MathSport International 2022

Conference

-- PROCEEDINGS --

9th International Conference on Mathematics in Sport

Reading, United Kingdom

11 – 13 July 2022

Organized by Department of Economics, University of Reading

ISBN: 9789464988833

Contents

This is a collection of papers presented at the 9th Mathsport International Conference, hosted by the University of Reading on July 11–13, 2022.

A Statistical Review of Referee Assignments for Chosen European Football Leagues

Tankut Atan*, Burak Çavdaroğlu** and Zühal Özcan Yavuz***

Department of Industrial Engineering, Bahçeşehir University,
Beşiktaş, İstanbul, Turkey, 34353
+ email address: sabritankut.atan@eng.bau.edu.tr
Department of Industrial Engineering, Kadir Has University,
Cibali, İstanbul, Turkey, 34083
+ email address: burak.cavdaroglu@khas.edu.tr
** Department of Industrial Engineering, İstanbul Technical University,
Maçka, İstanbul, Turkey, 34367
+ email address: ozcanz18@itu.edu.tr

Abstract

In some countries, referees' assignments to sports events and their decisions in these events are highly criticized especially when the teams involved think that the referees were unfair to them. We looked at the referee assignment statistics for several European football leagues in recent years. While providing several descriptive statistics, we also report results of certain hypothesis tests. Differences among the statistical results indicate that authorities in different countries have different referee assignment strategies.

1 Introduction

When there are perceived peculiarities about the assignments of referees to football games, the officials' decisions will be heavily criticized in the media by all stakeholders. Only the referees are not allowed to respond to even the heaviest accusations. This is certainly true in Turkey where the authors are from. The 2021-22 season of the Süper Lig, the highest football division in Turkey, had some interesting developments. There is a central committee (MHK) of ex-referees under the Turkish Football Federation which takes all referee-related decisions in Turkish football. On March 8, 2022 (just before the 29th round of games), they have decided to let 13 referees go. In the following weeks, there were only 14 referees available for calling the games in the Süper Lig. Around that time, the Premier League was making use of 22 referees (Bundesliga 23, La Liga 20, Ligue 1 25, and Serie A 44). One of the forced-out referees, Cüneyt Çakır, actually had a good chance of going to the World Cup in 2022 (not possible, though, if he did not referee any more games). No good reason for this decision was given. Interestingly, the decision was overturned only after a few weeks on March 26 by a higher committee where the referees had objected. Some days later the president of the MHK resigned lasting only 173 days. He was appointed there on October 19, 2021 after the previous head

had resigned at the beginning of the season. The president of the Turkish Football Federation, a prominent businessman, had mentioned that the original decision was backed by the whole federation. He also resigned without waiting until the end of the season. After these turbulent developments with behind-closed-doors discussions and many speculations, we decided to look at some other European leagues and their referee assignments, and make some comparisons.

2 Previous Work

In this paper, we conduct a statistical analysis of the data on referee assignments in Europe. This is mostly descriptive work, and we could not find any quantitative academic literature that is similar in nature to what we are reporting here. Therefore, we opt to give a brief literature review on the referee assignment problem (RAP) which can be skipped on first reading. The RAP deals with optimally assigning the referees of a tournament subject to some constraints. The objective may be to minimize the violations of soft constraints with obtaining a feasible solution being the main goal. There is also a version similar to the Traveling Tournament Problem where the total travel times of the referees is minimized (Traveling Umpire Problem).

While discussing the literature on sports scheduling in general, [16] also includes a review section on the RAP. [13] and [12] appear to be the first works on the RAP focusing on baseball tournaments, and considering referee travel times, limiting the number of games refereed for the same team, and balancing the number of times the referees are assigned to each team. [21] argued that multi-criteria decision-making techniques could be applied to referee assignments by considering the experts' opinions. [17] suggested solving an assignment problem using the points given by experts each week. Both of these approaches are similar to having a weekly committee meeting to make the referee assignment decisions. [8] utilized Room squares to make the assignment decisions. Although a different referee can be assigned to each game, this approach is not suitable for including other customized constraints. [10] and [9] defined a general RAP formulation by taking several practical constraints such as a referee's performance rating satisfying the required minimum rating of a game, and not assigning referees to games where they are not available. They also suggested a heuristic algorithm for solving the problem. Furthermore, [10] showed that the problem is NP-hard.

As one might expect, the assignment constraints differ depending on the location and type of the tournament. There is work looking at football referee assignments in Turkey, Chile, Italy, Portugal and the Netherlands [2, 19, 20, 26, 29]. [14] developed and implemented a methodology for assigning the tennis referees in the US Open. [11] is a contemporary work on assigning referees in the Argentine basketball league solving an integer linear model where the optimization criterion is the minimization of referee travel times. Referees may have to call for several games in one trip. The Traveling Umpire Problem deals with minimizing the referees' travel times while considering several isues such as a referee not being assigned to the same team's games within a certain duration and each referee being assigned to each team's games at least once. This problem is especially prevalent in baseball [23, 24, 25]. In baseball, the referees are assigned as a team. Solution approaches developed to solve this difficult problem include simulated annealing [25] and exact methods [22, 27, 28]. There is also work on improving the lower bounds in the exact approaches to speed the process [6, 7]. A recent work by [5] provides a matheuristic for the problem.

There is also some work on combining the RAP with tournament scheduling and solving an integrated problem. [4] combine the Traveling Tournament and Traveling Umpire problems. Using the first division in Turkish football as a case study, [3] provide an integer model for an integrated referee assignment and

tournament scheduling problem, and also develop a genetic algorithm to solve the proposed problem.

3 Comparisons

The data on the assigned referees in different leagues was obtained from worldfootball.net using Java and VBA. As expected, a significant amount of time was spent on cleaning the data after downloading. All the statistics reported here are based on the information contained in the downloaded data. Due to the large amount and limited public existence of the data related to referee assignments, it was not possible for us to verify the correctness of the data. Therefore, we remind the readers that some of the results may not be accurate since the recorded data may have contained some errors.

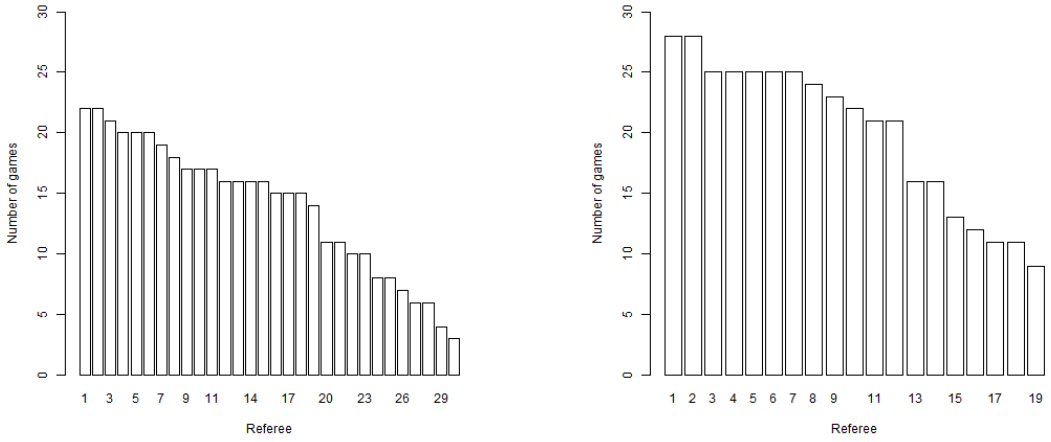
Table 1 provides the total number of referees used in the seasons in each league. The Serie A is using almost twice as many referees as in the other main leagues in Europe. Ligue 1 used a few more referees in the 2021-22 season compared to previous seasons because there were four referees from Portugal each calling only one game. Otherwise, the French are very consistent in using 21 to 23 native referees. No league, however, can compete with La Liga on consistency; the Spanish always use 20 referees. There was a referee from another country (Australia) in the 2021-22 season in the Premier League as well. He, however, refereed 10 games.

Table 1: Number of different referees used in a season

League	2021-22	2020-21	2019-20	2018-19	2017-18	2016-17	2015-16
Bundesliga	27	27	27	25	24	23	23
La Liga	**20**	**20**	**20**	**20**	**20**	**20**	**20**
Ligue 1	**25**	23	22	22	23	23	22
Premier League	22	19	20	18	21	19	19
Championship	37	40	36	35	36	32	**61**
Serie A	**45**	**45**	37	39	35	39	35
Süper Lig	34	30	24	**56**	23	23	39
1. Lig	**82**	46	58	79	**41**	73	73

The Süper Lig is also using relatively more number of referees. However, the assignments to individual referees are not very uniform as can be seen from Figure 1 (as a barchart) and Figure 2 (as a histogram) in comparison with the Premier League for the 2020-21 season. However, when compared to the Spanish, the Premier League is also over- and underutilizing some of its referees (Figure 3).

All leagues considered here are double round-robin tournaments. Thus, any two teams will face each other twice in the season. Table 2 gives the counts of the same referee being assigned to both of these games. La Liga and the Premier League are clearly doing these type of assignments more often. In the other investigated leagues, such appointments are far less. Ligue 1 seems to be avoiding it altogether especially in more recent seasons. The Süper Lig has produced the maximum number of same-referee appointments. Interestingly,the number of different referees used in the league was 56 in that season (2018-19). Another observation is that



(a) Süper Lig (b) Premier League

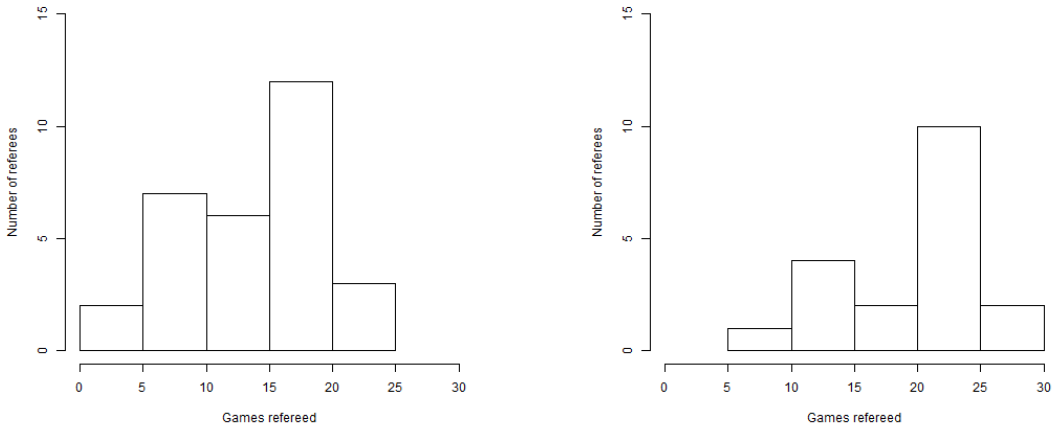
Figure 1: Number of games refereed, 2020-21 season

the second divisions of Turkey and England seem to have different decision-making processes as the numbers in the Championship and 1. Lig are much lower compared to the Premier League and Süper Lig.

Table 2: Number of assignments of the same referee to the games of the same opponents

League	2021-22	2020-21	2019-20	2018-19	2017-18	2016-17	2015-16
Bundesliga	8	6	6	10	5	3	8
La Liga	8	16	14	20	16	11	17
Ligue 1	2	1	**0**	2	6	7	5
Premier League	18	14	18	17	13	16	6
Championship	6	10	13	4	8	7	3
Serie A	2	11	7	9	5	3	5
Süper Lig	2	11	2	**31**	7	9	1
1. Lig	**0**	1	**0**	2	2	**0**	8

Table 3 gives the maximum number of games refereed by a referee in consecutive rounds of each league. In general, since the referees are evaluated weekly for their performance, being assigned to a match without a break can be an indicator of how good a referee is. The counts were collected by sorting the games by date, and then considering every $n/2$, games to be a round where n is the number of teams in the respective league. The Süper Lig had 21 teams in the 2020-21 season, therefore, we could not apply this definition of a round to that season. As a note, thinking of every 10 games as a round resulted in a maximum of 4 games. With counts of 12 and 13, some referees in the Premier League are the only ones who officiated more than 10 games consecutively. Some Süper Lig referees were also given relatively more consecutive games. This number is



(a) Süper Lig (b) Premier League

Figure 2: Histograms of games refereed, 2020-21 season

expected to be low where there are quite a bit of referees such as in the Serie A. La Liga and Ligue 1 have a similar number of referees compared to the Premier League with the same number of teams but they do not seem to allow (have) a referee to call for games without taking a break.

