features of this method, however due to its generalised training approach, this severely limits the model in specific use-cases, especially where accuracy is a desired factor. The benefits of a custom trained dataset would likely be the same as when applied to YOLO models.

Figure 3: (a) SAM2 + Grounding DINO; (b) Marigold depth estimation

The Marigold depth estimator was also trialled (Figure 3(b)) and while this model does have resolution limitations it did provide an extra layer to counter perspective distortion through the implementation of a bias system alongside the chosen computer vision model. It was hoped to assist in troubles caused from occlusion.

The statistical modelling forms part of the *a priori* estimations of horse velocities. The process runs semi-autonomously, with sectional times scraped and compiled over two years, meetings scraped, data joined on pre-raceday, velocities are estimated, probabilities established, and the resulting sectionals are then utilised in the homography models. Example outputs are shown in Figure 4(a) and (b).

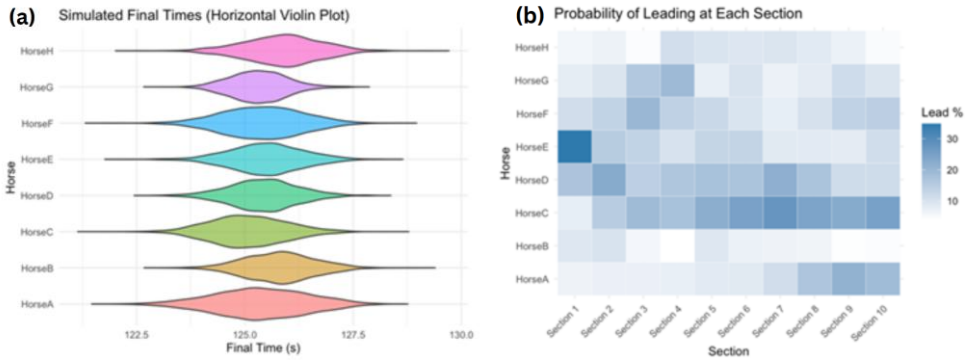


Figure 4. (a) Violin plot of simulated final times from LOESS estimations (b) Heatmap showing the probability of a leading horse at each section of the track

4 Discussion

Our aim was both to adapt CV models for horse detection and develop a scraper of historical speeds to package a system of accurate live speed estimation without the use of any wearables or and fixed gates – solely vision. The process of generation is detailed in Figure 5 and outlines several moving parts – the long-form builder and sectional database process (completed as soon as prior to the Raceday); the simulated race (pre Raceday); the CV pipeline of work; the centroid estimation via both CV and Homography, the fusion model, and the final product.

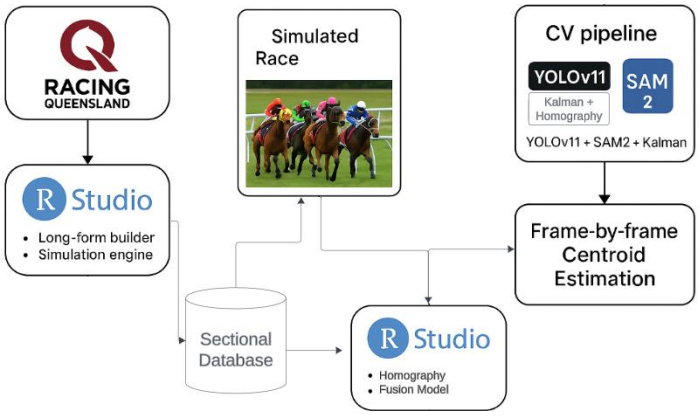


Figure 5. Schema of the entire process.

This process has become remarkably fast, with near real time ability of the CV models, and the meshing of modelling with the centroids the key to cleaning up the messiness of vision-based modelling. Much work is needed in refining the model, and whilst we have a lovely visual product, including removal of objects (i.e. other horses) for the purposes of protests, there remains some significant post modelling to do.

5 Conclusion

We have demonstrated that building a model for prediction and analysis of horse racing is now feasible in near-real-time through the synergy of statistical estimation, CV models, and homography. The next step is to use the centroids generated by the model to identify moments when a horse loses ground due to injury or interference, enabling improved technological approaches to ancient problems of protests in horse racing.

Current subjective methods for adjudication are enhanced with the inclusion of calculated velocities, and visual representations without occlusion. The steward's processes in determining contact between horses and impeded runs is clearly improved with value-based evidence. Face validity, particularly through side-by-side vision, will be enhanced with future developments of more accurate frame-by-frame velocities and expected velocities.

Additionally, we aim to model the perceived speed *without* contact versus contact as it happened through the race. We aim to back-fit the vision for those purposes and further, to utilise live features; watch the run lines; generate speed worms; and virtual race replication – all intended to be in a live format.

Acknowledgement

We wish to acknowledge the use of data from Racing Queensland and the contributions of Casey Cleland, Aniket Chopra, Matt Greenbury for their assistance in the development of the CV models.

References

[1] P. Colle, "What AI can do for horse-racing?," *arXiv preprint* arXiv:2207.04981, 2022. [Online]. Available: https://arxiv.org/abs/2207.04981

[2] W. W. Y. Ng, X. Liu, X. Yan, X. Tian, C. Zhong, and S. Kwong, "Multi-object tracking for horse racing," *Information Sciences*, Elsevier, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0020025523005364

[3] A. Bedford, E. Mealy, and A. Koay, "Modern Solutions to Ancient Problems: Artificial Intelligence and Computer Vision Technology in Horse Racing", MathSport Asia, 2024.

[4] L. Zhang, X. Yu, A. Daud, A. Mussah, and Y. Adu-Gyamfi, "Application of 2D homography for high resolution traffic data collection using CCTV cameras," *arXiv preprint* arXiv:2401.07220, 2022.

[5] N. Ravi, V. Gabeur, Y. T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, and E. Mintun, "SAM 2: Segment anything in images and videos," *arXiv preprint* arXiv:2408.00714, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2408.00714

Mathematical models for speed climbing applied to data collected on competitors in recent World Cup events

L. Benga[1] B. Hatch[2], and D. Sylvan[2]

Hunter College High School[1]
Hunter College of the City College of New York[2]
lucabenga@hunterschools.org, benjamin.hatch00@myhunter.cuny.edu, dsylvan@hunter.cuny.edu

Abstract

Speed climbing is one of the newest Olympic sports, debuting at the 2020 Tokyo Olympics. With many races decided by hundredths of a second, speed climbing quickly gained recognition as the fastest sport at the Paris 2024 Olympics. Speed climbing appeals to data scientists since it uses a standardized 15-meter wall, making it easy to compare times and strategies across a vast array of competitions and competitors. Surprisingly, however, there has been little rigorous analysis of a professional level race to the best of our knowledge. In this paper, we model data compiled from the 2023 World Cup events in Wujiang, China and Salt Lake City, USA, analyzing both numerical and categorical variables. Examples of quantitative variables include the reaction time displayed in the video for each athlete, along with the total time, or split times, obtained by running the recording for each athlete frame by frame and estimating the exact point at which each section is reached. An example of a binary variable is the skips strategy, which draws attention to the holds each athlete omits on their run. Another example of a categorical variable is the round designation - either round 1 or round 2 - which refers to the order of athletes' runs. We explored these variables extensively, built several general linear models for athlete performance and used model selection to determine the best predictive models. We found that reaction times are normally distributed and appear to be very weakly correlated from one race to another. Counter-intuitively, however, they appear to have minimal bearing on the race result, despite making up a portion of the overall time. Another interesting observation is that many athletes attempt a more aggressive skip strategy in their second run, omitting a greater number of holds. This is either because they either already recorded a viable time for qualification in Round 1 and can afford the risk, or because they felt the need for substantial improvement. In ongoing work, we have been focusing on expanding the analysis, using data from additional World Cup events for both men and women.

1 Introduction

Speed climbing uses a standardized wall that is 15 meters high and has a 5° overhang [1]. The wall includes 20 big holds, each with the same dimensions, along with a number of smaller foot chips. Figure 1 displays a schema of the wall. There are two routes (A on the left side, B on the right side), with no difference between

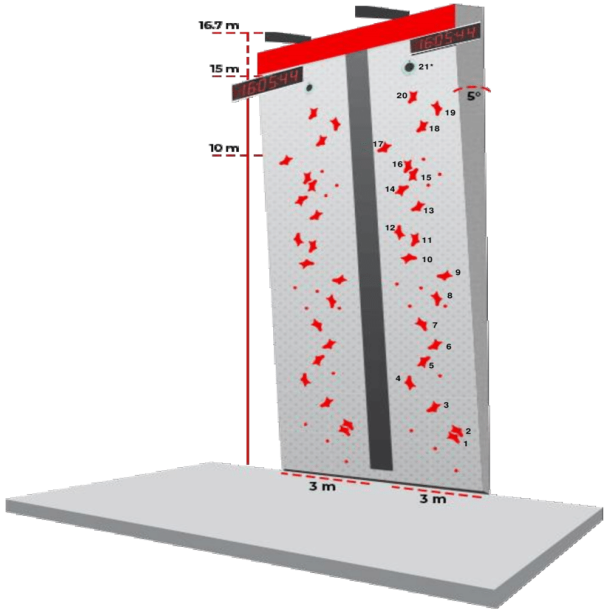


Figure 1: A diagram of a regulation speed climbing wall.

them. In qualifiers, each competitor has one attempt of each, with the best time of the two being used for their tournament ranking. The top finishers enter a finals round, which is a playoff bracket format. Over time, the main way of improving performance has been skipping more and more holds by performing a number of fast, dynamic movements. For example, most World Cup male competitors will jump directly from hold 3 to 5, bypassing 4 (often called the Tomoa skip, in reference to Japanese climber Tomoa Narasaki). They also usually connect hold 8 with 10, bypassing 9. Skipping holds carries the risk of falling, so not all competitors use the same technique; some may find it faster to still use some holds. One difference in the current World Cup circuit, for example, is skipping hold 14. Most competitors still use it (they often will use holds 11-14-16 in sequence), but a few, including the world record holder Sam Watson, go directly from 11 to 16. One very unique technique is done by Noah Bratschi. He does not do the Tomoa Skip (making him the only one to use Hold 4), but opts to use hold 12 and skips 14. There are three common strategies that we focus on: skipping both hold 12 and 14 (i.e., world record holder Sam Watson) which is the fastest approach, but also the highest risk. When runners need to make up time, they often use this strategy; skipping hold 12 but using hold 14 (i.e., Jinbao Long) which is the most typical approach; skipping hold 14 but using hold 12 (i.e., Liang Zhang). As in sprinting, a start faster than 0.1 seconds after the final beep is considered a false start and leads to elimination from the whole competition. Typical reaction times vary between 0.15 and 0.20, as the data shows. One topic of interest in this paper is whether reaction time is statistically significant to one's time or not.

Variable	Description	Type
X_1	Time to hold 20 in Route A	Numeric
X_2	Time to hold 16 in Route A	Numeric
X_3	Whether Route A is the first attempted in the competition	Binary
X_4	Time from hold 0 to hold 10 in Route A	Numeric
X_5	Whether hold 14 is used in Route A	Binary
X_6	Time to hold 10 in Route A	Numeric
X_7	Time from hold 0 to hold 10 in Route B	Numeric
X_8	Time to hold 16 in Route B	Numeric
X_9	Whether hold 12 is used in Route B	Binary

Table 1: Speed climbing variables

2 Data

A detailed dataset was compiled for the IFSC World Cup Wujiang 2024 qualification round, with two runs recorded for each athlete. The starting list, athletes' names, height (where available), bib number, and total time for both runs were obtained from the IFSC Results website, `ifsc.results.info/event/1354/`. All other variables were collected using a detailed video analysis of the competition's recording on IFSC's YouTube channel. They are displayed in Table 1.

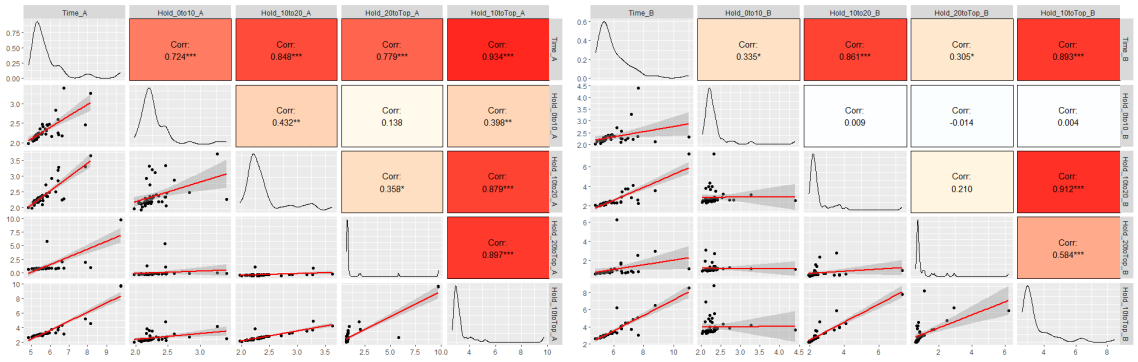


Figure 2: Pair plots of several split variables for Route A (left) and Route B (right).

Reaction time variables were displayed in the video for each athlete along with the total time. Skips strategy (known as the 'hold' variables) were observed and recorded from the video for each athlete and each run. Split times, referred to as the 'Time to' variables were obtained by running the recording for each athlete frame by frame and estimating the exact point when each section is reached. The margin of error is around 0.03-0.04 seconds based on the time gap between each frame. The corresponding distributions of times

are generally right-skewed with most times falling between 5 and 6 seconds. There is a rare chance that a runner falls from the course, costing them several seconds. One area of intrigue was whether falling had any correlation with their present performance in the race, that is if they were often doing poorly and rushing to get back to qualifying pace. Figure 2 displays a bird's eye view of the relationships between variables via pair plots between the various numeric features and the final times for each athlete on Route A and B. It is immediately evident that the time between holds 0 and 10 (**r** = .724 for Route A, **r** = .335 for Route B) is not nearly as relevant as that between holds 10 and 20. One of our first instincts was to check the relationship between the final time and reaction time was what was missing, depicted in Figure 3. However, the correlations consistently landed near 0 between reaction time and overall time, making it largely irrelevant to a good pace. Additionally, reaction time seems to have little consistency from run to run with a correlation of just 0.153. If we limit the analysis to what we define as 'good' runs, which pertain to attempts that finish in less than six seconds, we actually see slower times on average. We speculate that the additional mental burden of trying to react to the buzzer especially quickly may make one slower than when simply reacting naturally.

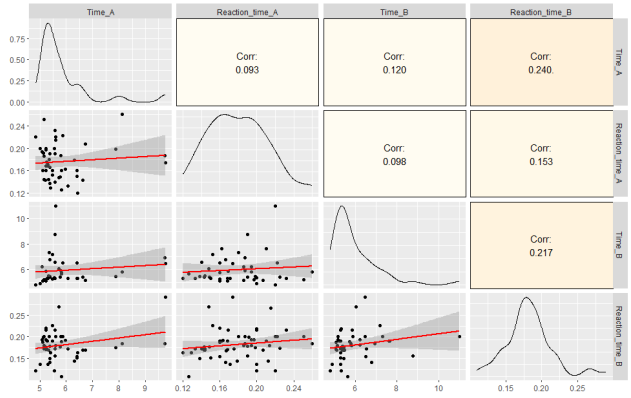


Figure 3: Pair plot of overall race time and reaction time variables for Route A (left) and Route B (right).

3 Linear models for speed climbing

To analyze the data previously described we use linear models based on the variables in Table 1 together with meaningful interactions between them. For a general description of the statistical methodology we refer to [2]. Here is, for example, a simple (overfiting) model for predicting Route A time.

$$\text{Model 1: } TimeA = 0.431 + 1.391X_1 - 0.388X_2$$
$$\text{Adjusted-}R^2 = .9507, \text{ RSE} = .1648$$

Given there are only 20 holds on the wall, it is clear that a racer's split to hold 20 will be statistically significant in determining the overall time on the route. As a result, it performs very well with an Adjusted-R^2 of .9507. Given the obvious target leakage, which occurs when data is used in a model that would not be available at the time of prediction, the model is not suitable for our purposes. Going forward, we remove

all split variables past hold 10 from the model fitting process to limit the issue. Model selection yielded the following model for predicting Route A time with only Route A splits.

$$\text{Model 2: } TimeA = 2.692 + .3624X_3 + 8.101X_4 - 2.688X_5 - 6.501X_6 + 1.208X_5X_6$$
$$\text{Adjusted-}R^2 = .6443, \text{ RSE} = .4428$$

As a result of our adjustment, this model only uses splits up to hold 10. It does include features related to hold 14, but only in relation to the runner's general strategy. Runners typically implement the same strategy every race, so the binary value can be known ahead of the buzzer going off. Given the earlier finding that the split to hold 10 is not significant, we have clearly improved our predictive ability using that variable by adding a couple extra features. Of particular relevance is variable X_3, which is a binary variable that indicates whether Route A was the first that a runner attempted in the competition. Generally, one has a slower time on the first route to ensure a valid score for qualification (competitors are ranked based on their fastest of the two times for the playoffs), while taking a more aggressive approach on the second route. This is supported by a greater number of falls during runners' second attempt in the data. An Adjusted-R^2 of .6443 indicates a strong fit. Concerning predicting Route A time with only Route B splits, the following model has been selected. This model is the most practical because it can be employed as a true prediction of Route A time simply based off Route B performance. Its Adjusted-R^2 of .5455 is the lowest of the three models but that is inevitable with the risk of falling always present and the two routes being completely different paths.

$$\text{Model 3: } TimeA = -28.9977 - 12.0108X_3 + 15.3040X_7 + 8.4712X_8 + 1.6232X_9 + 5.5963X_3X_7 - 3.7545X_7X_8$$
$$\text{Adjusted-}R^2 = .5445, \text{ RSE} = .5005$$

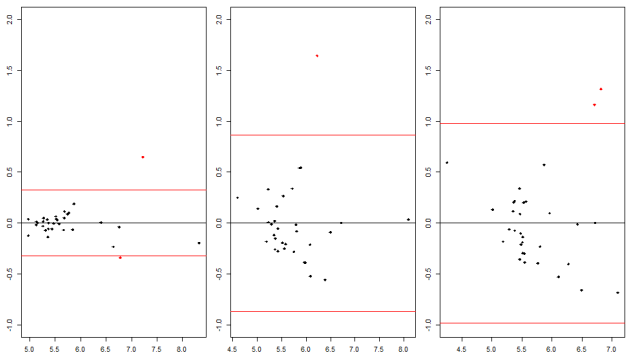


Figure 4: Residual plots for Model 1 (left), Model 2 (center), and Model 3 (right). Red lines mark 95% confidence bands and red points are significant outliers.

Regarding diagnostics, we found that the normality assumption was reasonably met in all models. Residual plots are shown in Figure 4. One outlier worth mentioning (falling outside the 95% confidence bands in all plots) is Leander Carmanns of Germany, who did well up to hold 16 before falling and losing several seconds of time. Unlike many other runners who fall, he opted to complete the race regardless. This allowed his run to remain in our model fit and create a large positive residual. Without incorporating falls into the dataset,

this sort of error would be very difficult to avoid.

4 Discussion

To summarize, we created a data frame of numerical and binary variables based on open source information on performance of several male speed climbing world champions and showed resulting best linear predictive models. We found that reaction time has minimal impact on total time and is inconsistent between runs. Runs under 6 seconds tend to have slower average reaction times, suggesting that mental burden of fast reaction has detrimental effect on total time. Moreover, first-half split time (holds 0 to 10) seem to yield lower correlation to total time than second-half split time (holds 10 to 20). Starting strong remains important, but finishing strong appears to be far more important. We also found that athletes take a more aggressive, but riskier skip strategy on the second run, yielding faster times but also more falls. In ongoing work we will gather more data to improve predictive models. We also aim to collect data on women climbers and will consider including more variables that may have an impact such as athletes' height and weight. We will also consider variance stabilizing transformations for some of the variables to improve resulting residual plots.
In ongoing work we will gather more data to improve predictive models. We also aim to collect data on women climbers and will consider including more variables that may have an impact such as athletes' height and weight.

References

[1] Lau, Emily (2021). Identifying physiological demands of Speed Climbing within a sample of recreational climbers. 10.13140/RG.2.2.18266.06089.

[2] R. Pruim (2011). Foundations and Applications of Statistics. An Introduction using R. American Mathematical Society, Providence, Rhode Island

[3] `ifsc.results.info/event/1354/`

[4] `https://www.youtube.com/@sportclimbing`

Identifying Extreme Representative Tennis Players and Match External Load in Male Grand Slam

Q. Brich*, M.Casals**, J. Cortés***, D. Fernández***, E. Baiget*

* Institut Nacional d'Educació Física de Catalunya, Spain. brich.pose@gmail.com; ebaiget@gencat.cat
** Institut Nacional d'Educació Física de Catalunya & Universitat de Vic-Universitat Central de Catalunya, Sport and Physical Studies Centre (CEEAF), Spain. marticasals@gmail.com
*** Department of Statistics and Operations Research, Research group in Biostatistics and Bioinformatics, GRBIO and Institute for Research and Institute for Research and Innovation in Health (IRIS), Universitat Politècnica de Catalunya - BarcelonaTech (UPC), Spain. jordi.cortes-martinez@upc.edu; daniel.fernandez.martinez@upc.edu

Abstract

This study explores extreme match demands and external load profiles in male Grand Slam tennis through cluster and archetypoid analyses. Data from 282 matches across the 2017 Grand Slam tournaments were examined to uncover distinct patterns in match characteristics and representative player profiles. Key variables included volume, intensity, and efficiency of play—such as points played, distance covered, shot count, hitting frequency, running speed, serve velocity, and first-serve success rates.

Clustering analysis identified four distinct match types, ranging from low-volume, low-intensity matches (primarily on grass courts) to high-volume, high-intensity matches (mostly on hard courts). Archetypoid analysis revealed diverse player profiles, representing extremes from high-volume, high-intensity defensive styles to low-volume, high-intensity offensive play.

These extreme representative player archetypes provide a nuanced understanding of external load demands and strategic diversity among elite male players. The findings offer practical implications for tailoring training and recovery strategies to specific match types and player styles. Future research using longitudinal data throughout the course of the match could further enhance our understanding of player adaptation and match dynamics.

1 Introduction

The Grand Slam tournaments—Australian Open, Roland Garros, Wimbledon, and US Open—represent the highest tier of professional male tennis (ATP, 2024). These two-week events feature 128 players competing in best-of-five-set matches across varied surfaces. To win a Grand Slam, a tennis player must endure over 20 sets and more than 200 games, highlighting the importance of physical preparation (ATP, 2024).

Recent years have seen a surge in tennis-focused research within sports science and analytics, driven by technologies like Hawk-eye and Foxtenn (Mecheri *et al.*, 2016; Baiget, Corbi and López, 2023). These tools have enabled detailed assessments of physical demands, especially in Grand Slams (Reid, Morgan and Whiteside, 2016; Kovalchik and Reid, 2017; Whiteside and Reid, 2017; Verhagen *et al.*, 2021). A central concept is "competition load", which combines exercise volume and intensity during a competition (Impellizzeri, Marcora and Coutts, 2019; Staunton *et al.*, 2022), and is often classified into external (physical workload) and internal (psychophysiological response) components (Impellizzeri, Marcora and Coutts, 2019; Impellizzeri *et al.*, 2023).

Grand Slam play involves high-intensity bursts with short rest intervals, governed by strict ITF timing rules (Fernandez, Sanz and Mendez, 2009; ITF, 2020; Pluim *et al.*, 2023). The best means for understanding external load in tennis include hitting and movement loads (Reid, Morgan and Whiteside, 2016; Kovalchik and Reid, 2017; Whiteside and Reid, 2017). For example, during the Australian Open (2012–2016), players averaged over 2,200 shots and nearly 10,000 meters of distance covered in the first four rounds, with frequent changes of direction (Fernandez, Sanz and Mendez, 2009; Kovalchik and Reid, 2017; Pluim *et al.*, 2023; Giles, Peeling and Reid, 2024).

While extensive descriptive data exist, much of the research examines isolated parameters. Emerging machine learning approaches now allow for integrated analyses, such as rally classification, serve optimization, and player profiling based on skill and consistency (Murray and Hunfalvay, 2017; Cui *et al.*, 2019; Fitzpatrick *et al.*, 2019; Giles *et al.*, 2023). However, no prior studies have applied unsupervised learning to map external load profiles in tennis. This study aims to fill that gap by analyzing player and match patterns using such methods.

2 Methods

2.1 Data collection

Point-by-point data from men's singles matches in the 2017 Grand Slam tournaments were sourced from Jeff Sackmann's GitHub repository (https://github.com/JeffSackmann/tennis_slam_pointbypoint), which compiles ATP data via web scraping. Aggregated datasets used in this study are available at https://github.com/jordicortes40/clustering_tennis. Only matches tracked by IBM Slamtracker or Infosys Oncourt technologies—using high-speed cameras, radar, and motion sensors—were included. Matches missing required variables, or those ending in walkovers or retirements, were excluded. A total of 282 matches involving 151 players met the inclusion criteria. Although Hawk-Eye data reliability is established, its assessment was beyond the scope of this study. At the end, we analyzed 282 men's singles matches from the 2017 Grand Slams, with hard courts being the most frequent surface (58%), reflecting their use in the Australian and US Opens.

2.3 Statistical analysis

Data from a single year were grouped into general match characteristics (e.g., surface, match outcome) and performance-specific metrics. These variables were further categorized into indicators of volume, intensity, and efficiency to assess external load and define player profiles. To account for variations in match length, all data were time-standardized. Additionally, a 20% Effective Playing Time (EPT) was considered to carry out this homogenization process. Exploratory data analysis was then conducted to examine the distribution of variables, detect outliers, and identify preliminary patterns in match and player characteristics. Cluster analysis was conducted to group matches based on external load characteristics. Clusterability was assessed using the Hopkins statistic (Hopkins and Skellam, 1954). Standard k-means clustering (Hartigan and Wong, 1979) using 10 random starting set of centroids was applied to identify match groups, with the optimal number of clusters determined via the elbow method. Archetypoid analysis (ADA) (Vinué and Epifanio, 2017) was used to extract real player profiles representing distinct external load patterns. All analyses were performed in R (v4.1.2) (R Core Team, 2020), using `kmeans` and `stepLArchetypoids3` (Epifanio, Ibañez and Simó, 2018) functions, and the `compareGroups` package (Subirana, Sanz and Vila, 2014).

3 Results

3.1 Match characteristics according to clustering approach

K-means clustering was applied to group matches by external load profiles. The Hopkins statistic (H > 0.99) confirmed the dataset's clusterability. The optimal number of clusters was four, as determined via the elbow method. Clusters were distributed as follows: Cluster 1 (n = 91), Cluster 2 (n = 86), Cluster 3 (n = 39), and Cluster 4 (n = 66).

Cluster 1 encompassed the least demanding matches, low in both volume and intensity, with an overrepresentation of grass courts. Cluster 2 had low volume but higher intensity, notably with a higher proportion of clay-court matches. Cluster 3 included high-volume, high-intensity matches, predominantly played on hard courts. Cluster 4 featured high volume but low intensity (except for M_SV), with a lower presence of clay surfaces. Efficiency variables (M_1stS, M_2ndS, M_DF) showed no significant variation across clusters.

3.2 Players representatives according to archetypoid analysis (ADA)

ADA revealed four distinct player profiles (H > 0.99), grouped into clusters: Cluster 1 (n = 29), Cluster 2 (n = 39), Cluster 3 (n = 52), and Cluster 4 (n = 32). Clusters were defined using volume (P_EPT, P_PP, P_ShC, P_Dist), intensity (P_HF, P_ARS), and efficiency (P_1stS, P_2ndS, P_DF) variables.

Cluster 1 represented the most physically demanding profile (moderate-to-high volume and high intensity), while Cluster 2 was the least demanding (low volume and intensity). Cluster 3 showed low volume with moderate-to-high intensity, and Cluster 4 involved the highest volumes but lower intensity. Serve efficiency metrics did not differ significantly among clusters.

Each player's profile was expressed as a combination of the four archetypoid profiles, represented by Darian King (black), Sam Groth (red), Luca Vanni (green), and Janko Tipsarevic (blue). For instance, Figure 1 displays the results of Cluster 4, where each player is represented by a pie chart showing their similarity to different archetypal profiles. The blue segment indicates similarity to the archetypoid represented by Janko Tipsarevic, with players showing larger blue areas being more similar to his playing style or performance characteristics.

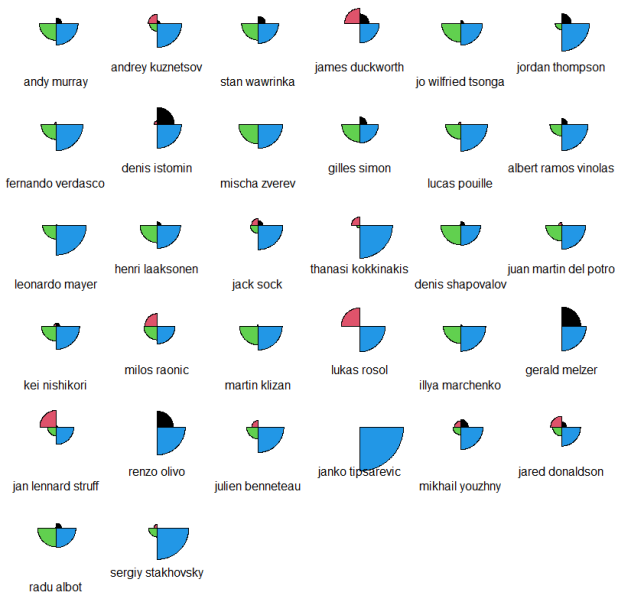


Figure 1. Players more similar to Janko Tipsarevic (Cluster 4).

4 Discussion

This study applied unsupervised learning techniques to characterize external load profiles in 2017 ATP Grand Slam matches, offering a more holistic view compared to prior research focused on isolated metrics. Using k-means and archetypoid analysis, we identified four distinct match and player profiles, differing in volume and intensity patterns. Matches ranged from low-volume, low-intensity (Cluster 1, often on grass) to high-volume, high-intensity (Cluster 3, mostly on hard courts), with two intermediate profiles (Clusters 2 and 4) mixing characteristics. Player clusters followed a similar pattern, with Cluster 1 being the most physically demanding and Cluster 2 the least.

4.1 Cluster comparison by match variables

External load was analyzed through intensity (HF, ARS, SV) (Reid, Morgan and Whiteside, 2016; Baiget and Iglesias, 2017) and volume (EPT, PP, ShC, Dist) (Reid, Morgan and Whiteside, 2016; Pluim *et al.*, 2023) metrics, which reflect hitting and movement loads (Reid, Morgan and Whiteside, 2016; Pluim *et al.*, 2023; Brich *et al.*, 2024). Cluster 1 featured low volumes and high SV values, indicating dominant players and shorter rallies, common on grass courts. Cluster 3, by contrast, involved longer, high-intensity rallies with low SV, mostly on hard courts, aligning with literature linking hard surfaces to higher hitting frequency (Baiget and Iglesias, 2017; Carboch *et al.*, 2019). Clusters 2 (low-volume, high-intensity) and 4 (high-volume, low-intensity) showed atypical patterns. Cluster 2, despite a higher proportion of clay matches, showed low match duration (M_PP), challenging assumptions that clay promotes matches with higher volumes. In contrast, Cluster 4 had longer matches but lower ShC, possibly due to fewer clay matches. These findings highlight the role of both surface and match dynamics (e.g., dominance, serve efficiency) in external load.

4.2 Cluster comparison by player's variables

Player profiles reflected similar external load dynamics. Cluster 1 included highly demanding players with strong defensive skills and low SV, leading to high movement and hitting metrics. Players such as David Ferrer and Alex de Miñaur exemplify this counterpunching style. In contrast, Cluster 2 players (e.g., Sam Groth, John Isner, Ivo Karlovic) relied on powerful serves and quick points, resulting in low loads. Clusters 3 and 4 included more versatile or hybrid players (e.g., Novak Djokovic, Rafael Nadal, Andy Murray, Kei Nishikori), blending offensive and defensive styles with varied external loads.

4.3 Study limitations

This study is limited by its reliance on 2017 data. However, by sharing the dataset and code, we encourage reproducibility and future updates with newer data. The use of estimated EPT (20%) may reduce precision, especially across tournaments with different time-keeping methods. Additionally, external factors—such as playing style, weather, and injuries—introduce variability not fully captured in the analysis. Additionally, the profiles created are based on only one year, meaning they do not reflect the players' overall profiles, but rather their state during that specific year. Finally, while various metrics were used, a universally accepted indicator of external load in tennis remains elusive, underscoring the complexity of workload quantification in this sport.

5 Conclusions

This study identified four distinct match profiles and four representative player archetypes based on external load variables from men's Grand Slam matches, using unsupervised clustering and archetypoid analysis. The resulting profiles revealed considerable variability in physical demands, shaped by factors such as surface type, rally structure, and player dominance. These findings emphasize the need for

individualized preparation and recovery strategies that reflect the specific external load patterns encountered by different players and match contexts. Moreover, this research highlights the utility of machine learning methods—particularly archetypoid analysis—in capturing the complexity of performance demands in elite tennis. Future studies should expand on this framework by incorporating longitudinal data from multiple seasons to better track how player profiles and match characteristics evolve over time.

References

1. ATP (2024) *ATP Stats*. Available at: https://www.atptour.com/en/.
2. Baiget, E., Corbi, F. and López, J. (2023) 'Influence of anthropometric, ball impact and landing location parameters on serve velocity in elite tennis competition', *Biology of Sport*, 40(1), pp. 273–281. doi: 10.5114/biolsport.2023.112095.
3. Baiget, E. and Iglesias, X. (2017) 'Maximal Aerobic Frequency of Ball Hitting: A New Training Load Parameter in Tennis', *Journal of Strength and Conditioning Research*, 31(1), pp. 106–114.
4. Brich, Q. *et al.* (2024) 'Quantifying Hitting Load in Racket Sports: A Scoping Review of Key Technologies', *International Journal of Sports Physiology and Performance*, pp. 1–14.
5. Carboch, J. *et al.* (2019) 'Match characteristics and rally pace of male tennis matches in three Grand Slam tournaments', *Physical Activity Review*, 7, pp. 49–56. doi: 10.16926/par.2019.07.06.
6. Cui, Y. *et al.* (2019) 'Clustering tennis players' anthropometric and individual features helps to reveal performance fingerprints', *European Journal of Sport Science*. Taylor & Francis, 19(8), pp. 1032–1044. doi: 10.1080/17461391.2019.1577494.
7. Epifanio, I., Ibañez, M. V. and Simó, A. (2018) 'Archetypal shapes based on landmarks and extension to handle missing data', *Advances in Data Analysis and Classification*, 12(3), pp. 705–735.
8. Fernandez, J., Sanz, D. and Mendez, A. (2009) 'A review of the activity profile and physiological demands of tennis match play.', *Strength and Conditioning Journal*, 31(4), pp. 15–26.
9. Fitzpatrick, A. *et al.* (2019) 'Important performance characteristics in elite clay and grass court tennis match-play', *International Journal of Performance Analysis in Sport*, 19(6), pp. 942–952. doi: 10.1080/24748668.2019.1685804.
10. Giles, B. *et al.* (2023) 'Differentiating movement styles in professional tennis: A machine learning and hierarchical clustering approach', *European Journal of Sport Science*, 23(1), pp. 44–53.
11. Giles, B., Peeling, P. and Reid, M. (2024) 'Quantifying Change of Direction Movement Demands in Professional Tennis Matchplay: An Analysis From the Australian Open Grand Slam', *Journal of Strength and Conditioning Research*, 38(3), pp. 517–525.
12. Hartigan, J. A. and Wong, M. A. (1979) 'Algorithm AS 136: A K-Means Clustering Algorithm', *Applied Statistics*, 28(1), p. 100. doi: 10.2307/2346830.
13. Hopkins, B. and Skellam, J. G. (1954) 'A New Method for determining the Type of Distribution of Plant Indivisuals', *Annals of Botany*, 18, pp. 213–227.
14. Impellizzeri, F. M. *et al.* (2023) 'Understanding Training Load as Exposure and Dose', *Sports Medicine*. Springer International Publishing, Online ahe. doi: 10.1007/s40279-023-01833-0.
15. Impellizzeri, F. M., Marcora, S. M. and Coutts, A. J. (2019) 'Internal and external training load: 15 years on', *International Journal of Sports Physiology and Performance*, 14(2), pp. 270–273. doi: 10.1123/ijspp.2018-0935.
16. ITF (2020) 'ITF Rules of Tennis'. Available at: http://www.itftennis.com/media/220771/220771.pdf.
17. Kovalchik, S. A. and Reid, M. (2017) 'Comparing matchplay characteristics and physical demands of junior and professional tennis athletes in the era of big data', *Journal of Sports Science and Medicine*, 16(4), pp. 489–497.
18. Mecheri, S. *et al.* (2016) 'The serve impact in tennis: First large-scale study of big Hawk-Eye data', *Statistical Analysis and Data Mining*, 9(5), pp. 310–325. doi: 10.1002/sam.11316.
19. Murray, N. P. and Hunfalvay, M. (2017) 'A comparison of visual search strategies of elite and non-elite tennis players through cluster analysis', *Journal of Sports Sciences*. Routledge, 35(3), pp. 241–246. doi: 10.1080/02640414.2016.1161215.

20. Pluim, B. M. *et al.* (2023) 'Physical Demands of Tennis Across the Different Court Surfaces, Performance Levels and Sexes: A Systematic Review with Meta-analysis', *Sports Medicine*. Springer International Publishing, 53(4), pp. 807–836. doi: 10.1007/s40279-022-01807-8.
21. R Core Team (2020) 'R: A language and environment for statistical computing'. Vienna: Foundation for Statistical Computing. Available at: https://www.r-project.org/.
22. Reid, M., Morgan, S. and Whiteside, D. (2016) 'Matchplay characteristics of Grand Slam tennis: implications for training and conditioning', *Journal of Sports Sciences*, 34(19), pp. 1791–1798. doi: 10.1080/02640414.2016.1139161.
23. Staunton, C. A. *et al.* (2022) 'Misuse of the term "load" in sport and exercise science', *Journal of Science and Medicine in Sport*. The Authors, 25(5), pp. 439–444. doi: 10.1016/j.jsams.2021.08.013.
24. Subirana, I., Sanz, H. and Vila, J. (2014) 'Building Bivariate Tables: The compareGroups Package', *R. Journal of Statistical Software*, 57(12), pp. 1–16.
25. Verhagen, E. *et al.* (2021) 'Tennis-specific extension of the International Olympic Committee consensus statement: Methods for recording and reporting of epidemiological data on injury and illness in sport 2020', *British Journal of Sports Medicine*, 55(1), pp. 9–13. doi: 10.1136/bjsports-2020-102360.
26. Vinué, G. and Epifanio, I. (2017) 'Archetypoid analysis for sports analytics', *Data Mining and Knowledge Discovery*, 31, pp. 1643–1677. doi: https://doi.org/10.1007/s10618-017-0514-1.
27. Whiteside, D. and Reid, M. (2017) 'External match workloads during the first week of australian open tennis competition', *International Journal of Sports Physiology and Performance*, 12(6), pp. 756–763. doi: 10.1123/ijspp.2016-0259.

Scoring probability maps on the basketball court through Spatial Point Pattern analysis

M.L. Carlesso* A. Cappozzo** A. Gilardi*** M. Manisera* P. Zuccolotto*

*Big&Open Data Innovation Laboratory (BODaI-Lab), University of Brescia, Italy
mirko.carlesso@unibs.it - marica.manisera@unibs.it - paola.zuccolotto@unibs.it
** Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Italy - andrea.cappozzo@unicatt.it
*** Department of Economics, Management and Statistics, University of Milano - Bicocca - andrea.gilardi@unimib.it

Abstract

Measuring shooting performances on the basketball court is crucial for understanding game dynamics and enhancing strategic decision-making. Accurate scoring probability evaluation offers insights that directly impact coaching decisions and players development. Spatial statistics and, in particular, point process analyses provide an ideal framework to accomplish these tasks. In this paper, we model the spatially-varying intensity of shots using classical point pattern methods, taking into account the outcome of each shot (i.e., made or missed). This approach lets us capture the spatial nature of shooting, going beyond traditional binary outcome models. By estimating the shot intensity at different locations, we derive scoring probabilities that reflect shooting performances across the court. Then, we create scoring probability maps, offering a clear visualization of shooting efficiency by location. These maps enable the coaching staff to better understand shooting dynamics and enhance their strategic planning. Our approach is validated through a case study using data from the Italian Basketball First League (LBA), provided by a professional club, ensuring high data quality and real-world relevance.

1 Introduction

Accurately assessing a basketball team's or player's offensive performance primarily requires to understand the probability of scoring on each shot attempt, a fundamental aspect often quantified through effective field goal percentage (EFG %) since the development of a formal setting for the analytical approach to basketball thanks to [4]. Analyzing scoring probability from a spatial perspective, *i.e.* considering the specific location on the court where a shot is taken, provides a much richer and more granular evaluation, as pointed out in the spatial analysis of professional basketball by [5].

The spatial approach allows for the creation of scoring probability maps, visually representing areas of higher and lower efficiency. These maps offer immediate visual insights into a player's or team's shooting efficiency. Recent research has significantly advanced methodologies for estimating these probabilities, with a wide variety of approached from the point of view of the statistical methods adopted. An interesting research line addresses the use of Bayesian methods: [3] propose a Bayesian joint model for the mark and the intensity of marked point processes, where the intensity is incorporated in the mark model as a covariate. [8] employ

Bayesian nonparametric learning for point processes, with a flexible modeling of the underlying spatial intensity of shot attempts built upon a combination of Dirichlet process and Markov random field. This allows a local spatial homogeneity when estimating a globally heterogeneous intensity surface. Furthermore, [7] resort to Bayesian hierarchical models to examine positional differences in shooting accuracy, acknowledging that players in different roles might exhibit varying spatial shooting profiles. Machine learning techniques such as CART, random forests, and extremely randomized trees have been applied for spatial performance analysis by [10], building upon earlier work on spatial performance indicators and graphs [9]. In these works, machine learning methods have proved able to effectively capture non-linear relationships between shot location and scoring probability. More recently, [2] explore the use of Indicator Kriging for generating scoring probability maps, providing an alternative able to account for spatial correlation and comparing its performance with machine learning methods. Another model-based approach has been investigated by [6], who propose a statistical framework for shot charts that explicitly considers the physical boundaries of the basketball court by means of Gaussian mixtures for bounded data.

This contribution refers to the field of spatial analysis, specifically to the domain of marked spatial point processes. Our primary aim is to estimate shot intensity and scoring effectiveness as a function of spatial variables such as distance and angle, generating interpretable shooting maps that reflect a team's scoring patterns. To this end, we use data from the 2022/2023 season of the Lega Basket A (LBA), Italy's top-tier professional basketball league. The insights provided by the analysis are intended to support coaching decisions and inform performance optimization strategies. The paper is organised as follows: in Section 2 we briefly define the statistical approach to our problem, in Section 3 we present and comment the results of the case study, while Section 4 concludes and outlines the next research lines.

2 Methods

We assume that the basketball shots performed by a given team during a complete season, hereby denoted as $x = \{(\boldsymbol{x}_1, m_1), \ldots, (\boldsymbol{x}_n, m_n)\}$, represent a (finite) realisation of a *multitype* point process X within a bounded spatial window $W \subset \mathbb{R}^2$ [1]. The term $\boldsymbol{x}_i = (x_{1i}, x_{2i})$, $i = 1, \ldots, n$ denotes the cartesian coordinates of the i-th shot within the basketball court whereas m_i denotes its mark, taking value in a discrete space M. In this paper, we consider binary marks, meaning that $M = \{0, 1\}$, where 1 indicates a successful shot (i.e., a made basket) and 0 is a miss.

Assume that there exists a function, say $\lambda(\boldsymbol{x}, m)$, satisfying the following condition

$$\mathbb{E}[N(A \times B)] = \int_A \sum_{m \in B} \lambda(\boldsymbol{x}, m)\, \mathrm{d}\boldsymbol{x}, \qquad A \subset W, B \subset M,$$

where $N(A \times B)$ denotes the number of locations $\boldsymbol{x}_i$ falling in the set A having mark in B. Such function is usually termed the *intensity* of the process and, broadly speaking, it describes the rate at which events of type $m \in M$ occur in the given region. Following a classical and pragmatic hypothesis in the spatial point processes literature, we assume that X is a spatially inhomogeneous marked Poisson point process on $W \times M$.

The log-likelihood function for a multitype Poisson point process is (up to a constant) equal to

$$\log L = \sum_{i=1}^{n} \log \lambda(\boldsymbol{x}_i, m_i) - \sum_{m \in M} \int_W \lambda(\boldsymbol{x}, m)\, \mathrm{d}\boldsymbol{x}.$$

Throughout this paper, we specify a semi-parametric log-linear model for $\lambda(\boldsymbol{x}, m)$ as a function of a series of covariates related to the characteristics of the basketball court:

$$\log \lambda(\boldsymbol{x}, m) = \beta_{0,m} + \beta_{1,m}(x_1) + \beta_{2,m}(x_2) + \beta_{3,m}(\text{distance}(\boldsymbol{x})) + \beta_{4,m}(\text{angle}(\boldsymbol{x})). \tag{1}$$

The term $\beta_{0,m}$ represents a mark-specific intercept, whereas x_1 and x_2 correspond to the Cartesian coordinates of the shot location. In addition, distance($\boldsymbol{x}$) and angle($\boldsymbol{x}$) respectively denote the Euclidean distance from location $\boldsymbol{x}$ to the basket and the shooting angle, measured in radians from $-\pi$ to $+\pi$. The notation $\beta_{j,m}(\cdot)$ for $j = 1,2,3$ highlights that the corresponding covariates were smoothed using mark-specific thin-plate spline transformation to capture potential non-linear relationships between their effects and the shot intensity [11]. For the angular component, namely $\beta_{4,m}(\cdot)$, we adopted a cyclic cubic spline to account for the periodic nature of the angle and ensure smoothness at the boundaries (i.e., between $-\pi$ and π). These semi-parametric transformations allow for greater flexibility in the log-linear intensity by incorporating smooth spatial trends and complex relationships with court features, while maintaining regularization through smoothness constraints and a certain degree of interpretability.

As already mentioned in the Introduction, a fundamental aspect in the statistical analysis of basketball data is the quantification of the scoring probability In the literature of spatial point processes, this quantity is usually named *(normalised) relative-risk function* or *probability distribution of one type* and it is defined as

$$\rho(m = 1|\boldsymbol{x}) = \frac{\lambda(\boldsymbol{x}, 1)}{\lambda(\boldsymbol{x}, 0) + \lambda(\boldsymbol{x}, 1)}. \tag{2}$$

The values of $\rho(m = 1|\boldsymbol{x})$ represent the conditional probability that, given that there exists an event at location $\boldsymbol{x}$, such point is of type 1 (i.e. a successful shot). More precisely, values of $\rho \simeq 0.5$ indicate that made and missed shots at location $\boldsymbol{x}$ are equally likely, and the absence of hot and cold spots. Values near 1 mean that virtually every shot taken there goes in (a perfect *hot spot*), whereas values approaching 0 signal that almost every shot is missed (a *cold spot*). Given a parametric model for $\lambda(\boldsymbol{x}, m)$, such as the one described in Equation (1), and a fit for the intensity function, say $\hat{\lambda}(\boldsymbol{x}, m)$, a parametric estimate of $\rho(m = 1|\boldsymbol{x})$, say $\widehat{\rho}(m = 1|\boldsymbol{x})$, can be obtained by replacing the intensity functions $\lambda(\cdot, \cdot)$ in Equation (2) with the corresponding fitted values.