Table 3: Maximum number of consecutive games refereed

League	2021-22	2020-21	2019-20	2018-19	2017-18	2016-17	2015-16
Bundesliga	3 (3 referees)	3 (Frank Willenborg)	3 (2 referees)	3 (Manuel Gräfe)	3 (6 referees)	4 (2 referees)	**5** (2 referees)
La Liga	4 (3 referees)	4 (Carlos Del Cerro Grande)	3 (David Medié Jiménez)	3 (2 referees)	**7** (Ricardo De Burgos Bengoetxea)	4 (2 referees)	4 (2 referees)
Ligue 1	4 (Pierre Gaillouste)	4 (Frank Schneider)	4 (**Jérôme Brisard**)	4 (**Jérôme Brisard**)	6 (François Letexier)	5 (4 referees)	4 (2 referees)
Premier League	5 (5 referees)	7 (Andre Marriner)	9 (**Martin Atkinson**)	13 (Anthony Taylor)	9 (Michael Oliver)	12 (Bobby Madley)	12 (**Martin Atkinson**)
Serie A	3 (8 referees)	4 (**Maurizio Mariani**)	4 (**Maurizio Mariani**)	3 (5 referees)	3 (5 referees)	4 (Carmine Russo)	3 (Andrea Gervasoni)
Süper Lig	9 (FA, YK, YU)		6 (Ali Şansalan)	5 (Cüneyt Çakır)	7 (HM, MK, ÜÖ)	9 (**Halis Özkahya**)	4 (**Halis Özkahya**)

In Table 4, some statistics for the days between the games of referees are reported for the 2021-22 season. Serie A has a large median as the total number of referees used is very high (45) compared to the other leagues. Some waiting times for the referees are very large (such as 282 in the Süper Lig) because a few referees in those leagues had only one or two games assigned to them. La Liga is utilizing its 20 referees every other week. The maximum time between two games is 17 days only. Ligue 1 is also using all of its referees fairly often.

Being a referee for the same home team could be a factor influencing a referee's decisions since those games will be played in front of the same crowd. In the investigated leagues and seasons, this happened at most 5 times (Table 5). One has to say that, in general 4's and 5's are rare in all of the leagues. It may be an indicator that no league wants this situation to occur very often. The 2017-18 season of La Liga is interesting because no referee called a game more than twice for the same home team. It is as if a deliberate attention was paid to this issue in that season. Table 5 also reports the averages in parentheses. One could say that

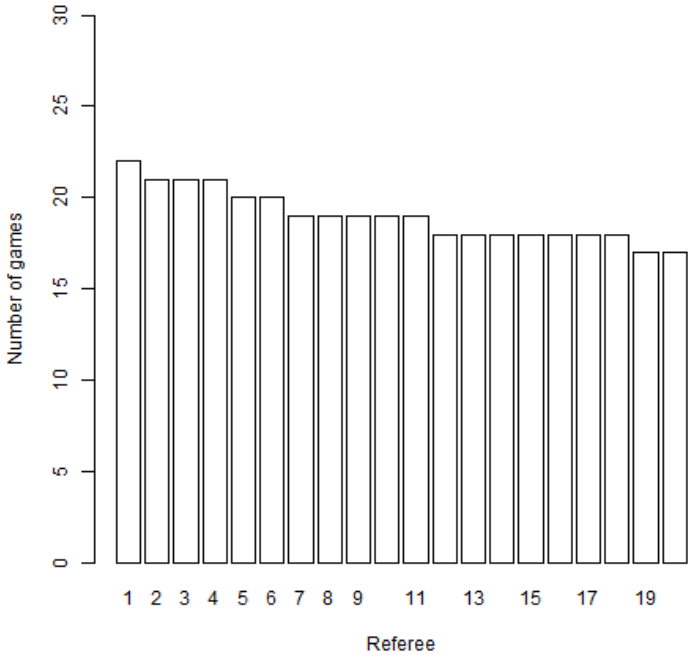


Figure 3: Number of games refereed, La Liga, 2020-21 season

the lower this average the more attention is being paid to this phenomenon. In the 2018-19 season, referees, on the average, called 1.37/19*100=7.2% of the possible home games of a team in La Liga whereas this percentage is 1.70/19*100=8.9% in the Premier League (Table 5). This higher percentage of the Premier League in general indicates that more home games of the same team can be assigned to the same referee in England. In fact, when a one-factor analysis of variance by league is conducted, not all means are found to be equal. The significant p values found by the Tukey HSD test are listed in Table 6.

Table 7 and Table 8 show the total number of games refereed for a team by each referee in the Süper Lig and Premier League between 2015-22. The tables only show the teams and referees which were present in all of the seasons. Several zeroes catch one's attention. This is normal if the referee is from the same city as the team such as in the case of Kevin Friend from Leicestershire. Kevin Friend apparently is also banned from calling for Bristol City as he is their fan [18]. In the Süper Lig, things get more interesting. Arda Kardeşler's lack of assignments to Trabzonspor games can be due to the fact that his brother has been a substitute goalkeeper for Trabzonspor since 2019 [1]. But, that still does not explain the time period before 2019. Özgür Yankaya had officiated Fenerbahçe's games in the past but his calls were not liked by the fans. The then-president of Fenerbahçe criticized him publicly after a match in 2015 declaring that 'he can never come to a Fenerbahçe game and even if he comes that he will not be able to leave the stadium' [15]. Cüneyt Çakır is arguably the best referee in Turkey; he was one of the referees chosen to officiate in the World Cup 2014. But, interestingly, he was assigned to only 10 games of Trabzonspor which is always a championship

Table 4: Average days between games of the referees in the 2021-22 season

League	Min	Median	Mean	Max
Bundesliga	13.95	19.71	28.70	183.00
La Liga	12.82	14.47	14.64	**17.00**
Ligue 1	13.71	15.61	16.00	24.00
Premier League	10.07	15.25	24.70	91.67
Serie A	10.50	25.90	45.46	273.00
Süper Lig	10.41	18.07	**55.95**	282.00

Table 5: Maximum (average) number of home games refereed for the same team

League	2021-22	2020-21	2019-20	2018-19	2017-18	2016-17	2015-16
Bundesliga	3 (1.30)	4 (1.37)	3 (1.35)	4 (1.37)	3 (1.40)	4 (1.45)	3 (1.41)
La Liga	5 (1.47)	4 (1.47)	4 (1.41)	3 (1.37)	**2** (1.28)	3 (1.33)	3 (1.34)
Ligue 1	4 (1.49)	3 (1.44)	3 (1.48)	3 (1.48)	3 (1.48)	4 (1.50)	4 (1.47)
Premier League	4 (1.48)	5 (1.60)	5 (1.70)	5 (1.70)	4 (1.58)	4 (1.64)	4 (1.71)
Serie A	4 (1.23)	3 (1.28)	4 (1.40)	4 (1.36)	4 (1.36)	3 (1.38)	3 (1.36)
Süper Lig	5 (1.45)	4 (1.34)	4 (1.35)	4 (1.20)	4 (1.51)	4 (1.52)	4 (1.38)

contender along with Beşiktaş, Fenerbahçe and Galatasaray referred to as the 'four greats'. His officiating is, apparently, not much liked by the Trabzonspor fans [30]. Serkan Tokat has been utilized as a video assistant referee in many games which are not reflected in Table 7. His total number in the reported six seasons is also not large enough for placing any meaning to zeroes. Volkan Bayarslan called even fewer games in those seasons. He also was assigned as a video assistant referee to many games.

We wanted to look at the assignments of referees to teams in more detail. Due to the irregular events in the Süper Lig in the 2021-22 season, we chose to compare the 2020-21 seasons of the Premier League and Süper Lig. The Liga Portugal was also chosen because of a mention in an academic publication saying that the referees of each week are chosen randomly there [19]. While we thought that it was unlikely to see a uniform distribution since referee assignment decisions are very much related to referees' ongoing performance, we still wanted to test this statistically. Not a single referee's distribution was found to be uniform in the Premier League and Süper Lig as can be seen from the very low p values of two different tests (Tables 9, 10). However, in contrast to the Premier League and Süper Lig, some referees' game allocations may be taken as uniformly distributed according to the Kolmogorov-Smirnov test (Table 11) in the Liga Portugal.

Table 6: Significant p values when mean percentages of home game assignments are compared across the leagues

Leagues	p Value
Bundesliga-La Liga	0.037
Bundesliga-Serie A	0.004
La Liga-Premier	0.001
Premier-Serie A	0.000
Serie A-Süper Lig	0.033

Table 7: Assignment counts between 2015-16 and 2020-21 seasons of the Süper Lig for common referees and teams

Referee/Team	Antalyaspor	Basaksehir	Besiktas	Fenerbahce	Galatasaray	Kasimpasa	Kayserispor	Konyaspor	Trabzonspor	TOTAL
Ali Palabiyik	13	14	14	22	20	10	8	15	21	137
Alper Ulusoy	8	8	10	9	6	6	8	7	8	70
Arda Kardesler	6	4	6	5	7	12	7	10	0	57
Cuneyt Cakır	12	17	24	19	24	8	10	11	10	135
Firat Aydinus	8	15	13	25	15	11	12	14	18	131
Halil Meler	7	12	13	17	12	10	15	8	15	109
Halis Ozkahya	11	10	14	8	15	9	9	15	15	106
Huseyin Gocek	9	8	8	9	9	10	11	11	12	87
Mete Kalkavan	10	12	16	18	19	13	12	11	8	119
Ozgur Yankaya	5	8	1	0	4	6	10	9	9	52
Serkan Cinar	3	4	11	4	3	2	4	6	5	42
Serkan Tokat	7	4	0	1	2	9	3	5	0	31
Suat Arslanboga	8	2	2	3	6	6	5	3	2	37
Umit Ozturk	10	8	10	10	4	9	11	9	6	77
Volkan Bayarslan	3	0	0	0	1	2	4	6	1	17
Yasar Ugurlu	12	13	10	10	8	9	13	1	17	93
TOTAL	132	139	152	160	155	132	142	141	147	

References

[1] Ahmet Ercanlar on Twitter, 2019. `https://twitter.com/ahmetercaniar/status/1189485165980114944`, last accessed on 2022-06-16.

[2] F. Alarcón, G. Durán, and M. Guajardo. Referee assignment in the Chilean football league using integer programming and patterns. *International Transactions in Operational Research*, 21(3):415–438, 2014.

[3] T. Atan and O. P. Hüseyinoğlu. Simultaneous scheduling of football games and referees using Turkish league data. *International Transactions in Operational Research*, 24(3):465–484, 2017.

[4] M. Bender and S. Westphal. A combined approximation for the traveling tournament problem and the traveling umpire problem. *Journal of Quantitative Analysis in Sports*, 12(3):139–149, 2016.

[5] R. C. Chandrasekharan, T. A. Toffolo, and T. Wauters. Analysis of a constructive matheuristic for the traveling umpire problem. *Journal of Quantitative Analysis in Sports*, 15(1):41–57, 2019.