3 Case study

Building on this theoretical framework, we now turn to the practical application using real-world basketball data. Our analysis starts from a detailed dataset of the LBA 2022/2023 season provided by Openjobmetis Varese, one of the sixteen teams in the Italian league. This dataset includes play-by-play information, where each row corresponds to a specific game event, capturing various characteristics of the action. For our purposes, we focus on shot events, with each entry detailing the shot's coordinates on the half-court, the player attempting it, the result (made or missed), and other relevant information. As already mentioned, we treat each shot as a point in the court space, aligning our analysis with the spatial point pattern framework. A key aspect of this approach is the definition of an observation window W, the region within which points are analyzed. Instead of using the entire half-court as our observation window, we developed a custom window tailored to typical shooting behavior in basketball. Specifically, we excluded regions where shots are generally attempted only in desperate situations, such as at the end of a possession. These excluded

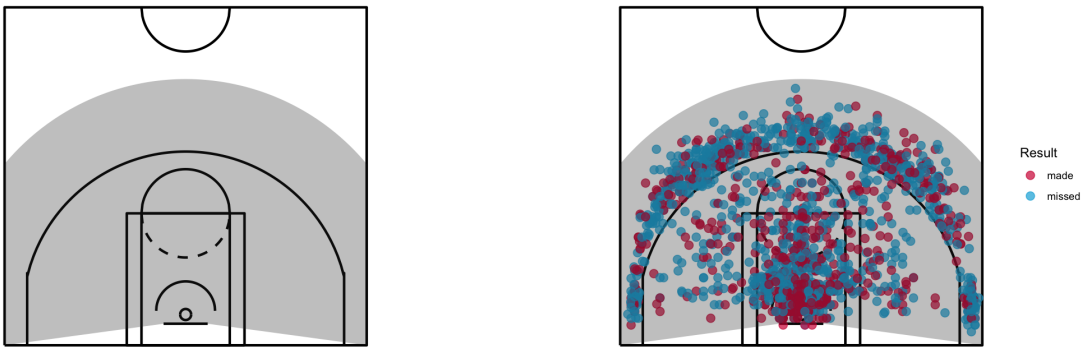


Figure 1: *Left*: The outermost rectangle denotes the half-court, whereas the grey polygon denotes the observation window. *Right*: Shotchart of made/missed shot - data of Tezenis Verona, season 2022/2023.

shots account for only 0.06% of all attempts, further reinforcing the spatial validity of this refinement. This spatial crop ensures that our analysis captures meaningful shooting patterns, avoiding distortions caused by low-probability areas. Figure 1 illustrates this setup. The left panel shows the refined observation window, while the right panel presents a scatter plot of all shots taken by the team Tezenis Verona, with each point colored according to its result. As seen in the scatterplot, shot distribution, and thus intensity, is not uniform across the court, reflecting modern basketball strategies that favor attempts near the basket or beyond the three-point line.

In the practical application of the theoretical framework, scoring probability maps are generated by estimating Equation (2) over a fine grid of points within the observation window. Specifically, for each point in this grid, the estimated probability of scoring, previously denoted as $\hat{\rho}(m = 1|\boldsymbol{x})$, is calculated using the fitted intensity functions for made and missed shots. This approach produces a detailed spatial map where each location displays the probability of a successful shot. An example of these maps is shown in the left panel of Fig.2. The map shows that shots taken near the basket are almost always successful, and that the probability of scoring decreases with distance in a non-linear way. It can also be seen how scoring probability depends on the shooting angle. In general, Tezenis Verona seems to have had better efficiency from the right side of the half-court, which represents useful information for coaching staff.

After generating this map using the parametric approach, we proceed to evaluate its quality by carrying out a residuals analysis. Similar to classical methods adopted in the spatial statistics literature, our evaluation procedure involves comparing this parametric map with a non-parametric estimate of the scoring probability derived using kernel smoothing techniques. More precisely, we construct a raw-residual map, where raw residuals are computed as point-wise differences between the parametric and the nonparametric estimate. As illustrated in the right panel of Fig.2, the residuals are generally close to zero, providing strong evidence that the parametric model effectively approximates the non-parametric approach. This result supports the use of a parametric approach, providing confidence in its reliability. This groundwork provides a valuable starting point for future developments, to be briefly discussed in the conclusion.

Figure 3 illustrates an additional application of the aforementioned methodology, displaying two shot maps for different teams based on the parametric intensity estimation approach. Focusing on the left panel of

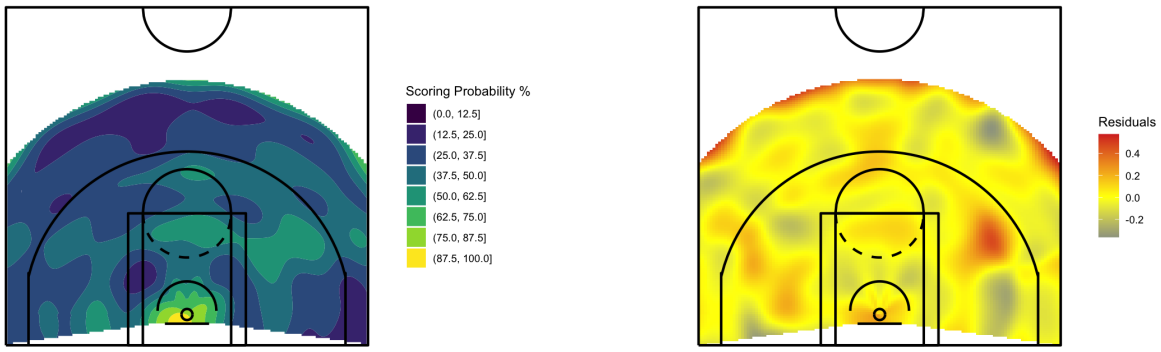


Figure 2: Scoring probability map produced via parametric estimation of the intensity function (left) and residuals map (right) - data of Tezenis Verona, season 2022/2023.

Fig. 3, which displays the map for Banco di Sardegna Sassari, we observe that the team shows higher shooting efficiency from the side areas compared to the central regions. In particular, there is a high-performance area in the mid-lower left corner. Such insights, when cross-referenced with the player profiles of Sassari, can yield actionable information for training or game preparation. Another notable feature in Sassari's map is a low-performance zone approximately 5-6 feet from the basket, which extends across all angles. In contrast, the map for Umana Reyer Venezia (right panel of Fig.3) shows stronger performance within the same 5-6 feet range from the basket, highlighting a difference in shot success between the two teams. However, Venezia struggles with mid-range shots from the wings, with visible low-performance areas on both sides. Similarly to Sassari, Venezia shows better efficiency on side three-point shots compared to central three-point attempts. This difference may be attributed to the nature of side threes, which are often catch-and-shoot opportunities, typically easier to execute than off-the-dribble shots attempted frequently from the top of the arc. These maps provide coaching staff with a valuable tool for analyzing shooting performance, either for their own team or for an opponent. Compared to traditional methods, such as scatterplots or shot charts that divide the court into predefined zones, these maps offer a more comprehensive and visually intuitive assessment, capturing subtle variations in performance across the court [2].

4 Conclusions

In this paper, we explored the use of spatial point processes to evaluate a team's shooting performance. Our analysis focused on estimating the intensity of made and missed shots, allowing us to generate shooting maps that display the probability of scoring from any location on the court. These maps offer valuable insights for coaching staff, supporting better in-game decisions and targeted training strategies. We adopted a parametric approach to model shot intensity, using shot coordinates, distance, and angle as predictors. The accuracy of these parametric maps was validated by comparing them to their nonparametric counterparts through residual analysis, demonstrating a strong alignment between the two. The encouraging results obtained validate the ongoing use of the parametric approach. Future work will focus on extending the framework through a unified training procedure encompassing all teams, where team-specific variations are modeled as spatially

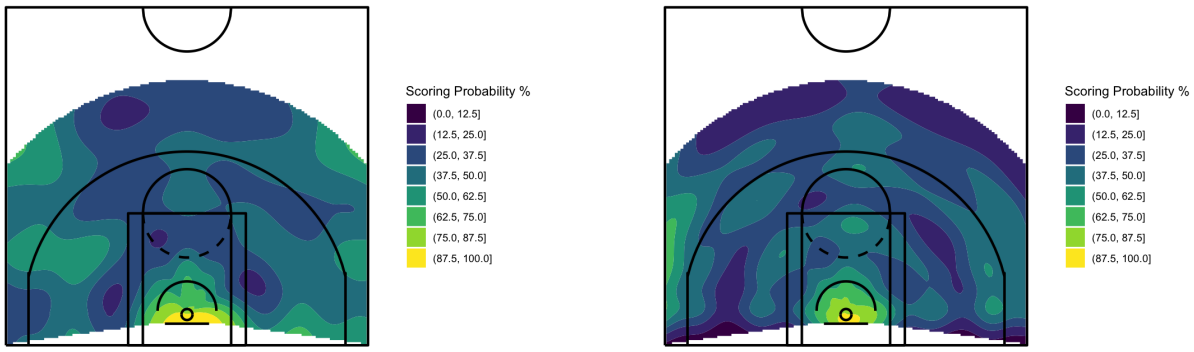


Figure 3: Scoring probability map produced via parametric estimation of the intensity function for Banco di Sardegna Sassari (left) and Umana Reyer Venezia (right) - data of season 2022/2023.

distributed random effects.

References

[1] Baddeley, A. and Rubak, E. and Turner, R. (2016) *Spatial point patterns: methodology and applications with R* CRC press.

[2] Carlesso, M.L., Cappozzo, A., Manisera, M. and Zuccolotto, P. (2024) *Scoring probability maps in the basketball court with Indicator Kriging estimation.* Computational Statistics, 1-21.

[3] Jiao, J., Hu, G. and Yan, J. (2021) *A Bayesian marked spatial point processes model for basketball shot chart*. Journal of Quantitative Analysis in Sports **17**, 77-90.

[4] Kubatko, J., Oliver, D., Pelton, K. and Rosenbaum, D. T. (2007). *A starting point for analyzing basketball statistics. Journal of quantitative analysis in sports* Journal of quantitative analysis in sports, **3**.

[5] Miller, A., Bornn, L., Adams, R. and Goldsberry, K. (2014). *Factorized point process intensities: A spatial analysis of professional basketball.* In International conference on machine learning, PMLR, 235-243.

[6] Scrucca, L. and Dimitris, K. (2025) *A model-based approach to shot charts estimation in basketball.* Computational Statistics, 1-18.

[7] Wang, F. and Zheng, G. (2022) *Examining positional difference in basketball players' field goal accuracy using Bayesian Hierarchical Model.* International Journal of Sports Science & Coaching **17**, 848-859.

[8] Yin, F., Jiao, J., Yan, J. and Hu, G. (2022) *Bayesian nonparametric learning for point processes with spatial homogeneity: A spatial analysis of NBA shot locations.* In International Conference on Machine Learning, PMLR, 25523-25551.

[9] Zuccolotto, P., Sandri, M. and Manisera, M. (2021). *Spatial performance indicators and graphs in basketball.* Social Indicators Research **156**, 725-738.

[10] Zuccolotto, P., Sandri, M. and Manisera, M. (2023) *Spatial performance analysis in basketball with CART, random forest and extremely randomized trees.* Annals of Operations Research **325**, 495-519.

[11] Wood, S.N., 2017. Generalized additive models: an introduction with R. Chapman and hall/CRC.

Tennis match outcome prediction using temporal directed graph neural networks

Lawrence Clegg* and John Cartlidge**

* School of Computer Science, University of Bristol, UK
** School of Engineering Mathematics and Technology, University of Bristol, UK
* *lawrence.clegg@bristol.ac.uk*, ** *john.cartlidge@bristol.ac.uk*

Abstract

We present the first application of a graph neural network for tennis match outcome prediction. Using MagNet, an existing spectral graph neural network for directed graphs, we construct temporal directed graphs by representing players as nodes and surface-specific historical match outcomes as edges. The model is trained and evaluated using a dataset of Grand Slam, ATP Masters 1000, and two ATP 500 events from 2007 to the conclusion of the US Open in September 2024. Following hyperparameter optimisation, a tuned model on the out-of-sample data achieves comparable predictive accuracy (66.0%) to the benchmark weighted Elo rating system (65.6%). Many recent advancements in tennis match prediction have focused on incremental improvements to the Elo rating system, such as incorporating margin of victory and surface-specific adjustments. Our research shifts this paradigm by demonstrating that graph neural networks, which inherently capture complex relational and temporal dynamics, offer a powerful alternative for pairwise comparison tasks such as tennis match prediction.

1 Introduction

The sport of tennis is well suited for predictive modelling due to several distinctive features. Its scoring structure is inherently hierarchical, advancing from points to games to sets, which lends itself to mathematical modelling. Furthermore, singles matches involve only two athletes, which avoids the roster-level complexities that arise in team sports where transfers and injuries can obscure model signals. The worldwide tournament calendar also ensures that the same players face one another repeatedly on multiple surfaces and across various event tiers, thereby generating a dense record of head-to-head outcomes that can be analysed for statistical inference.

In 2016, Kovalchik (2016) surveyed several tennis prediction methods and found that an Elo rating system approach developed by Morris & Bialik (2015) was the most effective match outcome predictor of tennis by both accuracy and log-loss metrics, when not considering a bookmaker consensus model proposed by Leitner et al. (2009). However, the survey did not consider graph theoretic methods, such as a PageRank approach by Dingle et al. (2013), and since the publication of the 2016 survey, there have been significant improvements in the utilisation of graph theory and graph representation techniques. Bayram et al. (2021) derived surface-specific player scores from three centrality indices (out–in degree difference, Hubs, and PageRank) and supplied them to several match-level classifiers; the best classifier, SVM+, reached 66% accuracy on 21,083 matches (2012–2020). Considering richer graph representations of historical tennis

matches, Arcagni et al. (2023) generated global ratings via the eigenvector centrality and used them in a logit model, achieving a Brier score of 0.194 on ATP matches (2016–2020), outperforming standard Elo and a margin-of-victory Elo variant by Kovalchik (2020). The notable performance of these methods highlights the increasing efficacy of graph-theoretic approaches in tennis prediction and indicates the possibility of further advancements in this area. Graph neural networks (GNNs) have gained prominence over the last decade in learning complex graph representations and they have been applied to outcome prediction in various sports. For instance, GNNs have been used to predict outcomes in American football and Counterstrike: Global Offensive (Xenopoulos & Silva 2021), association football (Mirzaei 2022), and basketball (He et al. 2022). However, the use of GNNs in tennis match prediction remains unexplored.[1]

Here, we present our initial findings of a graph neural network approach to predicting tennis matches, using surface-specific graphs and a spectral model tailored for directed graphs. Specifically we use the GNN developed by Zhang et al. (2021), named MagNet, and demonstrate its ability to achieve comparable out-of-sample prediction accuracy compared to the Elo rating system variant proposed by Angelini et al. (2022).

2 Method

2.1 Graph Representation

We interpret each tournament round (e.g., Wimbledon 2010 Quarter Finals) as a discrete time point, which we call a snapshot. For each snapshot, we construct a surface-specific graph using all the preceding matches played on that particular court surface (i.e., *grass* for Wimbledon). The choice of surface-specific graphs is motivated by the well-documented variation in player performance across different court surfaces, with researchers such as Fayomi et al. (2022) finding "surface on which a game is played on contributes significantly towards a player's performance". By maintaining separate graphs for clay, grass, and hard courts, we attempt to capture this surface-dependent dynamic of player performance. For each snapshot i and surface $S \in \{clay, grass, hard\}$, we construct the graph,

$$G_i^S = (V, E_i^S, \mathbf{X_i^V}, W_i^S) \tag{1}$$

We consider the set of all players in the dataset as the fixed node set V. We use both static player attributes and the dynamic graph-based metrics as node features $\mathbf{X}_i^V \in \mathbb{R}^{|V| \times d}$. The static player-node feature set comprises the *height*, *weight*, *date of birth*, and *handedness* of each player. We assume player weights to be static due to data limitations. Additionally, where we could not obtain these values for some players, we imputed the medians. For dynamic features, we use the node in-degrees and out-degrees, summarising the number of incoming and outgoing edges for each player-node, respectively. Each feature vector is ℓ_2-normalised.

The edge set $E_i^S \subseteq V \times V$ dynamically evolves for each new snapshot. Provided there is at least one match played between player u and player v on surface S prior to the snapshot i, we add a weighted directed edge between the two player-nodes. We use edge weights W_i^S to describe the historical dominance between players in their previous encounters. Each weight $w_{uv} \in W_i^S$ corresponds to an edge in the set E_i^S. To determine a weight w_{uv}, we first calculate a dominance score,

[1] A systematic Google Scholar search (30/03/2025) using combinations of GNN and tennis prediction terms (e.g., "graph neural network", "tennis prediction") yielded no relevant publications applying GNNs to tennis outcome forecasting.

$$D_k(u,v) = \frac{\sum_{j=0}^{k} g_j(u,v) e^{-\lambda(t_k(u,v)-t_j(u,v))}}{\sum_{j=0}^{k} e^{-\lambda(t_k(u,v)-t_j(u,v))}}, \tag{2}$$

where $g_j(u,v)$ is the fraction of games won by player u against player v in match j, $t_k(u,v)$ is the timestamp of the k-th match between player u and player v, and $\lambda > 0$ is an adjustable parameter that controls the rate at which the influence of older matches diminishes. In this base implementation, we set λ such that the contribution of a match from one year prior to t_k is scaled by a factor of 99% compared to the contribution of a match at t_k. This time-decayed weighted historical record of the proportion of games won provides a more accurate estimation of player skill, while also accounting for potential changes in player skill over time.

If $D_k(u,v) > 0.5$, we assign the direction of the edge as v,u, pointing from the historically weaker player to the stronger. If $D_k(u,v) < 0.5$, the edge is directed as u,v with weight $1 - D_k(u,v)$. Mathematically:

$$w_k(u,v) = \begin{cases} D_k(u,v) & \text{if } D_k(u,v) > 0.5, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

2.2 MagNet

To generate match outcome probability estimates from the constructed graph, we employ the spectral graph convolutional network (GCN) MagNet (Zhang et al. 2021).[2] We use MagNet to estimate the probability a directed edge u,v exists, which can be used to form match probability estimations for player u against player v. We employ a straightforward hyperparameter selection strategy for MagNet, setting the learnable parameter $q = 0.25$ to maximise the influence of edge direction and weights. The model uses a Chebyshev polynomial order of 1, a single MagNet convolutional layer, and 32 hidden channels, with a dropout rate of 0.3. Each snapshot training run consists of a maximum 75 epochs, with early stopping applied after 7 epochs of no improvement.

We compute the probability of player u winning a set against player v, denoted $\hat{p}_{uv}$ so that we can use the hierarchical structure of tennis scoring and account for both best-of-3-sets matches $\hat{P}_3$ and best-of-5-sets matches $\hat{P}_5$ under the assumption of independent and identically distributed sets. To mitigate the impact of player ordering (u,v) on probability estimates and to ensure a balanced prediction, we average the model outputs from both directed perspectives:

$$\hat{p}_{uv} = \frac{\hat{z}_{uv} + (1 - \hat{z}_{vu})}{2}, \tag{4}$$

where $\hat{z}_{uv}$ represents the estimated probability associated with the directed edge from v to u (indicating player u's dominance over v), and $\hat{z}_{vu}$ represents the estimated probability associated with the directed edge from v to u (indicating player u's dominance over v). Thus, final match outcome probabilities are calculated from:

$$\hat{P}_3 = \hat{p}_{uv}^2 + 2\hat{p}_{uv}^2(1 - \hat{p}_{uv}), \quad \hat{P}_5 = \hat{p}_{uv}^3 + 3\hat{p}_{uv}^3(1 - \hat{p}_{uv}) + 6\hat{p}_{uv}^3(1 - \hat{p}_{uv})^2 \tag{5}$$

Before predicting outcomes for each snapshot, the model undergoes retraining, using an Adam optimizer with a learning rate of 0.01 and a cross-entropy loss function to optimise the model's filter coefficients and parameters.

[2]MagNet is accessible via the PyTorch Geometric Signed Directed library (He et al. 2024).

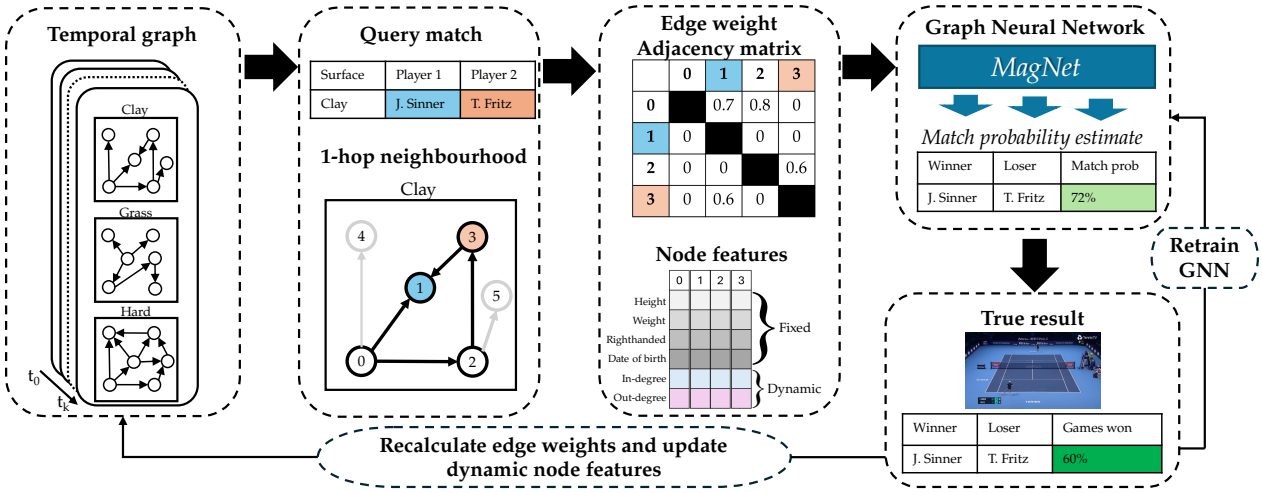


Figure 1: Overview of temporal-graph prediction methodology

Left-to-right: Surface-specific graphs G_k^S (see Equation 1) incorporate all observed match data on surface S up to timestamp t_k. When a match is queried, the 1-hop neighbourhoods of the two competing players within the relevant surface-graph are considered. The features of the neighbouring player-nodes are provided as input to the MagNet model, along with edge weights w_k (see Equation 3). Represented in an asymmetric adjacency matrix, they are uni-directional and point towards the player that, time-adjusted, has won more games. For example, the edge between players Sinner and Fritz has weight $w_k(3,1) = 0.6$, indicating that Sinner has a time-adjusted win rate of 60% of games played against Fritz. A forward pass through the MagNet architecture, followed by averaging (see Equation 4) and set-to-match transformation (see Equation 5) yields an estimated match win probability. Following the actual match outcome at time t_k, the graph structure and its associated attributes are updated to form G_{k+1}^S for matches taking place in the next timestamp. Edge weights w_{k+1} are recalculated using Equation 3. The static player features, consisting of height, weight, date-of-birth, and righthandedness, are held constant. If there has been a new edge added, or a change in direction of an edge, the two remaining dynamic player features (node in/out degrees) are updated. Finally, the MagNet model is retrained on the updated graph.

For training and evaluation, the dataset is processed chronologically. Initially, a graph is constructed from the first 65% of matches for model training, with the subsequent 15% of matches used to optimize the model via cross-entropy loss. A further 10% of matches are allocated as a validation set, where we implement early stopping to prevent overfitting. For the remaining 10% of snapshots, a walk-forward validation approach is adopted. Graphs are constructed using all preceding matches for each snapshot. After predictions are made, the matches from that snapshot are incorporated, updating the data distribution such that the initial graph eventually encompasses 75% of the data, while training and validation maintain proportions of 15% and 10%, respectively. A schematic overview of the model architecture is provided in Figure 1.

3 Results

We use match data from the two highest tiers of men's professional tennis (Grand Slams and ATP Masters 1000) from 15 January 2016 to 8 September 2024, sourced from tennis-data.co.uk. Since there is only one

Table 1: Model Performance and Betting Profitability: Overall and by Surface.

			Accuracy		Favourite-Kelly Staking				
Surface	Matches	Model	Acc	Brier	Staked	Return	Profit	ROI (%)	Sharpe
Clay	314	MagNet	**0.675**	0.216	59.07	59.21	0.14	**0.24**	**0.08**
		WElo	0.631	**0.213**	49.20	39.26	−9.94	−20.21	−7.29
		PS	0.736	0.179					
Grass	169	MagNet	**0.710**	0.209	44.15	49.27	5.12	**11.60**	**3.54**
		WElo	0.692	**0.199**	20.59	19.42	−1.17	−5.68	−2.44
		PS	0.746	0.173					
Hard	591	MagNet	0.638	0.221	82.29	91.23	8.94	**10.87**	**1.94**
		WElo	**0.658**	**0.210**	65.68	64.27	−1.41	−2.14	−0.62
		PS	0.699	0.193					
All	1074	MagNet	**0.660**	0.218	185.50	199.71	14.20	**7.66**	**1.81**
		WElo	0.656	**0.209**	135.47	122.94	−12.52	−9.24	−2.99
		PS	0.717	0.186					

* Note: Bold values indicate the best performance (highest Accuracy, lowest Brier Score, highest ROI, highest Sharpe Ratio) among the evaluated models for that surface/metric, excluding the PS (Pinnacle Sports) benchmark. 'All' summarises overall performance across surfaces. Matches is the number of matches in the out-of-sample test set for each surface. Kelly Staking metrics are rounded to two decimal places; Accuracy and Brier Score to three decimal places.

grass tournament in these two tiers (Wimbledon), we include two ATP 500 tournaments: Queen's Club Championships and the Halle Open, to ensure there are sufficient grass court matches for robust model training. The dataset contains information such as match date, competitors, games won by each competitor for each set played, tournament, surface, round of the tournament, and betting odds from several bookmakers; from which we select Pinnacle Sports to represent betting market accuracy and the available odds in our betting analysis. In total, our dataset contains 7110 matches played by 426 players with an out-of-sample test set comprising 1075 professional men's tennis matches beginning on 31 October 2023 and concluding with the 2024 US Open final, on 08 September 2024. We gather player-node features from tennisexplorer.com.

We assess our model's predictive strength using the classification accuracy of wins and the Brier Score, which reports the mean squared error between the predicted probabilities and the actual outcomes. The results are summarised in Table 1, where we provide performance values per surface. Our proposed graph model attains an overall classification accuracy of 66.0%, outperforming the Weighted Elo (WElo) approach proposed by Angelini et al. (2022), which attained 65.6%. When filtering by surface, the graph model shows superior performance on clay and grass courts, although its accuracy remains lower than WElo on hard courts. All results remain lower than the bookmaker odds' implied probability, denoted as "PS".

All models showed the highest performance on grass courts, likely because 67% of the matches were from Wimbledon, a Grand Slam tournament. We observed that Grand Slam matches were generally predicted with higher accuracy in the dataset (e.g., PS had an accuracy of 77% for Grand Slams compared to 66% for Masters 1000). This is partly due to the main draw matches in Grand Slams being played in a best-of-5 sets format, allowing stronger players more opportunities to demonstrate their superiority. The graph model also exhibits a highest overall Brier score compared to WElo, indicating weaker probability calibration despite its strong classification performance.

Table 1 also presents the outcomes of applying a betting strategy. Due to the model's weak probability calibration, we limit our bets to favorites, as determined by the model. We employ a modified Kelly stake size f^*, calculated as $f^* = \frac{\hat{p}(o-1)-(1-\hat{p})}{o-1}$ when the estimated win probability $\hat{p} > 0.5$, and $f^* = 0$ otherwise, where o denotes the decimal odds. This strategy reduces the impact of the model's calibration issues while still taking advantage of discrepancies between the model's predictions and market odds. Additionally, we

follow an approach by Boshnakov et al. (2017), by resetting the bankroll to 1 before each bet, ensuring that the final return on investment is unaffected by the sequence of bets. We assess risk-adjusted returns using the return on investment (ROI) along with the annualised Sharpe ratio, calculated as $S = (\bar{P}/\sigma_P) \times \sqrt{365.25}$, where $\bar{P}$ and σ_P are the sample mean and standard deviation of the total daily profits $\{P_d\}$, respectively.

The graph model's betting strategies consistently yield positive returns across all surfaces, unlike WElo. WElo's underperformance is expected, as it is a well-known incremental improvement on the highly popular Elo rating system by Elo & Sloan (1978). Our model produces the greatest returns on grass at 11.60% ROI, compared to 10.87% on hard courts and just 0.24% on clay. We applied a significance test, proposed by Wunderlich & Memmert (2020), to confirm our strategy's systematic profitability. By simulating 100,000 trials of 1074 random bets, we determined the probability (p_{bs}) that random wagering could match or exceed our observed ROI. Our Kelly strategy achieved a significant result ($p_{bs} = 0.012$ for 7.66% ROI).

In summary, despite a comparatively weaker Brier score, our model achieves consistent profitability by targeting mispriced bets in Pinnacle Sports odds, rather than maximizing predictive accuracy alone. Using the Kelly criterion to optimise bet sizes based on perceived "edge", our approach confirms the known divergence between statistical forecasting skill and effective betting profitability, as discussed by Wunderlich & Memmert (2020) and Hubáček & Šír (2023).

3.1 Live Testing During 2025 Clay Court Season

We published ex-ante predictions from our model for matches at the Monte Carlo Masters, Madrid Open, and Rome Masters, held in 2025.[3] In this live test, our model achieved an accuracy of 63.3% and a Brier score of 0.233. While bookmaker-implied probabilities demonstrated higher accuracy (67.2%) and a better Brier score (0.211), our model yielded a positive Kelly criterion ROI of 3.6%. For comparison, a Weighted Elo model attained an accuracy of 63.7%, a Brier score of 0.222, and an ROI of 2.2%.

4 Conclusion

In this paper, we have introduced the first application of graph neural networks to tennis match outcome prediction, using a novel temporal directed graph representation with informative edge weights. Our MagNet-based model achieved competitive classification accuracy, outperforming the WElo benchmark, though its probability calibration (Brier score) requires further improvement. Despite weaker calibration, the model demonstrated a notable ability to identify market inefficiencies. A modified Kelly staking strategy, focusing on favourites identified by our model, yielded statistically significant positive returns (7.66% ROI, $p_{bs} = 0.012$).

Our work contributes a new graph-based methodology to sports forecasting, showcasing the potential of GNNs to capture the relational and dynamic aspects of tennis. Key limitations remain, including the dependency on reliable historical odds data, as explored by Clegg & Cartlidge (2025), and performance variations across surfaces, particularly on hard courts. Future research could explore incorporating data from a wider range of tournaments to enrich the graph, balancing detail with computational tractability, and focus on enhancing model calibration. Overall, this study provides a foundational step, highlighting the promise of GCNs for advancing predictive analytics in tennis and other sports.

[3]For full live test predictions and results, see: https://github.com/Faxulous/tennisgnn_predictions

References

Angelini, G., Candila, V. & De Angelis, L. (2022), 'Weighted Elo rating for tennis match predictions', *European Journal of Operational Research* **297**(1), 120–132.

Arcagni, A., Candila, V. & Grassi, R. (2023), 'A new model for predicting the winner in tennis based on the eigenvector centrality', *Annals of Operations Research* **325**(1), 615–632.

Bayram, F., Garbarino, D. & Barla, A. (2021), Predicting tennis match outcomes with network analysis and machine learning, *in* T. Bureš, R. Dondi, J. Gamper, G. Guerrini, T. Jurdziński, C. Pahl, F. Sikora & P. W. Wong, eds, 'SOFSEM 2021: Theory and Practice of Computer Science', Springer International Publishing, Cham, p. 505–518.

Boshnakov, G., Kharrat, T. & McHale, I. G. (2017), 'A bivariate weibull count model for forecasting association football scores', *International Journal of Forecasting* **33**(2), 458–466.

Clegg, L. & Cartlidge, J. (2025), 'Not feeling the buzz: Correction study of mispricing and inefficiency in online sportsbooks', *International Journal of Forecasting* **41(2)**, 798–802.

Dingle, N., Knottenbelt, W. & Spanias, D. (2013), On the (page) ranking of professional tennis players, *in* M. Tribastone & S. Gilmore, eds, 'Computer Performance Engineering', Springer, Berlin, Heidelberg, p. 237–247.

Elo, A. E. & Sloan, S. (1978), *The rating of chessplayers: Past and present*, ARCO Publishing, New York, USA.

Fayomi, A., Majeed, R., Algarni, A., Akhtar, S., Jamal, F. & Nasir, J. A. (2022), 'Forecasting tennis match results using the bradley-terry model', *International Journal of Photoenergy* **2022**(1), 1898132.

He, Y., Gan, Q., Wipf, D., Reinert, G. D., Yan, J. & Cucuringu, M. (2022), GNNrank: Learning global rankings from pairwise comparisons via directed graph neural networks, *in* 'International Conference on Machine Learning', PMLR, pp. 8581–8612.
URL: *https://proceedings.mlr.press/v162/he22b/he22b.pdf*

He, Y., Zhang, X., Huang, J., Rozemberczki, B., Cucuringu, M. & Reinert, G. (2024), PyTorch geometric signed directed: A software package on graph neural networks for signed and directed graphs, *in* 'Learning on Graphs Conference (LoG)', PMLR.
URL: *https://proceedings.mlr.press/v231/he24a/he24a.pdf*

Hubáček, O. & Šír, G. (2023), 'Beating the market with a bad predictive model', *International Journal of Forecasting* **39**(2), 691–719.

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

Kovalchik, S. A. (2016), 'Searching for the GOAT of tennis win prediction', *Journal of Quantitative Analysis in Sports* **12**(3), 127–138.

Leitner, C., Zeileis, A. & Hornik, K. (2009), 'Is Federer stronger in a tournament without Nadal? An evaluation of odds and seedings for Wimbledon 2009', *Austrian Journal of Statistics* **38**(4), 277–286.

Mirzaei, A. (2022), Sports match outcome prediction with graph representation learning, Master's thesis, School of Computing Science, Simon Fraser University, CA, USA.
URL: *https://summit.sfu.ca/_flysystem/fedora/2022-08/input_data/22492/etd21919.pdf*

Morris, B. & Bialik, C. (2015), 'Serena williams and the difference between all-time great and greatest of all time', FiveThirtyEight. Accessed: 2025-04-03.
URL: *http://fivethirtyeight.com/features/serena-williams-and-the-difference-between-all-time-great-and-greatest-of-all-time/*

Wunderlich, F. & Memmert, D. (2020), 'Are betting returns a useful measure of accuracy in (sports) forecasting?', *International Journal of Forecasting* **36**(2), 713–722.

Xenopoulos, P. & Silva, C. (2021), Graph neural networks to predict sports outcomes, *in* '2021 IEEE International Conference on Big Data (Big Data)', IEEE, pp. 1757–1763.

Zhang, X., He, Y., Brugnone, N., Perlmutter, M. & Hirn, M. (2021), MagNet: A neural network for directed graphs, *in* 'International Conference on Neural Information Processing Systems', pp. 27003–27015.
URL: *https://dl.acm.org/doi/10.5555/3540261.3542329*

Prediction-based evaluation of back-four defense with spatial control in soccer

Soujanya Dash[1], Kenjiro Ide[1], Rikuhei Umemoto[1], Kai Amino[1], Keisuke Fujii[1]

[1] Graduate School of Informatics, Nagoya University, Nagoya, Aichi, Japan.
{dash.soujanya, ide.kenjiro, umemoto.rikuhei, amino.kai, fujii}@g.sp.m.is.nagoya-u.ac.jp

Abstract

Defensive strategies in soccer are crucial to preventing goal scoring opportunities and maintaining team structure. The defensive line (e.g., back four or back three) plays a vital role in these strategies. Despite its importance, evaluating the contribution of defensive line configurations remains an area of active research. This study hypothesizes that collective actions of the defensive line significantly contribute to a team's defensive success by maintaining defensive compactness. To test this hypothesis, we propose novel defensive indicators based on the predictive evaluation approach, including rule-based spatial control, defensive compactness, and pressure indices, handcrafted using event and tracking data. Rule-based spatial control penalizes defenders when attackers are near the penalty box and rewards the defenders positioned closest to the on-ball player. Statistical analysis reveals that rule-based spatial control served as a significant indicator for distinguishing defensive success and failure ($p < 0.05$), while defensive compactness did not have a significant impact in determining defensive success or failure ($p > 0.05$). These findings challenge conventional assumptions about compactness and emphasize the importance of spatial control.

1 Introduction

Soccer is a dynamic sport in which defensive transitions play a crucial role in shaping match outcomes. During a negative transition—when a team loses possession—the defensive objective shifts to either regaining control quickly or minimizing the opponent's advancement into dangerous zones. The last line of defense, typically consisting of the four outfield players closest to the goalkeeper, is responsible for reorganizing the team structure, limiting space for attackers, and preventing goal-scoring opportunities. Although commonly deployed in formations such as 4-4-2 and 4-3-3, the direct influence of this defensive line on transition success has received limited quantitative attention.

Recent work in sports analytics has highlighted the importance of spatial organization and collective defensive behavior in transition phases. Prior studies have explored defensive recovery patterns [3, 4], spatial influence surfaces [2], and structural breakdowns preceding goals [7]. Meanwhile, a growing body of research applies machine learning and spatial modeling to assess defensive effectiveness and zone control [8, 10, 9]. These approaches emphasize the significance of not only positional arrangement but also the dynamic interaction between defenders and attackers within contextually critical regions.

To address remaining gaps, this study introduces a set of handcrafted spatial metrics to quantify the collective behavior of the last defensive line during negative transitions. The proposed metrics include (i) *defensive compactness*, measuring cohesion among defenders, (ii) *pressure indices*, quantifying localized marking pressure, and (iii) *rule-based spatial control*, which penalizes defender inactivity near critical zones and rewards proximity to the on-ball attacker. These features are computed using synchronized tracking and event data from professional matches and evaluated for their ability to discriminate between successful and failed defensive sequences. We hypothesize that our three novel spatial metrics—Defensive compactness, Pressure Index, and rule-based space score—will effectively discriminate between successful and failed defensive sequences during negative transitions. By proposing a spatially grounded evaluation framework, this work contributes to a deeper understanding of back-line coordination and its role in shaping transition outcomes.

2 Methodology

2.1 Dataset

This study investigates the role of defensive line configurations during negative transitions in elite football using multimodal data from the 2023–24 LaLiga season. We analyze around 10 matches featuring RC Celta de Vigo, a team selected for its consistent use of a flat 4-4-2 formation, providing a stable tactical structure for analysis. The dataset combines high-resolution StatsBomb event data—widely used in academic and professional contexts for actions such as passes, tackles, and pressures [11, 5]—with SkillCorner tracking data, which captures continuous player and ball positions from broadcast video. Despite being vision-based, SkillCorner has proven effective in elite-level studies for modeling defensive shapes [1]. This integration enables frame-level alignment of spatio-temporal positioning and tactical events.

2.2 Preprocessing and Synchronization

2.2.1 Event-to-Tracking Synchronization

Synchronizing event and tracking data is challenging due to mismatches in timestamp resolution and recording offsets. Events often fall between discrete 25 Hz tracking frames, causing temporal misalignment and missing player positions at critical moments.

To address this, we used synchronization tools from the OpenSTARLab framework [12, 6], aligning each event to the nearest tracking frame without interpolation. A ± 1 second buffer around each transition ensured contextual completeness. We excluded frames missing full positional data for all 23 agents (22 players + ball), enabling reliable frame-level analysis.

2.2.2 Negative Transition Sequence Extraction

Negative transitions—defensive reorganizations following possession loss—were extracted in four steps. First, we filtered core events (passes, interceptions, tackles, clearances, and entries into danger zones). Second, we defined 5–10 event windows centered around each turnover, ending when the attacking team entered the penalty area or final third. Third, tactical filters were applied, requiring at least four defenders in the defensive third, attacker progression toward goal (via velocity), and confirmed entry into high-risk zones.

Finally, we excluded sequences with unclear possession, half-switch transitions, or missing agent data at the turnover frame.

This process yielded approximately 120 high-quality negative transitions from 10 matches. Each sequence is stored as a frame-indexed tensor:

$$(\mathbf{x}_t^{\text{pos}},\ \mathbf{v}_t,\ \text{possession\_label}_t,\ \text{team\_role}_t,\ \text{zone\_tag}_t) \quad \forall t\,,$$

where $\mathbf{x}_t^{\text{pos}}$ and $\mathbf{v}_t$ represent the 2D positions and velocities at time t.

2.3 Feature Engineering

After synchronization and filtering, we labeled each five-event sequence as either a defensive success (no entry into danger zones, shot attempt, or goal) or a defensive failure (any of those actions occurred). To explain these outcomes, we designed three rule-based spatial indicators that capture core back-four defensive principles:

1. Defensive Compactness. Quantifies the cohesion of the back-four relative to the nearest attackers:

$$\text{Compactness} = \lambda\, S_D + (1-\lambda)\, P_{DA},$$

where S_D is the convex-hull area of the four defenders and P_{DA} is the average distance from each of the three closest attackers to any defender. Lower values indicate a tighter, more coordinated line. $\lambda \in [0,1]$ is the weighing parameter. We set the weighing parameter $\lambda = 0.5$ to give equal importance to both defender compactness (SD) and attacker proximity (PDA), ensuring a balanced contribution from spatial tightness and attacker suppression. This choice reflects our hypothesis that both internal cohesion and external containment are equally vital during defensive transitions. Future work may explore optimizing λ via data-driven methods such as cross-validation or grid search.

2. Pressure Index. Counts how many attackers are under immediate pressure. Let $\mathscr{A} = \{a_1, a_2, a_3\}$ be the set of the three attackers closest to the last defensive line, and $\mathscr{D} = \{d_1, d_2, d_3, d_4\}$ be the set of the four last-line defenders (excluding the goalkeeper). For each attacker $a_i \in \mathscr{A}$, we compute the distance to the nearest defender $d_j \in \mathscr{D}$. If this distance is less than a fixed threshold $\delta = 3$ meters, we consider the attacker to be under immediate pressure. The Pressure Index is then defined as:

$$\text{Pressure Index} = \sum_{a_i \in \mathscr{A}} \mathbf{1}\left(\min_{d_j \in \mathscr{D}} \|d_j - a_i\| < \delta\right)$$

where $\mathbf{1}$ is an indicator function. This metric counts the number of attackers currently marked within a 3-meter radius by the last-line defenders. Based on empirical testing, we fixed the number of attackers to 3 and defenders to 4. This choice reflects the typical structure of back-four defensive lines and prioritizes evaluating pressure on the most immediate attacking threats during negative transitions.

3. Space Score. Space Score is our main contribution: this metric evaluates how the back-four defenders control four tactically critical zones. For each frame t, we identify the four last-line defenders (excluding the goalkeeper) and the three nearest attackers to the defensive line, including the on-ball player. Based on the zones that is occupied—central final third, penalty box proximity, wing pockets, and the 3 m ball-carrier radius—we calculate a weighted zone control score using:

$$C_z(t) = \frac{D_z(t) - A_z(t)}{D_z(t) + A_z(t) + \varepsilon}, \quad S_t = \sum_{z \in Z} w_z C_z(t),$$

where $D_z(t)$ and $A_z(t)$ are the number of defenders and attackers in zone z, and w_z is the tactical weight of zone z. A defender receives a high score when effectively marking an attacker or denying access to high-risk areas. Conversely, if an attacker is present in a critical zone without defensive coverage, the score decreases. At each frame, we compute the average of S_t across the four defenders, and then take the mean across the entire sequence to obtain the final Space Score:

$$\overline{S} = \frac{1}{T} \sum_{t=1}^{T} S_t.$$

To quantify the spatial importance of defensive presence, we define four spatial zones with fixed weights, prioritizing them according to their tactical risk.

The most critical area is the **Central Final Third**, which spans the last 35 meters of pitch length and the central 30 meters of width which includes key central attacking corridors and is weighted highest at 0.35. Next is the **Penalty Box Proximity**, defined as a 5-meter buffer surrounding the penalty area (16.5 × 40.32 m), capturing near-box congestion and is weighted at 0.30. The **Wing Pockets** occupy the outer 10-meter-wide lanes in the final 25 meters of the pitch which is known for wide attacks and low crosses, and are assigned a weight of 0.20. To avoid overlap with the penalty area, the Wing Pocket zones are shaped as six-sided polygons that taper inward near the box. Finally, a **Ball Carrier Radius** of 3 meters is defined around the ball location and is weighted at 0.15. It only contributes when it lies outside more critical zones. In cases where zones overlap, we retain only the maximum weight at each location. This ensures that a defender standing in overlapping zones is credited for occupying the most tactically dangerous space, rather than accumulating multiple scores.

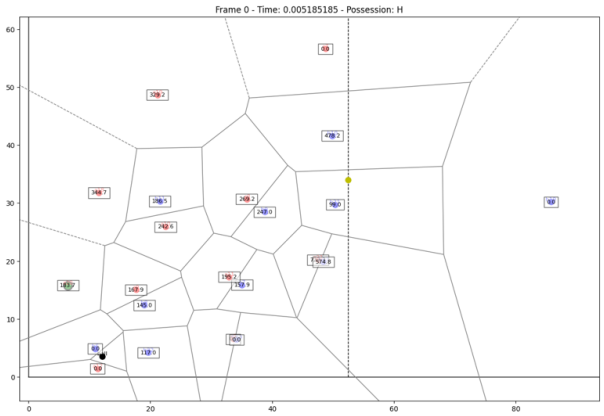


Figure 1: Illustration of Space Score computation.

2.4 Statistical Analysis

To assess whether our handcrafted features could discriminate between defensive success and failure, we conducted independent two-sample Welch's t-tests (unequal variances) for Defensive Compactness, Pressure Index, and Space Score. The null hypothesis (H_0) for each test was that there is no significant difference in the metric between successful and failed sequences, while the alternative (H_1) posited a significant difference. Prior to testing, we verified normality assumptions and, where violated, employed nonparametric Mann–Whitney U tests. We set the significance threshold at $\alpha = 0.05$, considering features with $p < \alpha$ as effective discriminators of defensive performance. Statistically significant results imply that the corresponding metric distinguishes between effective and ineffective defensive responses during negative transitions.

3 Results

We compared Defensive Compactness, Pressure Index, and Space Score between successful and failed defensive sequences using independent two-sample Welch's t-tests. Table 1 summarizes the test statistics.

Table 1: Welch's t-test results for each defensive metric

Metric	***t*-statistic**	***p*-value**
Defensive Compactness	0.085	0.934
Pressure Index (3 m radius)	0.503	0.621
Rule-based Space Score	4.599	**0.00035**

Among the three features evaluated, the Space Score showed a highly significant difference between successful and failed defensive sequences ($p \ll 0.05$), supporting its effectiveness in capturing context-aware defensive behavior. In contrast, Defensive Compactness did not yield a significant difference ($p = 0.93$), suggesting that line tightness alone does not reliably predict defensive outcomes during transitions. Similarly, the Pressure Index showed no significant effect ($p = 0.62$), indicating that simple proximity-based measures are insufficient without accounting for spatial and tactical context.

These results support our hypothesis that context-weighted spatial control is a key determinant of defensive success. The Space Score's design—penalizing unguarded incursions into high-risk zones and rewarding effective coverage—captures positional discipline more effectively than compactness or proximity-based metrics. Qualitative feedback from coaches confirmed that high Space Scores aligned with organized zone denial, while low scores reflected breakdowns in defensive structure. This emphasizes that the quality of spatial coverage, not just the number or tightness of defenders, plays a decisive role during negative transitions.

4 Conclusion

In conclusion, this study presents a spatial metric framework for evaluating the last defensive line in soccer. Among the three proposed novel indicators—Defensive compactness, Pressure Index, and the rule-based Space Score—only the Space Score showed a statistically significant difference between successful and failed defensive sequences. This highlights that while all three metrics introduce new spatial formulations,

only the zone-based control mechanism effectively captures context-aware defensive organization. These findings emphasize the tactical value of space-oriented defense and suggest directions for future metric refinement and predictive modeling. Future research should apply these metrics across broader contexts and integrate outcome prediction frameworks.

Acknowledgments

This study is supported by the JSPS KAKENHI Grant Number 23H03282. The author would gratefully acknowledge the support and feedback of a highschool soccer team coach, whose expertise guided the design and interpretation of the Space Score metric.

References

[1] M. Bassek, R. Rein, H. Weber, et al. An integrated dataset of spatiotemporal and event data in elite soccer. *Scientific Data*, 12:195, 2025.

[2] I. I. Bojinov and L. Bornn. The pressing game: Optimal defensive disruption in soccer. In *Proceedings of the MIT Sloan Sports Analytics Conference*, Cambridge, MA, 2016. MIT Sloan School of Management.

[3] C. A. Casal, M. Á. Andujar, J. L. Losada, T. Ardá, and R. Maneiro. Identification of defensive performance factors in the 2010 fifa world cup south africa. *Sports*, 4(4):54, 2016.

[4] C. A. Casal-Sanjurjo, M. Á. Andujar, A. Ardá, R. Maneiro, A. Rial, and J. L. Losada. Multivariate analysis of defensive phase in football: Identification of successful behavior patterns of 2014 brazil fifa world cup. *Journal of Human Sport and Exercise*, 16(3):503–516, 2021.

[5] E. e. a. Kassens-Noor. World cup soccer and ai: A match made in heaven. *arXiv preprint arXiv:2204.02313*, 2022.

[6] Y. Ogawa, R. Umemoto, and K. Fujii. Space evaluation at the starting point of soccer transitions. *arXiv preprint arXiv:2505.14711*, 2025.

[7] A. Tenga, I. Holme, L. T. Ronglan, and R. Bahr. Effect of playing tactics on goal scoring in norwegian professional soccer. *Journal of Sports Sciences*, 28(3):237–244, 2010.

[8] K. Toda, M. Teranishi, K. Kushiro, and K. Fujii. Evaluation of soccer team defense based on prediction models of ball recovery and being attacked. *PLoS One*, 17(1):e0263051, 2022.

[9] R. Umemoto and K. Fujii. Evaluation of team defense positioning by computing counterfactuals using statsbomb 360 data. In *StatsBomb Conference*, 2023.

[10] R. Umemoto, K. Tsutsui, and K. Fujii. Location analysis of players in uefa euro 2020 and 2022 using generalized valuation of defense by estimating probabilities. *arXiv preprint arXiv:2212.00021*, 2022.

[11] C. Yeung, R. Bunker, and K. Fujii. Unveiling multi-agent strategies: A data-driven approach for extracting and evaluating team tactics from football event and freeze-frame data. *Journal of Robotics and Mechatronics*, 36(3):603–617, 2024.

[12] C. Yeung, K. Ide, T. Someya, and K. Fujii. Openstarlab: Open approach for spatio-temporal agent data analysis in soccer. *arXiv preprint arXiv:2502.02785*, 2025.

Team Dynamics and Home Continent Advantage: Europe's Dominance in the Ryder Cup

Justin Ehrlich*, Hunter Geise, Collin Kneiss, and Charlotte Howland

*Syracuse University, Syracuse, New York email address: jaehrlic@syr.edu

Abstract

This study examines team dynamics in the Ryder Cup by addressing three questions: (1) whether teams exhibit a fixed-effect advantage where the whole outperforms the sum of individual parts, (2) whether players consistently over- or underperform relative to OWGR rankings, and (3) whether home-field advantage plays a significant role. The Ryder Cup, as a biennial U.S. vs. Europe event, offers a unique chance to evaluate the interplay of team ability, individual skill, and environmental context.

A new measure, "world golf ability," defined as the reciprocal of OWGR, was used to weight top players more heavily. Team ability was based on the median of this measure to limit outlier influence. Linear and GAM models were used to test relationships between team strength, location, and performance.

Results show a sizable team-level advantage for Europe—estimated at 2.94 points, even after accounting for individual ability and home advantage—suggesting stronger cohesion or preparation. No consistent pattern of players over- or underperforming relative to OWGR was found. A home-field edge of 2.04 points was also identified, likely driven by course familiarity, crowd support, and reduced travel strain. Overall, the findings highlight that team structure and leadership can meaningfully influence competitive outcomes.

1 Introduction

This study focuses on three key questions about performance in the Ryder Cup. First, it looks at whether either the Americans or Europeans have a cohesive, team-level advantage where their overall performance exceeds what would be expected based on their individual players. Second, it examines whether either team plays above or below their expected ability based on the Official World Golf Rankings (OWGR), helping to identify whether there are any meaningful differences between individual and team performance. Finally, it considers whether there is a home advantage in the Ryder Cup and whether playing on home soil gives either Team Europe or Team USA a significant edge. By addressing these questions, the study aims to better understand the factors that drive success in this unique team competition.

2 Background Literature

Sprengel (2022) suggests that as golfers learn in their early years, the types of courses they learn on affect how they perform on different types of courses. As Europeans grow and learn on their home

courses, they would tend to play better on these same courses compared to American courses due to their familiarity. The same can be thought for American golfers playing on American courses.. This forms the basis of Sprengel's explanation for the Ryder Cup's home-course advantage, where the home team had won 68% of the time at the time of his study. Using several models, including a logistic regression predicting match outcomes, Sprengel estimated that Team USA's probability of winning increased from 57.5% to 75.8% when playing on a U.S. course.

Nevill et al (1999) examineshow Europeans perform in home majors (the Open) compared to the other majors within the US. The authors found that throughout these tournaments, there was no real difference in performance when accounting for world ranking when the golfers performed in the US or in Europe. The authors' only true findingwas that those granted special permission (sponsor entrance, etc.) were shown to perform significantly better than their ranking at a home course, primarily in the US Open.

Using round-level data across a spectrum of handicaps, O'Brien (2024) finds that at all levels, golfers perform more consistently at their home courses than at courses they are unfamiliar with. Scratch golfers' differential in score increases by 2 (3.9 to 5.9) when playing an unfamiliar course whereas a 15-handicap golfer had their differential increase by even more (18.4 to 20.8).

Using Arccos data, Heath (2023) finds that golfer's playing outside of their home course for the first time had::

- 39% chance at shooting 2 strokes gained lesser or worse
- 55% chance at shooting 1 strokes gained lesser
- 11% chance at shooting 2 strokes gained or better
- 19% chance at shooting 1 strokes gained better

The OWGR system, conceived by Mark McCormack in collaboration with The Royal and Ancient Golf Club of St Andrews, was introduced during the 1986 Masters. This system was designed to enhance the selection of players globally, as international players began to challenge the notion that PGA Tour players were the pinnacle of talent. Prior to the OWGR's launch, the PGA Tour Money list was the primary method for ranking the top players. Since its inception, the OWGR ranking has become crucial in evaluating players for consideration in the Ryder Cup, an international tournament. This global ranking system assigns points to players based on their finishes in eligible tournaments over a rolling two-year period, with more recent results weighted more heavily. Each tournament has a "strength of field" rating that determines the number of points available, and a player's average points per event is calculated using a divisor (minimum 40 events, maximum 52). These rankings are updated weekly and offer a standardized way to compare players from different tours and regions (OWGR, 2025). While OWGR is not the only factor used for Ryder Cup selection, it is often a helpful benchmark for determining which players are in strong form heading into the event.

The players selected for both teams are a combination of automatic qualifiers and captain picks. The automatic qualifiers for the U.S. team are the top American players in a Ryder Cup-specific points system, which is based on earnings in PGA Tour events during a set qualification window, including bonus weight for major performances. The number of automatic qualifiers changes each year (Ritter, 2018);during the 2023 Ryder CupTeam USA had six automatically qualified members. The remaining six players are selected by the U.S. team captain, who often considers recent performance, course fit, and team chemistry when making final picks (PGA Tour, 2023).

On the European side, qualification up through 2023 involved a dual points list — the European Points List and the World Points List — with three players qualifying from each list in 2023. The

remaining six were captain's picks (PGA Tour, 2023), though, similar to Team USA, this number can vary from year to year. (Note: For 2025, Europe has moved to a single points list, but this change falls outside the scope of our data.)

Captains for both teams are chosen well in advance of the competition and often reflect not just career success but also leadership and Ryder Cup experience. For Team USA, the PGA of America appoints the captain, usually a respected veteran with Ryder Cup history either as a player or vice-captain. Similarly, the DP World Tour (formerly European Tour) selects the European captain, typically someone who has represented Europe multiple times and is well-regarded in the locker room. These captains are announced roughly two years before the event and are responsible for course scouting, roster decisions, pairing strategies, and overall team culture (Colgan, 2021).

3 Methodology

Data on Ryder Cup results, including participant names and their Official World Golf Rankings (OWGR), were obtained from each tournament's Wikipedia entry ("2023 Ryder Cup," 2024). We included only those tournaments for which complete OWGR data were available for all participants, resulting in a final sample covering the 1987 through 2023 Ryder Cups. A summary table for our data is presented in table 1.

Characteristic	**United States** N = 18[1]	**Europe** N = 18[1]
year		
Median (Min, Max)	2,005 (1,987, 2,023)	2,005 (1,987, 2,023)
host		
Europe	9 (50%)	9 (50%)
United States	9 (50%)	9 (50%)
points		
Median (Min, Max)	13.50 (9.50, 19.00)	14.50 (9.00, 18.50)
wg_ability_mean		
Median (Min, Max)	0.12 (0.06, 0.20)	0.10 (0.04, 0.19)
point_difference		
Median (Min, Max)	-1.0 (-9.0, 10.0)	1.0 (-10.0, 9.0)
wg_ability_median_difference		
Median (Min, Max)	0.04 (-0.01, 0.09)	-0.04 (-0.09, 0.01)
home	9 (50%)	9 (50%)
[1] n (%)		

Table 1: Summary Statistics

To assess each team's ability, we developed a metric called 'world golf ability,' calculated as the median reciprocal rank. The reciprocal rank (1/rank) assigns greater weight to top performers. Using the median helps mitigate the influence of outliers, reflecting that the Ryder Cup is primarily a team event where players at the extremes of the rankings have a limited impact on overall performance.

The wg_ability_median_difference is compared with point_difference in figure 1. Each team outcome is color-coded based on whether it was a home or away tournament, and the teams are differentiated using shapes. A linear best-fit line was also rendered showing each team at home or away. The vertical difference between the home and away for each team represents home-field advantage. The vertical disjoint difference between the same home and away lines between Team Europe and Team United States demonstrates any advantage or disadvantage that a team has overall, controlling for performance differences. Were there no difference, both team points should fit on the same home or

away line; however, there is an obvious drop in performance for team US, even though the wg_ability_median_difference is typically higher.

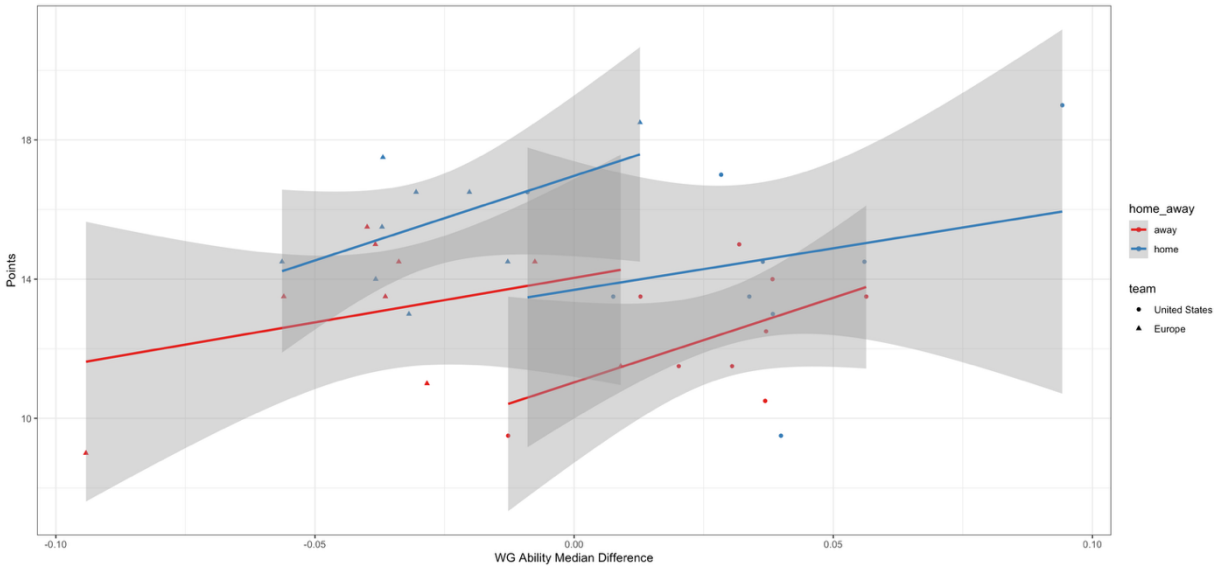


Figure 1: Home Field and Team Advantage

To understand how strong these relationships are, and to understand the marginal effect of home field advantage and any specific team advantage, a series of linear models were estimated. The general formula is shown in formula 1:

$$points=\beta0+\beta1\cdot wg\_ability\_median\_difference+\beta2\cdot(team:wg\_ability\_median\_difference)+\beta3\cdot team + \beta4\cdot home\_away +\epsilon. \quad (1)$$

Where: β0 is the intercept, β1, β2, β3, and β4 are the coefficients for the respective terms, and ϵ is the error term. Since there are two observations per tournament, we used robust standard errors clustered at the year level to account for within-year correlation.

4 Results

In table 2, the coefficients of the estimated linear models from formula 1 are shown. Model 1 is the base model, without differentiating the teams in any way, but does include home_away fixed effects. In this model, we can see a statistically significant home field advantage of 2.31 points. Model 2 adds an interaction between WG Ability Median Difference and Team, which results a coefficient that is not statistically significant. Model 3 adds team fixed effects, and finds that Team Europe has a significant advantage of 2.94 points. There is no reason to interact team with HFA, as HFA is by definition a team's home score minus their away score, and in a two-team league, this would be symmetrical for the other team and so there will be no difference between the two teams.

	Model 1			Model 2			Model 3		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	12.76	11.90 – 13.62	**<0.01**	12.80	11.90 – 13.70	**<0.01**	11.47	9.82 – 13.11	**<0.01**
wg ability median difference	2.16	-27.73 – 32.06	0.88	0.92	-31.55 – 33.39	0.95	31.16	-16.37 – 78.68	0.19
home away [home]	2.31	0.47 – 4.16	**0.02**	2.31	0.46 – 4.16	**0.02**	2.04	0.19 – 3.89	**0.03**
wg ability median difference × teamEurope				2.48	-4.08 – 9.04	0.45	2.48	-4.09 – 9.05	0.45
team [Europe]							2.94	-0.61 – 6.48	**0.10**
Observations	36			36			36		
R^2 / R^2 adjusted	0.241 / 0.195			0.241 / 0.170			0.374 / 0.293		

Table 2: Points Linear Models with Clustered Standard Errors

5 Discussion

The first question investigated whether Team Europe or Team USA possesses a cohesive team-level advantage—one in which the collective performance of the team significantly exceeds the sum of its individual contributions. Model 3 in Table 2 provides compelling evidence that Team Europe holds a distinct advantage, with an estimated 2.94-point edge over Team USA, ceteris paribus. This result highlights Europe's ability to consistently leverage team dynamics or strategies that lead to superior collective performance compared to their competitors.

The second question explored whether either team plays above (or below) their expected level based on the Official World Golf Rankings (OWGR). Using ranking differences to predict points, Model 2 in Table 2 reveals no significant difference in the slope between Team Europe and Team USA. This finding suggests that neither team consistently outperforms nor underperforms relative to individual player abilities as measured by OWGR, reinforcing the notion that overall outcomes are not dictated by deviations in individual performance.

The final question examined whether there is a home advantage in the Ryder Cup. Model 3 estimates that the home team scores 2.04 more points than when playing away. This implies that switching from an away to a home venue results in a shift from -2.04 to +2.04 in point differential, holding all else constant—a total swing of 4.08 points.

5 Conclusion

In conclusion, these findings suggest that Team Europe benefits from a cohesive, team-level advantage of 2.94 points over Team USA. While a team-level effect is clearly present, we do not find any evidence that either team consistently outperforms the other based on player ability, as measured by our novel world golf ability metric derived from OWGR. Additionally, home advantage was shown to offer a 2.04-point benefit to the host team, implying a 4.08-point swing in point differential when shifting from away to home.

These results highlight the importance of team cohesion, preparation, and strategy—factors that appear to give Europe a sustained edge even after accounting for player ability and location. No consistent pattern of individual over- or underperformance relative to OWGR rankings was found, reinforcing that outcomes are shaped more by collective factors than by isolated player differences.

Overall, the findings underscore the meaningful influence of team structure, leadership, and contextual elements on Ryder Cup outcomes. Future research could explore the mechanisms behind Europe's team-level advantage and further examine the sources of home-field benefit in international golf competition.

References

[1] Colgan, J. (2021) 'How are Ryder Cup captains decided? Inside the selection process', *Golf*, 26 September. Available at: https://golf.com/news/ryder-cup-captains-selection-2021/ (Accessed: 28 May 2025).

[2] Heath, E. (2023) *Do You Play Better At New Courses Or Your Home Club? What The Stats Say...*, *Golf Monthly Magazine*. Available at: https://www.golfmonthly.com/features/do-you-play-better-at-new-courses-or-your-home-club-what-the-stats-say (Accessed: 30 January 2025).

[3] Nevill, A.M. and Holder, R.L. (1999) 'Home Advantage in Sport: An Overview of Studies on the Advantage of Playing at Home', *Sports Medicine*, 28(4), pp. 221–236. Available at: https://doi.org/10.2165/00007256-199928040-00001.

[4] O'Brien, S. (2024) *Is there home field advantage in golf?*, *The Grint*. Available at: https://thegrint.com/range/post/is-there-home-field-advantage-in-golf (Accessed: 30 January 2025).

[5] OWGR (2025) *Official World Golf Ranking - Ranking Explained*, *OWGR*. Available at: https://www.owgr.com/how-the-ranking-works (Accessed: 28 May 2025).

[6] PGA Tour (2023) *How it works: Ryder Cup qualification*, *PGATour.com*. Available at: https://www.pgatour.com/article/news/latest/2023/07/03/how-it-works-ryder-cup-qualification-us-team-europe-marco-simone-rome-italy (Accessed: 15 April 2025).

[7] Sprengel, B. (2022) 'Golf's Fiercest Tournament: Estimating the Impact of Home Course Advantage in the Ryder Cup', *The Park Place Economist*, 29(1). Available at: https://digitalcommons.iwu.edu/parkplace/vol29/iss1/11.

Predicting the probability of breaking a world record

G. Fonseca*, F. Giummolè**, M. Lambardi di San Miniato*† and V. Mameli*

*Department of Economics and Statistics, University of Udine, Udine, Italy
**Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Venice, Italy
† email address: michele.lambardi@uniud.it

Abstract

Statistical analysis may help answer some intriguing questions in athletics, such as when the current world records will be improved. Sport records are extreme observations, which can be analyzed through extreme value theory. However, modeling is only one part of the problem, since estimation is also troubled by small sample issues. Here, we present some improved estimates of the expected time to break the record. The property needed for this task is probabilistic calibration. Bootstrap-based approaches can help assess and recover this property to improve predictions. We show that, thanks to improved estimates, the near future is richer in new records than suggested by the classical estimates.

1 Introduction

Although athletics is an ancient heritage, the practice of keeping records is relatively recent. In modern times, hype surrounds not only the athletes themselves, but also the ultimate limits of humankind. In contrast to this, it may seem that athletes have reached records that are hard to break, to the point that we may not witness any better performances in our lifetime. In this study, we aim to show that such predictions may be overly pessimistic, mainly due to the limitations of the modeling approaches commonly employed for this task. Several approaches exist to address these issues, as recently reviewed by [4]. By examining these methods, we argue that more valid predictive distributions, with heavier tails than those typically used in these cases, allow for greater predictive potential for future records.

Here, we analyze men's and women's annual records of several disciplines in athletics, such as sprint running and high jump. These data can be viewed as block maxima and analyzed by assuming a suitable extreme value distribution (EVD), as done in [1]. The model includes some unknown constants, called parameters, which make the model flexible enough to capture the truth. Nonetheless, a value for those parameters must be chosen (suitably) to make the model usable for prediction. The classical estimative approach requires replacing the parameters with a single estimate. The Bayesian approach implies using all possible values, weighted by a posterior distribution, so that the estimation uncertainty can be weighed in. However, the latter approach requires a prior distribution, which can significantly affect the analysis in small samples. Here we resort to the method outlined by [2], that allows to incorporate the uncertainty within the classical estimative framework via bootstrap.

Thus, we present improved estimates of the expected time it takes to break the world records and compare them with the Bayesian and classical estimates. Although current world records look hard to break, they should take less time than previously estimated by classical approaches.

2 Methodology

Let Y denote a random variable of interest, such as the annual world record in a given discipline. In year t, this variable is denoted by Y_t. In general, Y has an unknown distribution, which can be characterized by its cumulative distribution function (CDF) $F_0(\cdot)$ or its inverse, the quantile function (QF) $Q_0(\cdot)$. Typically, a parametric model would be assumed, indexed by a d-dimensional parameter vector $\theta = (\theta_1, \ldots, \theta_d)$, with generic CDF and QF denoted by $F(\cdot;\theta)$ and $Q(\cdot;\theta)$, respectively. Under appropriate conditions, extreme value theory suggests that annual best performances follow a Generalized Extreme Value (GEV) distribution. Based on insights from preliminary data analysis, in this work, we focus on a special case of the GEV family, known as EVD or Gumbel distribution, and defined as

$$F(q;\theta) = \exp(-\exp(-(q-\mu)/\sigma)), \quad Q(p;\theta) = \mu - \sigma\log(-\log p), \quad \theta = (\mu,\sigma), \quad \sigma > 0.$$

The density function $f(\cdot;\theta)$ is a regular model. The model should be flexible enough to contain the ground truth θ_0, such that $F(\cdot;\theta_0) = F_0(\cdot)$; however, to make prediction, one must resolve the parameter, obtaining a CDF $\hat{F}(\cdot)$ and a QF $\hat{Q}(\cdot)$ that are free from θ.

Provided some suitable smoothness assumptions, it looks reasonable to predict Y via

$$\hat{F}(\cdot) = F(\cdot;\hat{\theta}), \quad \hat{Q}(\cdot) = \hat{Q}(\cdot;\hat{\theta}), \tag{1}$$

for some consistent estimator $\hat{\theta} = \hat{\theta}(y)$, based on past data $y = (y_1, \ldots, y_n)$. This approach is known as the estimative method. The most efficient version of this approach uses the maximum likelihood estimator $\hat{\theta} = \arg\max_\theta \log L(\theta;y)$, where $L(\theta;y)$ is the likelihood function. With large samples, $(\hat{F},\hat{Q})$ should converge to (F_0,Q_0), on a point-by-point basis; however, in finite samples, the estimative approach can misbehave significantly: its bias and variance can be considerable.

Besides unbiasedness and efficiency, calibration is also a desideratum when dealing with probabilistic prediction. In particular, $\hat{F}$ produces calibrated probabilities if

$$\mathbb{E}\left[Q_0(\hat{F}(q))\right] = q,$$

while $\hat{Q}$ yields calibrated quantiles if

$$\mathbb{E}\left[F_0(\hat{Q}(p))\right] = p,$$

for all $q \in \mathbb{R}$ and $p \in]0,1[$. So, a calibrated CDF works as the inverse of the true QF, on average, whereas a calibrated QF works as the inverse of the true CDF, on average.

Both F_0 and Q_0 are calibrated but unavailable; the estimative approach can achieve the same behaviour asymptotically, but it may fail severely if n is small. Due to misbehaviour in finite samples, so-called miscalibration issues may occur. In contrast, the Bayesian approach naturally accounts for uncertainty by specifying a prior distribution $\pi(\theta)$, which is then combined with the likelihood to yield the posterior distribution, $\pi(\theta \mid y) \propto \pi(\theta)\, L(\theta;y)$. This posterior directly leads to the predictive distribution, which serves as the natural basis for making probabilistic forecasts:

$$\hat{F}_B(\cdot) = \int F(\cdot;\theta)\pi(\theta \mid y)\,\mathrm{d}\theta\,. \tag{2}$$

Unfortunately, this approach relies on choosing a prior distribution, which can significantly affect the results in small samples. For objectivity, we use an uninformative prior of the fiducial kind, which is naturally defined as $\pi(\theta) = 1/\sigma$ for location-and-scale models [3].

One may need to resort to the frequentist framework, then the sampling distributions of $\hat{F}$ and $\hat{Q}$ must be analyzed to correct miscalibration issues. Parametric bootstrap can be helpful in this assessment. Specifically, one can simulate a large number N of scenarios under the assumption that $\theta = \hat{\theta}$. In the generic s-th scenario, the dataset $y^s = (y_1^s, \ldots, y_n^s)$ is generated from $f(\cdot; \hat{\theta})$. From these synthetic datasets, one obtains N parametric bootstrap estimates, the generic estimate $\hat{\theta}^s = \hat{\theta}(y^s)$ being the result of the same estimation procedure as $\hat{\theta}$ but applied to the dataset y^s. After [2], the estimative approach can be improved by using calibrated probabilities (CP), obtained via the calibrated CDF $\hat{F}_{CP} = \hat{Q}_{CP}^{-1}$, with

$$\hat{Q}_{CP}(p) = \frac{1}{N} \sum_{s=1}^{N} Q\left(F\left(Q\left(p; \hat{\theta}\right); \hat{\theta}^s\right); \hat{\theta}\right), \tag{3}$$

and calibrated quantiles (CQ), obtained via the calibrated QF $\hat{Q}_{CQ} = \hat{F}_{CQ}^{-1}$, with

$$\hat{F}_{CQ}(q) = \frac{1}{N} \sum_{s=1}^{N} F\left(Q\left(F\left(q; \hat{\theta}\right); \hat{\theta}^s\right); \hat{\theta}\right). \tag{4}$$

As a remark, the two enhanced CDFs $\hat{F}_{CP}$ and $\hat{F}_{CQ}$ are distinct in general, see for instance Figure 1, which illustrates the case when $n = 10$ and $N = 10^4$.

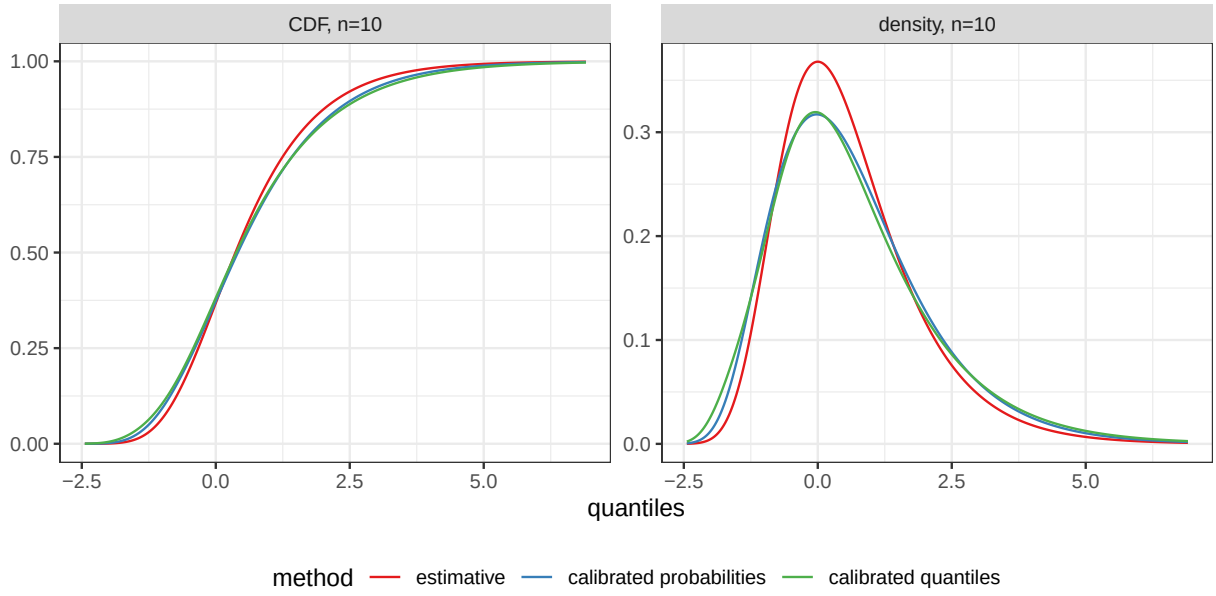


Figure 1: Predictive distributions: estimative and calibrated ones for the case $n = 10$, $\hat{\mu} = 0$, $\hat{\sigma} = 1$, and $N = 10^4$. The difference between estimative and calibrated CDFs decays approximately as $1/n$, see [2].

As a consequence, one cannot simultaneously calibrate both probabilities and quantiles. This result holds more generally, even with other approximately calibrated predictions, essentially because probabilities and quantiles are non-linearly related, making it difficult to simultaneously preserve calibration properties for both. However, in some cases, the aim of the analysis is delimited enough that only one of the two predictive distributions, probability- or quantile-based, is relevant for the purpose at hand. Our focus is on the expected time to break some world record w, so the parameter of interest is defined as

$$\psi = \psi(\theta) = \frac{1}{1 - F(w;\theta)}\,. \tag{5}$$

The estimative approach would produce the estimate $\hat{\psi} = \psi(\hat{\theta})$, which obeys the invariance principle, so it is the same as plugging the estimate (1); the Bayesian approach naturally yields the posterior mean estimate $\hat{\psi}_B = \int \psi(\theta)\pi(\theta \mid y)\,\mathrm{d}\theta$, analogously to (2); the calibrated probabilities approach, instead, would replace the unknown CDF with its enhanced estimate based on (3), and one can also consider (4) for a comparison, yielding estimates denoted by $\hat{\psi}_{CP}$ and $\hat{\psi}_{CQ}$, respectively.

3 Analysis

Our motivating example is the analysis of sports records in athletics. Although several disciplines exist in modern times, we focus on the long-standing ones, such as sprint running, hurdles, high jump, long jump, and javelin throw, which are reported below. We limit the analysis to men's and women's world records for analogous reasons. However, some changes in rules, technologies, and techniques have occurred over time, making only recent data relevant to current predictions: in particular, we consider only data from 2001 to 2024, as shown in Figure 2. These data form the training data vector y to estimate the parameter of interest in (5). The world record to be broken, denoted by w in the equation, is also available and has occurred before 2001 for many of the disciplines considered. The estimates are always stratified according to both disciplines and gender. Annual records can be retrieved from several online sources; however, the most reliable data can be freely accessed from the World Athletics website. This platform allows individual performances to be tracked at all internationally recognized competitions. Care must be taken with recent performances, as it can take time to validate them, and some violations are confirmed only the following year.

The original outcomes are either times, like in sprints and hurdles, or distances, like in jumps and javelin throws. We convert times into speeds, so all the outcomes are "the higher the better", to align with EV analysis. Thus, the annual records can be assumed to be well described by the EVD model.

For each discipline and gender, the EVD model was used in the classical, Bayesian, probability and quantile-calibrated fashions to estimate the time to break the record, as defined in Equation (5). In terms of sprint runs, the estimates are also stratified according to distances since the progressions of their records are essentially unrelated across them, although the groups of athletes can overlap. Estimates of the time it takes to break the record are reported in Figure 3. Many records look hard to break, as in the case of hurdles and long jumps, whereas we may expect some new records (relatively) soon in the case of sprints and on several distances.

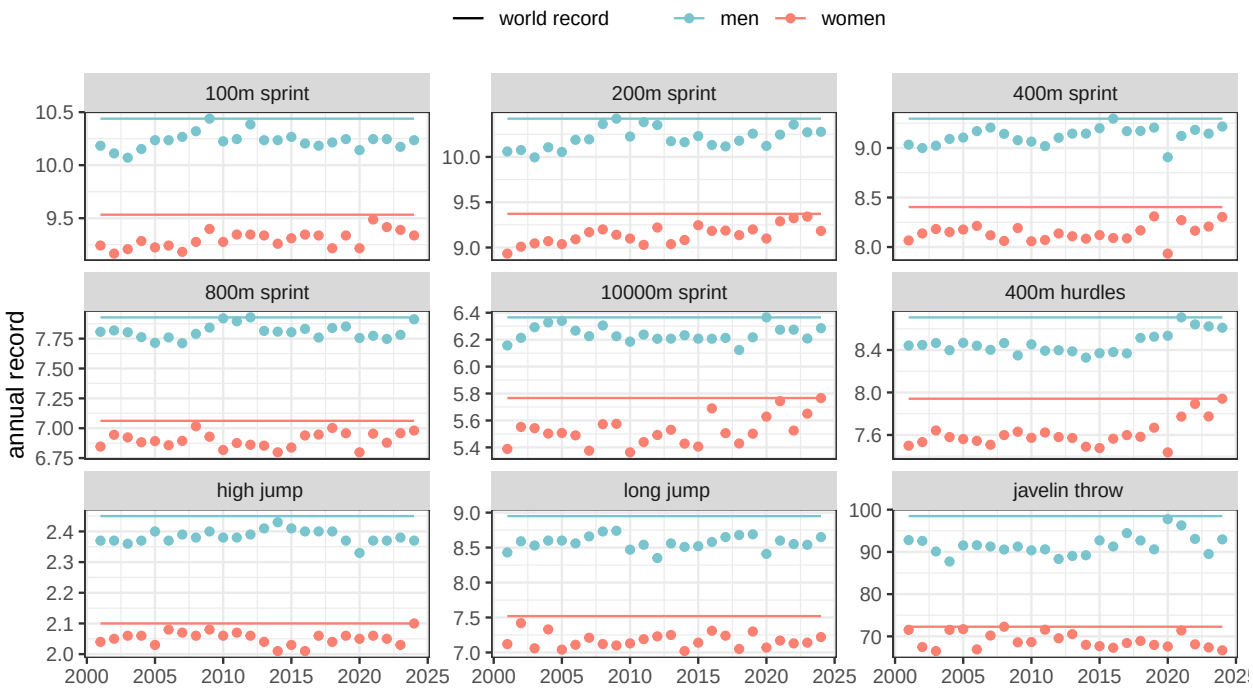


Figure 2: Annual and world records for both genders in the considered athletic disciplines.

4 Future work

Calibrated quantiles and probabilities can improve sports record analysis, when interest often lies in exceedance probabilities and return values. These approaches require only a slight modification of the classical estimative approach via simulation. Depending on the specific predictive task under investigation, this approach can yield more reliable assessments in the context of record performances. Although not shown here, the same rationale could be extended to records from aquatic disciplines, such as freestyle and butterfly.

Moreover, we analyze data as block maxima, but a more informative approach could be the peaks-over-threshold, which would use not just annual records but also individual performance data [1]. However, this approach would require additionally tuning a hyperparameter to improve the underlying generalized Pareto approximation, which is problematic to incorporate into a bootstrap simulation approach.

Although enhanced in different respects, the two improved estimates seem to provide similar predictions, at least for the target chosen in (5). This result must be related to the fact that both predictive distributions are evaluated only in their tails, which are both heavier compared to that of the estimative distribution. The result is encouraging, since it implies that new records will soon be available. However, the recent emergence of new disciplines makes this assessment only partially meaningful.

The proposed approach relies on explicit CDF $F(\cdot;\theta)$ and QF $Q(\cdot;\theta)$, but these may be mathematically intractable. If replicates are available, as they should be for the bootstrap approach to be viable, one may

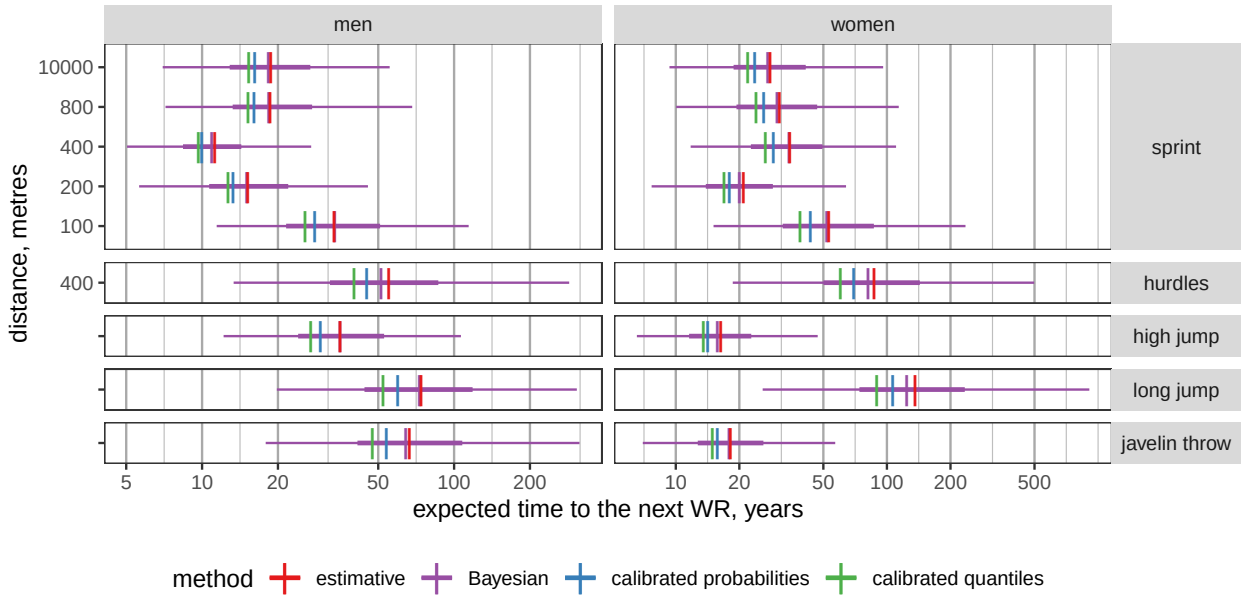


Figure 3: Estimative and calibrated predictions for the time to break the world record, for a few selected athletic disciplines. The 50% and 95% credible intervals are also reported for the Bayesian prediction as thick and thin segments, respectively.

approximate CFD and QF via kernel density applied to such replicates. This simplification may help spread the proposed technique as more complicated models are naturally supported. For instance, it would be natural to complement the EVD distribution with a copula model to include any potential serial correlation. Non-parametric extensions of the bootstrap component of the proposal would also be interesting.

Acknowledgments

This research is funded by PRIN 2022: Project prot. n. 2022R74PLE UGOV code PRIN_2022_MAMELI_DIES CUP G53D23001870006 funded by the European Union NextGenerationEU M4C2 inv 1.1.

References

[1] Einmahl, J.H.J. and Magnus J.R. (2008) *Records in athletics through extreme-value theory.* Journal of the American Statistical Association **103**, 1382–1391.

[2] Fonseca, G., Giummolè, F. and Vidoni, P. (2024) *Optimal prediction for quantiles and probabilities.* Statistical Papers **66**, 24.

[3] Hannig, J., Iyer, H., Lai, R.C.S., Lee, T.C.M. (2016) *Generalized fiducial inference: A review and new results.* Journal of the American Statistical Association **111**(515), 1346-1361.

[4] Tian, Q., Nordman, D.J., Meeker, W.Q. (2022) *Methods to compute prediction intervals: A review and new results.* Statistical Science **37**(4), 580-597.

Round-Robin Tournament Scheduling Under Total Game Attractiveness Objective

U. Güler* and T. Atan** and D. Günneç***

*Maastricht University, ugur.guler@maastrichtuniversity.nl
**Bahçeşehir University, sabritankut.atan@bau.edu.tr
***Özyeğin University, dilek.gunnec@ozyegin.edu.tr

Abstract

Tournament competitiveness plays a critical role in shaping the associated economy, influencing match attendance, viewership, merchandise sales, and related factors. Among various measures that can help increase tournament competitiveness, scheduling offers a cost-effective way for this purpose. Designing a tournament schedule with competitiveness in mind can significantly enhance a tournament's appeal. In this study, we present a new metric, the competitive difference, to measure this appeal and propose a mathematical model tailored for round-robin tournaments. While our numerical experiments involve single round-robin tournaments, the approach can be extended to multiple round-robin tournaments as well.

1 Introduction

The attractiveness of a sports league is a key consideration in tournament design, heavily influenced by the competitive balance of the league. *Competitive balance* refers to how evenly teams are matched. A league where teams show significant variation in playing strength is considered to have low competitive balance (or a higher degree of imbalance), whereas leagues with more evenly matched teams are seen as having higher competitive balance. Competitive balance is important because it directly impacts the uncertainty of match outcomes and the overall championship. Unpredictability in sporting events is generally thought of as enhancing spectator enjoyment.

We consider two types of uncertainty: match-level uncertainty, which refers to the unpredictability of individual match results, and seasonal uncertainty, which relates to the unpredictability of final standings and outcomes throughout the season. Leagues with a higher competitive balance typically maintain greater levels of both, thus keeping fan interest and engagement high. This is demonstrated, for example, by [18] for match-level uncertainty and by [22] and [1] for the seasonal uncertainty. Scheduling offers a straightforward solution for increasing tournament attractiveness without any modification of the tournament rules. With this in mind, we investigate a new approach to scheduling round-robin tournaments to increase the uncertainty of the outcome for each match thereby also enhancing seasonal uncertainty. To achieve this, we develop a mathematical model that schedules a single round-robin tournament (SRR), where every team plays against every other team exactly once, ensuring that teams of similar strengths face each other as often as possible.

In the remainder, we provide a brief literature review, present a mathematical model and give preliminary results.

2 Literature Review

Various criteria enhance the attractiveness of sports schedules, with key metrics including quality, competitive intensity, suspense, and tension. [7] reviewed research on tournament design, including attractiveness-related work.

Quality emphasizes strong team matchups. [17] analyzed balance in Dutch football, considering home advantage. [9] improved Chilean league schedules by prioritizing high-quality matches early. [13] adjusted Belgian league schedules to boost broadcast revenue, while [20] scheduled prime-time matches under fairness constraints.

Competitive intensity measures match balance. [15, 16] developed axioms for knockout tournaments, and [5] integrated quality and intensity to optimize seeding. [8] used a Swiss system model, introducing *unattractiveness* as the inverse of intensity, with the Colley rating method for dynamic updates.

Suspense, the incentive for teams to compete until the end, is widely used. [19] modeled Brazilian league playoffs to determine qualification certainty. [14] classified Belgian clubs by objectives (e.g., championship, relegation) and used simulations to analyze competitiveness. [11] linked decisive matches to attractiveness, while [10] highlighted irrelevant matches in large round-robin tournaments. [2, 3, 4] examined UEFA tournament incentives based on seedings.

Tension, distinct from suspense, refers to uncertainty in the final outcome. [12] defined lower bounds for decisive rounds and found that strong teams facing off in critical rounds heightens tension, which decreases with more draws.

3 Mathematical Model

Team strength can fluctuate significantly during the season. In this work, we introduce a new metric called the *competitive difference*, which incorporates the team strength variations based on the round in which teams compete. To calculate team strengths, we use the conventional point rating system, as it is common in traditional football leagues. The following definitions explain the notation and terminology that will be used later.

Definition 1 (Point Distribution). A *point distribution* is a tuple (p_w, p_t, p_l) where p_w represents points awarded for a win, p_t represents points awarded for a tie, p_l represents points received for a loss.

Definition 2 (Result Matrix). A *result matrix* P is an $n \times n$ matrix where n is the number of teams with each entry $p_{i,j}$ representing the points scored by team i in the match between teams i and j in an SRR tournament. Since a team cannot play against itself, the diagonal entries will be 0.

An example of a result matrix for a league with four teams and point distribution $(p_w, p_t, p_l) = (3,1,0)$ is given below:

$$\begin{bmatrix} 0 & 1 & 3 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 3 & 0 & 1 \\ 3 & 1 & 1 & 0 \end{bmatrix}$$

Definition 3 (Competitive Difference). The *competitive difference* $d_{i,j,r}$ is defined as the absolute difference in the ratings of two teams i and j just before their match in round r. That is,

$$d_{i,j,r} = |s_{i,r-1} - s_{j,r-1}| \tag{1}$$

where $s_{i,r-1}$ and $s_{j,r-1}$ are the points accumulated by team i and j, respectively, at the end of round $r-1$.

The competitive difference decreases when two teams with similar ratings compete. In other words, the closer the value $d_{i,j,r}$ (representing competitive difference) is to zero, the more evenly matched —and therefore competitive— the game is. Such matches are typically intense, as both teams are highly motivated to win, knowing that the outcome could have an immediate impact on their standings. While traditional rivalries can add to the intensity, the primary driver is the potential for a change in rankings. Ultimately, the overarching goal for any team in a tournament is to climb as high as possible in the rankings, with the ideal aim of securing the top spot.

In this study, we introduce a model designed to minimize the total competitive difference in an SRR tournament, with potential extensions to multiple round-robin formats. Although tournaments can involve additional considerations —such as home and away games— our focus is strictly on evaluating the proposed competitive difference metric and exploring its implications. Specifically, for a match between teams i and j, the competitive difference is calculated based on the points each team has accumulated up to round r. The remainder of this section presents the mathematical model, beginning with the relevant notation.

Sets

Let $i, j \in T = \{1, ..., n\}$ be the set of teams and $r \in R = \{1, ..., n-1\}$ be the set of rounds.

Parameters

P is the $n \times n$ result matrix of an SRR tournament. It is denoted as:

$$P = \begin{bmatrix} 0 & p_{1,2} & \dots & p_{1,n} \\ p_{2,1} & 0 & \dots & p_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n,1} & p_{n,2} & \dots & 0 \end{bmatrix}$$

Decision Variables

The binary variables $x_{i,j,r}$ are defined as:

$$x_{i,j,r} = \begin{cases} 1, \textit{if team i and team j have a match at round r} \\ 0, \textit{otherwise.} \end{cases}$$

To get rid of the absolute value in (1), we introduce two nonnegative variables $d^+_{i,j,r}$ and $d^-_{i,j,r}$ such that

$$d_{i,j,r} = d^+_{i,j,r} + d^-_{i,j,r}$$

and

$$d^+_{i,j,r} - d^-_{i,j,r} = s_{i,r-1} - s_{j,r-1.}$$

Here, $d^+_{i,j,r}$ and $d^-_{i,j,r}$ split the original absolute value into two parts: one for the positive deviation and one for the negative deviation. Next, we give a mixed-integer nonlinear mathematical model (MINLP) that minimizes the total competitive difference value.

$$\textbf{MINLP}: \quad \min \sum_{i\in T} \sum_{\substack{j\in T \\ i<j}} \sum_{r\in R} (d^+_{i,j,r} + d^-_{i,j,r}) \cdot x_{i,j,r} \tag{2}$$

s.t.

$$\sum_{r\in R} x_{i,j,r} = 1 \qquad \forall i,j \in T : i < j \tag{3}$$

$$\sum_{\substack{j\in T \\ j>i}} x_{i,j,r} + \sum_{\substack{j\in T \\ j<i}} x_{j,i,r} = 1 \qquad \forall i \in T, r \in R \tag{4}$$

$$d^+_{i,j,r} - d^-_{i,j,r} = \sum_{\substack{k\in T \\ k>i \\ k\neq j}} \sum_{w=1}^{r-1} p_{i,k} \cdot x_{i,k,w} + \sum_{\substack{k\in T \\ k<i \\ i\neq j}} \sum_{w=1}^{r-1} p_{i,k} \cdot x_{k,i,w} - \sum_{\substack{k\in T \\ k>j \\ k\neq i}} \sum_{w=1}^{r-1} p_{j,k} \cdot x_{j,k,w} - \sum_{\substack{k\in T \\ k<j \\ i\neq j}} \sum_{w=1}^{r-1} p_{j,k} \cdot x_{k,j,w} \qquad \forall i,j \in T : i < j, r \in R \tag{5}$$

$$d^+_{i,j,r}, d^-_{i,j,r} \geq 0, \quad x_{i,j,r} \in \{0,1\} \qquad \forall i,j \in T : i < j, r \in R \tag{6}$$

The objective function minimizes the total competitive difference in the tournament. Constraints (3) ensure that each pair of teams meets once and Constraints (4) enforce that each team plays one match in each round. These are the hard constraints of the SRR tournament. In constraint set (5), the expression $d^+_{i,j,r} - d^-_{i,j,r}$ stores the difference between the sums of points that the teams i and j have collected from previous matches that have occurred before round r, i.e. the competitive difference between teams i and j in round r should they play. Constraints (6) determine the domain of the variables.

The non-linear objective function can be linearized; we utilized the resulting linear model, MILP, in our numerical experiments.

4 Numerical Experiments

All computations were performed on an 11^{th} Gen Intel Core i7, 3.00 GHz processor with 16GB RAM and 8 cores. The MILP model was implemented in Python using Pyomo and solved with Gurobi 11.0.0.

We compared the best objective function values obtained by solving the MILP under a time limit with two heuristic approaches, using randomly generated result matrices. Win, loss, and draw probabilities were set at 38%, 38%, and 24%, respectively, reflecting the related average rates of the Big Five European leagues'

Table 1: Comparison of MILP model's results with results of other scheduling algorithms.

n	Canonical	Vizing	MILP	MILP-Canonical (%)	MILP-Vizing (%)
4	6.0	4.8*	4.8*	25.0	0.0
6	24.8	15.6	15.5*	60.0	0.6
8	58.8	36.4	26.8*	119.4	35.8
10	130	66	51	155.0	29.4
12	200	106	77	160.0	37.7
14	304	151	104	192.3	45.2
16	353	249	159	122.0	56.6
18	746	506	313	138.3	61.7
20	644	640	562	14.6	13.9

in 2022-2023. The MILP performance was evaluated with schedules generated via the canonical algorithm [6] and Vizing's algorithm [21], both implemented in Python.

Benchmark schedules were created by generating 10,000 schedules using Vizing's algorithm, shuffling each 10 times (yielding 100,000 total), and similarly shuffling the canonical schedule 100,000 times. Objective function values were computed for each schedule using the same result matrix, retaining the minimum value per algorithm.

Table 1 presents average objective function values from 100 result matrices for problems with $n \leq 20$ under a time limit of 10 hours. Gurobi found an optimal solution pretty quickly when $n \leq 8$. Results marked with an asterisk are optimal solutions. The MILP solution under a time limit significantly outperforms other heuristic results. For $n \leq 6$, Vizing's algorithm performs better than the canonical algorithm due to its ability to explore a larger solution space within 100,000 schedules. Due to the increased problem size, results reported by Gurobi after 10 hours suffer in quality when $n = 20$.

References

[1] Francesco Addesa and Alexander John Bond. Determinants of stadium attendance in Italian Serie A: New evidence based on fan expectations. *PLoS one*, 16(12):e0261419, 2021.

[2] László Csató. The UEFA Champions League seeding is not strategy-proof since the 2015/16 season. *Annals of Operations Research*, 292(1):161–169, 2020.

[3] László Csató. How to avoid uncompetitive games? The importance of tie-breaking rules. *European Journal of Operational Research*, 307(3):1260–1269, 2023.

[4] László Csató, Roland Molontay, and József Pintér. Tournament schedules and incentives in a double round-robin tournament with four teams. *International Transactions in Operational Research*, 31(3):1486–1514, 2024.

[5] Dmitry Dagaev and Alex Suzdaltsev. Competitive intensity and quality maximizing seedings in knockout tournaments. *Journal of Combinatorial Optimization*, 35:170–188, 2018.