[6] L. De Oliveira, C. C. De Souza, and T. Yunes. Improved bounds for the traveling umpire problem:

Table 8: Assignment counts between 2015-16 and 2020-21 seasons of the Premier League for common referees and teams

Referee/Team	Arsenal	Chelsea	Crystal Palace	Everton	Leicester City	Liverpool	M City	M United	Southampton	Tottenham Hotspur	West Ham United	TOTAL
Andre Marriner	18	19	20	14	13	23	25	17	18	19	14	200
Anthony Taylor	21	22	18	19	19	25	18	23	13	22	18	218
Craig Pawson	16	18	13	13	19	21	15	19	14	17	10	175
Graham Scott	8	6	7	8	10	4	4	4	13	11	10	85
Jonathan Moss	18	19	21	18	11	15	19	28	19	16	17	201
Kevin Friend	12	14	15	15	0	18	13	10	15	16	13	141
Lee Mason	10	8	6	14	12	6	11	9	11	5	8	100
Martin Atkinson	22	19	19	16	14	22	17	21	9	18	22	199
Michael Oliver	23	20	22	20	24	27	23	17	12	26	19	233
Mike Dean	17	14	13	6	23	7	21	27	13	21	17	179
Paul Tierney	8	7	6	11	8	13	12	12	11	9	8	105
Stuart Attwell	10	11	10	11	5	7	6	5	11	6	10	92
TOTAL	183	177	170	165	158	188	184	192	159	186	166	

A stronger formulation and a relax-and-fix heuristic. *European Journal of Operational Research*, 236(2):592–600, 2014.

[7] L. De Oliveira, C. C. De Souza, and T. Yunes. Lower bounds for large traveling umpire instances: New valid inequalities and a branch-and-cut algorithm. *Computers & Operations Research*, 72:147–159, 2016.

[8] J. H. Dinitz and D. R. Stinson. On assigning referees to tournament schedules. *Bulletin of the Institute of Combinatorics and its Applications*, 44:22–28, 2005.

[9] A. R. Duarte, C. C. Ribeiro, and S. Urrutia. A hybrid ILS heuristic to the referee assignment problem with an embedded MIP strategy. In *International Workshop on Hybrid Metaheuristics*, pages 82–95. Springer, 2007.

[10] A. R. Duarte, C. C. Ribeiro, S. Urrutia, and E. H. Haeusler. Referee assignment in sports leagues. In *International Conference on the Practice and Theory of Automated Timetabling*, pages 158–173. Springer, 2006.

[11] G. Durán, M. Guajardo, and F. Gutiérrez. Efficient referee assignment in Argentinean professional basketball leagues using operations research methods. *Annals of Operations Research*, pages 1–19, 2021.

[12] J. R. Evans. A microcomputer-based decision support system for scheduling umpires in the American baseball league. *Interfaces*, 18(6):42–51, 1988.

[13] J. R. Evans et al. Play ball!–the scheduling of sports officials. *Perspectives in Computing: Applications in the Academic and Scientific Community*, 4(1):18–29, 1984.

[14] A. Farmer, J. S. Smith, and L. T. Miller. Scheduling umpire crews for professional tennis tournaments. *Interfaces*, 37(2):187–196, 2007.

[15] Habertürk. Özgür Yankaya 3 yıl sonra Kadıköy'de, 2018. `https://www.haberturk.com/ozgur-yankaya-3-yil-sonra-kadikoyde-2251294-spor`, Last accessed on 2022-06-16.

[16] G. Kendall, S. Knust, C. C. Ribeiro, and S. Urrutia. Scheduling in sports: An annotated bibliography. *Computers & Operations Research*, 37:1–19, 2010.

[17] J. G. Lafuente. The best systems for appointing referees. In *Economics, Management and Optimization in Sports*, pages 101–120. Springer, 2004.

[18] Leicester Mercury. Referee is banned from doing Leicester City matches - but it's not why you might think, 2020. `https://www.leicestermercury.co.uk/sport/football/football-news/`

kevin-friend-leicester-bristol-city-4088537, Last accessed on 2022-06-16.
[19] R. Linfati, G. Gatica, and J. W. Escobar. A flexible mathematical model for the planning and designing of a sporting fixture by considering the assignment of referees. *International Journal of Industrial Engineering Computations*, 10:281–294, 2019.
[20] S. Mancini and A. Isabello. Fair referee assignment for the Italian soccer Serie A. *Journal of Quantitative Analysis in Sports*, 10(2):153–160, 2014.
[21] A. Scarelli and S. C. Narula. A multicriteria assignment problem. *Journal of Multi-Criteria Decision Analysis*, 11(2):65–74, 2002.
[22] T. A. Toffolo, T. Wauters, S. Van Malderen, and G. V. Berghe. Branch-and-bound with decomposition-based lower bounds for the traveling umpire problem. *European Journal of Operational Research*, 250(3):737–744, 2016.
[23] M. A. Trick and H. Yildiz. Benders' cuts guided large neighborhood search for the traveling umpire problem. *Naval Research Logistics*, 58(8):771–781, 2011.
[24] M. A. Trick and H. Yildiz. Locally optimized crossover for the traveling umpire problem. *European Journal of Operational Research*, 216(2):286–292, 2012.
[25] M. A. Trick, H. Yildiz, and T. Yunes. Scheduling major league baseball umpires and the traveling umpire problem. *Interfaces*, 42(3):232–244, 2012.
[26] F. Vallespi Soro. The referee assignment problem. B.S. thesis, Universitat Politècnica de Catalunya, 2019.
[27] T. Wauters, S. Van Malderen, and G. V. Berghe. Decomposition and local search based methods for the traveling umpire problem. *European Journal of Operational Research*, 238(3):886–898, 2014.
[28] L. Xue, Z. Luo, and A. Lim. Two exact algorithms for the traveling umpire problem. *European Journal of Operational Research*, 243(3):932–943, 2015.
[29] M. Yavuz, U. H. İnan, and A. Fığlalı. Fair referee assignments for professional football leagues. *Computers & Operations Research*, 35(9):2937–2951, 2008.
[30] Yeni Şafak. Trabzonspor'da Cüneyt Çakır endişesi, 2021. https://www.yenisafak.com/spor/trabzonsporda-cuneyt-cakir-endisesi-3723449, Last accessed on 2022-06-16.

Table 9: Number of times a referee is assigned to a team's games in the 2020-21 season of the Süper Lig

Referee/Team	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	AD*	KS**
1		1	1	1	4		1		1		1	1			3	1	3	1	1	1	1	0.001	0.000
2	1			1	3	2			3	3	1	1			3	1				1	2	0.001	0.001
3		1	1		2	1		2	2	1	1	1		1		1	1	2	1		3	0.001	0.001
4	3	1		1	2			1	2	3	1			1		2		1	1		1	0.001	0.001
5		1	3	1		1		1	1		2	1	3	1	1		1	1	2			0.001	0.000
6	1			2		1			1		2	2	1	1	3	2	1		2	1		0.001	0.005
7	1			2	2		1	2		2	1			1	2		1	1	2		1	0.001	0.005
8	1		1	1			2	1	1	1			2	1			2	1	1	3		0.001	0.000
9	1		1			2	1	1	2	1	1			1	1	1		1	1	2		0.001	0.009
10			2	1	1	1	1		1	1	2	1	1	1	1				1		2	0.001	0.009
11	1		2	2	1		1	1		2		1	1			1			1	3		0.001	0.000
12		1		2	3	1		1	2			1	1		1		1	1			1	0.001	0.000
13	2	1		1			2		1		2	1		1	1		2		1		1	0.001	0.001
14	2		2		1	1	1	1			1		1	1	1	1	1			1	1	0.001	0.002
15	1	2	1			1	1	1	1	2				2			1		1	1	1	0.001	0.005
16	1	2				2	1			1		2	1	1		1		1		2		0.001	0.000
17	1	1			1		1	1	1	1	1	2		1		2		1		1		0.001	0.002
18		2	1	1			2	1	1					1		1	1	3			1	0.001	0.000
19		2	1			3	1	1			1	1			2		1			1		0.001	0.000
20								1			1	2	1		1			1	3		1	0.001	0.000
21		2		2		1		1						1		1	1	1	1			0.001	0.000
22						1	1	2		1	1		1			1	1	1				0.001	0.000
23	2						1	1			1		1						1	3		0.001	0.000
24	1	1	1									1	2	1							1	0.001	0.000
25	1			1			1						1	1				1			2	0.001	0.000
26			1										1			1	2	2				0.001	0.000
27			1	1						1		1	1	1								0.001	0.000
28		1	1			2	1									1						0.001	0.000
29												1	1			1					1	0.001	0.000
30		1												1		1						0.001	0.000

* Anderson-Darling test's p-value when testing the goodness of fit to the uniform distribution.
** Kolmogorov-Smirnov test's p-value when testing the goodness of fit to the uniform distribution.

Table 10: Number of times a referee is assigned to a team's games in the 2020-21 season of the Premier League

Referee/Team	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	AD*	KS**
1	**5**	2	1	1	1	2	2		3		2	3	1		2			1		2	0.0005	0.001
2		1	1	4	1	2			3	4	1	2	3	1		1		3	1		0.0005	0.003
3	1		1	2	1	1	2	2	3		1	2			2	2	2		1	2	0.0005	0.081
4	1	2	1	1			1	2		1			3		1		4	3	2	3	0.0005	0.003
5		1	1	1	1		2	1	1	1	2		3	2	1	1	1	2	2	2	0.0005	0.036
6	3		1	1	1	2	3	1		1		3	3		1	1		1	2	1	0.0005	0.009
7			1	2	2	2	2		2		5			3	3	2	1				0.0005	0.001
8	2	1	2						1	1	4	1	1	4		1	2	1	1	2	0.0005	0.001
9	2	2	2		3	3	1	1		1				1	1	1	2		2	1	0.0005	0.036
10		2	2	2	1	1	1	2		1	2	2	1			1	1	1	2		0.0005	0.015
11	1		1	1	2	2		3	2			1	1	1	1	1	2		1	1	0.0005	0.002
12	1	1		1	2		1			1	2	1	1	2	3		2	2		1	0.0005	0.009
13	1	1	2		2			1	1	1		1	1		1	2	1		1		0.0005	0.015
14		3	1	1	1		2	1		2					1	3		1			0.0005	0.000
15	1	1		1	1	2		1	1	1				1		1	1		1		0.003	0.001
16		1				2	2	1	1	1					1			1	2		0.0005	0.000
17	1			1						2		2	1			1		2		1	0.0005	0.000
18		1	2					1		1		1		2				1		2	0.0005	0.000
19								2	1					2	1	1			1	1	0.0005	0.000

* Anderson-Darling test's p-value when testing the goodness of fit to the uniform distribution.
** Kolmogorov-Smirnov test's p-value when testing the goodness of fit to the uniform distribution.

Table 11: Number of times a referee is assigned to a team's games in the 2020-21 season of the Liga Portugal

Referee/Team	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	AD*	KS**
1	3	1		1	1	4		2	1	1	1		1	3		2	3	2	0.000	0.018
2	2	1	1	2	3	1	1	3	2	2		3		1		1	1	2	0.000	0.336
3	1	2		2	2	2	2	3	3	2		3	2	1	4	5	2	2	0.000	0.048
4	1	2	1	2		1		2	1		2		1					1	0.000	0.002
5	3	1	1	1	4	1	2	2	1	4	1	2	2	1	1	3	4	4	0.000	0.211
6	1	4	2	1	2			2	1	2	1	2	3	1			1	3	0.000	0.037
7	1	1	4	1	2	1	2	4		3	2		3	1	2	4	1	2	0.000	0.336
8	3	1	2		3	3	2	1	1	1	2	3	2	5	4	1	3	3	0.000	0.099
9		1	1			1		1		2	1		1						0.000	0.000
10								1			1								0.000	0.000
11		2	3	2				1	4	1	2	2	2	1					0.000	0.002
12	1	1	1	2	1	1	1		4	2	2	1	1	1		1			0.000	0.000
13	1	1	3	1	2		4	1	2	1	3	1	4	3	2	5	2	2	0.000	0.155
14	4	1	1	2	4	6	3	2	1		1	4	2	5	3	2	3		0.000	0.336
15	3	1	3	3	3	1	3		3	2	3	3		4	3		2	1	0.000	0.069
16		3	1	2		2	1	1	2	1	1	2	2	1	2	2	3	2	0.000	0.336
17		1	1	2		1		1		1			1						0.000	0.000
18	1	1		2		3	4	4	3	2	4		3		4	3	3	3	0.000	0.069
19	1	3	5	3	3	4	4	2	1	2	2	3	2	2	4	4	3	2	0.000	0.099
20	5	2		3	1	2	4	1		3	3	3	1	2	4	1	2	3	0.000	0.281
21	3	4	4	2	3		1		4	2	2	2	1	2	1		1	2	0.000	0.336

* Anderson-Darling test's p-value when testing the goodness of fit to the uniform distribution.