[6] Dominique De Werra. Scheduling in sports. *Studies on graphs and discrete programming*, 11:381–395, 1981.

[7] Karel Devriesere, László Csató, and Dries Goossens. Tournament design: A review from an operational research perspective. *European Journal of Operational Research*, 324(1):1–21, 2025.

[8] Zhi-Long Dong, Celso C Ribeiro, Fengmin Xu, Ailec Zamora, Yujie Ma, and Kui Jing. Dynamic scheduling of e-sports tournaments. *Transportation Research Part E: Logistics and Transportation Review*, 169:102988, 2023.

[9] Guillermo Durán, Mario Guajardo, Jaime Miranda, Denis Sauré, Sebastián Souyris, Andres Weintraub, and Rodrigo Wolf. Scheduling the Chilean soccer league by integer programming. *Interfaces*, 37(6):539–552, 2007.

[10] Marco Faella and Luigi Sauro. Irrelevant matches in round-robin tournaments. *Autonomous Agents and Multi-Agent Systems*, 35:1–34, 2021.

[11] Gery Geenens. On the decisiveness of a game in a tournament. *European Journal of Operational Research*, 232(1):156–168, 2014.

[12] Bas Gieling. Tension in round robin competitions. *Bachelor thesis, Eindhoven University of Technology. URL: https://pure. tue. nl/ws/portalfiles/portal/197521679/Thesis_BTW_Gieling. pdf*, 2022.

[13] Dries Goossens and Frits Spieksma. Scheduling the Belgian soccer league. *Interfaces*, 39(2):109–118, 2009.

[14] Dries R Goossens, Jeroen Beliën, and Frits CR Spieksma. Comparing league formats with respect to match importance in Belgian football. *Annals of Operations Research*, 194:223–240, 2012.

[15] Alexander Karpov. A new knockout tournament seeding method and its axiomatic justification. *Operations Research Letters*, 44(6):706–711, 2016.

[16] Alexander Karpov. Generalized knockout tournament seedings. *International Journal of Computer Science in Sport*, 17(2):113–127, 2018.

[17] Ruud H Koning. Balance in competition in Dutch soccer. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(3):419–431, 2000.

[18] Tim Pawlowski and Georgios Nalbantis. Competition format, championship uncertainty and stadium attendance in European football–a small league perspective. *Applied Economics*, 47(38):4128–4139, 2015.

[19] Celso C Ribeiro and Sebastián Urrutia. An application of integer programming to playoff elimination in football championships. *International Transactions in Operational Research*, 12(4):375–386, 2005.

[20] Celso C Ribeiro and Sebastián Urrutia. Scheduling the Brazilian soccer tournament with fairness and broadcast objectives. In *Practice and Theory of Automated Timetabling VI: 6th International Conference, PATAT 2006 Brno, Czech Republic, August 30–September 1, 2006 Revised Selected Papers 6*, pages 147–157. Springer, 2007.

[21] Celso C Ribeiro, Sebastián Urrutia, and Dominique de Werra. A tutorial on graph models for scheduling round-robin sports tournaments. *International Transactions in Operational Research*, 30(6):3267–3295, 2023.

[22] Nicolas Scelles, Christophe Durand, Liliane Bonnal, Daniel Goyeau, and Wladimir Andreff. Do all sporting prizes have a significant positive impact on attendance in an European national football league? Competitive intensity in the French Ligue 1. *Ekonomicheskaya Politika/Economic Policy*, 11(3):82–107, 2016.

The Split: Analysing Contest Design in the Scottish Premier League

Jessica K. Hargreaves* and Johan M. Rewilak**

*Department of Mathematics, University of York, York, YO10 5DD, UK: jessica.hargreaves@york.ac.uk
** Department of Sport and Entertainment Management, University of South Carolina, South Carolina, USA.

Abstract

This paper examines whether the policy to split the Scottish Premier League (SPL) into two after 33 games for post-season play generated negative externalities. Using a regression discontinuity (RD) design, it tests whether the policy reduced attendance for teams finishing in the lower half of the standings. The analysis uses data from 23 seasons (2000/01 to 2023/24, excluding pandemic-impacted seasons) in which the league has operated under this structure. The results show that teams just below The Split experience lower attendances compared to those just above, driven by the lost opportunity to play against the "top" teams such as Celtic and Rangers. This implies the new structure harmed a subset of clubs. Furthermore, this work highlights how large market teams subsidise smaller teams in sports leagues.

1 Introduction

The Scottish Football League (SFL) is a professional football competition similar to other open European football leagues, featuring multiple divisions with promotion and relegation. The Scottish Premier League (SPL) is the top division and, in the 2000/01 season, the SPL expanded from 10 to 12 teams [6]. The addition of these teams had the potential to create fixture congestion as, traditionally, SPL teams played each other four times in a round-robin format, twice at home and twice away, with 36 fixtures played in total. With the new format, teams would have to complete 44 games.

To avoid fixture congestion and to make the total number of fixtures comparable to other top leagues in Europe, the SPL altered its tournament design. It "split" the season into two: the "Regular Season" and the "Play-offs". During the Regular Season, teams play one another three times for a total of 33 matches. Then the league is "split" in two, creating two mini-leagues. Teams who finish in the **top six** places after 33 games play one another (one more time) in the "**Championship Play-off**" and the **bottom six** teams play one another in the "**Relegation Play-off**". This leads to a total of 38 games being played by each team, the same number of matches as other European leagues [8].

Figure 1 shows home attendances as a proportion of stadium capacity for all 12 teams in the 2023/24 SPL season. Celtic, Rangers and, to a lesser extent, Hearts, display stable attendance, nearly selling out all matches. In contrast, the other teams show fluctuating attendances. To investigate this further, Figure 2 shows home attendances for two SPL teams in 2023/24 (Dundee and Aberdeen). Attendance fluctuations are driven by the opposition, with higher attendances when Celtic/ Rangers (or a historic rival) visit. In particular, matched attendances (Pre- and Post-Split) are very similar.

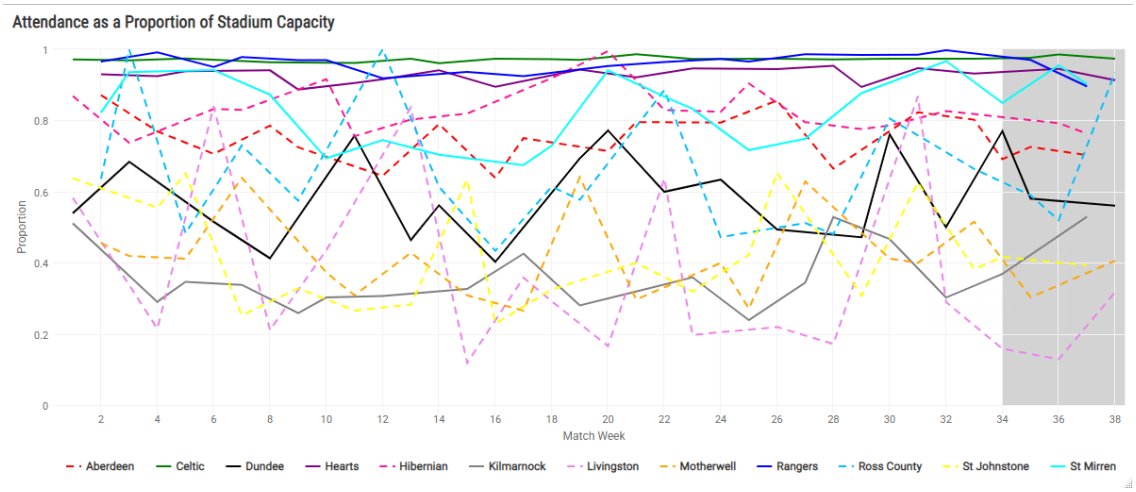


Figure 1: Attendances as a proportion of stadium capacity for all teams in the 2023/24 SPL Season. Solid lines indicate teams in the Championship Play-off. Dashed lines indicate teams in the Relegation Play-off.

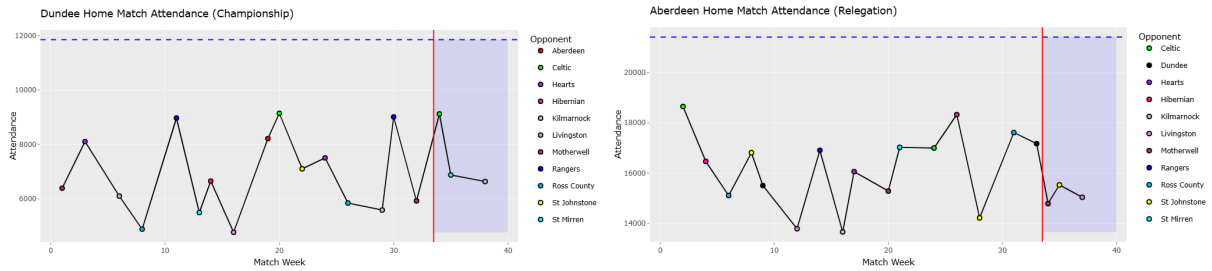


Figure 2: Home attendances for two example SPL Teams in the 2023/24 Season. The reported stadium capacity is shown as a horizontal dashed blue line. Blue background denotes play-off matches.

In this paper, we investigate whether The Split generated any negative externalities. Using a regression discontinuity (RD) design approach, we empirically test whether the "split" impacted teams above the cut-off more than those below in terms of (home) attendance and whether any home attendance differences may be explained via "superstar" effects.

2 Background and Motivation

In sporting contest design, to maintain sporting integrity, we require that participants utilise costly effort to achieve success [14]. Similarly, tournament designers often face multiple objectives when designing an optimal sporting contest and face difficult trade-offs that can create wrong incentives [4]. By using appropriate mechanisms – often financial – contest designers try to ensure that the competition is incentive compatible and teams take the correct actions [9].

The theory of superstars is well established [10, 11]. The superstar effect in sport has been extensively studied, with individual players or teams driving fan attendance [3, 13]. Since 1984-85, only Rangers and

Celtic have won the SPL title, making them dominant forces in the league. These “Old Firm" clubs drive fan interest, suggesting the league functions as a duopoly with a competitive fringe. As the Old Firm always play in the Championship segment of the SPL split, teams finishing in the top half will play an additional game against them. This could boost attendance and profits for those teams.

To investigate the impact of The Split, we adopt a Regression Discontinuity (RD) Design. The use of RD design in sports is increasing due to its ability to provide causal interpretations by estimating the local average treatment effect. It has been applied across a wide range of sports, including professional football, covering topics from contest design to on-field performance [9]. [7] and [8] use an RD design approach to investigate the impact of league design in the SPL on spectator attendance and club revenues.

3 Data and Methods

3.1 Data Description

We obtain data from the SPL and World Football websites[1]. The data spans from 2000/01 to 2023/24. Following [8], we exclude the 2019/20-2021/22 seasons, which were affected by restrictions due to the Covid-19 pandemic. Following other RD design studies in sport (e.g. [1], [9], [12]), we use **annual home attendance** data (i.e. one observation per club per season).

3.2 Methods

To comprehensively investigate the impact of The Split, we use Regression Discontinuity (RD) Design. For a detailed survey of the RD approach, we direct the reader to [5].

Following [7], the dependent variable is the natural logarithm of home match day attendance. We also construct three variations of this variable (see Section 4) to further develop the work by [7]. These include various changes in attendance Pre- and Post-Split, to capture like-for-like factors that vary solely due to the SPL split. These variables are normally distributed and may provide a better measure to examine the attendance effect of finishing in the Championship Play-off versus the Relegation Play-off.

The remaining data is manually constructed. The treatment variable is a dummy equal to one for teams finishing seventh to twelfth in a season, with teams finishing first to sixth coded as zero. In addition, the running variable is the position a team finishes Pre-Split, centred around zero, similar to other regression discontinuity studies [1, 9, 12].

An RD design should not require additional control variables but, in practice, the most relevant confounders are included [5]. This study includes team and season dummies. Club fixed effects are included as some teams have larger supporter bases than others and time fixed effects are used to capture season-wide shocks affecting all teams, such as Gretna’s liquidation and Rangers’ reformation as a new club.

Equation (1) outlines the RD design in its linear form. Subscript (i) represents individual teams and (t) indexes time. Club fixed effects are represented as (α_i) and period fixed effects (τ_t). The conditioning variables are shown in X.

$$Y_{i,t} = \alpha_i + \beta_1(Rank-7) + \beta_2(Rank-7) * Treat_{i,t} + \beta_3 X_{i,t} + \tau_t + \varepsilon_{i,t} \quad (1)$$

The RD design has three requirements [5]. In this context, they are as follows:

[1] urls: https://spfl.co.uk/league/premiership and https://www.worldfootball.net/

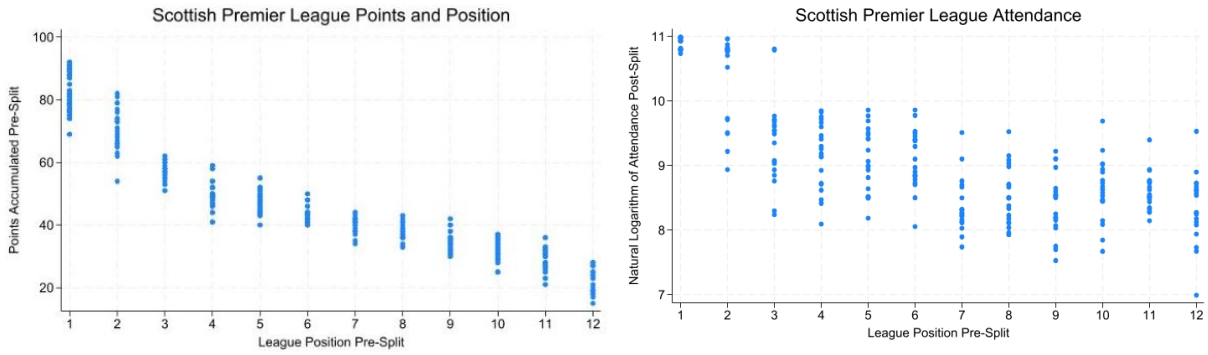


Figure 3: Points accumulated Pre-Split (**Left**) and Post-Split attendance (**Right**).

1. There is a threshold, with teams randomly assigned above and below this cut-off.
2. Teams on either side of the threshold are similar in characteristics, forming a good treatment and comparator group.
3. There is a significant jump in the dependent variable at the threshold.

After 33 games in the SPL, teams finishing seventh (sixth) or below (above) are placed in the lower (upper) half of The Split with 100% probability. Therefore, as teams finishing seventh and below are always treated, we observe a sharp RD design. Teams also have incomplete control over their allocation above or below the threshold, as they cannot influence other match results, referee errors or other random factors that may affect their position. Therefore, final allocations around the cut-off are random, satisfying condition 1.

As a team's final position is based on points from three rounds of round-robin matches, it is anticipated that teams are similar in terms of on-field performance around the cut-off. Indeed, the average points difference between teams finishing sixth and seventh is three- the number of points awarded for a win. Figure 3 shows the points accumulated Pre-Split for all teams in the league. Figure 3 shows a smooth curve for points accumulated Pre-Split, with no jump near the threshold, supporting the second assumption. However, Figure 3 also reveals a large jump between the top two and third place, indicating that Celtic and Rangers should be omitted from the analysis.

To examine Condition 3, a scatter plot of the data is presented in Figure 3, providing graphical evidence of a discontinuity between sixth and seventh place and thus supporting the econometric design. Furthermore, Figure 4 shows an RD plot, with a linear polynomial (Equation (1)), using three teams on either side of the threshold. The plot shows a clear "jump" in (log) attendance at the threshold.

4 Results

Firstly, we investigate the impact of The Split on home attendance using a local linear estimator and subsequently calculate the change in attendance following [2]. The results show that teams finishing below the cut-off face a statistically significant ($p < 0.01$) drop in attendance of approximately 30%, relative to those who finish above the threshold. This supports the findings of [7] who report a 24% drop in attendance for SPL teams in the Relegation Play-off compared to those in the Championship Play-off.

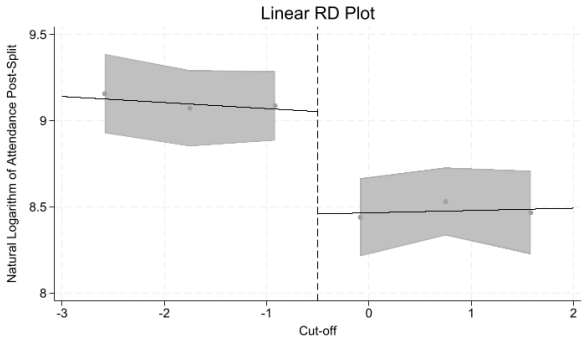


Figure 4: Regression discontinuity plot (linear regression) where the cut-off (zero) represents seventh place.

Y-Transform	Pre-Split - Post-Split	Last 5 Pre-Split - Post-Split	Matched By Pre/Post Split Fixture
Treatment	-1074.50*** (349.37)	-1154.90*** (439.05)	381.03 (257.09)
Club Fixed Effects	Yes	Yes	Yes
Time Dummies	Yes	Yes	Yes
Observations	126	126	126

Table 1: **Sensitivity Analysis.** Each column represents a separate regression. The dependent variable is the change in attendance Pre- and Post-Split (calculated in three ways). Standard errors are reported in parentheses and *** denotes statistical significance at the 1% level.

Furthermore, we conduct robustness tests using the change in attendance before and after The Split as the dependent variable. For all three tests, we focus on teams three places either side of the cut-off (positions 4-9), but alter how we calculate Pre-Split attendance. In the first test, we use all the data. However, in the second test we only use the last five home fixtures before The Split (subtracting the average Post-Split attendance from the average of the last five Pre-Split home games). This accounts for potential fan disengagement before The Split (if a team is destined to finish mid-table). Finally, we match any home fixtures against the teams played after The Split with the same respective fixtures before The Split (subtracting the average attendance). This overcomes issues of teams playing smaller sides before The Split and larger ones afterwards, as well as game-specific characteristics like derby matches that might influence the findings.

The results in Table 1 show that only in the final column, where fixtures Pre- and Post-Split are matched, is the treatment variable not statistically significant. Figure 2 illustrates this, highlighting that matches against the same opponent have similar attendance throughout the season. For all clubs, fixtures against Celtic/ Rangers (at any point in the season) attract the highest attendance. This suggests that a reason why attendance is lower for teams in the Relegation Play-off is that they miss out on these lucrative Old Firm fixtures.

5 Conclusion

In 2000, the SPL expanded from 10 to 12 teams. To avoid fixture congestion, it introduced a policy splitting the league into two halves after 33 matches, with teams in each half playing each other once more for a final

five matches. This study uses a regression discontinuity (RD) design and finds that “The Split" generated several externalities. In Section 4, we found that teams just finishing in the Relegation Play-off faced a 30% attendance drop compared to teams just qualifying for the Championship Play-off. However, when fixtures were matched Pre- and Post-Split, this negative effect disappeared. This suggests that the opposition, particularly the two superstar clubs Rangers and Celtic, drive these attendance differences.

References

[1] İ. Güner and M. Hamidi Sahneh. Dancing with the stars: Does playing in elite tournaments affect performance? *Oxford Bulletin of Economics and Statistics*, 85(1):1–34, 2023.

[2] R. Halvorsen and R. Palmquist. The interpretation of dummy variables in semilogarithmic equations. *American economic review*, 70(3), 1980.

[3] J. A. Hausman and G. K. Leonard. Superstars in the national basketball association: Economic value and policy. *Journal of Labor Economics*, 15(4):586–624, 1997.

[4] G. Kendall and L. J. Lenten. When sports rules go awry. *European Journal of Operational Research*, 257(2):377–394, 2017.

[5] D. S. Lee and T. Lemieux. Regression discontinuity designs in economics. *Journal of economic literature*, 48(2):281–355, 2010.

[6] L. J. Lenten. Unbalanced schedules and the estimation of competitive balance in the scottish premier league. *Scottish Journal of Political Economy*, 55(4):488–508, 2008.

[7] B. Reilly and R. Witt. The effect of league design on spectator attendance: A regression discontinuity design approach. *Journal of Sports Economics*, 22(5):514–545, 2021.

[8] B. Reilly and R. Witt. The effect of league design on club revenues in the scottish premier league. *Eastern Economic Journal*, 50(1):1–28, 2024.

[9] J. Rewilak. Dancing with the stars revisited: does dropping out of the champions league, into the europa league, impact domestic performance? *Managing Sport and Leisure*, pages 1–5, 2022.

[10] S. Rosen. The economics of superstars. *The American economic review*, 71(5):845–858, 1981.

[11] S. Rosen and A. Sanderson. Labour markets in professional sports. *The economic journal*, 111(469):47–68, 2001.

[12] J. D. Speer. The consequences of promotion and relegation in european soccer leagues: A regression discontinuity approach. *Sports Economics Review*, 1:100003, 2023.

[13] H. Sung and B. M. Mills. Estimation of game-level attendance in major league soccer: Outcome uncertainty and absolute quality considerations. *Sport Management Review*, 21(5):519–532, 2018.

[14] S. Szymanski. The economic design of sporting contests. *Journal of economic literature*, 41(4):1137–1187, 2003.

Optimization of the Tournament Format for the Nationwide High School Kyudo Competition in Japan

K. Hashimoto* and E. Konaka**

* Meijo University

** Meijo Univerisity. 1-501, Shiogamaguchi, Tempaku-ku, Nagoya, JAPAN. email address: konaka@meijo-u.ac.jp

Abstract

This study proposes an optimized format for the nationwide high school Kyudo tournament in Japan, addressing challenges in balancing fairness, educational value, and practical constraints. Kyudo's binary scoring system makes skill assessment difficult with limited attempts. Using historical data, we estimated participants' skill distributions and conducted simulations to evaluate tournament formats. The proposed format increases the preliminary attempts from 4 to 6 and removes semifinals, reducing standard deviation in total attempts while maintaining comparable ranking accuracy. This ensures fairness, sufficient opportunities for skill demonstration, and alignment with Kyudo's traditional and educational values, offering a robust framework for student-focused competitions.

1 Introduction

This study focuses on Kyudo, a target sport that has uniquely developed and been systematized as a competition in Japan, among target-type competitions that involve competing for shooting accuracy.

This study examines the National High School Kyudo Selection Tournament, a competition where high school students from all over Japan participate. Since it is a tournament for high school students (under 18), there is a wide variation in skill levels among participants, even though regional qualifiers are held.

Given the large number of participants and the limited number of attempts per person due to time constraints, the influence of luck on the competition results is relatively large. However, as it is a student-centered tournament, ensuring a minimum number of attempts for educational purposes is also essential. The objective of this study is to examine a tournament format that allows for enough experiences while also maintaining accuracy in skill assessment, even under the constraint of a limited total number of attempts.

Most previous studies in this sports research have focused on improving competitive skills or injury prevention, while there are few studies examining the appropriateness of tournament formats or scoring systems from a mathematical perspective. In archery, for example, the target diagram (Figure 1 right) with concentric circles scoring from 10 to 1 has remained largely unchanged since the 1930s [1]. In Kyudo, the score is a binary value indicating whether or not the target was hit, and the resolution for skil measurement of one shot is not sufficiently high. The target called *kasumi-mato* is shown in Figure 1 left. The target is colored by black and white bands, but the position where the arrow hits does not affect the score. However, there has been little academic discussion on whether this design is appropriate for skill quantification.

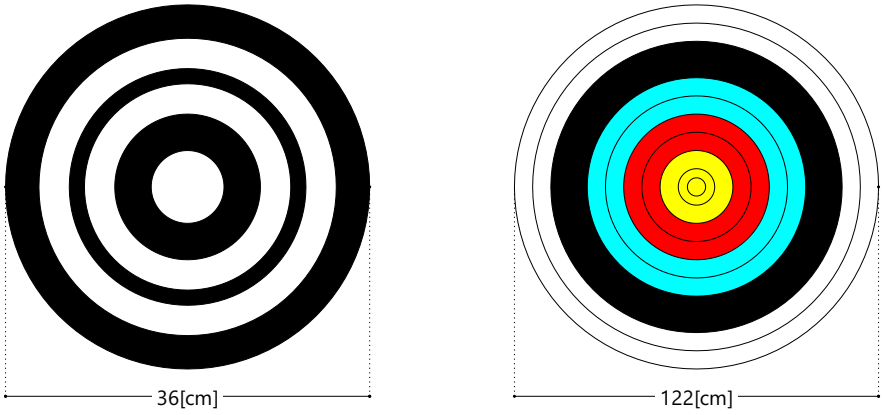


Figure 1: Targets. Left:*Kasumi mato*. *Mato* (target) used in *Kinteki* (short-range shooting) competition. Right:Target of outdoor target archery:

Since Kyudo has developed uniquely within Japan, international (English-written) studies on its tournament formats are extremely scarce. To the best of our knowledge, no prior research has addressed the tournament format of Kyudo from the perspectives of implementation cost and skill measurement performance, making this study novel.

The structure of this paper is organized as follows. Section 2 describes the composition of the data used and provides a basic analysis. Section 3 outlines the purpose of the analysis and the methods employed in this study. Specifically, it includes the estimation of participants' skill distribution based on past competition results (3.1), as well as the definition of evaluation functions related to skill estimation performance and implementation cost of the competition (3.2). Section 4 presents the analysis results and discussion. In conclusion, we discovered a new competition format that increases the minimum number of trials for each participant while maintaining comparable skill estimation performance and average total cost as the current format.

2 Data

The target competition for data collection is the All Japan High School Kyudo Selection Tournament. The data obtained from the web covers five editions and 3928 shots in total. For instance, the official website for the 41st tournament can be found at `https://kyudo-zenkoku.com/10-taikai/2021/senbatsu.html`

In the tournament, all participants first shoot four arrows as a preliminary round, and the number of hits is recorded. Participants who hit the target more than three times advance to the semifinals. The semifinals follow the same rule.At any stage, even if a participant hits the target three consecutive times, they still perform the fourth shot. Additionally, the number of hits from one stage does not carry over to the next. In the finals, unlike the preliminary and semifinal rounds, all remaining participants shoot one arrow at a time. If both hits and misses are present, the missed participants are eliminated, while those who hit the target advance to the next round. If all remaining participants miss, they are not eliminated and proceed to the next shot. This process continues until only one participant remains, who is declared the winner. This final

method is called "Izume" (shootout). In this paper, this tournament format is referred to as **"3/4–3/4–Izume"** and is labeled as the **"current format."** The purpose of this study is to investigate how the performance as a skill evaluation event changes when adopting a tournament format different from the current one. Additionally, the study aims to propose a better tournament format from multiple perspectives, including the guaranteed number of attempts per participant and the total number of attempts throughout the tournament.

Figure 2 shows the frequency distribution of the number of hits in the preliminaries and semifinals.

The average number of hits in the semifinals was $0.669 = 607/908$ for men and $0.625 = 455/728$ for women. Since the conditions for advancing to the semifinals are the same for both men and women (hitting at least 3 out of 4 arrows), it can be observed that the difference in skill level between men and women becomes smaller in the semifinals.

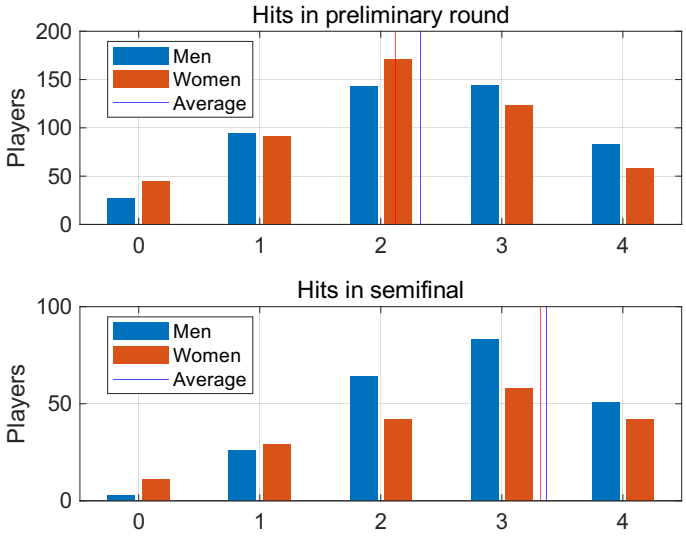


Figure 2: Hits in preliminaries and semifinals.

3 Objective of analysis and methods

The procedure of analysis in this paper is as follows:

- Assume that each tournament participant $i = 1, 2, \ldots, N$ has their hit rate $r_1, r_2, \ldots, r_N$.
  - The distribution suitable for generating $r_1, r_2, \ldots, r_N$ is identified based on the actual tournament results (Section 3.1).
- Specify the tournament format.
- Perform a tournament simulation using the specified format with the hit rates $r_1, r_2, \ldots, r_N$ (Section 4).
- Define evaluation indices reflecting the difference between $r_1, r_2, \ldots, r_N$ and the results (final ranking), as well as evaluation indices reflecting tournament costs (Section 3.2). Based on these indices, evaluate the tournament results (Section 4).

3.1 Estimation of player skill distribution

This paper makes the following assumptions:

- Assume that each tournament participant $i = 1, 2, \ldots, N$ has their hit rate $r_1, r_2, \ldots, r_N$.
- Hit rates $r_1, r_2, \ldots, r_N$ are samples generated from an appropriate distribution for each tournament. They are sorted in descending order without losing generality.
- Each participant's r_i remains constant during the tournament, and each shot is independent.

The hit rate is a continuous real number between 0 and 1. Based on Figure 2, it is assumed that the distribution has an unimodal peak. To satisfy these properties, the beta distribution Beta(α, β) is assumed.

The probability density function of the beta distribution is given by the following equation:

$$f(x|\alpha,\beta) = Cx^{\alpha-1}(1-x)^{\beta-1}, \quad (x \in [0,1]), \;\; C = \frac{1}{\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx}. \tag{1}$$

The parameters α and β are selected based on numerical experiments to generate results that most closely match the actual tournament results. The distance between two empirical cumulative distribution functions, i.e., the actual result and the simulated one, is measured using the KS-test statistic D_{KS} [2].

3.2 Performance indices of tournament formats

In sports tournaments, especially those for students, it is important that participants gain experience through their involvement. In the current format, the minimum guaranteed number of attempts for all participants is four. Increasing the number of attempts is desirable from the both viewpoints; participants gaining more experience and an skill assessment accuracy. With more shots, it is expected that the hit rate will more accurately reflect the participants' true skill. However, due to constraints such as time, the number of attempts has an upper limit. Additionally, in Kyudo, arrows are considered as pairs (referred to as Haya and Otoya), and this tradition is reflected in competition regulations [3, Article 15]. Therefore, preliminary or semifinal rounds cannot adopt formats with an odd number of arrows, such as five or seven (although Izume shootouts are conducted one arrow at a time, following different rules).

In this study, we propose two indices to evaluate the tournament format in terms of operational cost and accuracy of skill measurement: the "total number of shots in the tournament" and the "weighted distance between the tournament's actual ranking and the ranking based on hitting probability parameters r_i."

The "total number of shots in the tournament" is obtained by simply summing the number of shots, denoted as J_{shots}.

For the latter index, to assign more weight to rank discrepancies at the top level, we propose J_{rank} by the following equation:

$$J_{\mathrm{rank}} = \sum_{i=1}^{N} \left(\log_2 i - \log_2 \mathrm{Rank}(i)\right)^2, \tag{2}$$

where $\mathrm{Rank}(i)$ is the tournament ranking of the player with the i-th hitting probability parameter. This measure treats a discrepancy where the top-ranked player finishes second as equivalent to a case where the 20th-ranked player finishes 40th. A similar concept is used in professional tennis ATP ranking points [4].

Both J_{shots} and J_{rank} are better when smaller, and smaller standard deviation among tournament executions is also preferable. We will set up several tournament formats with a minimum number of attempts greater than four and calculate these indices through numerical simulations to examine whether formats comparable to or better than the current one can be developed.

4 Results and discussions

Based on a sufficient number of numerical simulations, the parameter values $(\alpha,\beta) = (5.1650,\ 3.7125)$ and $(6.3975,\ 5.6826)$ were obtained for men's and women's competitions, respectively.

Figure 3 shows the following for the male players; Blue dotted line: The frequency distribution obtained by simulating the number of hits out of four shots after generating the hit rate of 100 participants from the

estimated beta distribution. The simulation was performed 100 times. Red solid line: The actual relative frequency in the preliminary round of the five tournaments.

From the figure, we can see that the results are generally within the 100 simulations, and we can confirm that the estimated beta distribution can be considered an approximation of the actual distribution of the athletes.

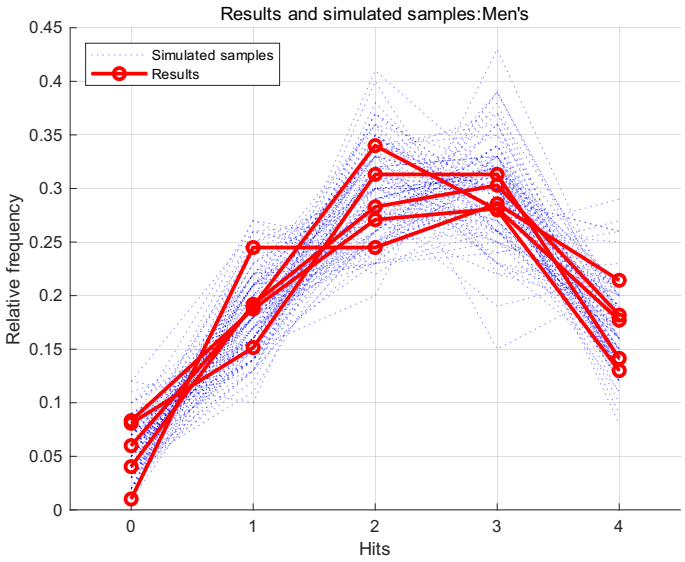


Figure 3: Results and simulations. Men.

4.1 Evaluation of Tournament Formats

In addition to the current format ("3/4–3/4–Izume"), we evaluated the following tournament formats: "6/8–Izume", "5/6–Izume", and "Izume".

The number of participants was set to 100 for both men and women, and each format was simulated 1,000 times.

Figure 4 shows scatter plots of the evaluation indices $(J_{\text{rank}}, J_{\text{num}})$ for men's tournament formats. The results for women's simulation are not present due to space limitation.

In this figure, the dots represent the results of individual simulations, while the solid line indicates a contour within which 95% of the simulations (950 runs) are estimated to fall. The white circles indicate the average value for each tournament format.

The following trends were observed from the results: Generally, the performance as a skill assessment improves as the minimum guaranteed number of attempts increases. However, the difference between simulations is considerably large.

Comparing the current format with the "5/6–Izume", the latter slightly outperforms the former in terms of skill assessment. This indicates that allowing all participants to take 6 shots with a slightly stricter criterion $(3/4 < 5/6)$ yields a more accurate skill estimation compared to conducting two rounds of 4-shot selection.

The "5/6–Izume" format has a similar average of total shots to the current format, while also exhibiting less variance. Small standard deviation in the number of attempts contributes to the simplification of tournament management. In a similar context, volleyball changed its rule from the "side-out system" to the "rally point system" to stabilize match duration[5], suggesting a certain benefit of the alternatives.

Other simulated formats, such as "Izume-only" and "6/8–Izume", either significantly differ from the current format in terms of minimum guaranteed shots, skill assessment performance, or total number of shots. Therefore, they cannot be proposed as possible alternatives.

Based on the above results, unless the use of a 4-shot unit is strictly required, we propose the "5/6–Izume" format as a possible alternative to the current format. This new format improves the minimum number of guaranteed shots from 4 to 6, slightly enhances skill assessment performance, and reduces the variance in the total number of shots, making tournament duration easier to predict.

Figure 4: Performance indices $(J_{\text{rank}}, J_{\text{num}})$ for Men's tournament formats

References

[1] CHRIS WELLS. A brief history of the competition formats used in international archery 1931-2020. https://www.worldarchery.sport/news/178443/brief-history-competition-formats-used-international-archery-1931-2020, July 2020. accessed 2024/12/12.

[2] Mathworks Inc. kstest2. https://jp.mathworks.com/help/stats/kstest2.html. Accessed 2024/12/12.

[3] https://www.kyudo.jp/pdf/documents/play_rules.pdf, 2016.

[4] ATP. ATP Rankings FAQ. https://www.atptour.com/en/rankings/rankings-faq. Accessed 2024/12/12.

[5] A Ureña, C Gallardo, J Delgado, R Calvo, and A Oña. Effect of the new scoring system on male volleyball. *The Coach*, 4:12–18, 2000.

Predicting International Success of Pace Bowlers in T20 Cricket

Ali Iltaf*, Richard Allmendinger**, Ali Hassanzadeh** and Richard Kingston**

University of Manchester, Manchester, United Kingdom
ali.iltaf@student.manchester.ac.uk
richard.allmendinger@manchester.ac.uk
ali.h@manchester.ac.uk
richard.kingston@manchester.ac.uk

Abstract

This study investigates the extent to which domestic T20 performance metrics can predict international success for pace bowlers in cricket. Using ball-by-ball data from over a decade of domestic and international T20 matches provided by the England and Wales Cricket Board (ECB), we engineer a comprehensive set of player-level features, including ball-tracking variables and outcome-based statistics. Success at the international level is evaluated using a Net Contribution metric adapted from the Duckworth-Lewis methodology. To identify key predictors, we apply feature selection techniques such as minimum redundancy maximum relevance (mRMR) and correlation clustering. Several regression models, including Random Forest and XGBoost, are trained and evaluated, with Random Forest achieving the best performance ($R^2 = 0.53$). Model interpretation using SHAP values reveals that a bowler's boundary percentage, dot ball percentage and percentage of their wickets taken that were caught are among the most influential features. These findings offer data-driven insights for selectors and talent scouts seeking to identify and fast-track promising pace bowlers from domestic leagues.

1 Introduction

T20 cricket is a short format of cricket designed to produce fast-paced games with an emphasis on scoring runs quickly. The England and Wales Cricket Board (ECB) manages the international cricket team as well as the domestic leagues, and it will always remain in the interest of the ECB to be able to identify new talent for the international team. The dynamics of the game at the T20 level do not translate perfectly to the international level, and it has been the case that players that have performed well in domestic leagues cannot uphold the same level of performance at the international stage.

This paper aims to aid this decision making process by asking: Can domestic T20 performance metrics predict a pace bowler's success in T20 Internationals? Performance is evaluated using a wide array of metrics which use ball-tracking statistics and match events to provide measures of bowlers' bowling ability and the outcomes of their bowling. Success is measured by the Average Net Contribution, which is the average difference in runs conceded and the expected number of runs scored over all deliveries bowled by a specific bowler in a match.

2 Background

There are several studies that attempt to predict player performance based on previous performance. Most of these studies use traditional metrics including strike rate and runs scored for batsmen and economy and wickets taken for bowlers. In order to differentiate players, it may also be required to go further into the details of player performance.

Asad et al. (2022) use commentary data to determine how many balls a batsman left, missed and hit to calculate the control of a batsman. This control measure was then used to calculate the 'Effective Runs', which is a metric proposed in the paper. Rupai et al. (2020) use pitch and weather data alongside ball tracking data to predict the outcome of each ball in a match. Mody et al. (2021) propose a formula for batting form, a pressure index and account for which team is the opposition. The study was in the context of the Indian Premier League (IPL) so the opposition team feature was a categorical variable which indicated one of the 8 IPL teams, but the opposition team cannot always be used as a factor if the teams are unknown. Mody et al. (2021) also uses classification to group players into scoring bands. The players in the highest rank are predicted to score the most runs. The problem with grouping players like this is that players of a different caliber can be grouped together in the same band, and it is made more difficult to make a comparative judgement of similarly ranked players.

A major factor in sports performance prediction is deciding what to use as a performance evaluation metric. Studies commonly used runs scored for a batsman or economy or wickets taken for a bowler as the target variable. The issue that arises with only using those metrics to profile performance is that it does not take into account the wholistic performance of the player. Lewis (2005) proposes two context-aware metrics called the Net Contribution and Resource Average. These metrics take into account the amount of wickets and overs remaining at each stage of the game and base the player's score on the aggregation of these resource contributions over every delivery. Lemmer (2002) proposes the Combined Bowling Rate (CBR), which is the harmonic mean of bowling strike rate (balls per wicket), economy rate (runs per over) and the runs per wicket, but this metric can only be applied to bowlers and is derived directly from other metrics which would be used as features. Thomson et al. (2021) propose a contextual batting score to measure batting and bowling performance when a team is batting second, but this metric can only be used to measure performance in the second innings.

Another aspect that is not seen in the literature is an interpretation of the models. In order to draw conclusions from the models to help inform decision-making, it is important that the driving factors for prediction for each model are considered. By making sure that the models are interpretable, it can also be checked that the models are making sense, and the relationship between the predictors and outcome have the desired relationship. For example, it would not make sense for a bowler's performance rating to be positively correlated with the bowler's economy (runs conceded per over).

3 Data

The data used for this study was provided by the ECB. This data includes ball-by-ball data on all professional T20 matches, including both international and domestic matches, from the start of 2010 up until 23rd October 2024. The data includes key details from the match the delivery was played in and more detailed data on each delivery, such as the information that can be found about the scorecard, shot and delivery types, foot

movement for the batsman, as well as some ball-tracking data.

Before analysis, the raw ball-by-ball match data was transformed to a player-level format, with rows representing individual players and columns capturing various performance metrics. The dataset was first divided into international and domestic subsets, with the former used for training and testing, and the latter reserved for prediction. To account for changes in performance across a player's career, statistics were further aggregated into age groups: 18–24, 25–28, 29–32, 33–36, and 37–42. The first and last groups span wider age ranges to ensure a sufficient number of matches for reliable statistics, minimizing distortion from outlier performances. After this aggregation, player records with missing values were removed, resulting in a final dataset of 630 players across 141 features.

4 Methods

4.1 Player Performance Evaluation

The Net Contribution metric in cricket evaluates a player's impact by measuring the difference between actual runs scored or conceded and the expected runs based on the Duckworth/Lewis (D/L) model, calculated on a ball-by-ball basis (Lewis, 2005). It incorporates match context—specifically, overs remaining and wickets lost—offering a more situational and comprehensive assessment of performance than traditional metrics. By integrating both run rate and wicket impact into a single value, it avoids inflation from performances against weaker opponents and remains calculable in all scenarios, unlike metrics such as bowling strike rate or CBR. Due to the proprietary nature of the original D/L formula (Duckworth and Lewis, 1998), this research uses a modified version proposed by McHale and Asif (2013). Overall, the Net Contribution metric offers a more nuanced understanding of player performance compared to traditional aggregate statistics (Lewis, 2008).

4.2 Feature Selection

With the feature set and target variable prepared, machine learning models can now be trained on the data. Given the presence of 141 features and high multicollinearity, feature selection is essential to retain predictive power while improving interpretability. Instead of using dimensionality reduction methods like PCA or t-SNE—which transform features into uninterpretable combinations—this study employs Minimum Redundancy Maximum Relevance (mRMR) and correlation clustering. These methods maintain the original features' meaning, which is crucial for understanding the relationship between features and performance. mRMR selects features that are highly relevant to the target while minimizing redundancy (Peng et al., 2005), improving efficiency without iterative model retraining. Correlation clustering uses hierarchical clustering with Spearman correlations and Ward's linkage to group redundant features, from which a single representative feature is selected per cluster for model training.

4.3 Model Training

Using the selected features, regression models were trained to predict Average Net Contribution, with five models evaluated: Linear Regression, Support Vector Regression, Decision Trees, Random Forests, and XGBoost. These models represent a range of techniques from simple linear to complex ensemble and

kernel-based approaches, allowing for a balanced comparison across different data characteristics. Data was normalized before training, and model performance was assessed using Mean Squared Error (MSE) and R^2. The best-performing model was then applied to domestic player data to predict performance and rank pace bowlers accordingly.

5 Results

Figure 1 shows that among the evaluated models, Random Forest achieved the best predictive performance, followed by XGBoost and then Linear Regression. Random Forest's robustness and ability to generalize well without intensive hyperparameter tuning made it particularly effective, especially given the noisy nature of the data. Although XGBoost is a powerful model, its performance may have been hindered by its sensitivity to hyperparameter settings, making it more prone to overfitting without careful tuning. Linear Regression performed reasonably well, likely due to the presence of some linear relationships in the data, while Decision Trees and SVR underperformed due to overfitting and sensitivity to noise or suboptimal parameters. Among the feature selection methods, correlation clustering performed poorly as it often grouped and excluded key predictors, resulting in uninformative feature sets. The mRMR methods (MID and MIQ) performed similarly across models, with MID working better for XGBoost and Decision Trees, and MIQ better for Linear Regression and SVR. For Random Forest, both mRMR methods produced nearly identical results, with MIQ slightly ahead. While omitting feature selection produced the best raw performance in most cases, the resulting models lacked interpretability, with many features showing zero or negative importance. This justified the use of interpretable feature selection techniques like mRMR.

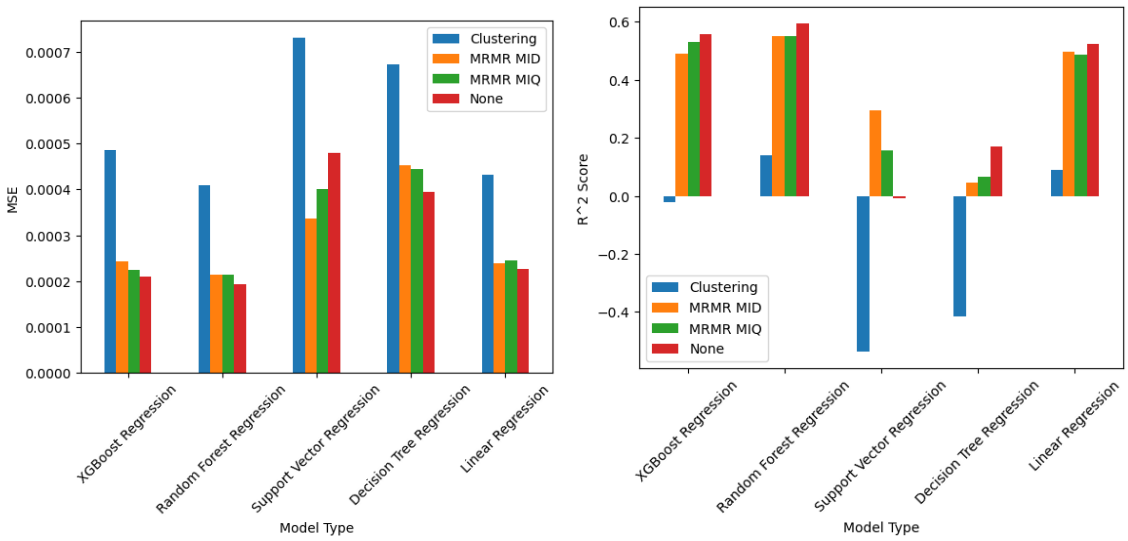


Figure 1: Mean Squared Error (left) and R^2 score (right) of each model type with each subset of features.

We select the Random Forest model using mRMR MIQ feature selection and inspect it more closely. Figure **??** shows a plot of the permutation importances of the features used in the model and Figure **??** shows a SHAP beeswarm plot, which ranks the features by their SHAP score and shows the relationship

between the feature and output for each sample. Both methods show that the three most important features are the Boundary %, the Dot Ball % and the % of wickets taken by the bowler that were caught. In both plots, the importance of these three features is significantly higher than the rest.

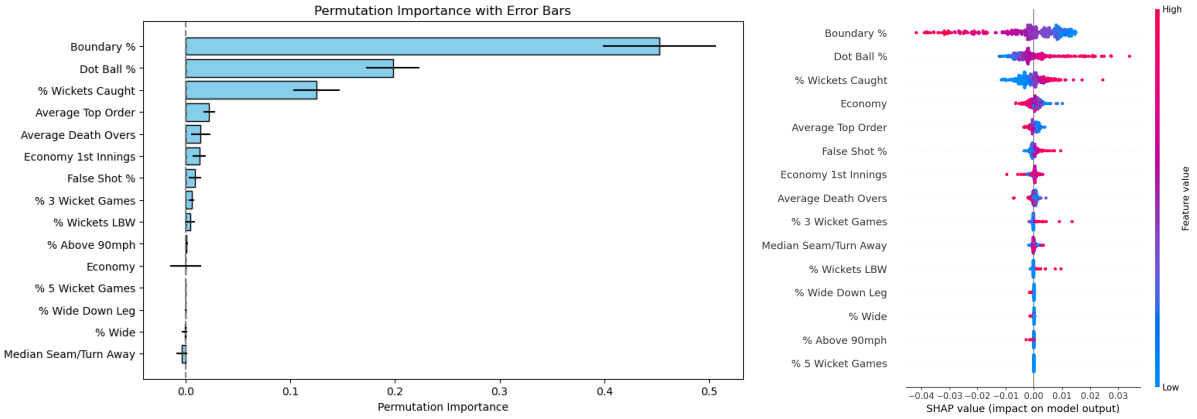


Figure 2: Permutation Importance plot (left) and SHAP beeswarm plot (right) of features used in the Random Forest model using mRMR MIQ feature selection.