** Kolmogorov-Smirnov test's p-value when testing the goodness of fit to the uniform distribution.

The Bogey Phenomenon in Sport

Rory Bunker*

* Graduate School of Informatics, Nagoya University, Nagoya, Aichi, Japan.

Abstract

Despite its existence being widely debated and discussed by fans, media, and commentators in sport, the bogey phenomenon has received minimal attention in the academic literature. Roughly speaking, the bogey phenomenon arises when a so-called bogey player (or team) performs consistently better than would be expected against a specific opposition, over a certain period. In this paper, a practical, data-driven bogey player identification method is described and demonstrated using publicly available data from professional men's tennis. For a set of historical matches between a given pair of players, betting odds are used to calculate implied match result probabilities, and actual match results are used to identify upset (unexpected) results. The Wald-Wolfowitz statistical runs test is then applied to the set consisting of the sequence of upset and non-upset results between the two players. Subsequently, a set consisting of the types of upset results, which is a subset of the original historical results set, is analysed to determine whether one player could be considered the bogey player of the other in the pair. The implied match result probabilities, obtained based on the betting odds, greatly simplify the analysis since odds already incorporate variables such as rankings, venue, and form, which therefore do not need to be explicitly incorporated into the method.

1 Introduction

The possible existence of bogey teams in sports has intrigued commentators and sports followers alike for many years, and has generated significant debate in sports forums and newspaper articles. Loosely speaking, a bogey team – or "Angstgegener" (translated as "feared opponent") as it is known in the German language – tends to habitually beat another specified team despite appearing to be the weaker side on paper. Although the concept of bogey teams been briefly mentioned in a small number of academic studies in the fields of education [5] and sociology [6, 12], and in theses in these fields [8, 18], it has been largely unexplored in the sports science and sports statistics fields.

There are various examples of bogeys that have been put forward. For example, in football, it has been suggested that Manchester United is a bogey team of Newcastle United[1], Tottenham won only one game out of 37 against Chelsea between 1990 and 2006, and Watford had not beaten Manchester City in the 30 year period since 1989[2]. However, media often use the "bogey" term without providing data as to whether such a streak in results is actually unexpected based on factors including the relative strengths of the teams. Some possibly bogeys are said to be tournament-specific, e.g., Portugal are considered England's bogey team at

[1] https://www.vavel.com/en/football/2015/08/20/manchester-united/528960-preview-manchester-united-vs-newcastle-united-winless-newcastle-travel-to-old-trafford.html

[2] https://www.theguardian.com/football/2019/sep/20/watford-manchester-city-bogey-teams-premier-league-football

World Cups[3], and England and Argentina are considered the respective bogey teams of Australia and France at Rugby World Cup tournaments[4]. However, others, including well-known football tipster Kevin Pullein, doubt the existence of bogey teams[5].

In this study, the focus is on tennis, which appears to have had much less attention in the media than soccer. However, tennis is simpler to analyse because there are only two possible outcomes as opposed to soccer's three (including draw).

Streaks – either in individual player actions (e.g., home runs in baseball, 3-pointers in basketball) or match winning streaks – and the "hot hand" phenomenon, which is also known as positive recency, have been considered in sports statistics. The existence of hot-hand type phenomenons assume that future outcomes can be determined (at least partly) based on the most recent outcomes, and that players or teams having successful streaks impacts upon their future successes [3, 4]. The most common techniques for such analyses have been Wald-Wolfowitz Runs Tests (WWRTs) and autocorrelation tests [11, 13, 14].

Given the lack of research attention and interest from sports fans and media alike, in this paper a novel bogey player identification method is proposed that combines the WWRT with an unexpected result identification method that uses bookmaker-provided betting odds as its key input. In particular, the method combines the results of applying the WWRT to the set of historical results between two given players, and if a statistically significant result is found, the second step involves examining the types of upset results (generated using the unexpected result identification method) to determine whether the bogey phenomenon has existed between the two players for some period. The method is demonstrated using three player-pair examples from professional men's tennis, from media and online sources.

The remainder of this paper is organised as follows. In section 2, we provide a definition for a bogey team/player. Then, in section 3, the materials and methods – the dataset, the unexpected result (upset) identification method, and the WWRT – are described. Following this, in section 4, the proposed method, which combines the upset identification method with the WWRT, is applied to three specific examples quoted in online sources. Finally, section 5 concludes and provides some potential avenues for further work.

2 Definition of the Bogey Phenomenon

It would be useful to specify a formal definition of a bogey team (in team sports) or a bogey player (in individual sports), as this appears to not have yet been provided in the literature. The bogey phenomenon could be defined as:

> *A team or player performing consistently better than would be expected against a specific opposition (the non-bogey) over a certain period, given factors including – but not limited to – their difference in the strengths, player availability, form, and home advantage.*

In team sports, this definition above implies that the bogey phenomenon exists between pairs of teams/players, similar to pairwise comparisons such as those used in the Bradley-Terry model, which has been used to predict tennis match results [10]. Note that the definition above also implies that there is a temporal element to the

[3]https://independent.ie/sport/rugby/world-cup/irish-news/idea-of-bogey-teams-is-flawed-thinking-but-it-can-have-an-effect-34116828.html

[4]https://www.theroar.com.au/2008/12/03/wallabies-avoid-bogey-team-in-2011
https://english.kyodonews.net/cities/news/2019/09/3db2cdcf4a36-rugby-france-look-to-overcome-world-cup-bogey-team-argentina.html

[5]https://www.theguardian.com/football/2007/apr/27/newsstory.sport11?CMP=gu_com

bogey phenomenon: a player might be a bogey player of another, but only for a certain period. For pairwise comparisons, one approach could be to use rankings or Elo ratings (e.g., [1, 7, 17]); however, although these measures incorporate the historical results of players, they do not incorporate variables that are external to the match, such as match venues, form, court surface, and so on.

Betting odds, which have been shown to perform well in predicting the outcomes of sports including tennis, e.g., in [16], are instead used in our method to derive probabilities for the expected outcomes of matches. The advantage of betting odds – especially when averaged across several bookmakers for reliability – is that they incorporate various factors including strength, form, home advantage, etc. Thus, the second-half of the bogey phenomenon definition specified above (following "given...") can be ignored, since these factors are largely already accounted for in the odds themselves.

3 Materials and methods

Dataset. In this study, we use publicly available data from professional men's tennis sourced from the website tennis-data.co.uk, the same data used by [1]. The match results of all ATP Tour matches from Masters, ATP Finals, and Grand Slams tournaments, as well as bookmaker odds, player rankings, and ATP points are available in this dataset. The dataset contains matches from July 4, 2005, to November 22, 2020. Initially, there were 38,868 matches in the dataset, and after passing it through the welo R package's clean() function (https://cran.r-project.org/web/packages/welo/index.html), the final dataset used for analysis consisted of 33,976 matches.

Unexpected Result Identification Method. In this subsection, the unexpected result "upset" identification method is described. Suppose there are two players, A and B, who have played each other in a total of T past matches. In a specific match, $t \in T$, the betting odds for A and B are denoted by $O_t(A)$ and $O_t(B)$, respectively, and the reciprocal of the odds indicates the implied probabilities of victory. Let the set of historical results between A and B be

$$HR_t(A,B) = \begin{cases} U & ((\frac{1}{O_t(A)} > \frac{1}{O_t(B)}) \wedge \text{ B beats A }) \vee ((\frac{1}{O_t(A)} < \frac{1}{O_t(B)}) \wedge \text{ A beats B }) \\ N & otherwise \end{cases}$$

for all $t \in T$, where U and N denote an upset result and non-upset result, respectively. Note that if there are two possible match outcomes as is the case in tennis, the two sets are equivalent, i.e., $HR_t(A,B) = HR_t(B,A)$. In the second stage, a set consisting of the type of upset results, which is a subset of the HR set above, is constructed:

$$UR_t(A,B) = \begin{cases} UL & (\frac{1}{O_t(A)} > \frac{1}{O_t(B)}) \wedge \text{ B beats A} \\ UW & (\frac{1}{O_t(A)} < \frac{1}{O_t(B)}) \wedge \text{ A beats B} \end{cases}$$

for all $t \in T$, where UL and UW denote an upset loss and upset win for player A, respectively, and N again denotes a non-upset result. Note that the order of A and B in this function is important, and represents whether the results are from A or B's perspective. In particular, UL stands for an upset loss (from A's perspective), UW stands for an upset win (from B's perspective), and the Ns denote a non-upset result. From B's perspective, the upset win and upset loss are swapped, i.e.,

$$UR_t(B,A) = \begin{cases} UW & (\frac{1}{O_t(A)} > \frac{1}{O_t(B)}) \wedge \text{ B beats A} \\ UL & (\frac{1}{O_t(A)} < \frac{1}{O_t(B)}) \wedge \text{ A beats B} \end{cases}$$

for all $t \in T$. The first part of the UR function indicates that based on the probabilities implied by the betting odds, B won despite A being expected to win in match t. Similarly, the second part indicates that that A won despite B being expected to win in match t.

Wald-Wolfowitz Runs Test (WWRT). In this paper, the WWRT, a statistical test that has been employed by several researchers in sports statistics previously (e.g., [2, 9, 15]), is utilised. As the name suggests, the WWRT considers runs, which are successions of symbols followed – or preceded by – different symbols. Related to the concept of runs are streaks: if a player has many lengthy streaks, they will have fewer runs. A player with many alternating wins and losses will have many runs. The WWRT considers the distributions of streaks of different lengths based on the number of runs, and compares it to the distribution that would be expected if successive outcomes were independent [19]. The test statistic of the WWRT, Z, which is asymptotically normally distributed, is:

$$Z = \frac{R - E(R)}{\sqrt{V(R)}}$$

where R denotes the number of runs. The expected value and variance of R are

$$E(R) = \frac{2nm}{n+m} + 1$$

and

$$Var(R) = \frac{2nm(2nm-n-m)}{(n+m)^2(n+m-1)}$$

respectively, where n and m denote the number of positive and negative values, respectively (in our case, the number of upset results and non-upset results).

Bogey Player Identification Method. The bogey player identification method combines the upset result identification method and WWRT, which were described above. The proposed method is depicted in Figure 1. First, the WWRT is applied to the set of upsets and non-upset results (HR) between two players, A and B. If

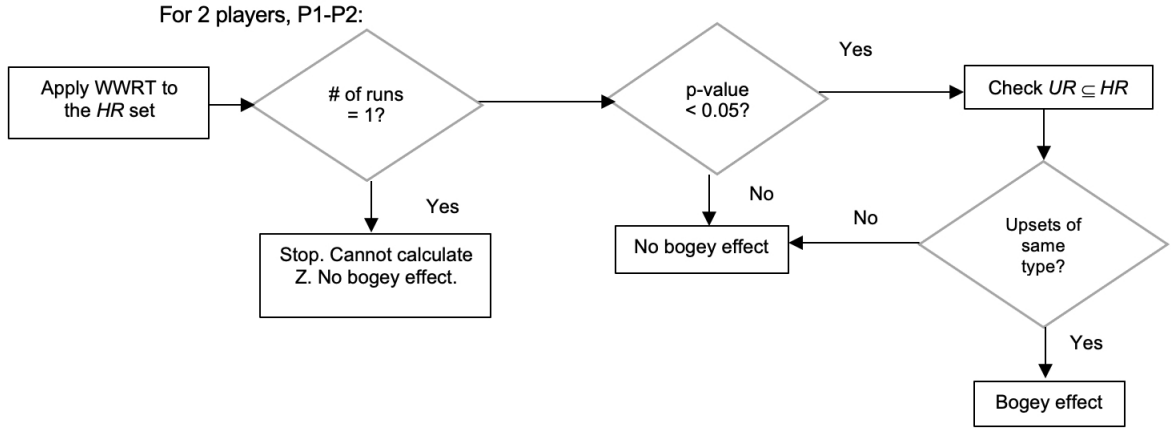


Figure 1: Illustration of the proposed bogey identification method.

this result is statistically significant (and the number of runs in the sequence is more than one), we proceed to the next stage, where the set of types of upset results (upset wins and upset losses), which is denoted UR,

is analysed. If the upsets in UR are largely of the same type for a certain period, we consider the bogey phenomenon to exist for that period (the WWRT can also be applied to UR).