Finally, using the MIQ Random Forest model, we predict the ouptut on the domestic data from the T20 Blast since 2016. Older tournaments are not included since we are interested in scouting recent performances. After removing retired players, we sort the players according to their predicted contribution, keeping only the most recent age group entry. The predicted top 10 bowlers for the England Cricket Team based on domestic performance are, in order of rank, Craig Overton, David Payne, David Willey, Benny Howell, Pat Brown, Tom Taylor, Jofra Archer, Matthew Waite, Paul Walter and Luke Fletcher.

6 Discussion

Out of the top 10 predicted pace bowlers, the fact that bowlers who have already made the international team, like Jofra Archer, Craig Overton and David Willey, shows that this method can predict which players are high performers. Other players who have played internationally have not played many games in the English domestic league, so they are not present.

The models show that there is a heavy emphasis on keeping the number of runs low, which is intuitive since T20 is a format that prioritises getting a large amount of runs quickly, rather than emphasising protecting the batter's wicket. Common knowledge suggests that the ability to bowl at high speeds and seam/swing bowling are crucial to breaking through to the international level. The models here show that these factors are not as important. This may be due to the fact that some metrics measure the bowlers actions, such as the line, length, and speed, whilst others measure the outcome of the delivery, such as the economy, strike rate, and boundary percentage. Further research should be undertaken on the causal relationship between these different types of metrics.

Acknowledgments

The data used for this research was provided by the England and Wales Cricket Board.

References

[1] Ahmad Al Asad et al. (2022) *Impact of a Batter in ODI Cricket Implementing Regression Models from Match Commentary.* In: 2022 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE), pp. 1–6. DOI: 10.1109/CSDE56538.2022.10089357.

[2] F. C. Duckworth and A. J. Lewis. (1998) *A Fair Method for Resetting the Target in Interrupted One-Day Cricket Matches.* In: The Journal of the Operational Research Society 49.3, Palgrave Macmillan Journals, pp. 220–227. DOI: 10.2307/3010471.

[3] H.H. Lemmer. (2002) *The combined bowling rate as a measure of bowling performance in cricket.* In: South African Journal for Research in Sport, Physical Education and Recreation 24.2, pp. 37–44. DOI: 10.4314/sajrs.v24i2.25839.

[4] A J Lewis. (2005) *Towards fairer measures of player performance in one-day cricket.* In: Journal of the Operational Research Society 56.7, pp. 804–815. DOI: 10.1057/palgrave.jors.

[5] Lewis, A.J. (2008) *Extending the range of player-performance measures in one-day cricket*, In: Journal of the Operational Research Society, 59(6), pp. 729–742. DOI: https://doi.org/10.1057/palgrave.jors.2602379.

[6] Ian G. McHale and Muhammad Asif. (2013) *A modified Duckworth–Lewis method for adjusting targets in interrupted limited overs cricket.* In: European Journal of Operational Research 225.2, pp. 353–362. DOI: 10.1016/j.ejor.2012.09.036.

[7] Khush Mody, D. Malathi, and J. D. Dorathi Jayaseeli. (2021) *An Artificial Neural Network Approach for Classifying Cricket Batsman's Performance by Adam Optimizer and Prediction by Derived Attributes.* In: 2021 Smart Technologies, Communication and Robotics (STCR), pp. 1–7. DOI: 10.1109/STCR51658.2021.9588836.

[8] Hanchuan Peng, Fuhui Long and Ding, C. (2005) *Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy*, In: IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), pp. 1226–1238. DOI: https://doi.org/10.1109/TPAMI.2005.159.

[9] Aneem-Al-Ahsan Rupai, Md. Saddam Hossain Mukta, and A. K. M. Najmul Islam. (2020) *Predicting Bowling Performance in Cricket from Publicly Available Data.* In: Proceedings of the International Conference on Computing Advancements., pp. 1–6. DOI: 10.1145/3377049.3377112.

[10] James Thomson, Harsha Perera, and Tim B. Swartz. (2021) *Contextual batting and bowling in limited overs cricket.* In: South African Statistical Journal 55.1, pp. 73–86. DOI: 10.37920/sasj.2021.55.1.6.

Quantifying and Comparing NBA Player Career Momentum Using Statistical Methods

Ross Lauterbach
CUNY Hunter College, New York, USA
rosslauterbach1@gmail.com

Abstract

Momentum is one of the most widely referenced yet poorly defined concepts in sports. In the NBA, commentators and fans routinely describe players as "heating up" or "catching fire," often attributing shifts in performance to an intangible momentum factor. Despite its prominence in narrative and analysis, momentum is a measure that has been hard to verify empirically. This paper introduces a statistical approach to capture player momentum throughout an NBA career using smoothed performance trajectories. By constructing game-by-game momentum data and powerful visualizations, we aim to identify sustained periods of elevated or diminished performance and quantify the uncertainty around them. We also take a deep dive into methods of calculation and modeling using momentum.

1 Introduction

While there have been numerous attempts to capture momentum at a specific point in time, particularly within individual games, few have sought to define and quantify it across a player's entire career. In the realm of basketball analytics, much of the momentum literature has focused on short-term phenomena such as hot streaks and game-to-game variability. A study by Gilovich et al. (1985) challenged the widely held belief in the "hot hand," arguing that perceived shooting streaks were simply cognitive illusions rather than statistical realities. More recent work, however, has re-evaluated this conclusion; Miller and Sanjurjo (2018) showed that earlier studies underestimated the likelihood of streaks due to statistical bias, providing evidence that hot-hand effects are both real and measurable.

Beyond in-game performance, researchers have also explored whether momentum carries over between games. Arkes and Martinez (2011) used an econometric framework to assess team-level momentum in the NBA, finding that recent success modestly improves the probability of future wins, even after controlling for team strength. These studies demonstrate a growing interest in quantifying momentum, but they remain largely confined to short-term patterns and team dynamics. This paper aims to extend that line of inquiry by shifting the focus to long-term, player-level momentum. Rather than capturing momentary flashes of brilliance, we propose a model that smooths game-by-game performance data to trace career-long trends. This enables us to identify sustained periods of elevated or diminished output and to quantify uncertainty around those trends. In doing so, we contribute a new tool for understanding player consistency and for empirically validating or challenging popular narratives about career arcs.

2 Data Description

The dataset consists of player-level game logs from NBA regular season games compiled from the official NBA API and also available on Kaggle. Each observation represents a single player's performance in a single game and includes a wide range of traditional and advanced statistics: points, assists, rebounds, steals, blocks, turnovers, shooting percentages, and more. Data was collected starting in 1980, when the 3-point line was introduced, all the way until the end of the 2022-2023 season, resulting in approximately 1.4 million entries. We filtered the dataset to include only regular-season games to avoid postseason variability and ensure comparability across players.

3 Calculation of Momentum

To construct a player's momentum curve, we indexed their games chronologically and computed a game number variable to serve as a time axis. We removed duplicate records and ensured consistent game tracking by identifying each player-game instance via a unique combination of player ID and game ID. A validation test was also run against publicly available data to ensure aggregates agreed with each other. This cleaned dataset serves as the foundation for our momentum calculations.

Figure 1 contains a scatter plot and correlation coefficient for each variable interaction. The variables that were initially under consideration for the momentum calculation were used: points, rebounds, assists, steals, blocks, turnovers, and various efficiency measures.

Figure 1: Triangular correlation plot of momentum variables.

Given the lack of extreme multicollinearity, we continued with our analysis as planned. However, the moderate correlation between output and efficiency led us to focus solely on player output as a measure of momentum. This specific weighting below was used given its extreme simplicity and the distribution of the data, but was overall an arbitrary choice based on input from basketball fans and analysts. The score aims to capture a player's all-around influence in a given game while only incorporating the 6 major simple performance metrics. We then smoothed these scores over time

using an exponentially weighted moving average (EWMA) to reflect recent performance trends while preserving long-term stability. This smoothed score serves as the core of our momentum metric. Initially, we also incorporated a team indicator representing recent team success, defined as a scaled 10-game rolling win count, to capture potential psychological or contextual effects on a player's performance. This led to the general momentum equation, which is the sum of the EWMA-based performance score and a weighted team indicator. The equation for the performance score, S_i is shown below.

$$S_i = \text{Points}_i + 2 \cdot (\text{Rebounds}_i + \text{Assists}_i) + 5 \cdot (\text{Steals}_i + \text{Blocks}_i - \text{Turnovers}_i)$$

The win indicator I_i is calculated as one fifth of the sum of the binary win/loss indicator W_k over the previous 10 games, with a subtraction of 5 to normalize it:

$$I_i = \frac{1}{5} \sum_{k=i-9}^{i} W_k - 5$$

where W_k is a binary indicator for win (1) or loss (0) at game k.

Finally, the momentum M is calculated by applying an exponential smoothing factor α and a decay parameter γ as follows:

$$M = (1-\alpha) \sum_{j=1}^{i-1} \alpha(1-\alpha)^{i-j-1} S_j + \gamma \cdot I_i$$

4 Momentum Optimization

While the primary goal of this metric is not predictive accuracy, we initially explored whether momentum could be tuned to enhance its alignment with future or current performance outcomes. The underlying hypothesis was that a player's momentum score could offer predictive value beyond its descriptive nature, potentially serving as a useful indicator of future game performance or for assessing a player's impact in the current game. To test this hypothesis, we focus on optimizing two key parameters in the momentum formula: the smoothing parameter, alpha, and the weighting factor of the team, gamma. The idea was to minimize the mean squared error (MSE) between the momentum score and either a player's next-game performance score or their current game plus-minus, which quantifies a player's overall contribution relative to the game outcome. By doing so, we aimed to validate the strength of the momentum signal in predicting a player's performance and to investigate whether the momentum metric could be improved by incorporating team performance factors alongside individual statistics.

The process involved systematically adjusting the parameters alpha and gamma and observing their impact on the predictive accuracy of the momentum score. After a series of trials, the optimal parameter values suggested an intriguing result: the best-performing momentum score excluded team context entirely, with a gamma value of 0, indicating that recent team success did not add any predictive power. This finding was particularly interesting, as it suggested that a player's momentum might be more directly tied to their own individual performance trajectory rather than the broader team performance. Additionally, the value of alpha that minimized the MSE was relatively low, at 0.1, signifying that a shorter smoothing window, one that emphasizes more recent games, was most effective in predicting future performance. Figure 2 shows a heatmap

of the optimization process, with darker colors indicating higher performance under the simple linear model with one predictor variable.

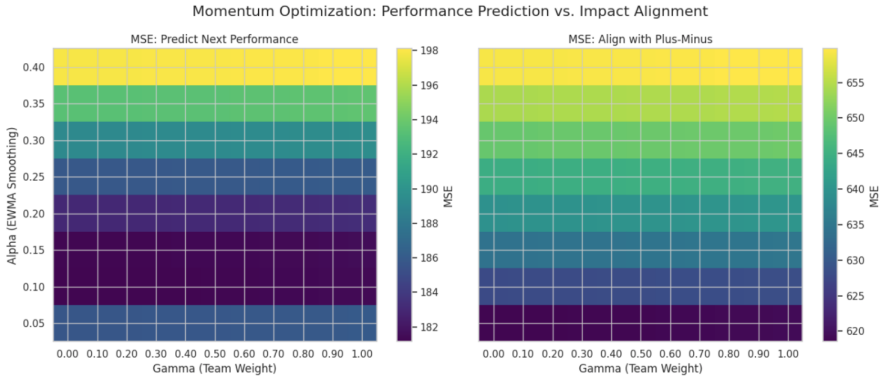


Figure 2: Heatmap of output from grid search momentum optimization using MSE.

5 Momentum Visualization

Now that we can calculate player momentum at a given point in time, we can track and visualize an athlete's performance throughout their career. In particular, momentum curves capture the fluctuations in a player's contributions during a season or throughout their career, providing a dynamic view of their consistency and impact. Rather than simply looking at traditional statistics, we can gain a better understanding of how a player evolves and maintains consistency throughout their career. The previously used methods are meant to filter out the noise of random fluctuations in performance and highlight underlying trends in the data. These smoothed trajectories reveal not only the magnitude of a player's performance but also the stability or volatility of their impact over time.

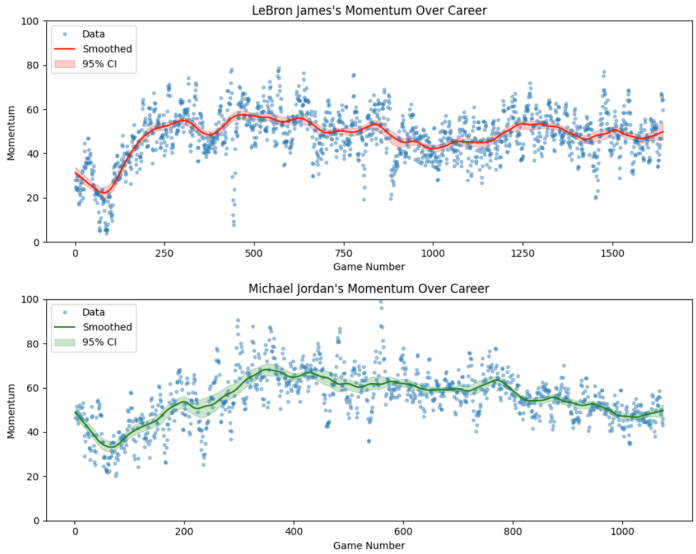


Figure 3: Momentum curves for Lebron James and Michael Jordan with 95% confidence bands.

6 Career Trajectory Clustering

The next objective was to classify players' careers by clustering them on the basis of the shape of their momentum trajectories, independent of the era in which they played. To do this, we created a fixed-length vector for each player by interpolating their momentum scores over the first 100 games of their career. Players with more than 300 games played were included. This interpolation allowed for a consistent representation of momentum between players, enabling an analysis of how their career trajectories evolved. Only players with at least 300 games were included in the analysis to ensure that each player had a sufficient sample size for comparison and to maintain consistency across the dataset.

However, during the interpolation process, some players had missing values in their momentum curves, often due to irregular game participation, injuries, or gaps in the available data. To ensure that the clustering analysis was based on clean and interpretable data, we excluded any momentum vectors that contained missing values. This filtering step was critical, as it prevented incomplete data from distorting the clustering results and ensured that the dimensionality reduction and clustering algorithms, such as principal component analysis (PCA) and k-means, operated on reliable and consistent input. By removing incomplete data points, we were able to minimize bias and instability in the clustering process, resulting in more accurate groupings of players based on the overall shape of their momentum trajectories. Figure 4 shows both the mean career trajectories in each cluster and the class distinction using PCA components.

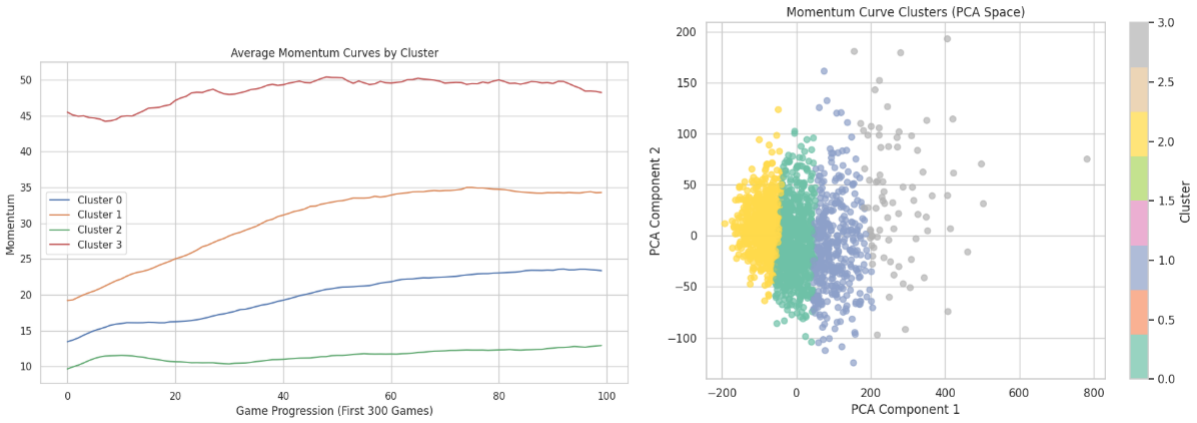


Figure 4: Plot of average career momentum within clusters and decision boundaries based on PCA components.

The benefit of momentum clustering is that it allows for comparison across different eras, something that has been difficult to do from an objective, data-driven standpoint. We can see that any player that is part of Cluster 3 is essentially an outlier in the projected 2D principal component space. Although claims of inflated statistics in the modern era may have some merit, members of Cluster 3 range from modern, durable stars like Shaquille O'Neal, Tracy McGrady, LeBron James, and Chris Paul, to pioneers of the game from the mid to late twentieth century such as Bill Walton, George McGinnis, Dave Cowens, and Wilt Chamberlain. Despite inconsistencies in output throughout different stages of the NBA, this gives us a way to compare performance trends throughout history. Because we ran k-means clustering on large vectors containing the first 300 games, we can visualize

the 'average trajectory' of players in a cluster. We can also make inferences about career trajectory. For example, we see that on average in cluster 3, star players will have a small but pronounced dip at the beginning of their career before figuring things out. Cluster one players, on the other hand, see a fast ascension to contributions but are never able to take the next step to stardom.

7 Conclusion

In conclusion, this study introduces a novel momentum metric to analyze NBA players' careers, focusing on sustained performance over time. Through careful data cleaning and interpolation of momentum scores, we ensured that the analysis was based on reliable and comparable data. The optimization of key parameters, such as the smoothing factor and team context weight, revealed that individual performance trends were the most significant predictors of future performance, while team success did not enhance short-term performance prediction.

The clustering of players based on their momentum trajectories provided valuable insights into career progression, grouping players across eras by the shape of their performance curves. This approach allows for meaningful comparisons of player careers, independent of historical context. In the future, it would be interesting to analyze the curvature of the momentum curves, as well as any insights that can be drawn from the gradient itself. Overall, this work offers a quantitative framework for evaluating momentum in player performance, laying the groundwork for future studies on career trajectories and player evaluation.

References

- Arkes, J., and Martinez, J.A. (2011). Finally, evidence for a momentum effect in the NBA. *Journal of Quantitative Analysis in Sports*, 7(3), Article 10.
- Gilovich, T., Vallone, R., and Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17(3), 295–314.
- Miller, J.B., and Sanjurjo, A. (2018). Surprised by the gambler's and hot hand fallacies? A truth in the law of small numbers. *Econometrica*, 86(6), 2019–2047.

The impact of physical parameters on match outcomes in Serie A. A preliminary analysis

A. Lucadamo* and M. Beato** and C. Savoia*** and D. Pompa**** F. Laterza***** and P. Troiani**** and M. Bertollo****

*DEMM, University of Sannio, Benevento, Italy: antonio.lucadamo@unisannio.it

**School of Allied Health Sciences, University of Suffolk, Ipswich, UK: m.beato@uos.ac.uk

*** The Research Institute for Sport and Exercise Sciences, Liverpool John Moores University: cristian.savoia@k-sport.tech

**** BIND Center, Department of Medicine and Aging Sciences, University "G. d'Annunzio" of Chieti-Pescara: dario.pompa@studenti.unich.it; pt1984@virgilio.it; m.bertollo@unich.it

***** Department of Neurosciences, Biomedicine and Movement Sciences, University of Verona: francesco.laterza@univr.it

Abstract

This study explores the link between external training load metrics and match outcomes in Serie A, using tracking data from the 2022/2023 and 2023/2024 seasons. Physical performance was analyzed by play phase, focusing on variables like sprinting, accelerations, and metabolic power. A Poisson regression with LASSO regularization addressed data complexity. Results show that high-intensity efforts in possession are associated with goals scored, while lower defensive output by opponents relates to goals conceded. These findings emphasize the role of explosive actions and match context, offering practical insights for training and future research.

1 Introduction

In recent years, the application of performance analytics in professional football has transformed how teams prepare, compete, and evaluate success. Among the most widely adopted tools in this evolving landscape are external training load metrics, which offer objective, quantifiable insights into the physical demands placed on players during both training and competition (McGregor et al. (2024)). These metrics are typically collected using GPS and other tracking technologies, enabling practitioners to monitor and manage player's training load with increasing precision (Dawson et al. (2024)). External load metrics encompass a range of variables that reflect different aspects of physical performance. Total distance covered is a foundational measure, representing the cumulative ground a player travels during a session or match. While it is useful for assessing overall training load, it does not offer the granularity needed to assess training intensity. To address this, analysts often examine high-speed running (HSR) and sprinting distances, which quantify more demanding efforts, typically defined as running above 19.8 km/h and 25.2 km/h, respectively (Gualtieri et al. (2023)). These metrics are particularly valuable (when analyzed in conjunction with tactical parameters) for understanding a player's involvement in sport specific actions such as pressing, counterattacks, and defensive transitions (Beato and Drust (2021)).

Beyond speed-based measures, metabolic power has emerged as another indicator of physical exertion (Polglaze and Hoppe (2019), Venzke et al. (2023)). It estimates the energy cost of movement, especially during accelerations, decelerations, and changes of direction—actions that are common in

football but not fully captured by traditional speed parameters. High-intensity metabolic power, often defined as exceeding 25.5 W/kg of body mass, provides a specific view of the high intensity demands placed on players. Similarly, distance covered during high-intensity accelerations (typically above 3.0 m/s^2) reflects the neuromuscular load associated with rapid bursts of movement (Silva et al. (2023)), which are critical in both offensive and defensive scenarios. This study focuses on Serie A, Italy's top-tier professional football league, known for its tactical characteristics, competitive intensity, and high physical demands. It is interesting to understand how external load metrics relate to match outcomes in Serie A. While previous research has often analyzed team performance in isolation (Savoia et al. (2024)), this study adopts a contextual and comparative approach, evaluating both the reference team and their opponents. This dual perspective is grounded in the understanding that football is a dynamic, interactive sport—performance is not only a function of a team's actions but also of the opposition's behavior and strategy. The primary aim of this study is to investigate whether selected external load variables are associated with match outcomes—specifically, scored and conceded goals—in Serie A. By incorporating data from both competing teams, the analysis seeks to identify patterns and relationships that may be obscured when examining teams in isolation. This approach acknowledges the context-dependent nature of physical performance in football, where the same level of exertion may yield different outcomes depending on the quality and style of the opposition. Ultimately, this research aims to contribute to a more nuanced understanding of match dynamics in elite football. The findings are intended to inform evidence-based decision-making for coaches, performance analysts, and sports scientists, offering practical insights into how physical performance metrics can be interpreted and applied within the broader tactical and competitive context of the game.

2 Data and methods

The data utilized in this study refers to the Italian Serie A football seasons 2022/2023 and 2023/2024. For each match, several outcome-related variables were recorded, including match outcomes (e.g., goals, cards, coach, date) and a wide range of physical metrics such as distances at different speeds, acceleration, deceleration, and metabolic power. Data were collected for both teams and categorized by phase of play (possession, non-possession, out-of-play), using the K-Sport Dynamix system (K-Sport World S.R.L., Pesaro, Italy) to process positional data.

In the first phase of the study, the analysis focused on a selection of variables, including, among others, total distance covered, distances covered during high-speed running and very high-speed running, metabolic power at high intensity, and distance covered during high-intensity acceleration. As stated before, the aim of this study is to determine whether these selected variables influence the match outcome. Both the values of the reference team and those of the opponents are considered, as it is believed that the importance of the variables lies not only in their values for the specific team but also in the context of the opposing team faced. The variables used are summarized in Table 1.

Table 1: Variable description

Variable Name	**Description**
D_S6_WB_team / D_S6_WB_opp	Mean distance covered with ball at speeds above 25 km/h (by the team / by the opponent team)
D_S6_NB_team / D_S6_NB_opp	Mean distance covered without ball at speeds above 25 km/h (by the team / by the opponent team)
D_S6_OP_team / D_S6_OP_opp	Mean distance covered during out-of-play phases at speeds above 25 km/h (by the team / by the opponent team)
D_A1_WB_team / D_A1_WB_opp	Mean distance covered with ball during high-intensity decelerations (< -3 m/s^2) (by the team / by the opponent team)

D_A1_NB_team / D_A1_NB_opp	Mean distance covered without ball during high-intensity decelerations (< −3 m/s²) (by the team / by the opponent team)
D_A1_OP_team / D_A1_OP_opp	Mean distance covered out-of-play during high-intensity decelerations (< −3 m/s²) (by the team / by the opponent team)
D_A8_WB_team / D_A8_WB_opp	Mean distance covered with ball during high-intensity accelerations (> 3 m/s²) (by the team / by the opponent team)
D_A8_NB_team / D_A8_NB_opp	Mean distance covered without ball during high-intensity accelerations (> 3 m/s²) (by the team / by the opponent team)
D_A8_OP_team / D_A8_OP_opp	Mean distance covered out-of-play during high-intensity accelerations (> 3 m/s²) (by the team / by the opponent team)
D_MPHI_WB_team/D MPHI_WB_opp	Mean distance covered with ball at > 25.5 W/kg power (by the team / by the opponent team)
D_MPHI_NB_team/D MPHI_NB_opp	Mean distance covered without ball at > 25.5 W/kg power (by the team / by the opponent team)
D_MPHI_OP_team/D MPHI_OP_opp	Mean distance covered out of play at > 25.5 W/kg power (by the team / by the opponent team)
Perc_ED_WB_team/ Perc_ED_WB_opp	Equivalent distance with ball (by the team / by the opponent team)
Perc_ED_NB_team/ Perc_ED_NB_opp	Equivalent distance without ball (by the team / by the opponent team)
Perc_AI_WB_team/ Perc_AI_WB_opp	Anaerobic index with ball (by the team / by the opponent team)
Perc_AI_NB_team/ Perc_AI_NB_opp	Anaerobic index without ball (by the team / by the opponent team)
AMP_WB_team/AMP_WB_opp	Average Metabolic Power with ball (for the team / for the opponent team)
AMP_NB_team/AMP_NB_opp	Average Metabolic Power without ball (for the team / for the opponent team)
D_20/25_Km/h_team/ D_20/25_Km/h_opp	Mean distance covered out of play at > 25.5 W/kg power (by the team / by the opponent team)

Since, as also indicated by the condition index, there is evidence of multicollinearity among the explanatory variables, and given that the response variables are count data, we employed a Poisson regression model with LASSO regularization. The correlation between the response variables was not explicitly modeled, as it is implicitly addressed by including covariates from both teams. In addition, a chi-squared test was conducted, which revealed no statistically significant dependence between the response variables. Parameter estimation is achieved by minimizing the penalized log-likelihood function $-\frac{1}{N}\sum_{i=1}^{n}\left(y_i(\beta_0+\beta^T x_i)-e^{\beta_0+\beta^T x_i}\right)+\lambda(\alpha\sum_{i=1}^{n}|\beta_i|)$.

3 Results

The results presented in Tables 2 through 4 indicate that the analyses conducted for the two seasons under consideration lead to similar conclusions. Specifically, certain variables show a significant effect in both the 2022/23 and 2023/24 seasons, with respect to both goals scored and goals conceded.

Table 2: Estimates from the Poisson regression with LASSO selection for scored goals (2022/23)

Variable	Estimate	Std.error	Statistic	p.value	signif
D_S6_WB_team	0.29648	0.08255	3.59150	0.00033	***
D_MPHI_OP_team	1.43807	0.35315	4.07211	0.00005	***
D_A1_WB_team	-0.37231	0.17074	-2.18048	0.02922	*
D_A1_NB_team	0.32150	0.15825	2.03162	0.04219	*
D_A1_OP_team	-0.88744	0.23698	-3.74486	0.00018	***
D_A8_WB_team	0.45380	0.15092	3.00692	0.00264	**
D_A8_NB_team	-0.40333	0.17108	-2.35754	0.01840	*
D_A8_OP_team	0.65638	0.21254	3.08818	0.00201	**
D_MPHI_OP_opp	-1.15358	0.35719	-3.22962	0.00124	**
D_A1_WB_opp	-0.28839	0.14333	-2.01206	0.04421	*
D_A1_OP_opp	0.84016	0.24697	3.40192	0.00067	***
D_A8_WB_opp	0.40889	0.16601	2.46298	0.01378	*
D_A8_NB_opp	-0.26946	0.15170	-1.77627	0.07569	.

*** p < 0.001; ** p < 0.01; * p < 0.05; . p < 0.1

Table 3: Estimates from the Poisson regression with LASSO selection for conceded goals (2022/23)

Variable	Estimate	Std.error	Statistic	p.value	signif
D_S6_WB_team	-0.19663	0.08666	-2.26914	0.02326	*
D_S6_OP_team	-0.47370	0.12168	-3.89318	0.00010	***
D_A1_WB_team	-0.31278	0.13053	-2.39628	0.01656	*
D_A8_WB_team	0.24483	0.14119	1.73406	0.08291	.
D_A8_OP_team	-0.27916	0.10666	-2.61737	0.00886	**
AMP_WB_team	0.26761	0.07381	3.62566	0.00029	***
D_S6_WB_opp	0.22562	0.08133	2.77427	0.00553	**
D_MPHI_WB_opp	-0.21933	0.09745	-2.25069	0.02440	*
D_MPHI_OP_opp	0.70142	0.16936	4.14151	0.00003	***

*** p < 0.001; ** p < 0.01; * p < 0.05; . p < 0.1

Table 4: Estimates from the Poisson regression with LASSO selection for scored goals (2023/24)

Variable	Estimate	Std.error	Statistic	p.value	signif
D_S6_WB_team	0.30701	0.05556	5.52532	0.00000	***
D_MPHI_OP_team	1.25845	0.21337	5.89793	0.00000	***
D_A1_WB_team	-0.23608	0.11638	-2.02853	0.04251	*
D_A1_OP_team	-0.52945	0.15228	-3.47684	0.00051	***
D_A8_OP_team	0.41214	0.14773	2.78973	0.00528	**
D_20/25_Km/h_team	-0.17818	0.07591	-2.34725	0.01891	*
D_20/25_Km/h_opp	0.20354	0.07929	2.56713	0.01025	*
AMP_NB_team	0.32468	0.16749	1.93846	0.05257	.
AMP_WB_team	-0.16443	0.07712	-2.13200	0.03301	*
Perc_ED_WB_opp	-0.19196	0.09406	-2.04080	0.04127	*
D_S6_WB_opp	-0.11952	0.05037	-2.37293	0.01765	*
D_S6_NB_opp	-0.15179	0.05900	-2.57277	0.01009	*

Variable	Estimate	Std.error	Statistic	p.value	signif
D_MPHI_OP_opp	-1.03957	0.21776	-4.77387	0.00000	***
D_A1_NB_opp	0.21720	0.12682	1.71264	0.08678	.
D_A1_OP_opp	0.50929	0.15401	3.30680	0.00094	***
D_A8_WB_opp	0.30468	0.12379	2.46125	0.01385	*
D_A8_OP_opp	-0.57811	0.15094	-3.82996	0.00013	***

*** p < 0.001; ** p < 0.01; * p < 0.05; . p < 0.1

Table 5: Estimates from the Poisson regression with LASSO selection for conceded goals (2023/24)

Variable	Estimate	Std.error	statistic	p.value	signif
D_S6_WB_team	-0.14133	0.06413	-2.20362	0.02755	*
D_MPHI_OP_team	-0.97747	0.23124	-4.22707	0.00002	***
D_20/25_Km/h_team	0.10718	0.06318	1.69652	0.08979	.
AMP_WB_team	0.13790	0.08372	1.64713	0.09953	.
D_S6_WB_opp	0.23793	0.05897	4.03435	0.00005	***
D_MPHI_OP_opp	1.39691	0.21335	6.54740	0.00000	***
D_A8_WB_opp	0.19320	0.10234	1.88775	0.05906	.

*** p < 0.001; ** p < 0.01; * p < 0.05; . p < 0.1

4 Discussion

The findings of this study provide valuable insights into the relationship between external training load metrics and match outcomes in Serie A, highlighting the importance of context-specific physical performance. The analysis revealed that certain high-intensity physical actions, particularly those performed in possession of the ball, are significantly associated with goal scoring. Specifically, greater distances covered while sprinting with the ball (above 25 km/h), higher volumes of high-intensity metabolic power output (above 25.5 W/kg), and distance covered during high-intensity accelerations (above 3.0 m/s²) in possession were all positively related to the number of goals scored.

These results reinforce the notion that explosive, high-intensity efforts in possession are critical to offensive success in elite football (Gualtieri et al. (2025)). Sprinting with the ball and accelerating at high intensities are often linked to decisive moments such as breaking defensive lines, creating space, or capitalizing on transitions (Beato et al. (2024), Chaize et al. (2024)). The association with high metabolic power further supports the idea that the energetic cost of these movements, particularly those involving frequent changes of direction and pace, is a key component of effective attacking play. Conversely, the analysis also found that goals were more likely when the opposing team exhibited lower physical output without the ball, particularly in terms of distance covered. This suggests that a lack of defensive intensity or pressing effort may create opportunities for the attacking team to exploit space and time on the ball. In this context, running less without the ball may reflect tactical passivity, fatigue, or poor defensive organization—all of which can contribute to conceding goals. Taken together, these findings emphasize the interactive and context-dependent nature of physical performance in football. It is not merely the absolute values of physical output that matter, but how they are expressed relative to the opponent's behavior. This supports the study's dual-team analytical approach, which considers both the reference team and their opponents to better understand match dynamics.

However, some limitations must be acknowledged. Many of the external load variables are interrelated. This collinearity may affect the precision of the statistical models and hinder interpretation

of individual variable effects. Therefore, this study should be viewed as a preliminary analysis. Future research should aim to refine the selection of variables, focusing on those most relevant to performance outcomes while minimizing redundancy. Additionally, the model used in this study does not account for the full complexity of match events, including tactical formations, player roles, or situational factors (e.g., scoreline, match phase). Integrating physical data with tactical and contextual variables could provide a more holistic understanding of performance.

In conclusion, this study highlights the importance of high-intensity physical actions in possession and the defensive implications of reduced off-ball effort. These insights can inform training design, match preparation, and in-game decision-making for coaches and performance staff. Future work should aim to build on these findings by refining the analytical framework and exploring how physical performance interacts with tactical and technical dimensions of the game.

References

[1] Beato, M. and Drust, B. (2021) *Acceleration intensity is an important contributor to the external and internal training load demands of repeated sprint exercises in soccer players.* Res Sports Med. **29(1),** 67-76.

[2] Beato, M., Drust, B. and Iacono, A.D. (2021) *Implementing High-speed Running and Sprinting Training in Professional Soccer.* Int J Sports Med. **42(4)**, 295-299.

[3] Beato, M., Youngs, A. and Costin, A.J. (2024) *The Analysis of Physical Performance During Official Competitions in Professional English Football: Do Positions, Game Locations, and Results Influence Players' Game Demands?* J Strength Cond Res. **38(5),** 226-234.

[4] Chaize, C., Allen, M. and Beato, M. (2024) *Physical Performance Is Affected by Players' Position, Game Location, and Substitutions During Official Competitions in Professional Championship English Football*. J Strength Cond Res. **38(12),** 744-753.

[5] Dawson, L., Beato, M., Devereux, G. and McErlain-Naylor, S.A. (2024) *A Review of the Validity and Reliability of Accelerometer-Based Metrics From Upper Back-Mounted GNSS Player Tracking Systems for Athlete Training Load Monitoring.* J Strength Cond Res. **38(8),** 459-474.

[6] Gualtieri, A., Rampinini, E., Dello Iacono, A. and Beato, M. (2023). *High-speed running and sprinting in professional adult soccer: current thresholds definition, match demands and training strategies*. *A systematic review.* Frontiers in Sports and Active Living, **5.**

[7] Gualtieri, A., Angonese, M., Maddiotto, M., Rampinini, E., Ferrari Bravo, D. and Beato, M. (2023) *Analysis of the Most Intense Periods During Elite Soccer Matches: Effect of Game Location and Playing Position.* Int J Sports Physiol Perform, **22**, 1-7.

[8] McGregor, R., Anderson, L., Weston, M., Brownlee, T. and Drust, B. (2024) *Intensity Gradients: A Novel Method for Interpreting External Loads in Football.* Int J Sports Physiol Perform. **19(8),** 829-832.

[9] Polglaze, T. and Hoppe, M. W. (2019). *Metabolic power: A step in the right direction for team sports.* International journal of sports physiology and performance*,* **14(3),** 407-411.

[10] Savoia, C., Laterza, F., Lucadamo, A., Manzi, V., Azzone, V., Pullinger, S. A., Beattie, C.E., Bertollo, M. and Pompa, D. (2024). *The Relationship Between Playing Formations, Team Ranking, and Physical Performance in the Serie A Soccer League.* Sports, **12(11), 286.**

[11] Silva, H., Nakamura, F.Y., Beato, M. and Marcelino, R. (2023) *Acceleration and deceleration demands during training sessions in football: a systematic review*. Sci Med Footb, **7(3),** 198-213.

[12] Venzke, J., Weber, H., Schlipsing, M., Salmen, J. and Platen, P. (2023). *Metabolic power and energy expenditure in the German Bundesliga*. Frontiers in Physiology*,* **14.**

Detection of front-door and back-door pitches in baseball and the characteristics that make them effective

Takumi Miura[1,*], Keisuke Fujii[1,2,**]

[1] Graduate School of Informatics, Nagoya University, Nagoya, Aichi, Japan.
[2] RIKEN Center for Advanced Intelligence Project, Osaka, Osaka, Japan.
[*]miura.takumi@g.sp.m.is.nagoya-u.ac.jp
[**]fujii@i.nagoya-u.ac.jp

Abstract

Front-door and back-door pitches (hereinafter referred to as "door-type pitches") in baseball refer to laterally breaking balls that move from outside to inside the strike zone. Door-type pitches often induce called strikes or weak contact, but they carry the risk of hard hits. However, there are no clear criteria for detecting door-type pitches and their effectiveness has not been verified. This study aims to clarify what requirements make door-type pitches effective in Nippon Professional Baseball (NPB). First, we used data from MLB to construct a machine-learning model that estimates the amount of pitch movement, allowing us to detect door-type pitches in NPB since NPB data did not include it. Next, we tested the effectiveness of door-type pitches and analyzed the relationship between the characteristics of pitchers and pitches and the test results. The results suggest that some pitches that induce field outs may be effective when used as door-type pitches and some slow front-door pitches may be ineffective. These findings can help players and coaches refine and evaluate the decision-making of their pitching strategies.

1 Introduction

In baseball, pitching strategy - the pitcher's choice of pitch type and pitch location to get the batter out - is an important element. Front-door and back-door pitches (hereinafter referred to as "door-type pitches") are two pitching strategies. Door-type pitches are laterally breaking balls that move from outside to inside the strike zone [6]. Pitches thrown close to the batter (inside corner) are called front door pitches, and those thrown far away (outside corner) are called back door pitches. The batter mistakes door-type pitches for pitches in the ball zone, and they can induce called strikes and weak contact caused by late swings. However, door-type pitches are more likely to be careless pitches because they are breaking balls thrown in the strike zone. Therefore, door-type pitches carry the risk of being thrown near the center of the strike zone, resulting in hard hits, and far from the strike zone, resulting in obvious balls or hit by pitches. It is difficult to judge whether the choice of door-type pitches is effective in getting batters out because door-type pitches have both advantages and disadvantages.

There have been many studies of baseball pitching strategies [1] [3]. However, there are no clear criteria for detecting door-type pitches, and their effectiveness has not been verified. The reasons for this are considered to be: (1) data indicating where the pitcher tried to throw the pitch is necessary to determine whether door-type pitches were thrown intentionally, (2) data on the amount of pitch movement is necessary to detect door-type pitches, and (3) statistical discussion is difficult because door-type pitches are not frequently used.

This study aims to detect door-type pitches using data from NPB and to clarify what requirements make door-type pitches effective. The contributions of this study are as follows: (1) it conducted a pioneering study of door-type pitches, for which there are few related studies, (2) it solved problems in research on door-type pitches by using data on catcher's stance positions in NPB, estimating the amount of pitch movement using machine learning models that use Major League Baseball (MLB) data, and testing the effect using permutation tests, which have fewer assumptions than traditional parametric tests, and (3) it clarified what requirements make door-type pitches effective in NPB.

2 Methods

2.1 Dataset

The NPB data used in this study come from 3 years (2021-2023) of official NPB games, which are the Central League, the Pacific League, and the Central/Pacific League exchange games, provided by Data Stadium Inc. This data includes catcher's stance positions for each pitch, but does not include the amount of pitch movement. This data was used to detect door-type pitches and test their effectiveness.

The MLB data used in this study come from 10 years (2015-2024) of MLB regular season data available at Baseball Savant (`https://baseballsavant.mlb.com`). This data includes the amount of pitch movement for each pitch, but does not include the catcher's stance positions. To obtain this data, the Python library "pybaseball" was used. This data was used to estimate the amount of pitch movement.

2.2 Estimation of the amount of pitch movement

This study requires data on the amount of pitch movement to detect door-type pitches; however, the NPB data lacked this information. Therefore, MLB data from 2015-2024 were used to estimate the pitch movement.

First, for pitchers with at least 100 pitches per season, we calculated feature scores and average amounts of lateral pitch movement by pitch type, excluding incomplete data. Due to the difference in classification between the NPB and MLB data, pitch types were grouped based on Table 1. Pitch types not listed were excluded due to the definition of pitch movement or small sample size. The feature scores (57 dimensions) consisted of the pitcher's handedness, speed, speed gap from the fastball, swinging strike rate, rate of pitches thrown in the strike zone, pitch type usage rate among all pitches (by batter handedness), and rate of pitches thrown in each of the 25 pitch location zones (by batter handedness). For left-handed pitchers, the amount of lateral pitch movement and the left-right relationship of each feature scores were reversed to simplify the distribution of feature scores and the estimated models. The amount of lateral pitch movement was defined as the difference from the 4-seam fastball. Therefore, pitchers without a 4-seam fastball are excluded. The amount of vertical pitch movement was not estimated as it was not used to detect door-type pitches. The sample sizes obtained by the above pretreatment are shown in Table 1.

Machine learning models (XGBoost, LightGBM, CatBoost) were constructed for each pitch type group, with the feature scores as input and the amount of lateral pitch movement as output. The best-performing model on the 2024 data was used for the analysis. The models were trained on 80 % of the MLB data from

2015-2023, with 20 % used for early stopping and hyperparameter tuning via Optuna, using RMSE as the loss function. The Python libraries XGBoost, LightGBM, and CatBoost were used.

The final model estimated the amount of lateral pitch movement for NPB pitchers by inputting their feature scores calculated from the entire NPB dataset to compensate for the smaller sample size due to the smaller number of games and fewer breaking balls.

2.3 Detection of door-type pitches

This study aims to evaluate the effectiveness of door-type pitches in getting batters out. Therefore, detection was based on the intentions of the pitcher and catcher, not the trajectory of the pitch. To the best of our knowledge, there are no previous studies that have detected door-type pitches. In this study, front-door/back-door pitches were defined using the following criteria, incorporating the estimated amount of pitch movement: (1) The catcher's stance position is at the inside/outside corner relative to the center of the strike zone; (2) The pitch breaks from inside/outside to outside/inside; (3) The catcher's stance position minus the amount of lateral pitch movement is at least 1.5 ball lengths (11.20 cm) away from the strike zone boundary. These criteria quantify the concept described in Chapter 1.

2.4 Evaluation of pitches

Pitch outcome evaluation was conducted in this study to analyze the effects of door-type pitches. NPB data were divided into 288 game situations based on count, outs, and runners, and the average runs scored from each situation until the end of the inning were calculated. This is referred to as the run expectancy 288 (RE288). The outcome of each pitch and change in RE288 were calculated. The average change in RE288 for each outcome is referred to as the linear weights (LWTS) [5], which serves as the evaluation index for pitches in this study. A lower LWTS indicates a more effective pitch.

2.5 Test of effectiveness of door-type pitches

The combinations of pitchers, pitch types, and batter handedness in NPB with at least 10 door-type pitches, at least 10 non-door-type pitches and at least 100 total pitches were selected for the test. The sample size for each pitch type is shown in Table 2. A one-tailed permutation test (1 % significance level) was used to compare the mean LWTS between door-type pitches and non-door-type pitches, approximated using 10,000 Monte Carlo permutations due to computational limitations. For each combination with a significant difference, pitch characteristics were investigated to clarify the criteria for the effective use of door-type pitches.

2.6 Evaluation of commanding ability

This study requires an index to evaluate pitch command in order to analyze how pitch command affects the effectiveness of door-type pitches. In this study, the area of the 95% confidence ellipse was used, assuming that the pitch error distribution between the catcher's stance positions and the actual pitch positions follows a two-dimensional normal distribution. The mean of the pitch error was not considered because the catcher could have adjusted the position of his stance by calculating the pitch error backward from the actual position where the catcher intended. In addition, in the study by Shinya et al.[4], the error between the target and actual pitching positions was assumed to follow a two-dimensional normal distribution, and the confidence ellipse was a tilted ellipse rather than a perfect circle. Based on the above, it was considered more appropriate to evaluate the variance of pitch errors under the assumption of a normal distribution, rather than the distance or mean value of the errors. Therefore, in this study, the area of the 95% confidence ellipse obtained from the

Table 1: The group of pitch types and the sample sizes.

NPB	MLB		
pitch type	pitch type	sample size [pitchers] 2015-2023	2024
Shoot Sinker	Sinker Screwball	2215	257
Changeup	Changeup	1877	206
Forkball	Split-finger Forkball	287	56
Cutter	Cutter	1003	145
Curveball	Curveball Knuckle Curve Slow Curve Slurve	1837	177
Slider	Slider Sweeper	2919	441

Table 2: The sample size for each pitch type on NPB

pitch type	sample size [pitchers] front-door	back-door
Shoot	13	16
Sinker	0	3
Changeup	0	8
Forkball	0	0
Cutter	10	53
Curveball	22	87
Slider	42	136

variance-covariance matrix of the pitch error distribution was used as an index to evaluate pitch command ability.

3 Results and Discussion

3.1 Estimation results

First, the RMSEs of the amount of lateral pitch movement estimated by the XGBoost, LightGBM, and CatBoost models using data for all MLB pitchers in 2024 were 8.46, 8.52, and 8.45 [cm], respectively, while the RMSEs for only pitchers not included in the training data were 8.63, 8.66, and 8.60 [cm], respectively. Scatter plots of the estimated and actual amounts of pitch movement by XGBoost and CatBoost for all data are shown in Figures 1 and 2. For the three models, CatBoost has the smallest RMSE, but it is not

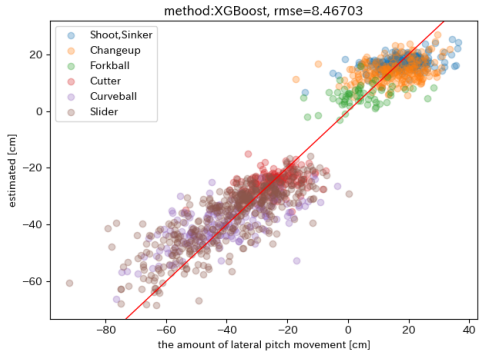


Figure 1: Scatter plot of the amount of lateral pitch movement and the estimated value by XGBoost.

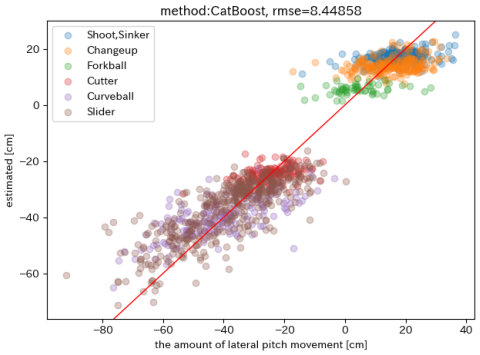


Figure 2: Scatter plot of the amount of lateral pitch movement and the estimated value by CatBoost.

much different from that of XGBoost, and in the scatter plots, XGBoost has a larger within-class variance of the estimated values than CatBoost. This was thought to better reflect differences in the amount of pitch movement by each pitcher. Therefore, the XGBoost estimation model was adopted.

Table 3: The combinations of pitchers and pitch type, and Cohen's d for which there was a significant difference in the mean LWTS between door-type and non-door-type pitches.

	effects	pitcher	pitch type	NP	door-type NP	Cohen's d
front-door	positive	Juri Hara	Slider	290	20	0.510
		Hiroya Miyagi	Slider	1087	90	0.270
	negative	Yoshinobu Yamamoto	Curveball	582	44	0.401
		Takahiro Nishimura	Slider	145	18	0.890
		Katsunori Hirai	Slider	745	46	0.457
back-door	positive	Frank Herrmann	Curveball	177	154	0.683
		Kenya Suzuki	Slider	360	227	0.310
	negative	Naoyuki Uwasawa	Curveball	493	416	0.309
		Yuki Nishi	Curveball	120	101	0.811
		Kodai Senga	Cutter	372	94	0.145

3.2 Pitchers and pitch types whose effectiveness of door-type pitches has a significant difference

The combinations of pitchers and pitch types for which there was a significant difference in the mean LWTS between door-type and non-door-type pitches, along with Cohen's d, are shown in Table 3. The following are two suggestions obtained as a result of this study. For pitchers who threw at least 100 pitches over three years, the mean, standard deviation, minimum, median, and maximum values of the confidence ellipse area for the pitch command index were 8441.8, 1211.0, 5505.1, 8430.0, and 13678.6 [cm^2], respectively.