The proposed method was implemented in Python, and the code is available on GitHub: `https://github.com/rorybunker/bogey-phenomenon-sport`.

4 Results

In this section, the proposed method is applied to three player-pair examples quoted in online/media sources, to test whether the bogey phenomenon exists between certain players. For all three examples, the level of statistical significance is set to 0.05, and p-values stated are the one-tailed values.

Andy Murray vs. Roger Federer. A Sydney Morning Herald article from 2013[6], dated January 25, 2013, suggested that Roger Federer was the bogey player of Andy Murray at Grand Slam tournaments. To test this, the start date parameter is set to be the minimum date in the dataset, and the end date is set to be the day before the article was written (January 24, 2013). The option in the code that considers Grand Slam matches only is also used. When applying the proposed method, $HR(P1,P2)_{GrandSlam} = HR(Murray,Federer)_{GrandSlam} = \{N,N,N\}$ was obtained, where the "GrandSlam" denotes that we consider only Grand Slam matches between P1 (Murray) and P2 (Federer). In this case, since there were only Ns in the sequence and, therefore, only one run in the sequence, the WWRT could not be run, and we conclude that the bogey phenomenon did not exist between Murray and Federer prior to January 25, 2013.

Andy Murray vs. Novak Djokovic. An online forum comment[7] on menstennisforums.com dated May 15, 2011 suggested that Andy Murray was the bogey player of Novak Djokovic, with the forum poster stating "I've always felt that Murray is Djokovic's bogey player". The proposed method obtains $HR(Murray,Djokovic) = \{U,N,N,N,U,U,U,U,N\}$, Z = -1.044, and p-value = 0.148. Therefore, over period prior to when the forum comment was posted, we conclude that there was no bogey phenomenon between Murray and Djokovic.

Kei Nishikori vs. Jo-Wilfried Tsonga. A Herald Sun article entitled "Jo-Wilfried Tsonga hopes to avoid Australian Open bogey Kei Nishikori" [8] was written on 12 January 2017. Setting the end date, therefore, to 11 January 2017, which retained 8 out of the 9 matches in the entire dataset, the historical result set was $HR(Nishikori,Tsonga) = \{U,U,N,U,N,U,N,U,N\}$, with Z-value = 1.854 and p-value = 0.032. Since the p-value is less than 0.05 in this case, we proceed to the next step in Figure 1, which is to check the upset result types in the UR set, $UR(Nishikori,Tsonga) = \{UW,UW,UW,UL,UL\}$. We can see from the first three items in this set that there were three consecutive upset wins to Nishikori over Tsonga (12 Oct 2011 to 10 Oct 2013). When WWRT was applied to $UR(Nishikori,Tsonga)$, a Z-value of -1.527 and a p-value of 0.063 was obtained, non-significant at the 95% level of significance, but significant at the 90% level. Although there were three upset wins to Nishikori between 12 Oct 2011 and 10 Oct 2013, by the time this news article was written, Tsonga had recorded two upset wins over Nishikori, on 29 Oct 2013 and 30 Oct 2014 (last two items of UR).

[6] https://www.smh.com.au/sport/tennis/its-murray-v-djokovic-20130125-2dcuv.html

[7] https://www.menstennisforums.com/threads/will-nadal-be-able-to-reach-the-rg-final.182597/?u=36369

[8] https://www.heraldsun.com.au/sport/tennis/jowilfried-tsonga-hopes-to-avoid-australian-open-bogey-kei-nishikori/news-story/2dfaab3f641494447405ae65fc7e7592

Table 1: Results of the three player pairs considered.

P1	P2	start_date	end_date	HR	Z	p-value	UR	Z	p-value
Murray A.	Federer R.	5 Jul 2005	up to 24 Jan 2013 (Grand Slams)	N, N, N	-	-	-	-	-
Murray A.	Djokovic N.	5 Jul 2005	up to 14 May 2011	U, N, N, N,U,U,U,U, N	-1.044	0.148	-	-	-
Nishikori K.	Tsonga J.W.	5 Jul 2005	up to 11 Jan 2017	U, U, N, U, N, U, N, U, N	1.854	0.032	UW,UW,UW,UL,UL	-1.527	0.063

5 Conclusions and further work

This paper described an approach to test for the bogey phenomenon between pairs of tennis players using an upset result identification model combined with the Wald-Wolfowitz Runs Test (WWRT). The proposed method was applied to some examples quoted in online sources to test whether the assertions that a specific player was a bogey player of another were accurate. Aside from statistical runs tests like the WWRT, other approaches including autocorrelation tests could be investigated. Yet another approach, recently used by Steeger, Dulin & Gonzalez [14] in the context of analyzing winning streaks in NHL Ice Hockey, is to employ an entropy-based method like that of [20]. In future work, whether the use of this type of entropy-based approach or autocorrelation tests changes the results could be investigated. The proposed method could be applied to other sports, e.g., Rugby, Ice Hockey, and Soccer, and the existence of the bogey phenomenon could be compared across sports, i.e., whether certain sports tend to have a greater number of bogey teams or players. The WWRT with two possible outcomes was applied in this study. However, if the k-category extension of the WWRT is employed, a set consisting of upset wins (UWs), upset losses (ULs), and non-upset results (Ns) can be tested directly. Finally, although only a small number of examples were tested here, if a large number of tests were to be performed simultaneously, a correction procedure such as the Bonferonni-Holm correction would need to be employed.

Acknowledgments

This work was supported by JSPS KAKENHI (20H04075) and JST Presto (JPMJPR20CA), of which principal investigator is Keisuke Fujii.

References

[1] G. Angelini, V. Candila, and L. De Angelis. Weighted elo rating for tennis match predictions. *European Journal of Operational Research*, 297(1):120–132, 2022.

[2] J. Arkes and J. Martinez. Finally, evidence for a momentum effect in the nba. *Journal of Quantitative Analysis in Sports*, 7(3), 2011.

[3] P. Ayton and I. Fischer. The hot hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & cognition*, 32(8):1369–1378, 2004.

[4] M. Bar-Eli, S. Avugos, and M. Raab. Twenty years of "hot hand" research: Review and critique. *Psychology of Sport and Exercise*, 7(6):525–553, 2006.

[5] T. Bruce. A spy in the house of rugby: Living (in) the emotional spaces of nationalism and sport. *Emotion, Space and Society*, 12:32–40, 2014.

[6] M. K. Chiweshe. Frenemies: understanding the interconnectedness of rival fan identities in harare, zimbabwe. *Soccer & Society*, 19(5-6):829–841, 2018.

[7] A. E. Elo. *The rating of chessplayers, past and present*. Arco Pub., 1978.
[8] J. Fürnkranz. Anwendung von data mining zu statistischen auswertungen und vorhersagen im sport. 2009.
[9] J. J. Koehler and C. A. Conley. The "hot hand" myth in professional basketball. *Journal of sport and exercise psychology*, 25(2):253–259, 2003.
[10] I. McHale and A. Morton. A bradley-terry type model for forecasting tennis match results. *International Journal of Forecasting*, 27(2):619–630, 2011.
[11] L. Peel and A. Clauset. Predicting sports scoring dynamics with restoration and anti-persistence. In *2015 IEEE International Conference on Data Mining*, pages 339–348. IEEE, 2015.
[12] E. Poulton. Mediated patriot games: The construction and representation of national identities in the british television production of euro'96. *International Review for the Sociology of Sport*, 39(4):437–455, 2004.
[13] M. Raab. Simple heuristics in sports. *International Review of Sport and Exercise Psychology*, 5(2):104–120, 2012.
[14] G. M. Steeger, J. L. Dulin, and G. O. Gonzalez. Winning and losing streaks in the national hockey league: are teams experiencing momentum or are games a sequence of random events? *Journal of Quantitative Analysis in Sports*, 17(3):155–170, 2021.
[15] R. C. Vergin. Winning streaks in sports and the misperception of momentum. *Journal of Sport Behavior*, 23(2), 2000.
[16] S. Wilkens. Sports prediction and betting models in the machine learning age: The case of tennis. *Journal of Sports Analytics*, (Preprint):1–19, 2021.
[17] L. V. Williams, C. Liu, L. Dixon, and H. Gerrard. How well do elo-based ratings predict professional tennis matches? *Journal of Quantitative Analysis in Sports*, 17(2):91–105, 2021.
[18] S. Wilms. *Mental Training im Sport: Möglichkeiten der Anwendung in der Sozialen Arbeit*. PhD thesis, Hochschule für angewandte Wissenschaften Hamburg, 2014.
[19] G. Wood. Predicting outcomes: Sports and stocks. *Journal of Gambling Studies*, 8(2):201–222, 1992.
[20] Y. Zhang, E. T. Bradlow, and D. S. Small. New measures of clumpiness for incidence data. *Journal of Applied Statistics*, 40(11):2533–2548, 2013.

Competitiveness and betting markets: A comparison between European and American football leagues

N. García Aramouni*

*Universidad Torcuato Di Tella, Buenos Aires, Argentina

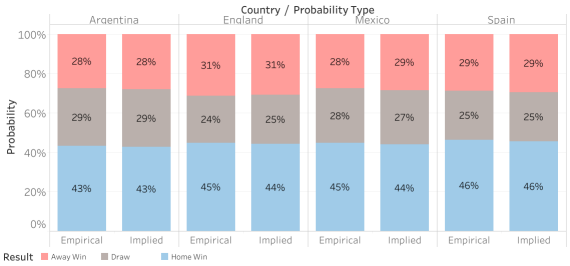
Abstract

Sport leagues may present significant differences across countries, even if they correspond to the same sport. These differences may be economic, related to their competitiveness, or other reasons. Whereas most of the specialized literature focuses on European competitions, studies considering other leagues are rather scarce. This paper uses information obtained from the betting market to compare two football leagues from Europe (England and Spain) and two leagues from America (Argentina and Mexico) regarding their level of competitiveness. We show that implied probabilities are more skewed in the European leagues, implying that they may be less competitive. Based on this analysis, we implement several simple betting strategies which, when evaluated using simulations, result in positive returns for the European leagues, but not for the American ones. This result opens the discussion on whether a greater skewness in the implied probabilities may represent an opportunity for investors.
Keywords: sports betting; market efficiency; sports results forecasting

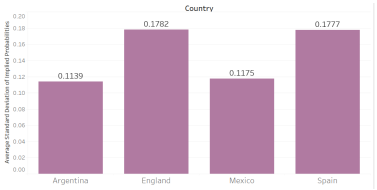
1 Introduction

The evaluation of the efficiency of sports betting markets has usually attracted the interest of academic researchers. Generally speaking, a set of odds is considered to be weak-efficient if they reflect objective probabilities so that no strategy generates positive expected return. Strong efficiency, on the other hand requires that no strategy would improve on expected returns from betting randomly. A significant portion of sports betting literature finds several violations of strong efficiency. Moreover, [5] finds that markets may not be weak-efficient either. Certain types of bias have generated interest among researchers, specially the favourite-longshot bias, which states that betting for the favourite generates greater returns than betting for the longshot [1] and the sentiment bias, which states that, contrary to theoretical models, betting on teams with a bigger fan base, usually generates greater returns (see [2, 3]). These phenomena show that sports betting markets should, at least, not be considered as strong-efficient.

Due to its economic power and to its international relevance, most of research regarding football/soccer betting markets has concentrated on European leagues. However, considering that American Continent national teams and players are among the best in the world, even when Latinamerican countries are less economically developed that nations in Europe, and that both financial and betting markets are less developed in Latin America, it is also important to analyze leagues in this continent as interesting conclusions may arise because of the economic or competitive differences.



(a) Empirical and Implied Probability by country and outcome



(b) Average standard deviation of implied Probabilities by league

Figure 1: Empirical and Implied Probability by country and outcome and Average standard deviation of implied Probabilities by league

In this paper, we try to evaluate differences in the sports betting markets of European and American football leagues, leveraging on data from football-data.co.uk. Our first analysis is descriptive, evaluating differences in the distribution of implied probabilities and differences in the distribution of outcomes. Results show that implied probabilities in the European football leagues are more skewed, which are an indicator of a lower lever of competitiveness.

With these results, we generate simple betting strategies, which are hard rules based on the distribution of implied probabilities of each match. preliminary results show that these strategies usually generate better profits that a *Naïve* strategy, which means betting on the most likely outcome. However, positive results are only obtained when applying these strategies on the European leagues, showing that the greater level of skewness may be an opportunity for bettors.

2 Descriptive Analysis

As we mentioned previously, we based our analysis on information available on football-data.co.uk, beginning in 2012 until the end of the 2020-21 season, and using average odds per bookmaker. For example, for the match in which Manchester United played at home against Liverpool on January 2013, the home win was quoted at 1.87, so a successful bet would result in a 0.87 return. Additionally, the draw was quoted at 3.64 and Liverpool's win was quoted at 4.06. Given this betting odds, we can calculate the implied probability of each outcome; however, considering that the sum of this implied probabilities don't sum up to 1 (due to the bookmaker's overround), we have to scale them. For example, given odds $\alpha_h, \alpha_d, \alpha_a$ representing a home win, a draw and an away win, we calculate the probability of a home win p_h in the following way (the probability of a draw or an away win can be calculated analogously):

$$p_h = \frac{1/\alpha_h}{1/\alpha_h + 1/\alpha_d + 1/\alpha_a}. \tag{1}$$

Having calculated the implied probability for each result and match, we compare the implied probability and the empirical probability by result and country.

In Figure 1a we can see that in general, implied and empirical probabilities are really close to each other, as the difference between them is never greater than 1 percentage point. However, some differences arise between leagues: England has the highest proportion of away wins, while Argentina has the highest draw ratio. Also, it seems that home teams are stronger in Spain, as they win roughly 46% of the times.

However, it is interesting to also evaluate the distribution of the implied probabilities in each league. To do this in a summarized manner, we evaluate the average standard deviation of the implied probabilities by league. This is done by calculating, by match, the standard deviation of implied probabilities, and then calculating the average by country of that metric.

Here, in Figure 1b, we can see that the standard deviation of odds is lower on American leagues than on European leagues. This is also consistent with the fact that, according to Figure 1a, Argentina and Mexico have a higher share of draws. This might be an indicator that Argentina's and Mexico's league are more competitive, and therefore, it might be more difficult to calculate these probabilities. If a league is more competitive, results will be more unpredictable, so it will be more complicated for bookmakers to assign a clear favourite. If this is the case, this would be consistent with the results found in [4], where it is shown that in multiple dimensions, such as distinct number of clubs that are champions, points difference between the champion and the runner-up, average point difference between the champion and the top 5 clubs and the champion's position in the previous tournament, European football leagues tend to be far less competitive than American ones. Moreover, according to [6], this difference in the league's level of competitiveness might be an indicator of a different level of market efficiency.

As we can see, these leagues seem to be different in how competitive they are. Logically, these differences also affect the betting market and make us believe that a particular strategy might have really different returns in different countries.

3 Preliminary experimental results

Based on the previous analysis, we will use these differences in this section, as we create hard-rules-based unitary betting strategies that will have different returns in each country. These strategies will not incorporate more information that the implied probabilities of each match to make the predictions and will try to leverage on the distribution of the probabilities to try to make *smart* decisions.

Given that implied probabilities are more skewed on Europe, intuitively, this could potentially generate one of two results:

- have a higher return in America, as there are more longshot wins. We could try to use that to generate higher profits, making a higher proportion of riskier bets;
- have a higher return in Europe. As probabilities are more skewed, a longshot win, even if it is improbable, will generate a greater return. Therefore, we should make a sensible number of riskier bets

In order to make our hard-rules-based strategies, we evaluate the implied probabilities in two ways, evaluating the standard deviation of implied probabilities and the difference between the most likely and the second most likely outcome:

- Evaluation of Standard Deviation: Given a threshold i, we will make a riskier bet if the standard deviation of probabilities is lower than i, and the *Naïve* bet if it is higher.

- Evaluation of Probabilities differences: Given a threshold i, we will make a riskier bet if the difference between the most likely outcome and the second most likely outcome is lower than i, and the *Naïve* bet if it is higher.

Both the standard deviation and the probability difference are two ways to evaluate the expected level of competitiveness a match has: a smaller standard deviation or difference means the match is expected to be more competitive. However, for our experimental results, it is interesting to evaluate both strategies. In a way, the selection of this threshold i might be a indicator of risk aversion: if a person doesn't like risk, this threshold will be low, but if this person likes risk, this threshold will be higher. When the level of risk (again, parameter i) is tolerable by the person, a riskier bet is made because the higher return compensates this risk. If this level of risk is not tolerable, we will carry out a *Naïve* bet. Also, we might use threshold i as a proxy for parity: if the standard deviation of probabilities (or the probabilities difference), both represented by i, is small enough then the teams that are playing the match have similar strength. Therefore, we could bet for a draw between both competitors. So, in conclusion, we will use this parameter i in two different ways: to allow riskier bets (using it as a proxy for risk aversion), and to allow draw bets (using it is a proxy for evaluating parity)

Considering that the number of possible "riskier" bets that we can try is endless, we highlight the ones that we tried in our experiments:

- Bet for the most likely outcome (a *Naïve* bet)
- Bet for the home team (a *Home* bet)
- If the standard deviation of probabilities for a particular match is smaller than i, bet for the second most likely outcome (we'll call this strategy $StDMidRisk$)
- If the standard deviation of probabilities for a particular match is smaller than i, bet for a draw (we'll call this strategy $StDMidRiskDraw$)
- If the standard deviation of probabilities for a particular match is smaller than i, bet for the least probable outcome (we'll call this strategy $StDHighRisk$)
- If the difference between the two most probable outcomes is smaller than i, bet for the second most probable outcome (we'll call this strategy $DiffMidRisk$)
- If the difference between the two most probable outcomes is smaller than i, bet for a draw (we'll call this strategy $DiffMidRiskDraw$)

First, we evaluate an scenario when we can select the best threshold i by strategy *ex-post* (we select one threshold by country and strategy, the one that gave the highest return, considering all matches between 2013 and the end of the 2020-21 season). Here, we can see the total return by strategy and country:

Figure 2 shows that these strategies, we can generate positive returns for both England and Spain, but not in American leagues. In particular, we are just getting positive results for 3 out of the 7 strategies. Additionally, in all countries we generate a return that is greater than the *Naïve* strategy. This differences ranges from 3.74 percentage points in Argentina to 4.59 percentage points in Spain. Even if we haven't generated a positive return in Mexico, the improvement over the *Naïve* strategy has been greater there than what we got in the English Premier League.

However, it is true that the results that we showed in Figure 2 could not be achieved as the threshold is optimized *ex-post*. In order to improve this, we optimize threshold i by strategy and country with information

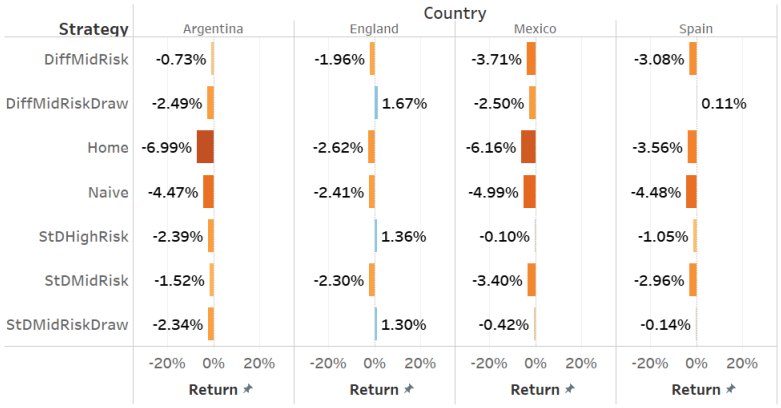


Figure 2: Total Return by Country and Strategy - Best Possible Results

of the previous year (this granularity was chosen for experimentation and stability reasons, although we could have chosen other level like, for example, monthly data). For example, with information of 2013, we select the best threshold i by strategy and country and apply that threshold to calculate the returns of 2014:

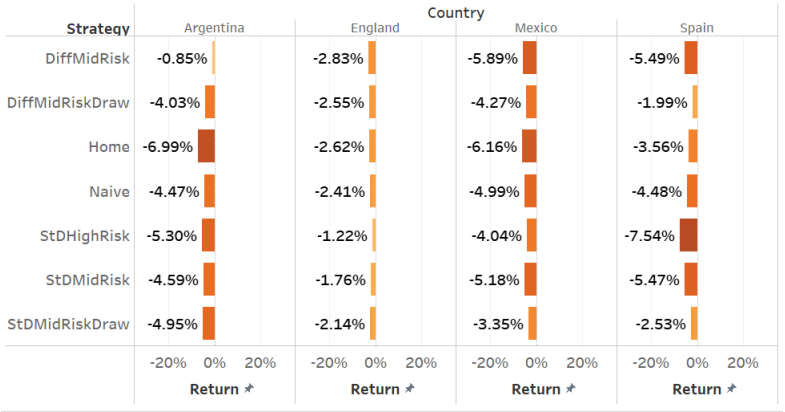
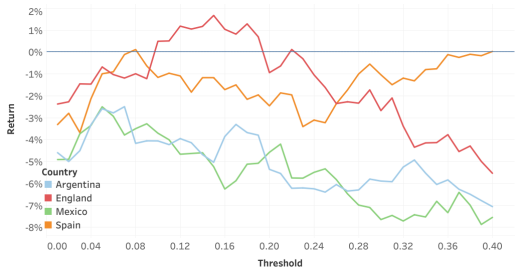


Figure 3: Total Return by Country and Strategy

Here, in Figure 3, we don't obtain positive returns, which might mean that results are weak-efficient. There are several reasons that may cause the drop in returns. First, considering the efficiency-market hypothesis of financial markets in general, we shouldn't be able to use information of the past to predict future behaviors as prices already incorporate all relevant information. Moreover, and putting into consideration the previous point, match and betting results should be independent between each other. If this is the case, there is no reason for us to believe that using the history to optimize parameters will give us the best possible results.

It is important to mention, however, that under this scenario the best returns are still being obtained on England and Spain. One possible reason for this is that, considering the mentioned greater skewness in



(a) Returns by Threshold and Country

Figure 4: Returns by Threshold and Country

implied probabilities, even when the home wins-draws-away wins distribution is similar in all countries, underdog wins have a greater reward in European leagues, which may represent a opportunity for investors. In Argentina and Mexico, even with a less developed market, the greater balance between teams generates a lower predictability in the results and a lower opportunity when betting for the underdog, which might contribute to the fact that we are always obtaining negative returns in Argentina and Mexico.

Lastly, we conduct some sensibility analysis, to evaluate the impact of threshold i under the strategy $DiffMidRiskDraw$:

The impact of the selection of the threshold is really different by country, as we can see in Figure 1. In Argentina and Mexico the pattern is similar, as the best potential results are achieved when this threshold is near 0.05, even when positive returns are never achieved. For England, positive returns are being obtained when this difference between the two most likely outcomes are between 0.1 and 0.2, but decreases afterwards. In Spain, results are very volatile; however small but positive results are obtained at particular values of the threshold.