3.2.1 Suggestions 1: Hara's slider

Hara is a right-handed overthrowing pitcher with a 4.07 the earned run average (ERA), which is calculated as $(earned\_runs \times 9)/innings\_pitched$, over 31 games. A slider is a pitch that breaks in the opposite direction of the pitcher's handedness. Therefore, if the batter is a right-handed hitter, it is likely to be a front-door pitch. His non-front-door slider had fewer swinging strikes (13.7%) and more field outs (15.9%) than the average (swinging strike: 14.7%, field outs: 11.2%), likely due to slower pitching speed(126.8 km/h, average: 129.8 km/h), larger speed gap from the fastball (17.1 km/h, average: 15.9 km/h) and greater amount of pitch movement (42.6 cm, average: 36.9 cm), making it less likely to be mistaken by the batter for another pitch type and thus increasing the probability of making contact, but more difficult to hit the sweet spot. Front-door sliders resulted in more fouls (35.0%), field outs (40.0%) and fewer balls (5.0%) compared to non-front-door (fouls: 11.5%, field outs: 15.9%, balls: 34.1%), likely due to the pitch error distribution. Shinya et al. [4] showed a correlation between the direction of the major axis of the confidence ellipse of the pitch error distribution and the angle of the pitcher's arm when pitching. This suggests that right-handed overthrowing pitchers are more likely to throw pitches in the upper right and lower left directions from their perspective. Therefore, if Hara tries to throw low outside, some pitches will be thrown into the obvious ball zone. On the other hand, if he tries to throw low inside, most of pitches will be inside the strike zone or near the border and are less likely to be taken. His lower pitch command (confidence ellipse area: 8835.7 cm^2 vs 7729.4 cm^2 average) reinforced this trend. Unlike typical door-type pitches, his front-door sliders didn't lead to more hits, likely because his high field-out rate reduced the risk of hard hits. This tendency was also observed for Miyagi's slider. These results suggest that some pitches that induce field outs may be effective when used as door-type pitches because the risk of hard hits, which is a disadvantage of the door-type pitch, is low.

3.2.2 Suggestions 2: Yamamoto's curveball

Yamamoto was a right-handed overthrowing pitcher with a 1.44 ERA over 75 games. A curveball is a slow pitch that breaks and falls in the opposite direction of the pitcher's handedness. Therefore, if the batter was right-handed, it was likely to be a front-door curveball. His non-front-door curveball was effective with a higher whiff rate (12.3%) than the average (9.2%), likely due to a higher pitching speed (124.1 km/h vs 117.9 km/h), a larger speed gap from the fastball (28.2 km/h vs 27.6 km/h), larger amount of pitch movement (44.5 cm vs 37.5 cm), and better pitch command (7924.7 cm^2 vs 9351.2 cm^2), and the probability that the batter could respond to the break was lower. In contrast, front-door pitches had more called strikes (45.5%) but also more hits (11.4%) compared to non-front-door ones (25.1% called strikes, 3.3% hits). This may be due to the intercept point and timing. The intercept point is the position where the bat and the ball meet. In general, it is said that strong batted balls can be hit by placing the intercept point a little closer to the pitcher on inside corner pitches [2]. Also, pitches that are slower than the fastball often cause the batter to swing at them earlier because batters usually adjust their timing with the fastball. This resulted in a tendency to hit the ball harder, which may have increased the number of hits. This tendency was also seen with Nishimura's and Hirai's sliders. These results suggest that some slow front-door pitches may be ineffective because they are more likely to be hit hard when the batter swings at them.

4 Conclusion

This study attempted to clarify the requirements for door-type pitch effectiveness in NPB by investigating the characteristics of door-type pitches that are effective or ineffective. The results suggest two insights described in Sections 3.2.1 and 3.2.2. The findings are intended to help players and coaches make decisions about using door-type pitches based on the suggestions. However, limitations include: (1) a lack of comparison with other pitch types that pitchers have, (2) a lack of consideration for the previous pitch and the characteristics of batters, (3) a lack of clarity regarding the quantitative requirements for effectiveness, and (4) a lack of verification of the results of estimating the amount of pitch movement.

Acknowledgments

The NPB data used in this study was provided by the "Research Organization of Information and Systems, The Institute of Statistical Mathematics" and "Data Stadium Inc.". This study is supported by JSPS KAKENHI Grant Number 23H03282.

References

[1] J. R. Bock. Pitch sequence complexity and long-term pitcher performance. *Sports*, 3(1):40–55, 2015.

[2] B. Clemens. Can "hard in and soft away" make your troubles go away? `https://blogs.fangraphs.com/can-hard-in-and-soft-away-make-your-troubles-go-away/`,(reference:2025-02-09).

[3] H. Nakahara, K. Takeda, and K. Fujii. Pitching strategy evaluation via stratified analysis using propensity score. *Journal of Quantitative Analysis in Sports*, 19(2):91–102, 2023.

[4] M. Shinya, S. Tsuchiya, Y. Yamada, K. Nakazawa, K. Kudo, and S. Oda. Pitching form determines probabilistic structure of errors in pitch location. *Journal of Sports Sciences*, 35:1–6, 01 2017.

[5] P. Slowinski. Linear weights, 05 2010. `https://library.fangraphs.com/principles/linear-weights/`,(reference:2025-05-14).

[6] WeeklyBaseballOnline. Explained with illustrations! what are front doors and back doors?, 03 2015. `https://column.sp.baseball.findfriends.jp/?pid=column_detail&id=001-20130617-09`,(reference:2025-05-01).

Football Analysis System using Computer Vision and Machine Learning

Nikhil Sushil Muneshwar*, Xing Liang ** and Gordon Hunter***

School of Computer Science and Mathematics, Kingston University, KT1 2EE, UK
*nikhilmuneshwar05@gmail.com
** x.liang@kingston.ac.uk
*** g.hunter@kingston.ac.uk

Abstract

Advanced software for analysing player performance and team tactics is now widely used in TV sports coverage, enabling pundits and coaches to provide detailed insights during or after matches. While systems like Hawk-Eye rely on high-frame-rate cameras and multi-view triangulation, our work presents a cost-effective alternative for tracking players, officials, and the ball in standard frame-rate soccer footage. Making use of YOLOv11, an object detection model derived from the GoogleNet Convolutional Neural Network Architecture, and enhanced through open-source transfer learning, our system reliably distinguishes between teams, referees, and the ball. By incorporating transformational geometry, optical flow, perspective transformation, we compensate for camera motion and generate player statistics such as speed and distance covered. Though less sophisticated than broadcast-grade systems, our method performs well on professional match footage, making it viable for lower-tier clubs, semi-professional teams, or fan channels with limited technological resources.

1 Introduction

Football analysis has undergone a paradigm shift in recent decades, evolving from rudimentary observational techniques to a sophisticated, data-driven discipline that integrates computer vision, machine learning, and artificial intelligence. At its core, football analysis involves the systematic evaluation of matches, players, and teams through qualitative and quantitative methods to uncover performance insights, optimize tactics, and enhance decision-making. Historically, analysts relied on manual notational methods—charting player movements and key events by hand—a process that was both time-consuming and prone to subjectivity. However, the advent of advanced tracking technologies, such as optical camera systems (e.g., Hawk-Eye [1], TRACAB [2]) and wearable GPS devices, has revolutionized the field, enabling real-time data collection on player positioning, sprint metrics, and ball trajectories precision.

Despite these advancements, a significant disparity persists in access to such technologies. Elite clubs and top-tier leagues invest heavily in proprietary systems, while smaller clubs, semi-professional teams, and grassroots organizations are often excluded due to prohibitive costs. For instance, installing a multi-camera tracking system like STATSports' Venue solution [3] can exceed (GBP) £ 100,000 annually, with additional expenses for maintenance and specialized personnel. This economic barrier exacerbates competitive imbalances, as resource-constrained teams lack the tools to analyse opponents, scout talent, or refine tactics with the same granularity as wealthier counterparts. While recent modern advancements in computer vision and neural networks have already begun to overcome some of these limitations. For example, Convolutional Neural Networks (CNNs) have achieved over 90% accuracy for automated player detection [4]. However, most state-of-the-art tools remain inaccessible to smaller clubs.

To bridge this inequity, we propose an affordable, AI-powered football analysis system that extracts high-fidelity insights from standard broadcast footage—*a ubiquitous and low-cost data source*. Unlike existing solutions that depend on expensive hardware, our approach centres on the optimisation of *YOLOv11*—a *lightweight yet powerful object detection model* tailored for football-specific applications, combined with *perspective transformation*, and *optical flow algorithms* [5] to track players, officials, and the ball while compensating for camera motion. The system democratises access to advanced analytics by offering: (1) Real-time player and ball tracking without reliance on sensor

arrays. (2) Tactical metrics (e.g., sprint speeds, positional heatmaps) derived from 2D video. (3) Cost efficiency, reducing dependency on capital-intensive infrastructure. Our system achieves a balance between speed (45 frames per second) and precision (88.3% tracking accuracy in preliminary tests), enabling *real-time analysis without costly specialised hardware*, and providing *a lightweight alternative to resource-heavy deep learning systems*.

Technical Challenges and Innovations
A core challenge in video-based analysis is distinguishing players from dynamic backgrounds, especially when jersey colours blend with the pitch (e.g., green kits on grass) or background. Early methods relied on colour thresholding, which failed under varying lighting conditions. Our system overcomes these issues by integrating advanced segmentation, tracking, and geometric mapping techniques, enabling robust and precise player tracking and performance analysis from broadcast footage. Specifically, we utilise:

- *Advanced segmentation*: K-means clustering + YOLOv11 to segment players from the pitch.
- *Tracking*: Optical flow to stabilize tracking during camera panning or zooming.
- *Geometric mapping*: Perspective homography to map 2D broadcast coordinates to a standardised pitch model, enabling metric-based analysis (e.g., distances covered).

2 Previous Related Work

The use of computer vision, machine learning (ML), and artificial intelligence (AI) in football analytics has seen significant progress over recent years, enabling advanced capabilities in player tracking, event detection, and tactical analysis. Early methods for object detection in sports relied on background subtraction and color-based segmentation, which often struggled under real-world conditions such as occlusion and variable lighting [6, 7]. With the rise of deep learning, convolutional neural networks (CNNs) like YOLO and Faster R-CNN have become the standard for accurate, real-time detection of players and the ball [8, 9].

For multi-object tracking, traditional techniques such as Kalman filters have been replaced by deep learning-based approaches such as DeepSORT [10] and ByteTrack [11]. These allow robust tracking of players over time, even with frequent occlusions and fast motion. This tracking data is essential for constructing heatmaps, movement patterns, and formation analysis.

Event detection in football—such as detecting goals, fouls, or passes—has benefited from temporal models like Long Short-Term Memory (LSTM) models and Transformers, trained on annotated video datasets to identify key match events and generate summaries [12]. Datasets such as SoccerNet [13] have played a critical role in supporting this research.

Several commercial systems have advanced the field. TRACAB, for example, uses a multi-camera setup to provide real-time 3D positioning of players [2], while Second Spectrum exploits ML to deliver tactical insights for teams and broadcasters [14].

Despite these advancements, challenges remain, including handling occlusions, generalizing across different camera angles, and processing unstructured video. This paper builds on prior work by proposing an integrated system that combines deep learning-based object detection, spatio-temporal tracking, and intelligent event classification to provide a comprehensive football analysis platform using standard resolution and frame rate TV footage.

3 Analysis of the Game

Our system detects, tracks, and identifies the players by their respective teams. It also maps the positions from the input captured by a camera during a broadcast game to the absolute positions on the football field which is viewed from a bird's eye view. Further details of the implementation can be found in Muneshwar [15].

The input videos are taken from a German Bundesliga game, which comes from a Kaggle competition sponsored by the Bundesliga [16].

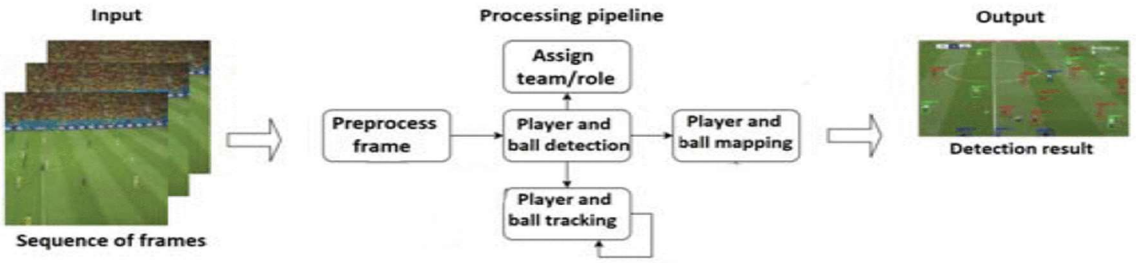


Figure 1: Processing "Pipeline" of our system

3.1 Ball and Player Detection

The object detection component of the system is based on the YOLO (You Only Look Once) architecture, particularly version 11 [5]. YOLOv11 builds upon previous versions with significant improvements in both performance and efficiency. It replaces the inception modules of GoogLeNet [17, 18] with a combination of 1×1 and 3×3 convolutional filters. The network consists of 24 convolutional layers followed by two fully connected layers, using Leaky Rectified Linear Units (ReLU) for activation functions but a linear activation function at the output neurons.
The model was initially trained on the ImageNet dataset [19] and then fine-tuned for object detection using the VOC Pascal datasets [20]. YOLOv11 outperforms its predecessors (YOLOv8, v9, v10) and other models such as Faster R-CNN [21] in metrics such as mean Average Precision at 50% Intersection over Union (mAP@0.5), mean Average Precision for Intersection Over Union between 50% and 90% (mAP@0.5–0.95), precision, recall, and inference speed [22].

3.2 Custom Training and Detection Optimization

To ensure precise object detection in football videos, the model was retrained using a custom dataset from Roboflow [23]. This fine-tuning focused exclusively on players, referees, and the football, removing distractions such as stewards and managers that had previously introduced noise. The training involved:

- Input image resizing to 640×640 pixels.
- 100 training epochs.
- Use of the best-performing weights for inference.

Post-training, the model effectively detects only relevant classes.
Performance:

- Achieves 92.1% mAP@0.5 on the test set.
- Faster than Faster R-CNN (13.5 ms/frame versus 63.8 ms/frame).

Training:

- Fine-tuned on the Roboflow Football-Players-Detection dataset (100 epochs, 640×640 pixel resolution).
- Limitation: Lower recall for the ball (22.9%) due to its smaller size and occlusion.

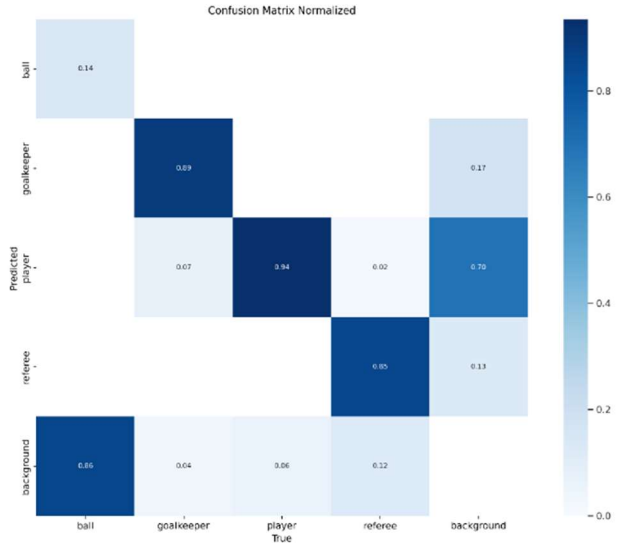


Figure 3: Confusion Matrix after training

```
Validating runs/detect/train/weights/best.pt...
Ultralytics 8.3.47 🚀 Python-3.10.12 torch-2.5.1+cu121 CUDA:0 (Tesla T4, 15102MiB)
YOLO11l summary (fused): 464 layers, 25,282,396 parameters, 0 gradients, 86.6 GFLOPs
                 Class     Images  Instances      Box(P          R      mAP50  mAP50-95): 100% 2/2 [00:01<00:00,  1.56it/s]
                   all         38        905      0.845      0.766      0.828      0.595
                  ball         35         35      0.812      0.229      0.403      0.189
            goalkeeper         27         27      0.686       0.97      0.959      0.739
                player         38        754       0.96      0.947      0.984       0.79
               referee         38         89      0.921      0.919      0.965      0.662
Speed: 0.1ms preprocess, 11.9ms inference, 0.0ms loss, 2.8ms postprocess per image
Results saved to runs/detect/train
```

Figure 4: Training Results Summary (R is the Recall)

3.3 Object Tracking with ByteTrack

Once objects are detected, YOLOv11 outputs are passed into a tracking system using ByteTrack [24], a powerful multi-object tracker that assigns unique IDs to objects across frames. This ensures continuity in tracking players, referees, and the ball.
Key techniques include:

- Bounding box centre and foot position calculations for spatial analysis.
- Batch detection for computational efficiency.
- Linear interpolation for handling missing ball detections, ensuring trajectory continuity.
- Visual annotations using OpenCV [25] for real-time feedback.

3.4 Team Assignment Using K-Means Clustering

To differentiate between teams, the project applied K-Means clustering to analyse jersey colours. The system crops player images and applies clustering to identify dominant colours in the top half of the jersey. These colours were used to train a K-Means model, which then assigns each player to a team.
Advantages of this approach are:

- Robust even in crowded or dynamic scenes.
- Works without explicit colour labelling.
- Computationally efficient and hardware friendly.

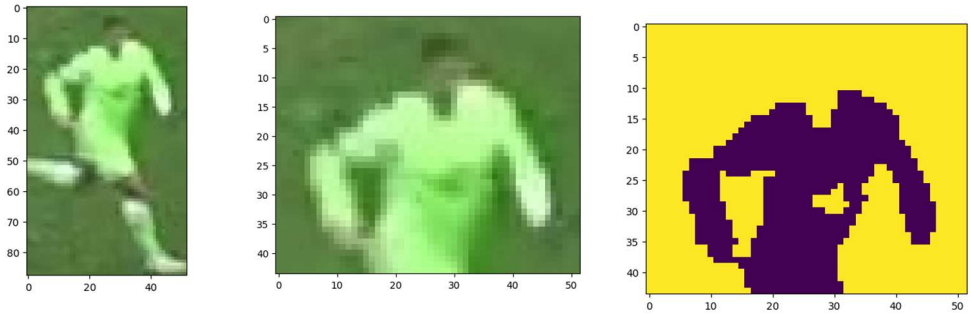


Figure 5: Feature Reduction and player silhouetting using K-means clustering;
(a) original image of player, (b) reduced features (torso only), (c) silhouette of torso

3.5 Ball Possession Assignment

The system determines ball possession by identifying the closest player to the ball using Euclidean distance calculations between the ball's centre and both feet of each player. A threshold maximum distance (70 pixels) filters which player is in possession. If no player meets the threshold, possession is temporarily unassigned. Possession statistics are stored frame-by-frame, enabling long-term analysis.

3.6 Camera Motion Estimation

To account for camera panning and movement, which can distort player speed and distance calculations, the system estimates camera motion using:

- Shi-Tomasi Corner Detection: Finds stable feature points [26].
- Lucas-Kanade Optical Flow: Tracks movement between frames [27].

Motion vectors were calculated and used to adjust player and ball positions, ensuring measurements were relative to the field, not to the camera's motion.

3.7 Perspective Transformation

Perspective distortion in wide-angle or high-angle footage can lead to inaccurate spatial measurements. A homography matrix was computed to transform the pixel coordinates of players and the ball into real-world field coordinates using four manually selected pitch points. This transformation ensures:

- Consistent player positions.
- Accurate distance and speed analysis.

3.8 Speed and Distance Estimation

Using the real-world coordinates from the perspective transformer, the system uses classical Newtonian dynamics to calculate:

- Distance: Between positions in consecutive frames using Euclidean distance.
- Speed: As distance divided by time (based on the frame rate of the video).

These metrics are overlayed on the video footage, positioned just above each player's feet for easy viewing:

- Speed in km/h.
- Distance travelled in metres.

4 Results, Discussion and Conclusions

The results demonstrate the effectiveness of the YOLOv11-based pipeline in achieving accurate and real-time football match analysis. After retraining the model with a custom Roboflow dataset, object detection performance significantly improved by focusing exclusively on players, referees, and the ball while eliminating irrelevant detections. The model achieved high precision (0.845) and recall (0.766), with excellent class-specific metrics, particularly for players (mAP@0.5: 0.984). Object tracking using ByteTrack maintained consistent identities across frames, and ball possession was accurately assigned using spatial proximity logic. Speed and distance metrics were successfully calculated using a perspective transformation and camera motion compensation, ensuring that all measurements reflected true on-field movements. Visual overlays, including possession statistics and player performance data, were effectively rendered on the video frames, resulting in a robust and interpretable analytical output suitable for performance evaluation and tactical analysis.

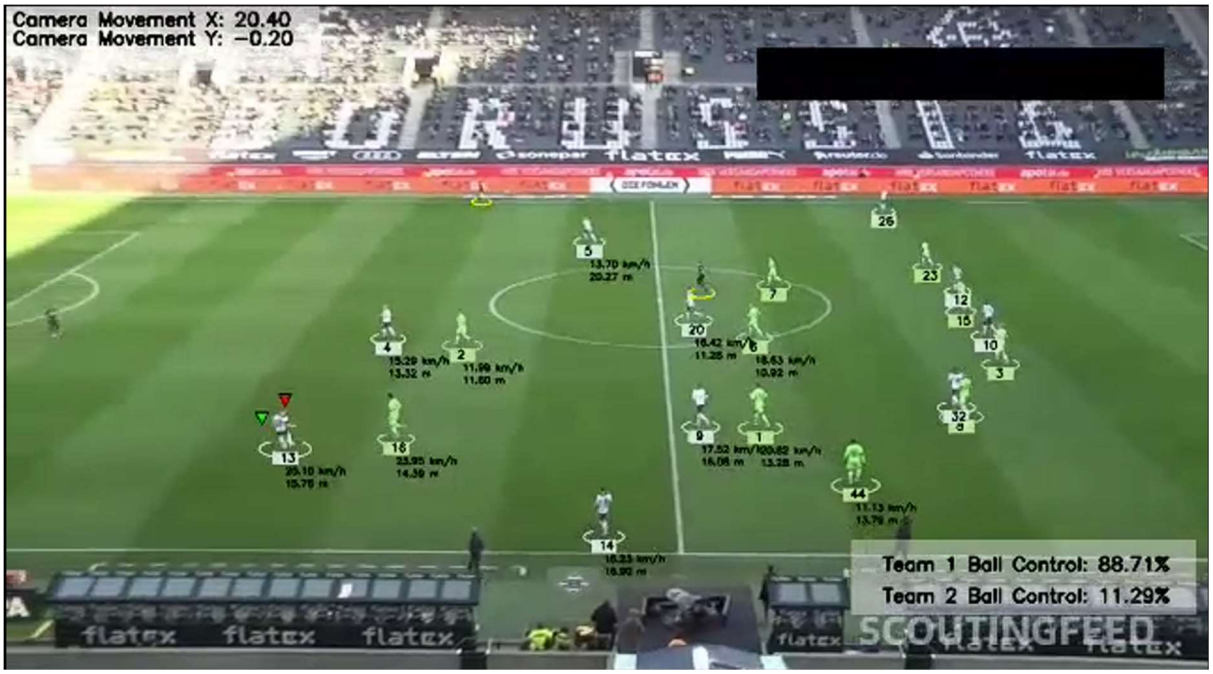


Figure 6:Still image from the final output Video

5 References

[1] Hawk-Eye Innovations (2025) https://www.hawkeyeinnovations.com/

[2] ChyronHego (2020) *TRACAB Player Tracking System*. Available at: https://chyronhego.com (Accessed: 27 May 2025).

[3] StatSport (2025) https://statsports.com/

[4] Bialkowski, A., Lucey, P., Carr, P., Yue, Y., Sridharan, S. and Matthews, I. (2014) "Large-Scale Analysis of Soccer Matches Using Spatiotemporal Tracking Data," 2014 IEEE International Conference on Data Mining, Shenzhen, China, 2014, pp. 725-730, doi: 10.1109/ICDM.2014.133.

[5] Ultralytics (2025) "YOLO v 11" https://docs.ultralytics.com/models/yolo11/

[6] Comaniciu, D. and Meer, P. (2002) 'Mean shift: A robust approach toward feature space analysis', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(5), pp. 603–619, https://doi.org/10.1109/34.1000236

[7] Zivkovic, Z. (2004) 'Improved adaptive Gaussian mixture model for background subtraction', Proceedings of the 17th International Conference on Pattern Recognition (ICPR), pp. 28–31. https://doi.org/10.1109/ICPR.2004.1333992

[8] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016) 'You only look once: Unified, real-time object detection', Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788. https://doi.org/10.1109/CVPR.2016.91

[9] Ren, S., He, K., Girshick, R. and Sun, J. (2015) 'Faster R-CNN: Towards real-time object detection with region proposal networks', Advances in Neural Information Processing Systems (NeurIPS), 28, pp. 91–99. Available at: https://papers.nips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html

[10] Wojke, N., Bewley, A. and Paulus, D. (2017) 'Simple online and realtime tracking with a deep association metric', 2017 IEEE International Conference on Image Processing (ICIP), pp. 3645–3649, https://doi.org/10.1109/ICIP.2017.8296962

[11] Zhang, Y. et al. (2022). ByteTrack: Multi-object Tracking by Associating Every Detection Box. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds) *Computer Vision – ECCV 2022*. Lecture Notes in Computer Science, vol 13682. Springer, Cham. https://doi.org/10.1007/978-3-031-20047-2_1

[12] Giancola, S., Amine, M., Dghaily, T. and Ghanem, B. (2018) 'SoccerNet: A scalable dataset for action spotting in soccer videos', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1711- 1721, https://doi.org/10.1109/CVPRW.2018.00217
[13] SoccerNet (2025) SoccerNet dataset https://www.soccer-net.org/
[14] Genius Sports (2022) "Genius Sports' Second Spectrum tracking technology approved by FIFA Quality Programme for EPTS" https://www.geniussports.com/newsroom/genius-sports-second-spectrum-tracking-technology-approved-by-fifa-quality-programme-for-epts/
[15] Muneshwar, N.S. (2025) "Football Analysis System using computer vision and Machine Learning", MSc project dissertation, Kingston University, U.K.
[16] Michalczyk, J., Maggie, Janetzke, M., Mücke, M.M., Holbrook, R. and Dick, U.(2022) DFL – "Bundesliga Data Shootout", https://www.kaggle.com/competitions/dfl-bundesliga-data-shootout
[17] Szegedy, C., Wei, L., Yangqing. J., Sermanet, P., et al. (2015). "Going deeper with convolutions". 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. pp. 1–9. arXiv:1409.4842
[18] Zhang, A., Lipton, Z., Li, M., Smola, A. J. (2024). "8.4. Multi-Branch Networks (GoogLeNet)". Dive into deep learning. Cambridge New York Port Melbourne New Delhi Singapore: Cambridge University Press. ISBN 978-1-009-38943-3.
[19] ImageNet (2021) https://www.image-net.org/
[20] PASCAL VOC project (2014) "PASCAL (Pattern Analysis, Statistical modelling And Computational Learning) Visual Object Classes Homepage" http://host.robots.ox.ac.uk/pascal/VOC/
[21] Ren, S., He, K., Girshick, R. & Sun, J. (2016) "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks" https://arxiv.org/abs/1506.01497v3
[22] Diwan, T., Anirudh, G. & Tembhurne, J.V. (2023). Object detection using YOLO: challenges, architectural successors, datasets and applications. Multimedia Tools and Applications, 82, 9243–9275. https://doi.org/10.1007/s11042-022-13644-y
[23] Roboflow (2022). "Football-Players-Detection Dataset", Roboflow Universe https://universe.roboflow.com/roboflow-jvuqo/football-players-detection-bzlaf.
[24] Zhang, Y. et al (2024) "ByteTrack: Multi-Object Tracking by Associating Every Detection Box", https://arxiv.org/abs/2110.06864v3
[25] OpenCV (2025) https://opencv.org
[26] Shi, J. and Tomasi, C. (1994). "Good Features to Track". Proceedings of 9th IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Springer. pp. 593–600.
[27] Lucas, B.D. and Kanade, T. (1981) "An iterative image registration technique with an application to stereo vision", Proceedings of Imaging Understanding Workshop, pp 121-130, https://cseweb.ucsd.edu/classes/sp02/cse252/lucaskanade81.pdf

Evaluating Soccer Player Movements Using the Attacker-Defender Model

Takuma Narizuka* and Issei Yamazaki**

*Faculty of Data Science, Rissho University, Kumagaya, Saitama, Japan, narizuka@ris.ac.jp
**Meiji Institute for Advanced Study of Mathematical Sciences, Meiji University, Nakano, Nakano-ku, Tokyo

Abstract

The present study investigates the attacker-defender (AD) model proposed by Brink et al. (2023), a motion model that describes the interactions between a ball carrier (attacker) and the nearest defender during ball possession. The model is based on the equations of motion for both players, incorporating resistance, goal-oriented force, and opponent-oriented force. It generates trajectories based on physically interpretable parameters. Although the AD model reproduces real dribbling trajectories well, previous studies have explored only a limited range of parameter values using small datasets. This study aims to (1) enhance parameter optimization by solving the AD model for one player with the opponent's actual trajectory fixed, (2) validate the model's applicability to a large dataset from 306 J1-League matches, and (3) demonstrate distinct playing styles of attackers and defenders based on the full range of optimized parameters.

1 Introduction

With the widespread availability of tracking and event data [1], the generation of short-term player trajectories has received increasing attention in football (soccer) analytics. Machine learning-based approaches [2] focus on predictive accuracy, while physics-based models [4] prioritize interpretability. The present study adopts a physics-based motion model, incorporating fundamental principles of player behavior.

Among one-dimensional motion models, the Keller model [3] has been widely used to analyze sprinting, where velocity evolves as $dv/dt = -v(t)/\tau + f$. However, the analysis of soccer player movements necessitates the use of two-dimensional models. A notable example is the Fujimura-Sugihara model [4], an extension of the Keller model, which has been validated through tracking data [5].

In a recent study, Brink et al. [6] proposed the Attacker-Defender (AD) model, which incorporates the interaction between a ball carrier (attacker) and the nearest defender into the Fujimura-Sugihara model. The model generates trajectories based on physically interpretable parameters and reproduces a wide range of dribbling movements. However, previous studies have primarily explored a limited range of parameter values using small datasets, and the model's applicability to large datasets remains underexplored.

The present study has three primary objectives. First, we enhance the parameter optimization process of the AD model. Specifically, we propose a method that solves the model for one player with the opponent's actual trajectory fixed. Second, we validate the model's applicability using a significantly larger dataset than those employed in previous studies. We analyze tracking and event data from 306 J1-League matches provided by DataStadium Inc., Japan. Finally, we quantitatively identify characteristic ball carriers (attackers) and their nearest defenders based on the optimized parameters of the AD model. By expanding the range of analyzed parameters, we provide new insights into the playing styles of attackers and defenders.

2 Methods

2.1 Dataset

We used data from 306 matches in the Japan Professional Football League (J1 League) during the 2023 season, which was provided by Data Stadium Inc., Japan. The dataset contains the absolute position coordinates (x, y) of all players, recorded every 0.04 seconds. The dataset also includes event data, such as timestamps for ball possession, passes, and shots.

For the analysis of dribbling situations, all instances of dribbling were extracted and the corresponding tracking data for the ball carrier and the nearest defender were identified. We further selected only the dribbling events that satisfied the following criteria: (i) the nearest defender did not change during the dribble, (ii) the duration of ball possession exceeded 0.5 seconds, and (iii) the linear distance traveled by both players exceeded 5 meters. Consequently, 30,788 instances of dribbling were obtained for parameter optimization, which is significantly larger than the 1,573 dribbling events analyzed in the previous study [6].

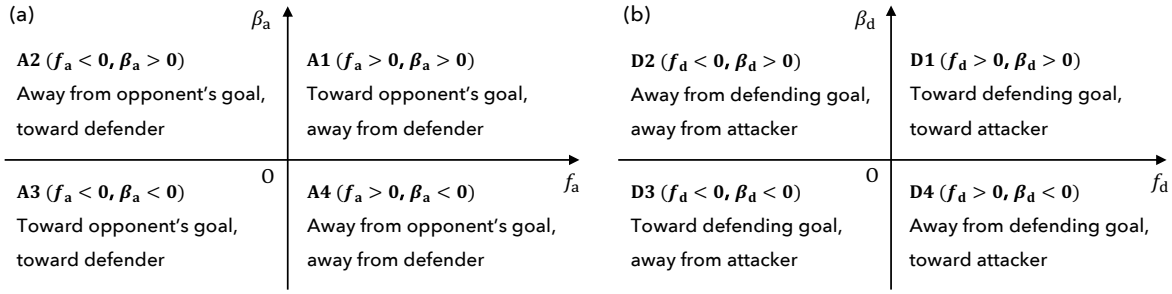


Figure 1: Expected attacker and defender motions based on the signs of (a) f_a and β_a, and (b) f_d and β_d.

2.2 Attacker-Defender Model

The Attacker-Defender (AD) model describes the movement of an attacker and a defender during a single dribbling event. Let $\vec{v}_a$ and $\vec{v}_d$ denote the velocity vectors of the attacker and defender, respectively. The equations of motion are as follows [6]:

$$\frac{d\vec{v}_a}{dt} = -\frac{1}{\tau_a}\vec{v}_a(t) + f_a \frac{\beta_a \vec{e}_{ag}(t) - \vec{e}_{ad}(t)}{\left\| \beta_a \vec{e}_{ag}(t) - \vec{e}_{ad}(t) \right\|}, \tag{1}$$

$$\frac{d\vec{v}_d}{dt} = -\frac{1}{\tau_d}\vec{v}_d(t) + f_d \frac{\beta_d \vec{e}_{dg}(t) + \vec{e}_{da}(t)}{\left\| \beta_d \vec{e}_{dg}(t) + \vec{e}_{da}(t) \right\|}. \tag{2}$$

Here, the first term on the right-hand side represents the resistance force proportional to velocity. The second term represents the driving force, directed toward both the goal and the opponent. Thus, the total force exerted by the attacker and defender is determined by the balance of these three components.

In the model parameters, f_a and f_d represent the maximum driving forces of the attacker and defender, respectively, while τ_a and τ_d denote the time constants required to reach their maximum speed. The parameters β_a and β_d control the relative weights of the goal-oriented and opponent-oriented driving forces.

The trajectories of the attacker and defender were obtained by numerically solving the coupled equations of motion with appropriate initial conditions. Note that these trajectories depend on the six parameters in the model. Figures 1(a) and (b) illustrate the classification of expected attacker and defender motions based on the signs of f_{a} and β_{a}, and f_{d} and β_{d}, respectively. While the previous study [6] limited its analysis to regions A1 and D1, the present study expands it to all four quadrants.

2.3 Parameter Optimization

Let $\vec{r}_{\mathrm{p}}(t)$ and $\vec{r}'_{\mathrm{p}}(t)$ denote the actual and simulated trajectories of player $\mathrm{p} \in \{\mathrm{a}, \mathrm{d}\}$, respectively. Following [6], the trajectory error is defined as

$$\varepsilon_{\mathrm{p}} = \frac{1}{T} \frac{\sum_{t=0}^{T-1} \|\vec{r}_{\mathrm{p}}(t) - \vec{r}'_{\mathrm{p}}(t)\|}{\sum_{t=1}^{T-1} \|\vec{r}_{\mathrm{p}}(t) - \vec{r}_{\mathrm{p}}(t-1)\|}, \tag{3}$$

where T is the number of frames in a single dribbling event. The error values ε_{a} and ε_{d} are functions of the model parameters. To obtain the optimal parameters for each dribbling instance, we minimize the total error defined as $\varepsilon = \varepsilon_{\mathrm{a}} + \varepsilon_{\mathrm{d}}$. Parameter optimization was performed for each dribbling event using the COBYLA algorithm from the SciPy library in Python. The parameters were constrained as follows: $f_{\mathrm{a}}, f_{\mathrm{d}} \in [-11.3, 11.3]$, $\tau_{\mathrm{a}}, \tau_{\mathrm{d}} \in [0.9, \infty]$, $\beta_{\mathrm{a}}, \beta_{\mathrm{d}} \in [-10, 10]$. For each event, N random initial parameter sets were generated, and the one yielding the lowest error ε was selected.

3 Results

3.1 Improving Parameter Optimization

We propose a new method that solves the equations of motion for the attacker and defender independently, using the actual trajectory of the other player. This enables the independent parameter optimization for attacker and defender. Consequently, the dimensions of parameter space are reduced, thereby significantly decreasing the computational cost.

To estimate optimized parameters of 30,788 dribbling events, Latin hypercube sampling was employed to generate $N = 100$ initial parameter sets. The set with the lowest error was selected for each event. We then retained events where both ε_{a} and ε_{d} were below 0.1. Consequently, 75% of events met this criterion using the original method, and 80% using the improved one. The proposed approach thus not only reduced the computational cost, but also increased the number of successful parameter estimates.

3.2 Attacker Characterization

A total of 30,788 attacker trajectories were categorized into four quadrants based on the signs of the optimized parameters, f_{a} and β_{a}. Table 1 lists the most frequently appearing attackers in each quadrant.

In region A1 ($f_{\mathrm{a}} > 0,\ \beta_{\mathrm{a}} > 0$), the AD model suggests that the attacker tends to move toward the opponent's goal and away from the defender. Two typical patterns were observed: maintaining a safe distance while advancing the ball, and drawing the defender in before releasing the ball (Fig. 2, A1). Center backs engaged in building-up play were frequently observed in this region.

In region A2 ($f_a < 0,\ \beta_a > 0$), the attacker is modeled to move away from the opponent's goal and toward the defender. The observed situations included maintaining possession under pressure from behind, and dribbling directly toward a nearby defender (Fig. 2, A2).

In region A3 ($f_a < 0,\ \beta_a < 0$), the parameter values imply movement toward both the opponent's goal and the defender. Typical cases involved direct one-on-one confrontations or forward progression toward the defender (Fig. 2, A3). The dribbling in this region corresponds to actions near scoring chances, which are crucial for evaluating attackers.

Finally, in region A4 ($f_a > 0,\ \beta_a < 0$), the model generates movement away from both the opponent's goal and the defender. The observed patterns included forward acceleration under rear pressure and backward dribbles that drew in the defender (Fig. 2, A4).

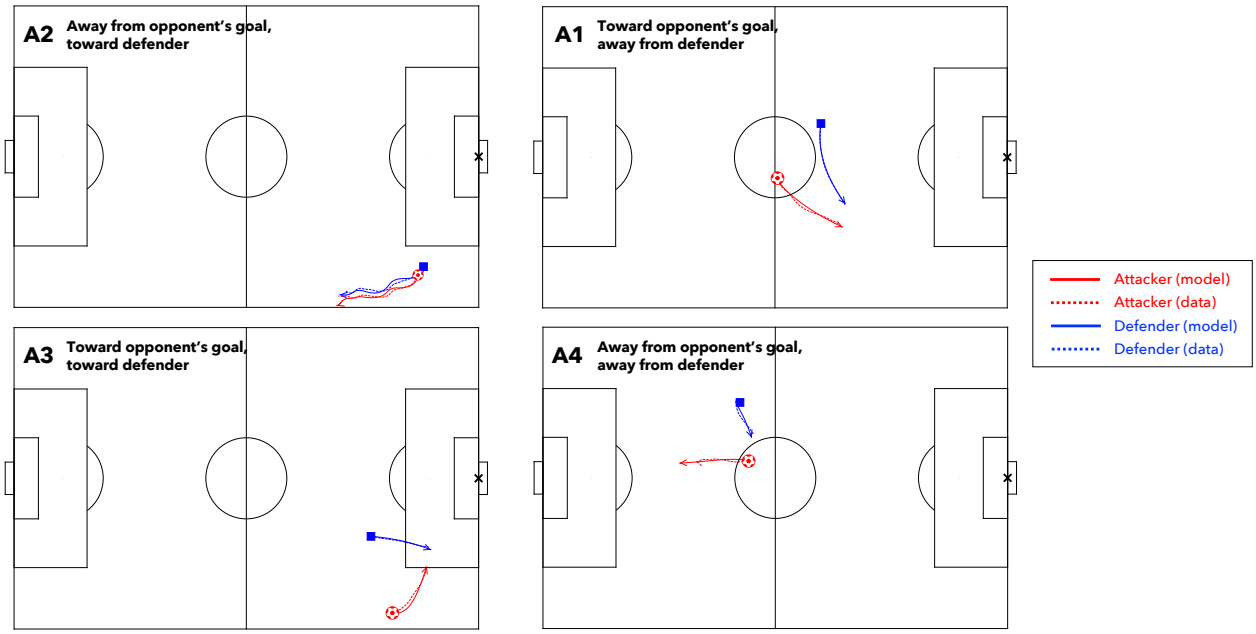


Figure 2: Typical attacker trajectories in regions A1–A4. The red circle and the blue square indicate the starting positions of the attacker and defender, respectively. Dashed and solid lines represent the actual and simulated trajectories, respectively. In all cases, the direction of attack is from left to right.

Table 1: Examples of the most frequently appearing attackers in regions A1–A4. The values in parentheses indicate the player's position and the number of occurrences.

A1	A2	A3	A4
A. Scholz (DF, 280)	A. Ienaga (FW, 48)	T. Okubo (MF, 26)	J. Schmidt (MF, 26)
M. Høibråten (DF, 247)	M. Høibråten (DF, 35)	J. Alano (MF, 25)	K. Watanabe (MF, 16)
T. Deng (DF, 240)	A. Scholz (DF, 31)	Élber (FW, 23)	S. Kawahara (MF, 15)
D. Okamura (DF, 190)	T. Kaneko (MF, 28)	K. Watanabe (MF, 18)	D. Pituca (MF, 14)
Eduardo (DF, 187)	D. Pituca (MF, 28)	R. Hatsuse (DF, 18)	H. Mae (MF, 12)

3.3 Defender Characterization

We also analyzed defender behavior across the four quadrants by identifying the most frequently appearing players and typical movement patterns (see Table 2).

In region D1 ($f_{\mathrm{d}} > 0,\ \beta_{\mathrm{d}} > 0$), the model generates movement toward both the defending goal and the attacker. We observed a pressing action in which the defender retreated toward his goal while guiding the attacker toward the sideline. Center forwards of defending teams were frequently observed in this region.

In region D2 ($f_{\mathrm{d}} < 0,\ \beta_{\mathrm{d}} > 0$) the trajectories show movement away from both the opponent's goal and the attacker. This region often included defensive retreat to prevent direct dribbling toward the goal.

In region D3 ($f_{\mathrm{d}} < 0,\ \beta_{\mathrm{d}} < 0$), the trajectories show movement away from the opponent's goal but toward the attacker. Typical defensive behavior included dealing with dribbles coming in from the side.

Finally, in region D4 ($f_{\mathrm{d}} > 0,\ \beta_{\mathrm{d}} < 0$), the trajectories also indicate movement away from the opponent's goal and toward the attacker. This region included cases where defenders challenged attackers near the penalty area, typically under high-risk situations. Because dribbling events in this region often results in scoring chances, it is crucial for evaluating defenders.

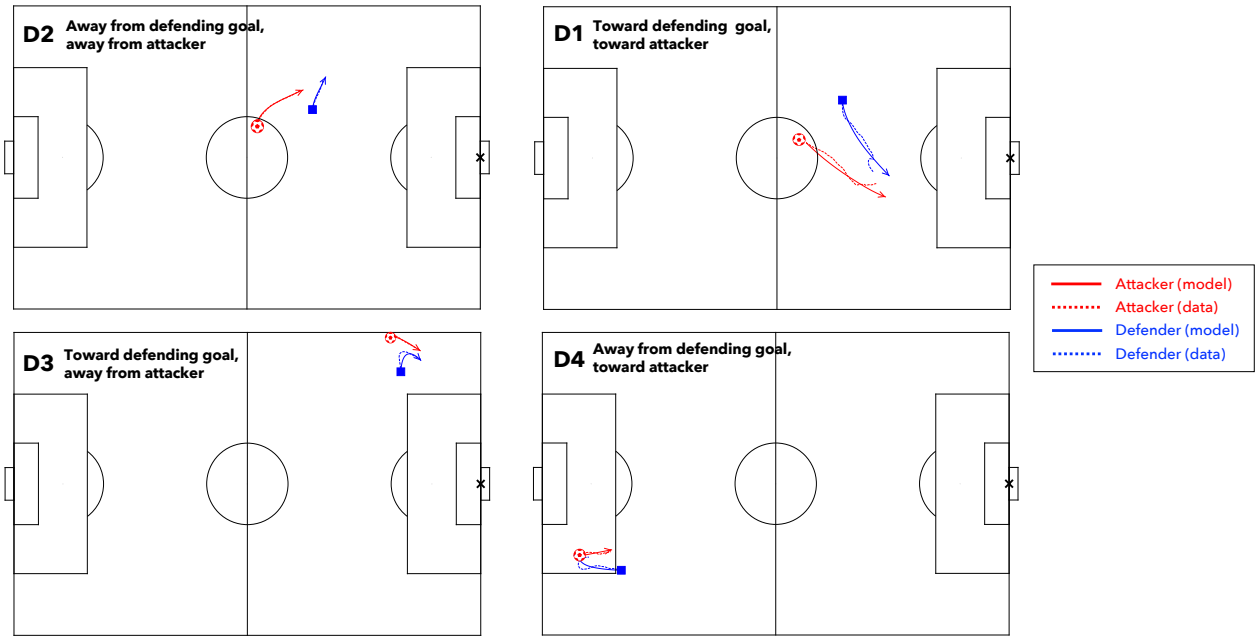


Figure 3: Typical defender trajectories in regions D1–D4. The red circle and the blue square indicate the starting positions of the attacker and defender, respectively. Dashed and solid lines represent the actual and simulated trajectories, respectively. In all cases, the direction of attack is from left to right.

4 Conclusions and Future Work

The present study enhanced the parameter optimization of the Attacker-Defender model and validated its applicability using a large dataset. The equations of motion for the attacker and defender were solved inde-

Table 2: Examples of the most frequently appearing defenders in regions D1–D4. The values in parentheses indicate the player's position and the number of occurrences.

D1	D2	D3	D4
M. Hosoya (FW, 342)	M. Hosoya (FW, 37)	M. Hosoya (FW, 17)	M. Hosoya (FW, 28)
K. Junker (FW, 249)	T. Kikuchi (MF, 22)	R. Hatsuse (DF, 13)	I. Jebali (FW, 19)
L. Ceará (FW, 247)	Y. Yamagishi (FW, 17)	I. Jebali (FW, 12)	Y. Muto (FW, 18)
Y. Yamagishi (FW, 226)	K. Nagai (FW, 17)	G. Sakai (DF, 12)	J. Alano (MF, 17)
K. Sucuki (FW, 210)	Y. Suzuki (FW, 16)	Lukian (FW, 12)	L. Ceará (FW, 16)

pendently, using the actual trajectory of the other player. This modification reduced computation cost and increased the number of dribbles accurately reproduced by the model. In addition, among 30,788 dribbling events, player trajectories were categorized into four quadrants based on the optimized parameters. Each quadrant exhibited distinct playing styles, and the most frequently appearing players were identified.

Future studies will focus on refining the error function and conducting more detailed analyses of dribbling events. The optimized parameters often approached the boundaries of the parameter space, indicating a need for further refinement of the error function. For example, incorporating penalty terms based on player velocity and acceleration could be considered. Additionally, analyzing dribbling events in specific contexts (e.g., actions on each side of the field) or for specific player roles (e.g., midfielders or forwards) may yield more detailed insights.

Acknowledgements

The authors are grateful to DataStadium Inc., Japan, for providing the data used in this study. This Research was supported in part by the Data-Centric Science Research Commons Project of the Research Organization of Information and Systems, Japan, a Grant-in-Aid for Early-Career Scientists (No.23K16729) from the Japan Society for the Promotion of Science (JSPS).

References

[1] Bassek, M., Rein, R., Weber, H., Memmert, D. (2025) *An integrated dataset of spatiotemporal and event data in elite soccer* Scientific Data **12**, 195.

[2] Brefeld, U., Lasek, J., Mair, S. (2019) *Probabilistic movement models and zones of control* Machine Learning **108**, 127–147.

[3] Keller, J.B. (1973) *A theory of competitive running* Physics Today **26**, 43–47.

[4] Fujimura, A., Sugihara, K. (2005) *Geometric analysis and quantitative evaluation of sport teamwork* Systems and Computers in Japan **36**, 49–58.

[5] Narizuka, T., Takizawa, K., Yamazaki, Y. (2023). *Validation of a motion model for soccer players' sprint by means of tracking data* Scientific Reports **13**, 865.