4 Conclusion

In this work we analyze results in the football betting market in Argentina, Mexico, England and Spain over the last decade. We show that implied probabilities are more skewed on European leagues, potentially indicating a lower level of competitiveness. However, we can use the distribution of the implied probabilities in each match to generate hard-rules-based betting strategies. The greater skewness in European odds has helped us get positive results in those countries. In Argentina and Mexico, as underdog wins are not rewarded highly (reflected by the lower standard deviation of probabilities), we have not generated a strategy that got us positive results there. Therefore, we have shown differences both at a descriptive and at a economic level between leagues.

References

[1] M. Cain, D. Law, and D. Peel. The favourite-longshot bias and market efficiency in uk football betting. *Scottish Journal of Political Economy*, 47(1):25–36, 2000.

[2] A. Feddersen, B. R. Humphreys, and B. P. Soebbing. Sentiment bias and asset prices: Evidence from sports betting markets and social media. *Economic Inquiry*, 55(2):1119–1129, 2017.
[3] D. Forrest and R. Simmons. Sentiment in the betting market on spanish football. *Applied Economics*, 40(1):119–126, 2008.
[4] N. García Aramouni. Analítica sports lab: Usamos los datos para medir cuál es la liga más competitiva del mundo. *Analitica Sports*, 2019, available at https://analiticasports.com/analitica-sports-lab-usamos-los-datos-para-medir-cual-es-la-liga-mas-competitiva-del-mundo/ (last accessed on 9-Oct-2021).
[5] J. Goddard and I. Asimakopoulos. Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting*, 23(1):51–66, 2004.
[6] A. Oikonomidis, A. C. Bruce, and J. E. Johnson. Does transparency imply efficiency? the case of the european soccer betting market. *Economics Letters*, 128:59–61, 2015.

A Markov game model for determining tactical changes in an association football match

N. Hirotsu*, Y. Masui*, Y. Shimasaki* and M. Yoshimura*

* Juntendo University, 1-1 Hiragagakuendai, Inzai, Chiba, Japan
e-mail address: nhirotsu@juntendo.ac.jp

Abstract

In this study, we model an association football match as a Markov game. We discretise the pitch into nine zones, and define the states of the Markov game according to the zone of the pitch in which the ball is located, the team in possession, and the score. We estimate transition rates between states based on the frequencies of transitions in J-league matches. The game is formulated as a two-player zero-sum game, in which both teams maximise their probabilities of winning the match as the payoffs by means of their tactical changes. A simple example of optimal tactical changes is demonstrated in the case that two teams can make tactical changes during a match, and how this approach can be used for determining tactical changes.

1 Introduction

In the field of sports, a lot of conflict scenes occur in a game. The application of game theory to sports enables the quantification and analysis of the scenes in which tactics are employed. For example, Hirotsu et al. (2010) analyzed tactics for attacking and blocking in volleyball. Chiappori et al. (2002) developed a game tactics model for penalty kicks between kickers and goal keepers, and analyzed how each tactic uses three options: left, right and middle.

Not only focusing on specific scenes occurred in the game, we can model a whole game as a stochastic process, i.e. as a stochastic game or a Markov game. In Markov game, players make decisions with a set of actions, each of which gives a different expected payoff in each state for the players. A single-player Markov game is known as a Markov Decision Process (MDP)(e.g., Van Roy et al., 2021), where one player or team seeks to maximize his/her or its payoff. It has been used to determine optimal in-play strategy given a particular state within a game.

Recently, in the context of reinforcement learning in artificial intelligence, application of MDP to sports are widely expanding. In basketball, under the situation of the availability of the optical tracking data, Cervone et al. (2016) propose a framework to estimate the expected number of points obtained by the end of a possession. In ice hockey, Schulte et al.(2017) model a play sequence as a Markov game using play-by-play data for valuing player actions, locations, and team performance.

In terms of association football, the sequences of plays being in possession to the end of the sequence such as shoot or no shoot are modelled in the framework of MDP that models the behaviour of the team possessing the ball in order to gain insight during a match (e.g. Van Roy et al.,2021). With the use of tracking data of players and the ball, Fernandez et al.(2021) propose a framework for evaluating the expected possession value of a possession which represents the likelihood of a team scoring or conceding the next goal, by producing visually-interpretable probability surfaces from a series of deep

neural network architectures. Liu et al (2020) evaluate all types of actions from play-by-play event data, utilizing a deep reinforcement learning model to learn an action-value Q-function to measure a player's overall performance by the aggregate impact values of his actions over all the games in a season.

A Markov game is a generalisation of a MDP which allows for multiple players with interacting or competing to maximize their payoffs. In terms of Markov game application to sports, Kira et al. (2019) developed the Markov game as a model of optimum tactics such as sacrifice bunts and base stealing in baseball, to calculate the increase in the probability of wins for opposing teams employing tactics at the best timing. Hirotsu and Wright (2006) and Hirotsu et al. (2009) model tactical changes of formations in an association football match as a zero-sum game and a non-zero-sum game. Luo et al. (2020) formulate the ice hockey as a Markov game, and propose inverse reinforcement learning by combining Q-function learning with inverse reinforcement learning to provide a player ranking method based on play-by-play events.

In this paper, we formulate an association football match as a two-player zero-sum Markov game. This is an extension of Hirotsu and Wright (2006, 2009), by considering the location of the ball on the pitch as nine zones together with the change of possession of the ball, following Hirotsu et al.(2022). We present a Markov game formulation not only for maximizing the probability of home team winning the game, but also away team minimizing it during the game by their tactical changes. We provide a numerical example of the effect of the change of transition rates based on the averaged data of the J-league. We set transition rates between states, and make a change of the transition rates, which is assumed to be caused by tactical change of the teams. We estimate the effect of the change on the probability of winning the match and the timing of the optimal tactical change.

2 Markov game formulation

An association football match game can be seen as progressing through a set of stochastic transitions occurring due to a change of possession of the ball or the scoring of a goal. A Markov process model can be used to appropriate the progress of the match as an approximation. Hirotsu et al. (2022) propose a Markov process model in which the pitch discretised into nine zones and define the states as follows:

State H_G: Home team scores a goal;
State H_I: Home team is in possession of the ball and the ball is located in the "I" zone (I=1,…, 9);
State A_I: Away team is in possession of the ball and the ball is located in the "I" zone (I=1,…, 9);
State A_G: Away team scores a goal.

Here, the "I" zone (I=1,….9) on the pitch is defined in Figure 1. There are two states for the goal scoring (states H_G and A_G) and 18 states relating to the location and team's possession of the ball. We make the following definitions, and show them in Figure 2.

The transition probabilities between them are defined in Table 1. In this table, a_{H1G} is interpreted as the transition rate from state H_1 to H_G (i.e. scoring a goal from the "1" zone by home team). The probability of a transition from H_1 to H_G and a transition from H_1 to H_2 in the next small time dt is expressed by $a_{H1G}dt$ and $a_{H1H2}dt$, respectively. Other transitions are also expressed in a same manner. The probability of remaining in state H_1 is thus $1 - (a_{H1G}+a_{H1H2}+\ldots+a_{H1A1})dt$. Here, a_{ij} (i,j= $H_1,H_2,\ldots,A_1$)

is defined as the transition rate from state *i* to state *j*. The 18 (=9×2) states except H_G and A_G are identified by the combination of the following factors:

- Location of the ball (9 possibilities);
- Possession of the ball (2 possibilities).

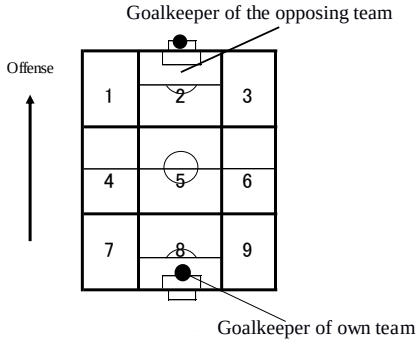


Figure 1: The areas on the pitch

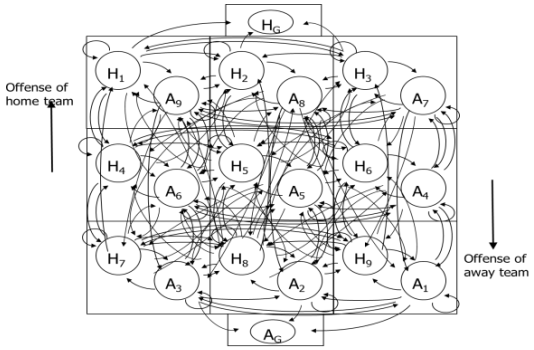


Figure 2: Image of the Markov process model

Table 1: Definition of transition probabilities in a football match

Transition	Probability	Remarks
$i \rightarrow H_G$	$a_{iG}\,dt$	Transition from possession to scoring a goal for home team from state i (i=H_1,H_2,…,H_6)
$i \rightarrow j$	$a_{ij}\,dt$	Transition from state i to state j (i,j=H_1,H_2,…,A_1)
$i \rightarrow A_G$	$a_{iG}\,dt$	Transition from possession to scoring a goal for away team from state i (i=A_1,A_2,…,A_6)

Further, introducing the number of goals by which the home team leads as states, that is,

- Number of goals by which the home team leads (21 possibilities, assuming that the number of runs by which either team may lead will never exceed 10);

378 (=9×2×21) different states are defined in the course of a game. Here, we look at the winning from the perspective of the home team. Given this specification, Equation (1) determines the probability of winning from each state:

$$\frac{d\mathbf{w}(t)}{dt} = \mathbf{P} \cdot \mathbf{w}(t) \tag{1}$$

where $\boldsymbol{w}(t)$ is a 378×1 vector, each entry of which corresponds to the probability of home team winning from a position of leading by l goals with time t remaining, starting from state i (i = H_1,H_2,H_3,…,A_1). **P** is a 378×378 matrix, which represents the transition between the 378 states. We will present the matrix formulation of Equation (1) in the conference. We can numerically solve them with the boundary conditions at the end of the game such that $w_{H1}(l|0) = w_{H2}(l|0) = \ldots = w_{A1}(l|0) = 1$ if $l > 0$ and 0 if $l < 0$. In this paper, we set $w_{H1}(l|0) = w_{H2}(l|0) = \ldots = w_{A1}(w|0) = 0.5$ only if $r = 0$ in the case of drawing.

By solving Equation (1), we can simultaneously obtain the probability of home team winning from a position of leading by l goals with time t remaining in each of the 378 states. This approach makes it possible to develop a Markov game formulation to identify the situations to enforce alternative transition rates between state by a tactical change. That is, we can derive the equations corresponding to the situation that the home team and the away team take tactics k (=0,1,2,…,K) and h (=0,1,2,…,H),

respectively, and compare these probabilities among the alternative tactic k and h, and then take the maximum of them in each state. We can formulate this procedure as follows:

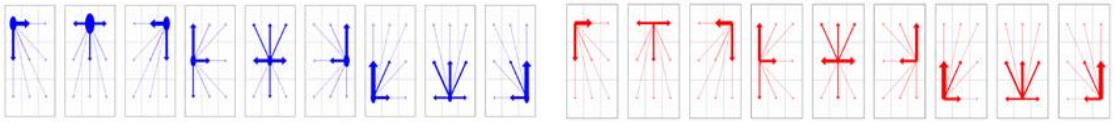
$$\frac{d\mathbf{w}(t)}{dt} = \max_{k,k' \in \{0,1,\cdots,K\}} \min_{h,k' \in \{0,1,\cdots,H\}} \begin{cases} \mathbf{P}^{kh} \cdot \mathbf{w}(t) & : No\ tactical\ change \\ \mathbf{P}^{k'h} \cdot \mathbf{w}(t) & : Tactics\ k \to k'\ (Tactical\ change\ from\ k\ to\ k' \neq k) \\ \mathbf{P}^{kh'} \cdot \mathbf{w}(t) & : Tactics\ h \to h'\ (Tactical\ change\ from\ h\ to\ h' \neq h) \\ \mathbf{P}^{k'h'} \cdot \mathbf{w}(t) & : Tactics\ k \to k'\ and\ h \to h'\ (Tactical\ change\ from \\ & \qquad to\ k' \neq k\ and\ h\ to\ h' \neq h \end{cases} \quad (2)$$

where maximisation is taken from possible different tactics k and k' (=0,1,2, $\cdots$,K) for the home team and minimisation is from possible different tactics h and h' (=0,1,2, $\cdots$H) for the away team. We note that under the assumption that both teams do not make a tactical change in the small time dt simultaneously, we can find the Nash equilibrium and obtain the solution of Expression (2) as a pure strategy.