[6] Brink, L., Ha, S.K., Snowdon, J., Vidal-Codina, F., Rauch, B., Wang, F., Wu, D., López-Felip, M.A., Clanet, C., Hosoi, A.E. (2023) *Measuring skill via player dynamics in football dribbling* Scientific Reports **13**, 19004.

Optimizing professional sports league games based on spectators and traveling

K. Nurmi*, J. Kyngäs** and Arto I. Järvelä***

* Satakunta University of Applied Sciences, Pori, Finland, cimmo.nurmi@samk.fi
** Satakunta University of Applied Sciences, Pori, Finland, jari.kyngas@samk.fi
*** Finnish Hockey League, Helsinki, Finland, arto.i.jarvela@liiga.fi

Abstract

The quality of the professional sports leagues schedules has become increasingly important, as the schedule has a direct impact on revenue for all parties involved. Most importantly, the schedule influences the number of spectators in the stadiums and the traveling costs for the teams. The Finnish Major Hockey League decided to promote one team to the league for the season 2024-2025. This means that there would be 16 teams in the league, and this causes problems with the formerly used base schedule. A new approach had to be considered where every team should meet every other team at least once at home and once away, for the sake of sportsmanship. The rest of the games would be decided based on the number of spectators and traveling. This paper presents a new format, where the number of times the teams play against each other is based on maximizing the total expected number of spectators and on minimizing the total traveling.

1 The Finnish Major Ice Hockey League

Professional sports leagues are huge businesses. The leagues involve significant investments in players, broadcast rights and merchandising. The quality of the schedules has become increasingly important, as the schedule has a direct impact on revenue for all parties involved. The general sports scheduling problem involves scheduling the games between the teams by determining the date and the venue in which each game will be played (see e.g. [1]). Sports scheduling involves four basic problems: Finding a schedule

1) with the minimum number of breaks [2]
2) which minimizes the travel distances, called the traveling tournament problem [3]
3) with the minimum number of breaks and, at the same time, take additional requirements and requests into account is known as the constrained minimum break problem [4]
4) which considers the break minimization and the travel issues simultaneously as well as many additional criteria and constraints, called the constrained sports scheduling problem [1].

Ice hockey is the biggest sport in Finland both in terms of revenue and in number of spectators. The Finnish Major Ice Hockey League (officially Finnish Hockey League) involves, for Finnish standards, significant investments in players, broadcast rights and merchandising. Finding the best schedule of games is a difficult task with multiple decision makers, constraints, and objectives involving logistics, economical and fairness issues [5].

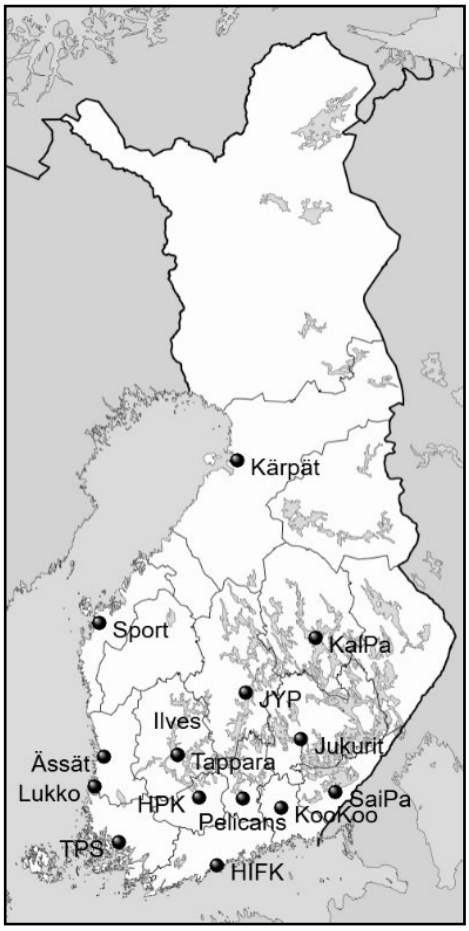


Figure 1. The 15 teams on the map of Finland.

We have generated the league schedule for the last fifteen years. The first ten years have been analyzed in detail in [6]. Since the 2015-2016 season, the League has had 15 teams. Six of the teams are located in "big" cities (over 100,000 citizens) and the rest in smaller cities. Team Kärpät is quite a long way up north and teams Sport and KalPa slightly separate from the other teams (see Fig. 1).

The format played in the league is somewhat eccentric (see detailed description in [6]). The basis of the regular season is a quadruple round robin tournament resulting in 56 games for each team. In addition, the teams are divided into five groups of three teams in order to get a few more games to play. The teams in the groups are selected based on the traveling time. These teams play a double round robin tournament resulting in 4 games for each team totaling 60 games for each team. The number of games totals 450.

The six best teams of the regular season proceed directly to the quarter-finals. Teams placed between 7th and 10th play preliminary playoffs best out of three. The two winners take the last two quarter-final slots. Teams are paired up for each playoff round according to the regular season standings, so that the highest-ranking team plays against the lowest-ranking, and so on. The playoffs are played best out of seven. The winner of the playoffs receives the Canada Bowl, the championship trophy of the League.

The four most important factors in generating the schedule are the following:

1) The interest of media and fans must be increased by special interest games. These games include local rival games, games between so-called big teams and back-to-back games.
2) Traveling issues must be considered. Some combinations of two games on consecutive days are forbidden due to long distances. Furthermore, some of the teams do not prefer to return home after each away game but instead prefer to have two away games in a row.
3) The number of Friday and Saturday games are maximized.
4) It is important to have as few breaks as possible. The fans do not like long periods without home games and the consecutive home games reduce gate receipts.

Three further important factors must be considered:

5) There must be at least two weeks between two games with the same opponents, and the games should be played on different venues.
6) The difference in the number of games played between different teams in any stage of the tournament should be minimized.
7) The so-called rest difference should be minimized.

The last criterion was added last year due to the everlasting conversation by some coaches and many fans. Most of the teams play two consecutive games on Fridays and Saturdays, but some teams have a rest day on Friday before the Saturday game. The argument is that the other team has the advantage of not playing on Friday. However, we have found no statistical evidence of that. Based on the last ten seasons, some teams are more victorious when they have played on Friday, and some if not.

Finally, there are several other goals and constraints considered:

A. The number of rounds is minimized.
B. The number of special interest games in preferred rounds is maximized.
C. The number of away tours for some teams is maximized.
D. Team Tappara and Team Ilves cannot play at home in the same round.
E. A break cannot occur in the last round.

The sports scheduling problem is not just NP-hard (see e.g. [7]), but also extremely challenging practical problem. It has been argued that scheduling the Australian Football League (AFL) is the most challenging mathematical problem in world sport [8]. We believe that scheduling the Finnish Major Ice Hockey League is at the same level of difficulty. Three major constraints complicate the scheduling:

1) Due to the travel distances between some venues, certain combinations of a home team playing the next day an away game against some opponents are not allowed. Similarly, certain combinations of a home team playing on the previous day away against some opponents are not allowed. The number of forbidden pairs is as high as 52.
2) Most of the teams cannot play at home on certain days because their venues are in use for some other events, e.g. concerts.
3) Some of the teams cannot play on certain days because they also play in the European Champions Hockey League (CHL).

2 Optimizing the games based on spectators and traveling

Generating the League schedule is a long process to ensure satisfaction of League owners, team owners, broadcast right holders, merchandisers and fans, while at the same time competing between them for their preferences. The League continuously looks for improvements in its schedule format and the schedule itself. We have continuously brainstormed to make improvements (see [5] and [6]) especially on two categories:

A. How to make a more interesting season for the broadcasting company, sponsors, and fans.
B. How to cut down the expenses of the teams.

The two main factors influencing these were stated in Section 2:

1) The number of special interest games.
2) The total traveling of the teams.

How can we maximize the first and minimize the second? We started to brainstorm from a scenario where the teams could play against each other based on interest. The most obvious way to measure the interest is the number of spectators. Some games are of big interest while others are not.

The games between the big teams are almost always of interest. These teams also interest the fans at smaller teams' home venues. It is obviously pleasurable to win a game against them at home. Furthermore, the games between some smaller teams are of interest, but only if they are local rivals. Teams from eastern Finland do not seem to interest the clubs from other part of the country. For some reason or another, there are one or two teams that do not generally interest other teams. In conclusion, we should maximize the number of games with the highest expected number of spectators.

The teams travel 5700-14500 km per year. The longest single travel is 1350 km back and forth between teams Kärpät and TPS. About 35% of the travels are more than 600 km and only 8 % under 200 km. The mean is about 500 km. Most teams make these travels by bus or train, so they need to allow plenty of time for travel. The longest travels usually mean that teams stay overnight during the travel or in the city where they are playing next. All this is exhausting and costs a lot of money. In conclusion, we should minimize the number of games where the traveling distances are the longest.

For double round robin schedules, a balanced schedule is such that the two games between every pair of teams occur in opposite venues. For multi round robin schedules, a common requirement is that the schedule is both balanced and that for every pair of teams, the deviation between venues of games for that pair be no more than 1. It is a common belief that only balanced sport schedules are fair. This is most likely true, but there could be several reasons why the schedules are generated unbalanced, such as extra local rival games, venue restrictions or capacities, traveling distances, and the length of the season compared to the number of teams. For example, in the USA and Canada the traveling distances are quite long, and almost all the major series schedules are unbalanced. As another example, the Australian Football League has such a number of teams and rounds that is impossible to have a balanced schedule.

As stated earlier, the scheduler should find the best possible schedule where colliding business issues, attractiveness issues and fairness issues are all optimized. In case of the Finnish Major Ice Hockey League, the business issue is the most important one. Therefore, the optimization task is as follows:

1) Each team plays 30 times at home and 30 times away.
2) Each pair of teams must play at least once at home and once away (hard constraint).
3) The schedule must include at least twenty-four full 7-games rounds (hard constraint).
4) The maximum number of home games against the same team is four (hard constraint).
5) Maximize the total expected number of spectators (importance weight 2).
6) Minimize the total traveling (importance weight 1).

The optimization task now has two most likely competing goals: Maximize the number of games with the highest expected number of spectators and minimize the number of games where the traveling distances are the longest. To the best of our knowledge, this is the first time such an optimization problem is introduced. We solved the problem using the PEASTP metaheuristic (see e.g. [8]). Table 1 shows the optimized number of times each team plays against each other. The optimization results show that the effect of the difference is a 5% increase in the number of spectators, and a 10% decrease

in traveling. This effect is significant. Note that the team pairs are not necessarily equally interesting to each other, for example HIFK vs. HPK and JYP vs. SaiPa.

Table 1. The optimized number of times each team plays against each other.

H/A	HIF	HPK	Ilv	Juk	JYP	Kal	Koo	Kär	Luk	Pel	Sai	Spo	Tap	TPS	Äss
HIF	0	1	1	1	1	1	4	4	1	3	4	1	3	4	1
HPK	3	0	4	2	1	1	1	1	2	4	2	1	4	3	1
Ilv	1	4	0	1	2	1	1	1	4	4	1	1	4	1	4
Juk	1	1	1	0	4	4	4	3	1	3	4	1	1	1	1
JYP	1	1	2	4	0	4	4	4	1	1	1	4	1	1	1
Kal	1	1	1	4	4	0	2	4	1	1	4	4	1	1	1
Koo	3	1	1	4	3	4	0	1	1	4	4	1	1	1	1
Kär	4	1	1	1	4	4	1	0	4	1	1	4	1	1	2
Luk	1	4	4	1	1	1	1	1	0	1	1	4	2	4	4
Pel	4	4	1	4	1	1	4	1	1	0	4	1	2	1	1
Sai	1	1	1	4	4	4	4	2	1	4	0	1	1	1	1
Spo	1	1	4	1	2	2	1	4	4	1	1	0	1	3	4
Tap	4	4	4	1	1	1	1	1	1	1	1	2	0	4	4
TPS	4	4	1	1	1	1	1	2	4	1	1	1	4	0	4
Äss	1	2	4	1	1	1	1	1	4	1	1	4	4	4	0

3 The impact of the new format on the optimization factors

Section 1 presented the main criteria in generating the League schedule. The criteria were divided into three categories:

- Four most important factors
- Three further important factors
- Several other goals and constraints considered.

We solved the two schedules using the PEASTP metaheuristic [8], which we have used to solve the actual League schedule for the last fifteen years. Table 2 shows the average number of hard and soft constraint violations of 36 test runs for the new and for the current format. The current schedule had no hard constraint violations in any of the 36 runs. The new schedule included one hard constraint violation in two of the runs. The difference in hard constraints was not significant, but it shows that the new schedule is a bit harder to optimize.

It was somewhat surprising that all the local rival games landed on weekends in both formats. This is especially surprising for the new schedule, since it includes 56 local rival games, which is 20 more than in the current format. The number of violations for each of the soft constraints were about the same for both schedules, except for the number of long travels, and the time between the games between the same teams.

The traveling violations decreased from 21 to 12, which is nearly in line with the decreased number of possible violations. There are 27 travels that are not allowed on consecutive calendar days. There are 54 such games in the current schedule, and 36 in the new schedule. The current schedule is slightly better in average, but these games can be effectively reduced by optimizing away tours for two or three teams.

The biggest difference is in the goal of scheduling at least two weeks between the games between the same teams. The number of violations in the new schedule is almost three times higher. However, this is not so surprising, since the new format includes 69 team pairs which play eight times against each other during the season. In summary, this is the “price” we should pay to maximize the total expected number of spectators and to minimize the total traveling. In conclusion, the results showed a

5% increase in the number of expected spectators, and a 10% decrease in traveling, without a loss in overall quality.

Table 2. The average number of violations of 36 runs for the new and current format.

Criteria	New	Current
Hard constraints		
- A team plays only one game per round	0	0
- A team cannot play at home in giving rounds (114 restrictions)	1/36	0
- A team cannot play at home on two consecutive calendar days	0	0
- Teams Tappara and Ilves cannot play at home in the same round	1/36	0
- At least twenty-four full 7-games rounds	0	0
Goals and soft constraints		
- Minimize the number of breaks		
- number of 4 breaks at home	0.2	0.0
- number of 4 breaks away	0.1	0.1
- number of 3 breaks at home	2.1	1.6
- number of 3 breaks away	7.8	7.5
- Minimize the number of long traveling	12	21
- Maximize the number of local rival games on weekends	7.5	6.0
- Maximize the number of games on requested weekdays	7.0	7.0
- Minimize the difference in the number of games played	13	12
- Minimize the rest difference	10	12
- At least 2 weeks between the games between the same teams	83	29

References

[1] Nurmi, K., Goossens, D., Bartsch, T., Bonomo, F., Briskorn, D., Duran, G., Kyngäs, J., Marenco J., Ribeiro, C.C., Spieksma, F., Urrutia, S. and Wolf-Yadlin, R.: A Framework for Scheduling Professional Sports Leagues. In: Sio-Iong, Ao. (ed.) IAENG Transactions on Engineering Technologies, vol. 5, pp. 14-28 (2010).

[2] Schreuder, J.A.M.: Combinatorial aspects of construction of competition Dutch Professional Football Leagues, Discrete Applied Mathematics 35, pp. 301–312 (1992).

[3] Easton, K., Nemhauser, G. and Trick, M.: The traveling tournament problem: description and benchmarks. In Proc of the 7th. International Conference on Principles and Practice of Constraint Programming, pp. 580–584 (2001).

[4] Rasmussen, R. and Trick, M.: Round robin scheduling - A survey, European Journal of Operational Research 188, pp. 617-636 (2008).

[5] Nurmi, K., Kyngäs, J. and Kyngäs, N.: Lessons Learned in Scheduling the Finnish Major Ice Hockey League. In Proc of the 7th International Conference on Mathematics in Sport (2019).

[6] Nurmi, K., Kyngäs, J. and Järvelä, A.I.: Ten-year Evolution and the Experiments in Scheduling a Major Ice Hockey League. In: Hak, D. (ed.) An in Depth Guide to Sports, pp 169-207 (2018).

[7] Easton, K., Nemhauser, G. and Trick, M.: Sports scheduling. In: Leung, J.T. (ed.) Handbook of Scheduling: Algorithms, Models and Performance Analysis, pp. 1-19 (2004).

[8] Kyngäs, N., Nurmi, K. and Kyngäs, J.: Crucial Components of the PEAST Algorithm in Solving Real-World Scheduling Problems, Journal of Lecture Notes on Software Engineering 1(3), pp. 230-236 (2013).

The Right Way to Synchronize Tracking and Event Data: Using Domain Knowledge to Optimize Algorithms

G.A. Oonk*, D. Grob**, and M. Kempe***

*University of Groningen, University Medical Center Groningen, Groningen, The Netherlands;
Feyenoord Rotterdam N.V., Rotterdam, The Netherlands: G.A.Oonk@umcg.nl
** University of Groningen, University Medical Center Groningen, Groningen, The Netherlands
*** University of Vienna, Department of Biomechanics, Kinesiology, and Computer Sciences in Sports, Vienna, Austria

Abstract

Valuable new insights can be obtained by combining tracking and event data in soccer analysis. However, how to synchronize the two data streams, is rarely discussed. Non systematic errors in the timestamps, and synchronizing with cost functions result in suboptimal synchronization, which hinders further analysis. Within this proceedings we will introduce a computationally optimized implementation of the Needleman-Wunch algorithm, by using domain knowledge about the game. The optimized version is over 70 times more efficient in terms of time constraints and memory usage. On top of that, we show that the properly synchronized approach translates back to practice with better performing *xG* models. Taken together, this implementation is a training-free, high-quality synchronization algorithm, with low computational cost that solves existing issues. On top of that, all data and code used for this proceedings is fully open-sourced and available in the DataBallPy package.

1 Introduction

In computer and data sciences, merging multiple data sources is an invaluable skill. By combining data streams, new dimensions of information are added to observations. When this is done right, the added dimensionality can be used to find new patterns in the data. However, when merging different data sources suboptimally, noise is added to the observations, leading to an unnecessary complex task to find valuable patterns in the data. Hence, it is worth making an effort to properly merge multiple data streams together.

In the analysis of football (soccer) matches, the added benefit of combining the two main data sources (event- and tracking data) is widely recognized. However, how this should be done is often not explicitly described [1, 2]. Event data captures the type of action (e.g., pass or shot), which player is involved, and the when and where of the action. Tracking data captures the location of all players and the ball over time at 25 frames per second [3]. By merging them together, one knows, for example, when a striker takes a shot (event data) and how the defenders and keeper are positioned at that moment in time (tracking data). This information is essential for for numerous metrics, like Expected Goal models (*xG*) that predict the probability of a shot resulting in a goal.

The complexity of synchronizing event- and tracking data is multi-factorial. First, the event- and tracking data are often captured by different data providers, resulting in unaligned timestamps in the provided data. Second, event data is captured manually leading to random errors in the provided timestamps and event locations. Last, there is no ground truth definition on when different types of events are happening. Passes and shots can intuitively be assigned at the moment the ball is touched, but one-on-one dribbles often take more than 0.04 seconds, the time of a single frame of the tracking data.

Some methods have been proposed to synchronize football tracking and event data. The most popular method is to minimize a cost function for a specific event. The cost/scoring functions range in complexity from simply minimizing the time difference in the provided timestamps in the event- and tracking data (which are known to contain random errors), to event specific cost functions[4]. However, the cost/scoring function approach aligns every event to the tracking data independently of all other events. This results in mash up of the event order in specific periods in the match, which is suboptimal. A different approach proposes to use machine learning or learnable parameters to align the tracking and event data [5, 6]. Consequently, this requires a significantly sized labeled dataset, which is difficult since no gold standard exists yet, and data availability is scarce.

The use of the Needleman-Wunch algorithm (NWA) was proposed as a learning-free approach that keeps the order of the events in place[7]. The algorithm uses scoring functions and aligns them to the tracking data with knowledge about the alignment of other events (See section 2 for an elaborate explanation of the algorithm). Therefore, this algorithm could solve the deficiencies of using a cost functions approach. However, the main concern raised for the algorithm is its computational performance[4]. To use information about the alignment of other events, the NWA computes scoring values for all event/frame combinations. Consequently, the NWA scales bilinear while cost/scoring functions alone scales linear, increasing the computation time tri-fold according to previous research.

In the current paper we aim to implement the Needleman-Wunch algorithm that is optimized to synchronize tracking and event data by exploiting domain knowledge and information from the data. We will show that the optimized implementation has an over 70-fold time reduction in processing time and used memory, without information loss. Taken together, this implementation is a training-free, high-quality synchronization algorithm, with low computational cost that solves the issue relating the order of the event. On top of that, all data and code used for this paper is fully open-sourced and available[8].

2 The Needleman-Wunch Algorithm

The NWA was originally developed in bioinformatics to align two amino-acid sequences[9]. Due to the evolutionary process, amino acids could be added or removed from the sequences, making it difficult to align them with the original sequence. The NWA was especially developed to find the optimal alignment, given that insertions and deletions in the sequence are possible. To account for these mutations, the NWA makes use of gap penalties: an alignment with a gap in a sequence is allowed, but it comes at a cost. These gaps are useful in football data since not all tracking data frames are coupled with an event, but all events must be coupled with a tracking data frame.

The algorithm starts by creating a function matrix F that is of shape $M+1$ by $N+1$ where M and N are the number of elements in sequence x and y respectively. The first column and row are filled with the cumulated gap penalties for the first and second sequence. From there, F is dynamically filled based on

formula 1.

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + S(x_i, y_j) \\ F(i-1,j) - p_y \\ F(i,j-1) - p_x \end{cases} \tag{1}$$

Here, S is the scoring function and p_x and p_y are the penalties for inserting a gap in x and y respectively. To keep track which of the three options was used in the formula, the pointer matrix P is filled simultaneously. Eventually, this will be used as a map to trace back the optimal alignment between the two sequences. The values in P can be represented as arrows. For instance, if the first option in formula 1 was the highest value, the information was used from $F(i-1,j-1)$ indicating a diagonal arrow ($\nwarrow$). This indicates that x_i and y_j are matched. When creating a gap in x (leaving value in y_j unmatched), a left arrow is added ($\leftarrow$). Last, a top arrow ($\uparrow$) is added when a gap is created in y (leaving x_i unmatched). Finally, when P is filled, one can trace back from bottom right to top left the optimal alignment between the sequences (for an example case see the Appendix).

3 The Football Specific Needleman-Wunch

In football we do not have amino-acid sequences, but we have tracking and event data. The challenge is to align an event to the right tracking data frame. To translate the above formula to this football context, we need to redefine the variables. Let x be the tracking data of length M, and y the event data of length N. A key contributor to the worse computational performance is the fact that the scoring function S needs to be computed for all event/frame combinations. In the current section we explore two methods to decrease the number of computations without loosing useful information for the synchronization.

3.1 Dead Ball Batching

Both the event and tracking data indicate when the game status shifts from in-play (alive) to out of play (dead) or vica-versa. We can use this information to create smaller batches. For instance, the first sequence starts at a kick-off and ends when a throw in is awarded. All frames and events that are within that sequence are now, independent of all frames and events in the rest of the matches, synchronized in the NWA. If the ball is dead 99 times a match, this approach will create 100 batches. Assuming that a match has 135000 frames (25 fps for 90 minutes) and 1200 events, the naive approach would result in $135000 * 1200 = 1.62e8$ computations. The dead ball batching approach results in $100 * (1350 * 12) = 1.62e6$ computations, which is, not surprisingly, a 100-fold decrease.

3.2 Reducing Computational Load

Although batching already significantly reduces the computing cost, some some batches still represent sequences of over 5 minutes of play. These sequences now take up considerable amount of computational time. Within the NWA, especially looping over the function matrix F takes considerable time. Since this matrix is filled dynamically, this process can not be vectorized. However, a matrix with the alignment scores based on S can be vectorized. Assuming that all events should be aligned to a frame in the tracking data, we can make

calculated estimations of which values in P will be accessed during the trace back in P. If only these values are compute in F, the number of values that need to be filled in in F is reduced, and thus the computation time and memory usage.

Consider filling matrix B of shape M by N with the scoring values returned from S. Instead of fully filling F, a set of frames are considered based on idx_{start} and idx_{end} for every event (y_i) which can be obtained from formulas 2 and 3. Two exceptions are the first event, where idx_{start} is always 1, and the last event where idx_{end} is always N. Using this approach, we (1) make sure the arrows in P overlap so that there are no gaps in P, and (2) that only the matrix is filled where we consider a likely possibility for a match between the event and frame ($B[i,:] > \alpha$), where α is the threshold the for which we consider a likely match between the event and a tracking frame.

$$idx_{start}(y_i) = \min \begin{cases} idx_{end}(y_{i-1}) - 1 \\ \min(where(B[:,y_i] > \alpha)) \end{cases} \tag{2}$$

$$idx_{end}(y_i) = \max(where(B[:,y_i] > \alpha)) \tag{3}$$

4 Experimental Setup and Results

Seven open-source matches from the German Bundesliga are used to validate this method[10]. Three methods are compared: (1) timestamp synchronization (a base cost function approach), (2) NWA, and (3) optimized NWA. To compare the computational cost of the approaches the execution time and peak memory usage is compared. Since no ground truth labels exist, the distance between the ball in the tracking data and the event location in the event data is used to assess the validity of the approach. In general, the both NWA approaches show a lower difference between the event and ball location compared to the timestamp approach. On top of that, the optimized NWA approach uses over 70 times less memory and computation time compared with the non-optimized NWA approach (Table 2).

From the event data, only passes, shots, and dribbles were considered since they are most interesting to synchronize to a specific tracking data frame. The scoring functions S are specific per event. They all include information about the time difference between the tracking and event data, and the distance between the ball and the event location. However, also event specific features are included, for instance, for shots the shooting direction is included. All event specific scoring functions can be found in the open-source python package DataBallPy (v0.6.0) and manually overwritten by users if they wish to do so[8].

	Timestamps	NWA	Optimized NWA
Peak Memory Usage (MiB)	2262.26	5012.43	49.32
Computational Time (s)	1.26	112.19	1.36
Distance (m)	9.50 +- 8.09	2.36 +- 2.33	2.56 +- 3.24

Table 1: The performance of the different algorithms to synchronize the tracking and event data of a single game. All values are averaged over the seven games (+- standard deviation).

5 Importance of Proper Synchronization

To showcase the importance of proper data synchronization, we will train two Expected Goals (*xG*) models. Both models will be trained on three features: the distance to the goal, the shot angle, and the available shot angle based on earlier research (Figure 1) [11]. All these three features have been shown to be important predictors of whether a shot results in a goal. The model itself is a logistic regression and the features were scaled using a standard scaler. To create a more stable dataset, only shots in active play and by foot were considered. So free-kicks and headers were excluded. The *xG* models were trained on a different dataset consisting of 6638 shots. The only difference between the two models is that one was trained on the frames of the timestamp synchronization, and the other of the optimized NWA synchronization.

In general, the two metrics perform equal in discriminating between misses and goals, depicted by similar scores in the ROC AUC score (Table 3). However, from the Brier Score we can conclude that the calibration of the predicted probabilities, or how close the predicted probabilities are to the true labels, is a lot better for the model trained on the NWA synchronized data compared to the timestamp synchronized data. Since the main goal of *xG* models is to assign a probability of a shot resulting in a goal, the Brier Score is a more important evaluation metric for this context[12].

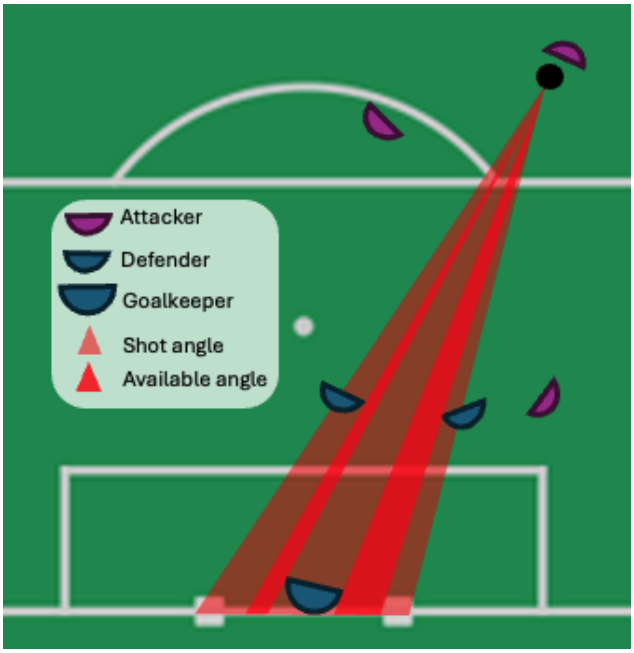


Figure 1: Graphical representation of the shot angle and the available shot angle used for training the *xG* models

The coefficients of the logistic regression show that for the model trained in the NWA synchronized data, besides the distance and shot angle, also the available shot angle is used (Table 4). The available shot angle is especially dependent on a proper synchronization. The absolute coefficient weight of the available angle is a lot lower for the model trained on the timestamp synchronized data, but also the direction is off. One would expect that a greater available shot angle would result in a greater probability to score, but this is not

	Timestamps *xG*	Optimized NWA *xG*
Brier Score	0.096	0.082
ROC AUC	0.722	0.729

Table 2: The performance of the different *xG* models. For the Brier Score, lower values represent a better prediction. For ROC AUC, higher values represent a better prediction

the case for the model trained on the timestamp synchronized data.

	Timestamps *xG*	Optimized NWA *xG*
Goal Distance	-0.85	-0.61
Shot Angle	0.84	0.31
Available Shot Angle	-0.02	0.09

Table 3: The model coefficients of the *xG* models trained on differently synchronised data.

6 Limitations and Future Directions

The current approach has some downsides, which mostly relates to faulty event or tracking data. The dead ball batching relies on information on whether the ball is in play or not. If this information is not, or incorrectly, given by data providers, events can be synchronized in the wrong batch, leading to a faulty synchronization. This is only a problem with datasets where the non-standard error is big (> +- 5 seconds). For smaller errors events are generally inserted in the correct batch. To handle this issue, one could try to synchronize the edge events (e.g. at the beginning or start of a batch) to multiple batches and see where the result of the scoring function S is the highest. Second, because we do not fully fill the pointer matrix P, the algorithm can get stuck. However, in practice this only happens when an event is wrongly labeled, for instance when a pass is labeled to the wrong player. This approach actually detects these wrongly labeled events which makes it possible to manually get them out of the data, or re-assign them to the correct player.

7 Conclusion

The current study showed that utilizing domain knowledge can increase performance of algorithms by over 70 times in both memory and time constraints. By optimizing the Needleman-Wunch algorithm for synchronizing soccer tracking and event data, making concessions by using only the timestamps or cost functions is not necessary anymore. Especially since all the code is open sourced (DataBallPy). Using the proper synchronization translated back to a better performing *xG* model. With this proceedings, we hope to highlight the importance of proper synchronization, and open the discussion to feasible approaches to do so.

A Appendix

A.1 Simple Needleman-Wunsch Example

Let us consider two sequences: $x = ABCD$ and $y = ABD$. For simplicity, lets say that $p_x = p_y = 1$ and $S(x_i, y_i) = \begin{cases} 1 & \text{if } x_i = y_i \\ -1 & \text{otherwise} \end{cases}$. By filling in F and simultaneously P by using Formula 1, the optimal path can be obtained by walking back through P from bottom right to top left (Table 1). The NWA rightfully matches the A's, B's, and D's (indicated by the diagonal arrow), and a gap in y (indicated by the horizontal arrow).

		A	B	C	D
	0	-1	-2	-3	-4
A	-1	1	0	-1	-2
B	-2	-2	2	1	0
D	-3	-3	1	1	2

(a) Example Function Matrix F

		A	B	C	D
	↖	←	←	←	←
A	↑	↖	←	←	←
B	↑	↑	↖	←	←
D	↑	↑	↑	↖	↖

(b) Example Pointer Matrix P

Table 4: Example Function (a) and Pointer Matrix (b)

References

[1] M. Brechot and R. Flepp, "Dealing with randomness in match outcomes: How to rethink performance evaluation in european club football using expected goals," *Journal of Sports Economics*, vol. 21, pp. 335–362, 10 2020.

[2] F. R. Goes, M. Kempe, L. A. Meerhoff, K. L. B. data, and undefined 2019, "Not every pass can be an assist: a data-driven model to measure pass effectiveness in professional soccer matches," *liebert-pub.com*, vol. 7, pp. 57–70, 10 2019.

[3] D. Linke, D. Link, and M. Lames, "Football-specific validity of tracab's optical video tracking systems," *PLoS ONE*, vol. 15, 2020.

[4] M. Van Roy, L. Cascioli, and J. Davis, "Etsy: A rule-based approach to event and tracking data synchronization," in *Machine Learning and Data Mining for Sports Analytics* (U. Brefeld, J. Davis, J. Van Haaren, and A. Zimmermann, eds.), (Cham), pp. 11–23, Springer Nature Switzerland, 2024.

[5] H. Biermann, F. G. Wieland, J. Timmer, D. Memmert, and A. Phatak, "Towards expected counter - using comprehensible features to predict counterattacks," *Communications in Computer and Information Science*, vol. 1783 CCIS, pp. 3–13, 2023.

[6] F. Vidal-Codina, N. Evans, B. E. Fakir, and J. Billingham, "Automatic event detection in football using tracking data," *Sports Engineering*, vol. 25, 10 2022.

[7] M. Kwiakowski and A. Clark, "The right way to synchronise event and tracking data."

[8] "databallpy — databallpy documentation."

[9] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, pp. 443–453, 3 1970.

[10] M. Bassek, R. Rein, H. Weber, and D. Memmert, "An integrated dataset of spatiotemporal and event data in elite soccer," *Scientific Data*, vol. 12, no. 1, p. 195, 2025.

[11] H. Karim and L. Marwane, "The kos angle, an optimizing parameter for football expected goals (xg) models," *International Journal of Computer Science in Sport*, vol. 22, pp. 49–61, 8 2023.

[12] J. Davis, L. Bransen, L. Devos, A. Jaspers, W. Meert, P. Robberechts, J. V. Haaren, and M. V. Roy, "Methodology and evaluation in sports analytics: challenges, approaches, and lessons learned," *Machine Learning*, vol. 113, pp. 6977–7010, 9 2024.

Evaluating the Improved Linear Model (and its successor?) with regards to the expanded College Football Playoff

John A. Trono

Saint Michael's College, Colchester, VT 05439 (USA): jtrono@smcvt.edu

Abstract

Now that the College Football Playoff (CFP) has increased the number of invited teams from four to twelve, this article will compare how well the original model's weights performed in the first year of this expanded championship (2024). This article also includes the performance of newly generated sets of weights using the updated criterion of attempting to match this significantly larger group that the CFP committee now selects (as its dozen championship playoff participants).

1 Introduction

Even though the National Collegiate Athletic Association (NCAA) has attempted to decrease the controversy regarding which teams will compete for its yearly, college football championship by deciding to invite an additional eight teams, starting with this past year (2024), the fans will probably now clamor for even *more* teams/playoff games because there will always be (in some fans' minds) certain injustices concerning some small number of teams being unfairly excluded from this opportunity to be crowned as the best team (at the conclusion of that season). Thankfully, this article will *not* be concerned with if twelve is the most appropriate number of teams to be invited to have an opportunity to earn this recognition, or, if the *best* twelve teams were actually invited (by the CFP selection committee).

After briefly reviewing the basic approach of the Improved Linear Model (ILM), and how well it has performed prior to these eight additional teams being invited to compete, the scope of this particular methodology will be widened to work with the larger number of invited teams. Several new sets of weights for this model will be generated and how well these – and the original set of – weights performed in 2024 will be presented.

2 Background

As mentioned in Trono (2020), the power rating system described by Carroll, Palmer and Thorn (1988) is an iterative process that determines a reasonable strength of schedule (SOS) value, that converges to an intuitively accurate set of SOS values for the teams – and season – in question. A team's power rating is the sum of the OD and SOS computed values, where OD represents the average, overall difference between the number of points a team scored minus the number of points that their defense allowed.

It was also noted in Trono (2020) that the power rating system matched 16 of the 24 teams chosen by the CFP committee, from 2014-2019, when using the full margin of victory (MOV). If the set of

game scores was modified so that only wins and losses mattered, i. e. setting the MOV to be at most one point, then the power rating system, when excluding MOV (which will now be referred to as no-MOV), matched a somewhat different collection of 20 (of the 24 teams chosen). This strong affinity between those two sets of power ratings and the CFP committee's selections led to the investigation of a linear model incorporating these two team ratings as well as each team's number of losses.

Using the first four years (2014-2017) as the training data set, one million sets, of three random weights – in the range zero to one, were generated. From these preliminary results, modifications were made to this initial linear model which simply adds the product of those two power ratings and the first two random weights, and then subtracts from that sum the product of the losses for that team by the third random weight. The first modification was to increase the range of these of random weights for the no-MOV power rating since there is much less variation in those values than the power ratings produced from using the actual scores. With this in mind, the random weight that would be multiplied by the power rating using the full MOV remained in the zero to one range; however, the randomly generated weight for the no-MOV rating value was multiplied by one hundred, and the penalty for each loss was also multiplied by ten. These small changes increased the number of very accurate random weight trios by almost a factor of 500. (More specifically: the number of original weight sets (11) that matched 14 of the 16 top four teams in the training data set was significantly increased (5119) when the subsequent one million weight sets were evaluated.)

The other modification was to separate both of the team's power ratings into their constituent components (OD and SOS), thereby increasing the weight set size from three to five: two weights for each rating – and the team loss penalty weight. Incorporating these two modifications produced what is now referred to as the Improved Linear Model (ILM), and working with the same training data set, another million random weight sets were generated again and the most accurate ones correctly matched 14 of the 16 teams chosen by the CFP committee as well as matching the exact position (anywhere in the range from the #1 team to the #4 team) 9 times. (Those positions are very important since the #4 teams plays the #1 team in one semi-final contest, leaving team #3 to play team #2 in the other semi-final.)

Since there were roughly 40 sets (of these five weights) that produced this level of performance, the weights which also generated the highest, overall Spearman Correlation Coefficient (SCC) with the committee's ranking of the top 25 teams (for the training data) was the set chosen. There was also one weight set (in the million that were generated) that had 10 exact matches, though only 13 of the 16 top four invited teams were matched (from 2014-2017). However, the SCC for the top 25 teams (SCC-25) was significantly lower for this set of weights (0.77288) than the one for 9 exact matches, and 14 of the 16 top four teams (0.83923), which is also why that latter set of weights was chosen. (These two quantities will now be notated as 10,13 and 14,16.) In the two years after the training set, i. e. 2018 and 2019, the 10,13 weights only had one exact match each year as well as matching only three of the top four in both, i. e. 12,19 in total, whereas, the chosen weight set matched all eight teams (six exactly), bringing its overall performance to 15,22 over the CFP's first six years. (For just 2018 and 2019: 10,13 → 2,6; 9,14 → 6,8. In fact, the 9,14 weights also matched the top eight teams exactly in 2018 – which is the first year after the training data set: 2014-2017.)

Because of the very restrictive, athletic scheduling policies that were being enforced during COVID, 2020 will not be included in any analysis in this report since there were not many games played between teams in different conferences, which would definitely impact any computed power ratings. However, in those other first nine years of the CFP, the ILM has exactly matched the position for 18 of the 36 teams, and 33 correct matches overall. (Excluding the first four years of training would result in the performance level of 9,18 for those 20 teams who appeared in the committee's top four teams from 2018-19 and 2021-2023.)

By retrospectively examining how the original ILM weights performed with regards to the committee's top twelve teams, in those nine years, its performance level was 31,96 (for those 108 invited teams). This also assumes that these rankings – as generated by the CFP selection committee – would not be any different regardless if four, or twelve, teams were being invited to compete in the CFP. The question then was: would it be possible to randomly generate other sets of weights that might perform better than these original ILM weights in subsequent years?

3 New Weights for Matching the CFP Committee's Top Twelve

Unlike the original ILM, that was described in Trono (2020), the results that were produced, during the Monte Carlo generation of the much larger number of set (roughly 15 billion) of these five random weights, to attempt to match the CFP committee's top twelve teams (over nine years), had several interesting criteria to consider when deciding which set of weights were the most accurate. After examining this newly generated, larger collection of weight sets, there were quite a few sets that matched 100 of the 108 teams that appeared in the top twelve positions over the nine years in question (2014-2023; excluding 2020 – due to COVID scheduling related anomalies). The number of exact matches in these instances ranged from 29 up to 33 (for those specific weight sets). However, many sets of weights had 99 matches and also matched 38 of the top twelve teams selected exactly. A significantly larger group matched 41 teams exactly, but only matched 96 of the 108 top twelve teams. Those latter two weight set results sum to 137 in all (38 + 99 = 41 + 96) when combining those two counts; 138 is the largest such combination found – with 40 exact matches and 98 matches overall. (The weight sets with the highest SCC-25 values would also be chosen in these cases, when more than one set of a certain performance level was generated.)

Any weight set with five more exact matches, and one fewer overall match (than the set it is being compared against), appears to be somewhat *better*; therefore, the conjecture is that 38,99 could be more likely to generate a more accurate ranking than the weight set that produced the 33,100 level of performance – if only one weight set had to be selected/used. With regards to the two other weight sets, the previous study by Trono (2020), using only those four years (2014-17) of training data, seemed to imply that overfitting to realize the largest number of exact matches could lessen the overall weight performance as seen in 2018 & 2019 with regards to the 10,13 versus 9,14 example (as mentioned previously). Therefore, it seems wise to still favor the chosen weight set that produced 38,99[1]- over the ones yielding 41,96 and 40,98 – to possibly avoid the same behavior. Including the original ILM weights, that brings our total to five different weight sets to evaluate the 2024 season with.

Besides using the SCC-25 value, to select one set, from amongst the many sets of weights producing the same matching performance, it would also be interesting to evaluate the weight set that actually achieved the highest SCC-25 value over the nine years of training data as well. (The name SCC will be associated with this weight set in subsequent tables.) Also, since the previous study, Trono (2020), only investigated matching the top four teams that were deemed worthy of being invited (by the CFP committee), a new formula was devised to provide another objective, quantitative measure to maximize a weight set's performance – besides the number of exact and overall matches in the top twelve CFP selections. This formula assigns a rewarded point value for each rank; the closer the prediction is to the CFP rank, the larger the portion of this point value that will be earned. In this formula, it seemed more important to match the higher rank teams, and so exactly matching the CFP #1 team would be worth

[1] There were two weight sets that produced the same SCC-25 value, so those five pairs of weights were averaged, and this new set generated the identical yearly SCC-25 values as well as matching the same results: 38,99.

100 points whereas exactly matching the #12 would only be worth 34 points. The values in the second row of Table 1 below decrease in a linear fashion, starting with subtracting eleven and then ten – all the way down to a one point difference between the full reward for matching the teams ranked eleventh and twelfth (by the CFP committee).

The value earned per rank would be reduced by the square of the difference between the weight set's prediction and the actual CFP rank, and this strategy will be referred to as F_Sqr. The other strategy, F_*, will reduce the value earned as well – but in this case, the reduction is multiplicative in nature; for each position that the prediction is further away from the actual CFP rank, the point value reduction is another 10%. So, for example, if the prediction for the team ranked as the fourth team is off by three, the value awarded for the prediction produces an award of 61 points (70 – 9) when applying the F_Sqr approach, whereas F_* would reduce the award by 30% yielding 49 points (70 * 0.7). The bottom two rows in Table 1 illustrate the reduction that would occur for the differences appearing in the top row in said table. These two measures can also be used as a criterion for choosing a weight set and so F_Sqr and F_* will be used both as the name for an evaluation measurement as well as the name of the weight set that maximized that quantity for all generated weight sets.

Table 1 – Point and update values, with regards to how close a set of weights is (to CFP top twelve).

	1	2	3	4	5	6	7	8	9	10	11	12
Value	100	89	79	70	62	55	49	44	40	37	35	34
F_Sqr	-1	-4	-9	-16	-25	-36	-49	-64	-81	-100	-121	-144
F_*	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.0	1.0

How well all eight models performed (in 2024) appears in Table 2 It was a little surprising that the original ILM weight set achieved the highest SCC-25 value. It seems like the weights associated with the 38,99 model had the best results since it was the only one to have eleven teams in *its* top twelve along with three exact matches. This 38,99 model had an only slightly smaller SCC-25 value (than the original ILM) as well as it also being associated with the highest F_Sqr, F_* and SCC-12 values (of the eight models, which is not so surprising given its strong performance in 2024). The Appendix includes several other tables – one of which provides the yearly breakdowns (2014-2023) for all eight mentioned weight sets with regards to the aforementioned performance measures.

Table 2 – Actual results for the 2024 NCAA season.

2024	Orig	33,100	38,99	40,98	41,96	SCC	F_Sqr	F_*
Results	1,11	1,10	3,11	2,11	1,10	2,10	2,10	2,10
SCC-25	.8796	.8231	.8738	.8642	.8565	.8262	.8273	.8304
F_Sqr	587	588	619	586	582	576	591	591
F_*	555.0	525.1	553.9	530.7	524.6	559.0	563.8	563.8
SCC-12	.6259	.6294	.7663	.6538	.6399	.5874	.6399	.6399
SCC-4	-0.1	-0.1	-1.8	-1.6	-1.4	0.0	0.0	0.0

When comparing these model's rankings, a curious observation was that in all eight, the CFP committee's #4 team (Penn State) and #10 team (SMU) always appeared as the 7th and 8th rated team for each model. (Oregon was the CFP committee's #1 team and they were also ranked #1 by all eight models as well.) The columns in Table 3 are abbreviated so that all eight models would fit, and so only the number of exact matches is listed there as opposed to how those models' names appear in Table 2.

The sum of all eight rankings is also listed as well as the position that said sum would place each team in an *overall ranking* (of all models).

Table 3 – Lists the integer, positional ranks for the eight models and a cumulative, overall ranking.

CFP	Team	Orig	33	38	40	41	SCC	F_Sqr	F_*	Sum	Rank
1	Oregon	1	1	1	1	1	1	1	1	8	1
2	Georgia	3	3	5	6	6	2	2	2	29	3
3	Texas	4	4	3	4	4	4	4	4	31	4
4	Penn State	7	7	7	7	7	7	7	7	56	7
5	Notre Dame	2	2	2	2	2	3	3	3	19	2
6	Ohio State	5	5	4	3	3	5	5	5	35	5
7	Tennessee	15	13	11	11	11	14	13	13	101	13
8	Indiana	6	6	6	5	5	6	6	6	46	6
9	Boise St	11	15	14	14	14	15	15	15	113	14
10	SMU	8	8	8	8	8	8	8	8	64	8
11	Alabama	12	9	9	10	9	10	9	9	77	9
12	Arizona St	9	11	12	12	13	9	10	10	86	10
13	Miami(F)	10	12	10	9	10	12	12	12	87	11
14	Mississippi	18	17	16	16	15	18	17	17	134	17
15	South Carolina	13	10	13	13	12	11	11	11	94	12
16	Clemson	16	16	17	17	17	16	16	16	131	16
17	BYU	14	14	15	15	16	13	14	14	115	15
18	Iowa St	17	18	18	18	18	17	18	18	142	18

4 Miscellaneous

Even though individual desktop computers are faster now than when the one million Monte Carlo weight sets were generated back in 2020, generating 15 billion sets would tie up a computer for quite some time. Therefore, there were essentially 15 executions of the program, each one generating one billion sets. Each run of this program would produce too much output if each set of weight's results were placed into a file, so only those matching the highest number of exact or total matches – or the highest total of those two values, or, the highest SCC value that had been found so far – were saved. The lowest SCC value typically appears as the first output value in these files, as computed from the first set of randomly generated weights, and only one such weight set had a SCC-25 value that was less than zero (-0.4850), and even *that* set matched 81 of the 108 invited teams in the nine years of training data – along with thirteen exact matches. (Due to the six page limit, for papers in this conference's proceedings, an additional eight tables are included in an expanded appendix that will eventually appear in an online version of this article, though you may have to read some of the content on the ILM webpage - that is located at https://www.smcvt.edu/directory/john-trono/improved-linear-model/ - to find a link to that version, which will hopefully appear online there later in 2025.) In retrospect, with regards to matching the top twelve ranked teams by the CFP committee (for the 9 years of training data), the MOV power rating values performed at the 12,83 level (12 exact matches and 83 overall matches); the no-MOV was demonstrably more accurate (28,91). As one would hope, since it is a combination of both components in these two rating values, the original ILM weight set did exhibit an

improvement over those results (31,96), and in 2024, MOV achieved an 0,9 outcome, no-MOV achieved 1,11 – and the latter was also the same level of performance as the original ILM.

5 Summary

Seven new sets, of five weights each, have been uncovered via a Monte Carlo evaluation process, and how accurate they were in 2024 has been compared with the original ILM set of weights. Several new objective measures have been introduced, for evaluation purposes; however, the number of (exact and) overall matches between a model's prediction and the CFP committee's ranking is still the objective to be maximized. At least for 2024, it appears that the aforementioned conjecture that the 38,99 model would be the most accurate model was confirmed, and it will be interesting to see if such behavior continues over the coming years with regards to this committee-based, invitation process.

Appendix

Please remember that for Table 5 below, the original ILM was only trained on four years of data (2014-2017); all seven other models appearing (in the column headings) were trained on all 9 years.

Table 4 – Five significant digits for each of the eight models' five weights.

	Orig	33,100	38,99	40,98	41.96	SCC	F_Sqr	F_*
OD	.30913	.39373	.70289	.40242	.88582	.28330	.30983	.33398
SOS	.83785	.98945	.98127	.46362	.91212	.94684	.85494	.95948
OD'	85.995	40.499	53.145	19.448	32.798	63.137	41.983	52.089
SOS'	49.288	23.542	23.973	15.300	33.371	30.693	14.301	19.934
Losses	.44386	2.5212	.70209	2.1111	4.8128	.07248	.22124	.09605

Table 5 – Results matching the CFP committee's top twelve teams: exact, overall.

	Orig	33,100	38,99	40,98	41,96	SCC	F_Sqr	F_*
2014	4,11	2,10	4,10	4,10	4,10	3,11	2,10	2,10
2015	5,9	4,11	3,11	4,11	5,11	3,10	2,11	3,11
2016	3,11	2,11	2,11	3.11	3,11	3,11	2,11	2,11
2017	2,11	3,11	3,11	4,11	5,11	1,11	1,11	2,11
2018	8,10	7,11	7,11	8,11	7,11	6,10	6,10	6,10
2019	2,10	5,11	6,10	5,10	6,9	5,11	5,11	5,11
2021	3,10	4,11	5,11	4,11	3,11	1,10	4,11	4,11
2022	3,12	5,12	5,12	5,12	4,11	4,11	7,12	7,12
2023	1,12	1,10	3,12	3,11	4,11	2,12	1,12	1,12
Totals	31,96	33,100	38,99	40,98	41,96	28,96	30,99	32,99

References

[1] Carroll, B., Palmer, P., Thorn, J. (1988) *The Hidden Game of Football*. Warner Books.
[2] Trono, J. (2020) *An Accurate Linear Model for Predicting the College Football Playoff Committee's Selections*. SMC Tech Report (SMC-2020-CS-001), last accessed on 5/20/2025, https://www.smcvt.edu/wp-content/uploads/2021/08/AccurateLinearCFPModel.pdf.

The trade-off between model flexibility and accuracy of the Expected Threat model in football

Koen W. van Arem[1*], Jakob Söhl[1], Mirjam Bruinsma[2] and Geurt Jongbloed[1]

[1] Delft University of Technology, The Netherlands
[2] AFC Ajax, The Netherlands
* k.w.vanarem@tudelft.nl

Abstract

With an average football (soccer) match recording over 3,000 on-ball events, effective use of this event data is essential for practitioners at football clubs to obtain meaningful insights. Models can extract more information from this data, and explainable methods can make them more accessible to practitioners. The Expected Threat model has been praised for its explainability and offers an accessible option. However, selecting the grid size is a challenging key design choice that has to be made when applying the Expected Threat model. Using a finer grid leads to a more flexible model that can better distinguish between different situations, but the accuracy of the estimates deteriorates with a more flexible model. Consequently, practitioners face challenges in balancing the trade-off between model flexibility and model accuracy. In this study, the Expected Threat model is analyzed from a theoretical perspective and simulations are performed based on the Markov chain of the model to examine its behavior in practice. Our theoretical results establish an upper bound on the error of the Expected Threat model for different flexibilities. Based on the simulations, a more accurate characterization of the model's error is provided, improving over the theoretical bound. Finally, these insights are converted into a practical rule of thumb to help practitioners choose the right balance between the model flexibility and the desired accuracy of the Expected Threat model.

1 Introduction

An average football (soccer) match contains around 3,000 on-ball events [1]. This data can provide new meaningful insights for football clubs. It can, for instance, be used to study players for scouting or opponent analysis. This can help practitioners make better informed decisions. Mathematical models can be used to extract more complex information from these on-ball events [2], which can give football clubs an advantage over the competition.

However, advanced mathematical models introduced by researchers might not be adaptable or practical enough for coaches and teams [3]. It is important to have models that can be explained in such a way that they can be understood by practitioners and utilized in practice [3]. To achieve this, models to distill information from the in-game data should also be assessed on how explainable and interpretable they are [4]. This means that explainable models are of added value because they offer an accessible option to retrieve complex information from in-game data.

The Expected Threat [5, 6] model quantifies the quality of an action, which gives detailed information about offensive player quality, and it is praised for its interpretability [7]. Nonetheless, practitioners still face one key design choice when applying the Expected Threat model. The Expected Threat model estimates the probability of scoring, and to do this, it divides the pitch into different in-game states. The number of in-game states, M, describes the model flexibility and can be chosen. A model with more game states can distinguish between more situations [8]. On the other hand, more states decrease the accuracy of the Expected Threat model because more probabilities are estimated with the same amount of data. Increasing the amount of data is often infeasible in practice, because of the additional costs. This trade-off between model flexibility and accuracy hinders the application of the otherwise accessible Expected Threat model in practice. The aim of this study is to provide a rule of thumb for practitioners to manage the trade-off between model flexibility and accuracy of the Expected Threat model.

2 Expected Threat

The Expected Threat model [5, 6] considers football as a Markov chain and is based on the idea that good actions increase the probability of scoring a goal within the possession chain. The pitch is divided into M squares, and the state of the game is defined as the square where the ball-possessing player is. For each state s, the model then calculates the probability of scoring, denoted by $xT(s)$. The quality of an action is defined as the difference before and after the action: $\Delta xT(s_{\text{before}}, s_{\text{after}}) = xT(s_{\text{after}}) - xT(s_{\text{before}})$.

When a player has ball possession, there are two ways to score a goal: either directly score a goal or move the ball to another state with a dribble or pass and score from there. To score a goal, the player has to decide to shoot, and the player has to score the shot. If the current game state is denoted as s, the probability of this happening is $P(shot|s) \cdot P(goal|shot, s)$. Because $P(goal|shot, s)$ is the quantity described by Expected Goal (xG) models, this can also be denoted as $P(shot|s) \cdot xG(s)$. If the player decides not to shoot, a goal can be scored by moving the ball to each other game state s' and scoring from there. The probability of scoring via the game state s' can be written as $T_{s\to s'} \cdot xT(s')$, where $T_{s\to s'}$ is the probability of transitioning from s to s'. This means that the probability of scoring from game state s is

$$xT(s) = P(shot|s) \cdot xG(s) + \sum_{s'} T_{s\to s'} \cdot xT(s'). \tag{1}$$

In practice, the values of $P(shot|s)$, $xG(s)$, and $T_{s\to s'}$ are estimated by counting the occurrences of these events in the data set. When these are estimated, the only unknowns in (1) are the values $xT(s)$. Because (1) holds for each state s, it gives a system of equations, which is generally solved using an iterative algorithm. In this way, the probability of scoring from state s is estimated.

Due to randomness in the training data, errors are made in the estimation of $P(shot|s)$, $xG(s)$, and $T_{s\to s'}$. These estimation errors cause errors in the estimated xT-values. For practitioners, it is important to have a bound on these errors. The model error in this study is defined as the maximal difference between the true and the estimated xT-values, denoted by $||xT - \widehat{xT}||_\infty$. The distribution of this error depends on the number of training points N and the number of game states M and can be used to describe the trade-off between model flexibility, described by M, and model accuracy.

Using the properties of the Expected Threat model, it is possible to derive probabilistic bounds on the error of the model. For this purpose, the transition matrix is the matrix with the transition probabilities $T_{s\to s'}$ as entries. A summary of our theoretical results is described in the following theorem.

Theorem 1. *Let $g \in \mathbb{R}^M$ be defined by $g_s = P(shot|s) \cdot xG(s)$ and let T be the transition matrix. Assume that $||T||_\infty < 1$ and that for the estimated transition matrix $||\hat{T}||_\infty < 1$. Moreover, assume that the estimates of the quantities in the Expected Threat model are obtained by taking averages of N independent Bernoulli random variables. Then the following bounds hold with probability at least $1-\alpha$:*

$$||\widehat{xT} - xT||_\infty \leq \frac{1}{1-||T||_\infty}\left(M\sqrt{\frac{\log(2M^2/\alpha)}{2N}} + \sqrt{\frac{\log(2M/\alpha)}{2N}}\right) \leq \frac{2}{1-||T||_\infty}\left(M\sqrt{\frac{\log(2M^2/\alpha)}{2N}}\right). \quad (2)$$

More specifically, the term $M\sqrt{\frac{\log(2M^2/\alpha)}{2N}}$ corresponds to the error in estimating T and $\sqrt{\frac{\log(2M/\alpha)}{2N}}$ to the error in estimating g.

This theorem shows that the error in estimating the xT-values is of the order $O(M\sqrt{\log(M)}/\sqrt{N})$. However, the results also suggest that, in practice, implemented models might fall under a finite sample regime where the error in g is still larger, and where a faster decay of the error can be observed.

3 Methods

To find the error distribution of the Expected Threat model in practice, simulations were performed based on the Markov chain underlying the model. The data used for this simulation study is obtained from the openly available Statsbomb data set [9]. All available games from the Premier League, Ligue 1, Serie A, La Liga, and the Bundesliga were used. The events that did not describe passes, dribbles, errors, clearances, or shots were filtered out. This resulted in a data set of approximately 4,000,000 events, which is equivalent to around 6.5 full seasons of one league.

In this research, the maximal model error was studied for Expected Threat models with different discretization grids, and thus for different flexibilities M, which are described in Table 1. For each grid size M, one Expected Threat model was calculated with the Statsbomb data. These models were assumed to be the ground truth for elite male leagues within the scope of this study. Because the data set is relatively large, this is a reasonable assumption.

Each of these ground truth models describes a Markov chain, which can be used to resample a new data set. For each ground truth model, new data sets of different sizes N were resampled, which were then used to train resampled Expected Threat models. The model error was then obtained by calculating the maximal absolute difference between the ground truth xT-values and the xT-values based on the resampled data. This process was repeated 1,000 times for each combination of M and N as described in Table 1. This created 48,000 data points describing the model error, the grid size M and the number of data points N.

The goal of this research is to obtain more insight into the trade-off between the model flexibility governed by M and the model accuracy, described by the model error. Because situations with $M\log(M)/\sqrt{N} \geq 15$ resulted in errors too large for practice, these were filtered out. This makes it possible to study the distribution of the errors in this setting, which is interesting for practitioners

Using the simulated data, the distribution of the maximal model error could be studied. To do this, the following lognormal model was assumed to describe the maximal model error:

$$||\widehat{xT} - xT||_\infty = C\frac{M^\alpha}{(\sqrt{N})^\beta}e^\varepsilon, \quad \text{where } \varepsilon \sim N(0,\sigma^2). \quad (3)$$

n_x	n_y	$M = n_x \cdot n_y$	N
16	12	192	100,000
32	24	768	130,000
40	30	1200	170,000
48	36	1728	240,000
56	42	2352	370,000
64	48	3072	630,000
			1,300,000
			4,000,000

Table 1: Grid configurations with corresponding M values and the sample sizes N of the simulations.

The model error, which is a maximal absolute difference, is known to be positive. Additionally, the theoretical results indicated that both the mean and spread of the error are small if M is small and N is large. This makes the lognormal model a reasonable assumption. Moreover, this model describes the powers of the variables M and N, which are unknown because the theoretical results only gave an upper bound.

If $c = \log(C)$, (3) is equivalent with

$$\log(||\widehat{xT} - xT||_\infty) = c + \alpha \log(M) - \beta \log(\sqrt{N}) + \varepsilon, \quad \text{where } \varepsilon \sim N(0, \sigma^2). \tag{4}$$

This formulation gives a linear model with normal residuals. Therefore, ordinary least squares (OLS) can be applied to estimate c, α, β, and σ^2.

4 Results

Dep. Variable:	$\|\|\widehat{xT} - xT\|\|_\infty$	**R-squared:**	0.835
Model:	OLS	**Adj. R-squared:**	0.835
No. Observations:	23000	**Log-Likelihood:**	-9297.8
Df Residuals:	22997	**AIC:**	1.860e+04
Df Model:	2	**BIC:**	1.863e+04

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
c	-1.8758	0.026	-72.138	0.000	-1.927	-1.825
α	0.9898	0.003	330.569	0.000	0.984	0.996
β	1.0416	0.004	238.837	0.000	1.033	1.050
Variance residuals	0.1314					

Table 2: Summary of the OLS model fitted on the log maximal model error for the data points with $M\log(M)/\sqrt{N} < 15$.

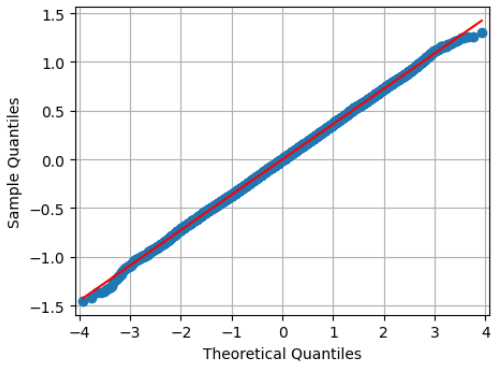


Figure 1: A QQ plot of the residuals of the fitted OLS model.

The summary of the ordinary least squares applied to (4) is given in Table 2. It shows that the adjusted R^2 is 0.835, which indicates that the model is able to explain most variance of the errors with M and N. Additionally, Figure 1 shows the QQ plot of the residuals of the lognormal OLS model. It is visible that the residuals indeed seem to be well-described by the lognormal OLS model. Thus, it can be concluded that the model in (4) provides a good description of the distribution of the maximal error of the Expected Threat model. With the found values for α, β and σ^2, the distribution of the model can be described by

$$||\widehat{xT} - xT||_\infty = e^{-1.8758} \frac{M^{0.9898}}{(\sqrt{N})^{1.0416}} e^{\varepsilon}, \quad \text{where } \varepsilon \sim N(0, 0.1314). \tag{5}$$

The values found for α and β are close to 1, although significantly different according to the confidence intervals in Table 2. This means that the maximal error of the model is of order $O(M^{0.9898}/(\sqrt{N})^{1.0416})$, which indicates that the error, in practice, is of a lower order than established by the theoretical results.

5 Rule of thumb

With the distribution of the model error, it is possible to give guidance to practitioners on how to use the Expected Threat model. If they have an existing model, the distribution of the error in (5) can be used to describe the distribution of the error of their model. For example, consider the Expected Threat model by Singh [6], which has a 16×12 grid and one season of the Premier League as training data. This corresponds to $M = 16 \cdot 12 = 192$ game states and roughly $N = 620{,}000$ training data points. In the experience of consulted experts, errors smaller than 0.03 are acceptable for scouting purposes. The distribution of the maximal error of this model is visualized in Figure 2. It indicates that there is a 62.09% chance of having an error that is lower than 0.03. This means that there is a reasonable chance of this model having an acceptable error.

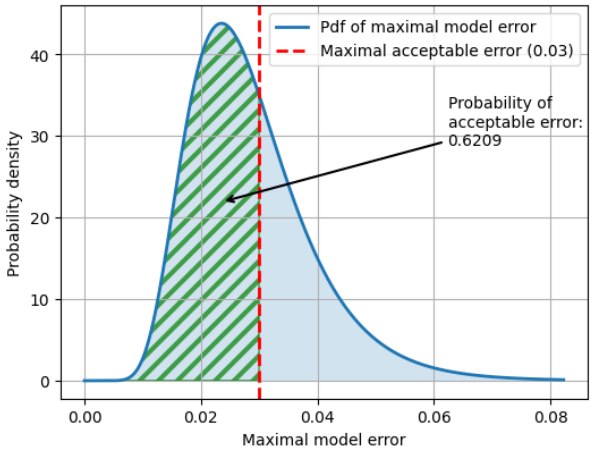


Figure 2: The distribution of the maximal model error of the Expected Threat model in [6], where $M = 192$ and $N = 620{,}000$.

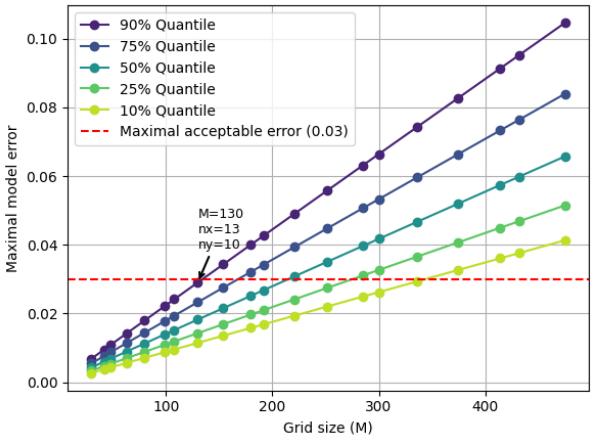


Figure 3: The quantiles of the maximal model error of an Expected Threat model with $N = 2{,}480{,}000$ training data points for different grid sizes M.

When training a new model, the number of game states M has to be chosen. It is desirable to have a low model error with a high probability. On the other hand, a more flexible model can better distinguish between in-game situations. To balance this trade-off, a reasonable idea would be to choose the most flexible model with an acceptable statistical error. Through consultation with experts, it was established that the Expected Threat model is sufficiently reliable for scouting purposes when the maximal model error is smaller than 0.03 with a 90% probability. This can be reformulated as the following rule of thumb: *select the most flexible model (highest M) such that the maximal model error is smaller than 0.03 with a 90% probability.*

To illustrate this rule of thumb, suppose a practitioner wants to train an Expected Threat model on a data set of 2,480,000 data points. This corresponds to data of 4 league seasons. Figure 3 shows different quantiles of the error for values of M for this number of data points N based on (5). The results show that the maximal M with a 90% quantile smaller than 0.03 is $M = 130$, which corresponds to a 13×10 grid. This means that the rule of thumb gives that a 13×10 grid yields the most flexible model with an acceptable model error. In this way, the rule of thumb provides guidance to practitioners on how to deal with the trade-off between model flexibility and accuracy, which makes the accessible Expected Threat model even more easily applicable for practitioners.

Acknowledgements

The authors would like to thank StatsBomb for making the StatsBomb Open Data publicly available.

References

[1] Yorke, J. (2022) *StatsBomb 360: Exploring Line Breaking Passes.* Accessed: 14-5-2025 at `https://statsbomb.com/articles/soccer/statsbomb-360-exploring-line-breaking-passes/`.

[2] Rein, R. and Memmert, D (2016) *Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science.* SpringerPlus **5**, 1410.

[3] Herold, M., Goes, F., Nopp, S., Bauer, P., Thompson, C., and Meyer, T. (2019) *Machine learning in men's professional football: Current applications and future directions for improving attacking play.* International Journal of Sports Science & Coaching, **14**(6), 798-817.

[4] Davis, J., Bransen, L., Devos, L., Jaspers, A., Meert, W., Robberechts, P., Van Haaren, J. and Van Roy, M. (2024) *Methodology and evaluation in sports analytics: challenges, approaches, and lessons learned.* Machine Learning **113**, 6977–7010.

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

[6] Sing, K. (2018) *Introducing Expected Threat xT.* Accessed: 5-11-2024 at `https://karun.in/blog/expected-threat.html`.

[7] Van Roy, M., Robberechts, P., Decroos, T., and Davis, J. (2020) *Valuing On-the-Ball Actions in Soccer: A Critical Comparison of xT and VAEP.* In Proc. of the AAAI-20 Workshop on Artifical Intelligence in Team Sports.

[8] Van Arem, K., Bruinsma, M. (2024) *Extended xThreat: an explainable quality assessment method for actions in football using game context.* In Proc. ISEA 2024 - The Engineering of Sport 15.

[9] StatsBomb (2024) *Open Data.* Accessed: 15-10-2024 at `https://github.com/statsbomb/open-data`.

Multisport YODA: Cognitively-Driven AI Adaptation for Cross-Sport Psychometric Profiling and Analytics

Sadanand Venkataraman[†], Sundharakumar K.B[†], Bharathi Malakreddy A[‡], Hema A Murthy[§†], Santhi Natarajan[†]

[†]Cognition Lab, Shiv Nadar University Chennai
[‡]Research & Consultancy Cell, BMS Institute of Technology & Management, Bengaluru
[§]Indian Institute of Technology Madras

Abstract

Mental and cognitive attributes are pivotal in sports, often shaping competitive success. This paper builds on Your Offence and Defence Analysis (YODA), an established football psychometric framework, proposing its methodical extension to multiple sports. We detail an approach utilizing Large Language Models (LLMs) as an assistive technology to adapt YODA's football-centric scenarios for new sporting contexts, piloted with cricket. This LLM-assisted adaptation shows the need to be critically guided by an expert validation protocol to ensure psychometric integrity and contextual authenticity. The YODA framework, which maps player responses in simulated scenarios to primary traits and sub-traits, remains the core engine. This work outlines the adaptation pathway, emphasizing how YODA's foundational strengths can be broadened across diverse athletic domains. While this paper focuses on scenario adaptation, with automated scoring as future work, it envisions Multisport YODA not only accelerating mental performance analytics but also contributing to the development of more specialized, domain-aware AI, such as Sports-Specific Language Models.

1 Introduction

The evolving landscape of sports performance analytics increasingly acknowledges the impact of cognitive and psychological factors on athletic success. Quantifying physical prowess is established; however, assessing elite performers' mental attributes remains complex[8], with existing psychometric tools often being sport-specific. Existing psychometric tools are often sport-specific. This specificity restricts their broader applicability, often necessitating extensive manual redevelopment and validation efforts to adapt them to the unique contextual demands of different sporting environments.

Our prior work introduced YODA, a psychometric instrument evaluating football players' cognitive profiles via simulated game scenarios [5]. YODA's strength lies in its bottom-up approach, mapping player reactions to psychological traits, providing deep insights into their mental framework. The scenarios within YODA, however, were calibrated for football. Extending such a robust tool to other sports is key for holistic athlete assessment.

This paper proposes "Multisport YODA," an extension of this YODA framework. We explore Open Source LLMs as an *accelerator* for adapting scenarios from football to cricket. While LLMs (e.g., Llama

3[6], piloted here) offer text adaptation capabilities, current general-purpose models struggle with nuanced sports contexts without expert guidance[7]. They provide a valuable starting point, but are not a solution.

This paper details leveraging LLMs as assistive tools within YODA, emphasizing technology's role in *facilitating* YODA's extension. The expert validation protocol for psychometric integrity and contextual authenticity is still needed. This paper concentrates on scenario adaptation; automated scoring is future work. We aim to demonstrate a feasible, YODA-centric methodology for versatile cognitive assessment, potentially informing domain-specific sports AI. Subsequent sections outline the YODA framework, AI-assisted adaptation, expert validation, illustrative outcomes, and implications.

2 YODA: A Foundational Football Psychometric Framework

YODA posits that athletes' on-field decisions under stress reflect their mental framework and cognitive processing. YODA uses a scenario-based assessment, designed and validated in football [3], to evaluate these attributes. Its robust architecture and evaluative capacity form a strong foundation for a multisport cognitive analytics system[5].

2.1 Core Psychometric Principles

YODA immerses players in realistic football scenarios(N=48), prompting responses based on likely thoughts, feelings, and actions. Responses (Likert scale) map to 14 primary psychological traits. Key traits include Drive and Determination, Game Sense, and Coachability [5, 3]. YODA also derives 58 sub-traits for a detailed psychological map of football players. The Big Five personality traits are also integrated for holistic understanding. This multi-layered approach offers a comprehensive view of a player's psyche, beyond skill evaluation, to uncover cognitive and personality drivers.

2.2 Overview of the Football Cohort

YODA's development and validation involved a substantial, diverse football cohort. This yielded rich data, establishing baseline cognitive patterns and refining the tool. The primary cohort (N=118 assessments) is detailed in Table 1.

Table 1: Composition of the Primary YODA Football Assessment Cohort.

Team/Group Description	**Category**	**No. of Players**
BMS Institute of Technology & Management, Bengaluru	College Men's Team	53*
TKM College of Engineering, Kollam	College Men's Team	45*
Crescent FC	Youth Players†	6
Ramaiah School of Law	College Men's Team	2
Christ School of Law	College Men's Team	2
Other Individuals	Individual Players†	10
Total		**118***

*A portion of these assessments were conducted in participants' native languages and are pending full translation and comparative quantitative evaluation.

†These cohorts include both men and women, with ages ranging from sub-15 youth to senior collegiate players.

This cohort (Table 1) forms a comprehensive baseline, with diverse experience and roles. Prior YODA analyses show patterns like prominent 'Learnability' and 'Coachability' in subgroups [4]. This data and YODA's depth provide critical grounding for its multisport extension, positioning it as a foundational technology for sports psychometry.

3 Methodology: Extending YODA via AI-Assisted Adaptation

Extending YODA requires transforming football scenarios to reflect other sports' unique contexts while preserving psychometric integrity. We detail using an Open Source LLM as an *assistive tool* for adapting scenarios from football to cricket, emphasizing subsequent expert validation.

3.1 Preserving Psychometric Integrity

Effective cross-sport adaptation demands more than superficial terminological changes. Each sport possesses a distinct "cognitive fingerprint." To maintain YODA's validity, adapted scenarios must be contextually authentic and elicit comparable psychological responses to the originals. The objective is functional psychometric equivalence[1], probing traits like "Game Sense" with similar depth in cricket as in football.

3.2 LLM as an Assistive Tool for Scenario Transformation

Multisport YODA uses an adaptable LLM to aid scenario transformation. This pilot uses Llama 3[6] via Ollama. This offers research advantages:

- Llama 3 offers strong open-source performance for language understanding and generation [6].
- Ollama simplifies local LLM deployment and management, offering ease of use and **portability**.
- This balances performance, support, and accessibility for local deployment.

The Llama 3 model is prompted to re-engineer YODA's football scenarios into cricket-relevant situations, guided by constructed prompts. The LLM receives the football scenario, its YODA traits, and response-trait mapping. The LLM then generates a cricket-equivalent scenario intended to elicit similar cognitive responses. Prompts guided the LLM by providing the original scenario, target traits, and response mappings, asking it to create cricket-equivalents preserving psychological demands and trait assessment.

3.3 Proposed Expert Validation Protocol

While Llama 3 generates initial adaptations, robust psychometric validity and contextual authenticity require rigorous expert validation. General-purpose LLMs lack the nuanced sports understanding of expert coaches and psychologists[7]. The human-in-the-loop validation process is illustrated in Figure 1.

An expert panel of experienced cricket coaches and sports psychologists will assess scenarios based on: 1) **Realism and Authenticity (R)**, 2) **Relevance to Intended Trait(s)**, 3) **Clarity and Unambiguity**, 4) **Cultural and Contextual Appropriateness**, and 5) **Psychometric Equivalence (Cognitive Load - C)**. Experts will review each scenario, providing qualitative feedback and quantitative ratings (0-5 scale) for

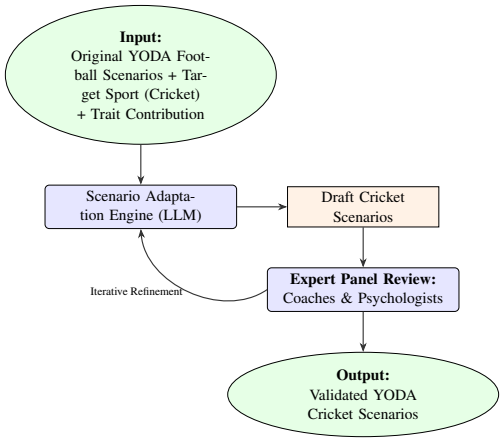


Figure 1: Flowchart of the LLM-assisted scenario adaptation for Multisport YODA

Table 2: YODA scenario adaptation from football to cricket using a Llama 3.

Trait(s) Assessed	**Original YODA Football Scenario**	**LLM-Adapted Cricket Scenario (Initial Draft)**
Responsibility, Team Spirit	(B3) You make a mistake that leads directly to the opposing team scoring a goal. Your teammates look to you, and there's still time to try and turn things around. In this situation, how likely are you to apologize to your teammates and focus on making up for it during the rest of the game?	(B3) You're given a crucial over in the final innings, but you've lost your wicket on the first ball of the over. The opposition is now in a strong position to win the match. Your team's captain asks you if you can try and turn this around by taking some aggressive shots in the remaining balls of the over. In this situation, how likely are you to take an aggressive approach with your bat, potentially putting pressure on the opposing team?
Team Sense, Decision-Making Under Pressure	(I1) It's a two-on-one situation against the opposing goalkeeper, and you're in possession. You have a chance to score yourself, but passing could lead to a simpler goal. In this situation, how likely are you to take the shot yourself rather than pass to your teammate?	(I1) You're batting with a partner, and the opposing team's fielder is standing very close to you, making it difficult to play a safe shot. You need to decide whether to hit a powerful drive down the ground or try to play a more defensive shot that might be easier for your partner to score from.

Realism (R), Cognitive Load (C), and Terminology/Contextual Accuracy (T). An overall quality score q_j is computed:

$$q_j = \frac{\sqrt{R^2 + C^2 + T^2}}{\sqrt{75}} \in [0,1]. \tag{1}$$

Scenarios with high consensus q_j (e.g., ≥ 0.8) and positive qualitative endorsement form the final battery. Others undergo iterative refinement via manual rewriting or new LLM prompts, ensuring quality.

4 Illustrative Adaptation Outputs

This section presents illustrative YODA scenario adaptations from football to cricket by Llama 3, generated *prior* to the expert validation protocol (Section 3.3). These examples show the LLM's initial transformation capability and highlight areas where expert input is indispensable for psychometric validity and contextual relevance. The LLM was tasked to maintain the original scenario's core psychological assessment goal. Table 2 shows examples.

The LLM-generated drafts (Table 2) show initial contextual mapping (e.g., "defensive error" to "misfield"). However, these examples show why expert validation is a critical, non-negotiable phase. Key challenges needing expert refinement include:

- **Capturing Nuanced Tactical Depth**: Adapting "breaking a compact defense" to "scoring against a defensive field" is too generic for cricket; specifics for format/game situation (e.g., T20 powerplay vs. Test match field) are needed. Current LLMs struggle with this.
- **Reflecting Authentic Player Interactions**: Leadership and team interaction after errors vary significantly in cricket by context. LLM adaptations may be plausible but superficial, missing specific psychological stressors.
- **Ensuring Psychometric Equivalence**: Maintaining comparable cognitive load is a fine-grained psychometric task beyond current LLMs.

These observations reinforce that while LLMs aid drafting, YODA's integrity relies on human experts' deep domain knowledge.

5 Discussion: YODA, Analytics, and the Need for Sport-Specific AI

Adapting YODA scenarios with Llama 3 shows Open Source LLMs are promising *assistive tools* for initial drafting in multisport psychometric instrument development. They handle foundational contextual mapping and linguistic transformations, potentially accelerating development. However, for complex applications like sports cognitive profiling, current general-purpose LLMs are not holistic solutions. Multisport YODA's strength stems from its psychometrically grounded framework; technology is supportive, and expert judgment is paramount.

5.1 The Limits of General LLMs in Specialized Sports Contexts

Pilot adaptation, while showing LLM drafting utility, highlights general-purpose model limitations in nuanced sports psychometry. Ensuring validity for cricket demands rigorous expert feedback. Subtle tactics, authentic sporting voice, and specific psychological pressures often exceed current LLM capabilities. These models, trained on broad text, lack ingrained sports knowledge and contextual reasoning vital for specialized assessments. LLMs struggle with deep sports understanding, confirming they are not yet turnkey solutions for intricate sports analytics[7]. Expert validation is thus an indispensable core component of psychometric engineering.

5.2 YODA as a Stepping Stone Towards Sport-Specific Cognitive Frameworks

Challenges reinforce the need for frameworks that understand sport and player cognition to quantify mental attributes. This LLM limitation opens research avenues relevant to MathSport's ethos of advancing quantitative and computational approaches. YODA's detailed trait mapping and scenario methodology provides a foundational structure for sport-specific understanding[3, 4]. The quantitative expert validation (Section 3.3, Equation 1) offers a mathematical approach to assessing adapted scenario quality. A validated Multisport YODA can yield accessible, standardized cognitive assessments, crucial for performance modeling and talent identification.

6 Conclusion and Future Directions

This paper presented a YODA-centric methodology for extending a validated football psychometric framework, piloting Llama 3 LLM for adapting scenarios to cricket as an assistive drafting tool. Illustrative outputs show LLM's initial transformation capability but underscore that robust expert validation is paramount for psychometric and contextual integrity. While LLMs aid drafting, expert human input is indispensable for refining these into valid assessments. YODA remains the foundational psychometric engine, with technology as a guided enabler for its expansion.

The gap between general LLM capabilities and sports application needs suggests a potential need for **Domain-Specific Language Models** (e.g., SSLMs)[2, 5] for sports. These require training/fine-tuning on curated sports-specific corpora (psychometric data, tactics, expert annotations). Multisport YODA, by gathering validated data, could be a **stepping stone** for these specialized AI tools. YODA's psychometrically grounded approach, prioritizing deep cognitive understanding, can help build foundational datasets for next-gen sports AI. The vision for Multisport YODA is a adaptable system enhancing mental performance analytics. By making cognitive tools more accessible and tailored, YODA aims for deeper insights for development, identification, and coaching.

Future work prioritizes implementing and analyzing the expert validation protocol (Section 3.3), then quantitative psychometric evaluation of validated cricket scenarios, and other team sports(hocky, and basketball). Developing and validating an **Sports-Specific Language Models *(SSPM)*[2]** automated scoring engine is a key subsequent objective.

References

[1] Ravenda, F., Bahrainian, S.A., Raballo, A., Mira, A., & Kando, N. (2025). Are LLMs Effective Psychological Assessors? Leveraging Adaptive RAG for Interpretable Mental Health Screening through Psychometric Practice. *arXiv:2501.00982*.

[2] Chen, Z., Li, C., Xie, X., & Dube, P. (2024). OnlySportsLM: Optimising Sports-Domain Language Models with SOTA Performance under 1 Billion Parameters. *arXiv:2409.00286*.

[3] Venkataraman, S., et al. (2024). Decoding the Psyche: Engineering Psychological Profiles in Football through YODA Analysis. *Proc. ISEA 2024*.

[4] Venkataraman, S., et al. (2024). Unveiling the Mental Game: Leveraging YODA Psychometry in Football Performance Analysis. *Proc. MathSport Australasia 2024 Conf.*

[5] Venkataraman, S., Sundharakumar, K.B., Malakreddy A, B., & Natarajan, S. (2024). YUVA-SQ: A Cognitive Scouting Model for The Beautiful Game. *Proc. ICITIIT 2024*, 1–6. https://doi.org/10.1109/ICITIIT61487.2024.10580784

[6] Touvron, H., et al. (2024). Llama 3: Open Foundation and Fine-Tuned Chat Models. *arXiv:2404.10719*.

[7] Xia, H., Yang, Z., Wang, Y., et al. (2024). SportQA: A Benchmark for Sports Understanding in Large Language Models. *arXiv:2402.15862*.

[8] Smith, R.E., Smoll, F.L., Cumming, S.P., & Grossbard, J.R. (2006). Measurement of Multidimensional Sport Performance Anxiety in Children and Adults: The Sport Anxiety Scale-2. *Journal of Sport & Exercise Psychology, 28*(4), 479-501.

Performance Evaluation and Ranking of Drivers in Multiple Motorsports Using Massey's Method

R. Yamaguchi and E. Konaka

*Meijo University. 243426041@ccmailg.meijo-u.ac.jp

** Meijo Univerisity. 1-501, Shiogamaguchi, Tempaku-ku, Nagoya, JAPAN. email address: konaka@meijo-u.ac.jp

Abstract

Motorsports are competitions based on the skill of driving a vehicle, and various championships are held with different vehicles and regulations. A feeder series structure is formed between some championships. Most drivers first participate in lower-level championships, and those who achieve good results challenge higher-level championships. However, many drivers also move to and participate in championships with significantly different vehicles, regulations, or regions. Due to this situation, each driver experienced different pathways, and it is not easy to evaluate and compare their previous achievements. This research proposes a method for quantitatively evaluating the achievements of all drivers who have participated in the target championship using Massey's rating method, which is a quantitative ability evaluation method.

1 Introduction

Motorsports are competitions based on the skill of driving a vehicle, and various championships are held with different vehicles and regulations. In this study, we focus on championships that use formula cars, for reasons such as the diversity of the levels and regions of the championships held, and the fact that Formula 1 (F1) is widely regarded as the pinnacle of global motorsports.

A feeder series structure is formed between some championships. Most drivers first participate in lower-level championships, and those who achieve good results challenge higher-level championships.

Figure 1 shows the eight championships considered in this study. At the highest level is F1, which is held all over the world, mainly in Europe, and F2 and F3 are its feeder series. Formula E is a championship that uses electric cars, different from F1 to F3.

In Japan and North America, Super Formula and Indy Car are the top-level championships, respectively, with lower-level championships for each.

While the specifications of the cars used in each championship, the regions, the regulations, and various other factors differ, there are connections among the championships. Some drivers move between championships, and higher-ranked drivers except F1 will be awarded Super License Points that are necessary for participating in F1.

Figure 2 shows the movement of drivers from 2021 to 2023, expressed as a directed graph. The nodes represent championships, and the edges represent the direction and number of drivers moved within the period.

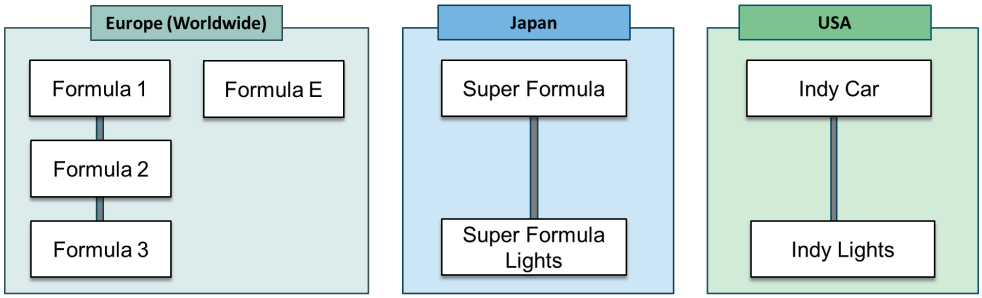


Figure 1: Championships of Formula Cars

Since the entire motorsports world is structured in this mannar, each driver has a different history of participation. Not only do the regulations differ for each championship, but the abilities of the drivers who participate also differ.

Therefore, it is extremely difficult to directly compare past achievements between drivers with different histories.

1.1 Super License Points

F1 is not only a formula car championship but also one of the most popular sports series in the world, and it is the highest peak of motorsports.

The Super License shall be issued to drivers who can participate in F1[1]. The drivers who meet multiple conditions shall be issued the license[2]. One of the conditions is that the driver must have earned 40 or more Super License Points in the championships they have competed in over the past three years. Super license points are allocated to some motor sports championships, and the top-ranking drivers in each championship can earn them.

Table1 shows the Super Licence Points awarded to the top drivers in the eight championships targeted in this study.

Table 1: Allocation of Super License Points[2]

Category	Short Name	1st	2nd	3rd	...
Formula 2	F2	40	40	40	...
Formula 3	F3	30	25	20	...
Formula E	FE	30	25	20	...
Super Formula	SF	25	20	15	...
Super Formula Lights	SFL	15	12	10	...
Indy Car	Indy	40	30	20	...
Indy Lights	IndyL	15	12	10	...

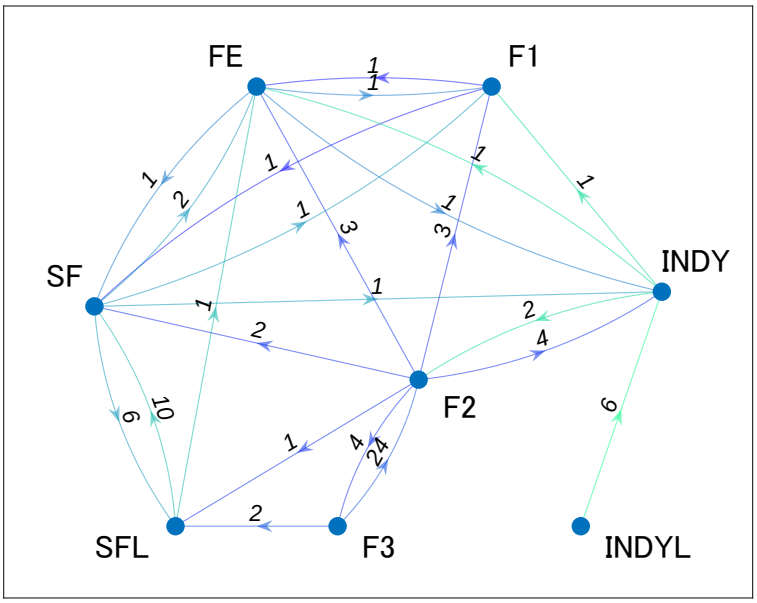


Figure 2: Driver move among championships

These values were set by the FIA (Fédération Internationale de l'Automobile), the governing body of the championships. The table reflects their assumption of the hierarchy between the championships. For instance, F2 is the highest-level championship in the table; F3 and FE are at a similar level of competition, and Super Formula is one level below them. IndyCar is at the level between F2 and F3.

1.2 Objective

The main objective of this research is to quantify the achievements of drivers participating in multiple championships and to order them as a single unified ranking. The ranking would be helpful for enthusiasts, team owners, and drivers themselves.

Super license points appear to be usable for quantifying the achievements of drivers participating in championships below F1, but there are some problems. Since 40 points are required for issuing the Super License, the maximum points that each driver could earn in one championship is 40. The top three drivers in F2 all earn 40 points. Moreover, only the top drivers in each championship can earn the points, therefore it is not suitable for creating an overall ranking. The biggest problem is that active F1 drivers are not awarded super license points.

Within the scope of the authors' research, there are no point systems or rankings defined by the FIA across multiple championships other than Super License Points.

The main objective of this research is to develop a method for quantitatively evaluating the achievements of all drivers participating in a championship using Massey's rating method, which is the most basic quantitative evaluation method.

Massey's method is a rating method for one-on-one competitive sports. In this research, we propose a

method by extending it to a race in which multiple vehicles compete for position while driving at the same time.

2 Data

Race results of eight championships in Fig.1 (F1, F2, F3, FE, SF, SFL, Indy, and IndyL) from 2021 to 2023 seasons were collected from the website [3].

The collected data include the race date, round, driver name, and finishing position. Drivers who participated only in practice sessions are excluded from the dataset. Retired drivers are ranked after those who finished the race, in descending order of the number of laps completed.

Table 2 lists the number of rounds and drivers included in the data set.

Table 2: Data set

Championship	Period	Rounds	Drivers
F1	2021-2023	66	29
F2	2021-2023	71	48
F3	2021-2023	58	79
FE	2021-2023	47	33
SF	2021-2023	26	32
SFL	2021-2023	53	32
Indy	2021-2023	50	55
IndyL	2021-2022	34	25
Total (unique)	2021-2023	405	275

3 Method

The core idea of the proposed method is as follows. To evaluate driver performance, the results of drivers who have moved between championships are used as a reference to compare the performances of other drivers. Specifically, this study proposes a modified version of Massey's rating method[4] for evaluating driver performance.

3.1 Massey's method

Massey's rating method estimates score differentials by solving an overdetermined system using the least squares method. The fundamental principle of Massey's method can be expressed as follows: if y_k represents the score differential in match k, and r_i and r_j denote the ratings of teams i and j, respectively, then

$$r_i - r_j = y_k + \varepsilon_k, \tag{1}$$

where ε_k denotes estimation error.

This equation implies that the rating difference between two teams ideally predicts the score differential between them. Since the score differentials are known, the system forms an n-dimensional system of m linear equations when there are m matches and n teams:

$$\mathbf{Xr} = \mathbf{y} + \boldsymbol{\varepsilon} \tag{2}$$

The solution $\mathbf{r}$ of this equation is the value that minimizes $E^2 = \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}$, called *the least-squares solution*, which is obtained using the following equation.

$$\mathbf{r} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \tag{3}$$

3.2 Proposed method

As described earlier, Massey's method assumes head-to-head competitions. In this study, however, we aim to evaluate driver performance based on their finishing positions in final races, rather than point differentials. Therefore, we replace the point differential with a position differential derived from the race results.

Furthermore, to place greater emphasis on differences among top positions rather than lower-position drivers, we define the position differential using the logarithmic difference of rankings.

Specifically, let c_i and c_j denote the finishing positions of drivers i and j, respectively. Then, the score differential y_k between drivers i and j is defined as:

$$y_k = -\log \frac{c_i}{c_j} \tag{4}$$

The comparison equation (1) is formulated for all pairs of drivers whose finishing positions are determined in the same race of the same championship. For example, in a race where 22 drivers start and all finishing positions are recorded, $22 \times 21/2 = 231$ equations are constructed.

The resulting ratings $\mathbf{r}$, computed through the procedure described above, are interpreted as performance scores for each driver. It is important to note that, due to the nature of the computation, the absolute value of a rating has no meaning. Only the differences between the ratings of two drivers are meaningful.

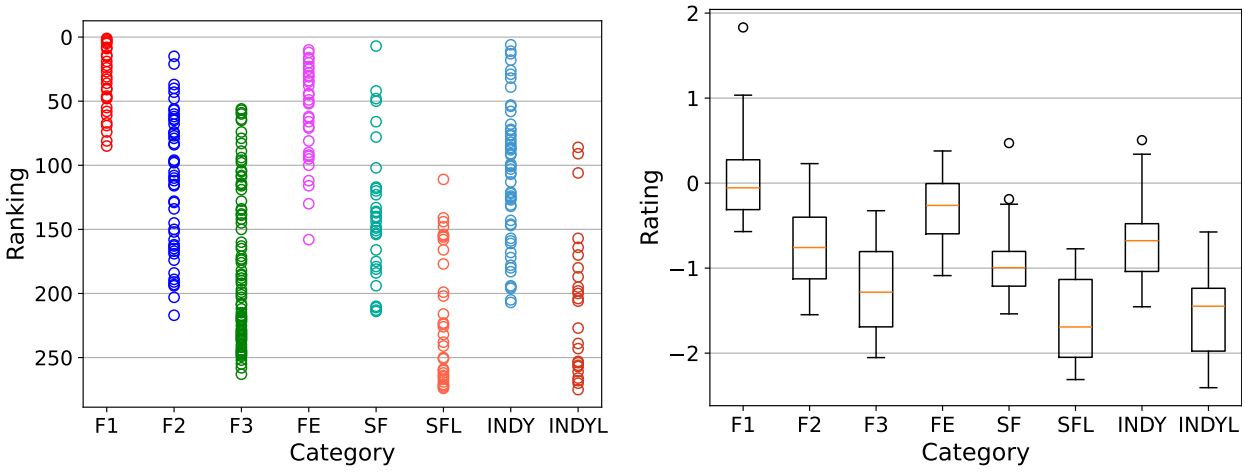
4 Results and discussions

Figure 3 presents the results of the rating calculations for the 275 drivers analyzed in this study. Panel (a) shows the driver rankings, while panel (b) displays the corresponding rating values. Drivers who participated in multiple championships have data points for each championship.

The driver rankings based on the proposed rating values reveal the hierarchical structure among the championships, which follows a feeder system. For example, Formula 1 (F1) drivers are followed by those from Formula 2 (F2), with Formula 3 (F3) drivers ranked below them. A similar hierarchical pattern is observed between the pairs of championships: Super Formula (SF) and Super Formula Lights (SFL), as well as IndyCar (INDY) and Indy Lights (INDYL).

The overlapping distributions among these championships reflect the fact that drivers who are promoted to higher-level championships do not necessarily finish at the bottom.

The proposed method allows us to compare multiple championships that do not have a direct hierarchical relationship. Specifically, F2 and Indy are equivalent, and SF is slightly lower.



(a) Overall ranking (b) Rating values

Figure 3: Rating distribution

When compared with the Super License Points system (Table 1), notable differences emerge between the two evaluation methods. For example, in the point-based system, Formula 3 (F3) and Formula E (FE) are assigned equal value, while Super Formula (SF) is rated lower than both. In contrast, the order based on the proposed rating values is FE, followed by SF, and then F3.

One possible explanation for this discrepancy lies in the typical career trajectory of drivers across championships: it is not uncommon for Formula 1 (F1) drivers to participate in FE, whereas such movements to F2 or F3 are rare.

Further investigation into the underlying causes of these inconsistencies remains a topic for future research. In particular, comparing the predictive performance of the proposed method and the Super License Points system, in terms of their accuracy in forecasting race outcomes, would be a meaningful direction for future work.

References

[1] FIA. 2024 FORMULA ONE SPORTING REGULATIONS. 2024. `https://www.fia.com/sites/default/files/fia_2024_formula_1_sporting_regulations_-_issue_1_-_2023-09-26.pdf`, accessed 2025/2/18.

[2] FIA. APPENDIX L TO THE INTERNATIONAL SPORTING CODE. 2024.

[3] Motor Sport magazine,. The Motor Sport Database. `https://www.motorsportmagazine.com/database/`, accessed 2025/2/10.

[4] A.N. Langville and C.D. Meyer. *Who's #1?: The Science of Rating and Ranking*. EBSCO ebook academic collection. Princeton University Press, 2012.