3 Example

3.1 Transition rates

We now present a numerical example. The play-by-play data used in this study was provided by Data Stadium Inc. Figure 3 shows the estimates of transition rates of the league average which is extracted as intercept as main factors for the transition rates between states which is identified most suitable log-linear model presented in Hirotsu et al.(2022). This means that the average number of transition from each location to the other location can be estimated by taking away of the effects of home advantage and strength of each teams. Figure 3(a) shows transition rates of an average team in the case of keeping its possession. The higher the transition rate is, the wider the arrow of the corresponding transition is. In terms of the transition between the same states, the higher the transition rate is, the larger the circle is. Figure 3(b) shows transition rates of an average team in the case of not keeping its possession. For example, the transition rate from H_2 to H_1 appears as 3.081 times/min. in the second picture from the left in Figure 3(a).



(a) Case of keeping its possession (b) Case of not keeping its possession

Figure 3: Average transition rates estimated by the J-League data

3.2 Setting the transition rates for tactical change

Based on the estimates of the average transition rates, we set up the transition rates of the hypothetical game between the average home team and the average away team. As an advantage of using the Markov game, we can calculate the effect of the change of transition rates on the probabilities of winning the match. We set the case that the home team have home advantage addition to the average transition rates

shown in Figure 3 as a base. The effect of home advantage is estimated in Hirotsu et al.(2022) and we introduce this effect into each transition rate. For example, the effect of home advantage in the transition from H_2 to H_1 is estimated 1.08, and the transition rate for the home team is set as 3.35(=1.08*3.081) times/min.

Here, as an example concretely to see the sensitivity of the transition rate, we manipulate the transition from zone "4" to "1" for the home team and from "7" to "1" for the away team. We change the transition rates by the amount of its 1SD and 0.5SD, respectively, and see the effect of the change of transition rates. That is, we set four cases: (1) both teams take base tactics ((0,0)), (2) the home team increase the transition rate from zone "4" to "1" by 1SD ((1,0)), (3) the away team increases the transition rate from zone "7" to "1" by 0.5SD ((0,1)), (4) both teams take the tactics which increases these transition rates ((1,1)). The SD can be obtained by the annual data of J-League, and in terms of the transition from H_4 to H_1 1SD=1.418 times/min. When the home team takes Tactic 1, the unsuccessful pass will increase. Thus, the transition rate from H_4 to A_9 is also set to increase by its 2SD. (i.e. from H_4 to A_9, 2SD=0.814 (=2×0.412) times/min.) In a similar manner, Tactic 1 for the away team is assumed to increase the number of long pass from zone 7 to 1 by 0.5SD. (i.e., the transition rate from A_7 to A_1 increases by 0.5SD.) As the unsuccessful long pass will increase, transition rate from A_7 to H_9 also increase by 2SD. (i.e. from A_7 to A_1, 0.5SD=0.1781 (=0.5×0.3562) times/min.; from A_7 to H_9, 2SD=1.0626 (=2×0.5313) times/min.)

3.3 Calculation result

Now we can calculate the probability of the home team winning using expression (2). We show the change of formations during the game under the condition that they always make their best decisions at time *t* remaining. Table 2 shows a whole image of the optimal tactics and timing of their tactical changes, representing the case where the number of goals by which either team leads is not to exceed 2. Both teams start off with Tactic 0 (i.e. (0,0)). If the home team falls behind by 2 goals with more than 57 min. remaining, or by 1 goal with less than 57 min. remaining, it should make a tactical change from 0 to 1. Otherwise, if the away team falls behind by 2 goals with less than 18 min. remaining, or by 1 goal with less than 7 min. remaining, it should make a tactical change from 0 to 1. This result looks reasonable because Tactic 1 is the offensive tactics for scoring goals in order to get the scores level after falling behind.

Table 2: Optimal tactics and timing of the tactical changes in the case where both teams make their best decisions

	Lead in goals for the home team				
Remaining Time	-2	-1	0	1	2
90–57 min.	(1,0)	(1,0)	**(0,0)**	**(0,0)**	**(0,0)**
57–18 min.	(1,0)	(1,0)	**(0,0)**	**(0,0)**	**(0,0)**
18– 7 min.	(1,0)	(1,0)	**(0,0)**	**(0,0)**	(0,1)
7– 0 min.	(1,0)	(1,0)	**(0,0)**	(0,1)	(0,1)

4 Conclusion

In this paper, we have modelled an association football match as a Markov game. Discretising the pitch into nine zones, we have defined the states of the Markov process model according to the zone of the pitch, the team in possession and the score. Based on the Markov process model, we have derived the formulation for getting the probability of winning the game with goal differences, and formulate the game as a two-player zero-sum Markov game. We have estimated transition rates between states in a hypothetical game between the average home and the average away teams using annual data of J-league. As an example of optimal tactical changes is demonstrated in the case that one and two teams can make tactical changes during a game, by setting the change of transition rates from state H_4 and A_7 by 1SD and 0.5SD, and obtain the reasonable result of the optimal tactic and timing of the tactical changes.

References

[1] Chiappori, P.A., S. Levitt and T. Groseclose (2002). *Testing mixed strategy equilibria when players are heterogeneous: The case of penalty kicks in soccer.* Amer. Econ. Rev., 92: 1138-1151.

[2] Cervone, D., D'Amour, A., Bornn, L. and Goldsberry, K. (2016). *A multiresolution stochastic process model for predicting basketball possession outcomes*. J. Am. Stat. Assoc., 111: 585–599.

[3] Fernandez, J., Bornn, L., and Cervone, D. (2021). *A framework for the fine-grained evaluation of the instantaneous expected value of soccer possessions*. Machine learning. ISSN 0885-6125.

[4] Hirotsu, N. and M. B. Wright (2006). *Modelling tactical changes of formation in association football as a zero-sum game.* J. Quant. Anal. Sports, 2: Article 4.

[5] Hirotsu, N., M. Ito, C. Miyaji, K. Hamano and A. Taguchi (2009). *Modeling tactical changes in association football as a non-zero-sum game.* J. Quant. Anal. Sports, 5: Article 2.

[6] Hirotsu, N., M. Ito, C. Miyaji, K. Hamano, and A. Taguchi (2010). *A game theoretic analysis of tactics in the phase of reception attack in volleyball*. Int. J. Comp. Sci. Sports, 9: 30-44.

[7] Hirotsu, N., K. Inoue, K. Yamamoto and M. Yoshimura (2022). *Soccer as a Markov process: modelling and estimation of the zonal variation of team strengths.* IMA J. Mang. Math., dpab042, https://doi.org/10.1093/imaman/dpab042.

[8] Kira, A., N. Kamiyama, H. Anai, H. Iwashita, and K. Ohori (2019). *On dynamic patrolling security games*, J. Oper. Res. Soc. Japan, 62: 152-168.

[9] Liu, G., Luo, Y., Schulte, O., and Kharrat, T. (2020). *Deep soccer analytics: learning an action-value function for evaluating soccer players*. Data Mining and Knowledge Discovery, 34, 09. doi: 10.1007/s10618-020-00705-9.

[10] Luo, Y., Schulte, O., and Poupart, P. (2020). *Inverse reinforcement learning for team sports: Valuing actions and players*. In Bessiere, C. (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI-20, pp. 3356–3363. International Joint Conferences on Artificial Intelligence Organization, 7. doi: 10.24963/ijcai.2020/464. URL https://doi.org/ 10.24963/ijcai.2020/464. Main track.

[11] Schulte, O., Khademi, M., Gholami, S., Zhao, Z., Javan, M. & Desaulniers, P. (2017). *A Markov Game model for valuing actions, locations, and team performance in ice hockey*. Data Min. Knowl. Discov., 31: 1735–1757.

[12] Van Roy, M., Robberechts, P., Yang, W.-C., De Raedt, L., and Davis, J. (2021). *Learning a Markov Model for Evaluating Soccer Decision Making*, Reinforcement Learning for Real Life (RL4RealLife) Workshop in the 38th International Conference on Machine Learning.

Modelling bookings in association football

J. Hargreaves* and B. Powell**

*Department of Mathematics, University of York, York, UK + email address: jessica.hargreaves@york.ac.uk
** Department of Mathematics, University of York, York, UK + email address: ben.powell@york.ac.uk

Abstract

In this report, we describe preliminary investigations into, and opinions on, the subject of statistical analysis of bookings in association football. More specifically, we consider whether more developed methodology for modelling goals can be re-purposed to predict the number of bookings in a given match, and, if so, how. A more thorough description of the topic and a more rigorous analysis of our modelling experiments will appear in forthcoming work.

1 Introduction

1.1 Executive summary

Preliminary modelling suggests that the process by which bookings are made is intrinsically more noisy and less predictable than that for goals. With a low signal-to-noise ratio and a potentially large number of team-specific parameters to infer, over-fitting is a major problem. This leads to poor out-of-sample predictions, which undermine the practical utility of our models. A natural recourse in such situations is to apply shrinkage/variance-reduction methods and/or to constrain models using additional prior knowledge. Our working conclusion is that out-of-the-box shrinkage methods, namely the L_1 coefficient-penalizing Lasso, do help alleviate the over-fitting problem but that further improvements require a more thoughtful consideration of the factors driving bookings.

We note that the models we employ in these investigations are penalized generalized linear models (GLMs), generally not including interaction effects. However, it is possible that this important, and generally well understood, class of models is not the best tool to use for the problem under consideration. Next-generation machine learning models such as artificial neural networks or random forests may allow us to identify and exploit covariate interaction effects that the GLMs are blind to. While potentially valuable, we suspect that these flexible, highly parameterized models will be even more susceptible to over-fitting. Further work in this direction is deferred for future studies.

1.2 Background and Motivation

Since the seminal paper by [13], a substantial statistical literature has developed on the topic of models for association football (soccer) matches and, in particular, probabilistic models for match scores. Maher's model assumes that the numbers of goals scored by each team in a football match follow independent Poisson processes and that the rates at which the teams expect to score goals are functions of the ability of the two teams to attack and defend. Subsequent papers, most notably [5], developed the Maher model in a variety of directions. For example: additional important work on an alternative bivariate Poisson model is presented in

[10]; dynamic models are presented in [15] and [12]; and [1] demonstrate improvements in model fit with the use of a Weibull based count model.

The Dixon-Coles model, whose influence on the literature has been especially strong, is essentially a Poisson regression model with one important structural difference. While marginally Poisson, the number of goals scored by each team within a match are, for a certain range of values, modelled as being probabilistically dependent. The dependence is motivated on both theoretical and empirical grounds, which the authors explain in their paper. Covariates for the model include dummy variables encoding offensive and defensive strengths for each team and another encoding whether a team plays at their home ground. The regression coefficients are implicitly time varying, with their values being estimated using data that is weighted according to an envelope/kernel function applied to match dates. In Section 3 we present first attempts at applying modified versions of the Dixon-Coles model to bookings for each team in a match.

2 Exploratory data analysis

Our principle data set for this study, which is publicly available at `football-data.co.uk`, consists of UK Premier League matches between 2011 and 2019. These data inform the majority of the findings below, and the modelling exercise in Section 3. The period has been intentionally chosen to avoid the effects of the COVID–19 pandemic, whose destabilizing effects are partially explored in Section 2.1.

2.1 COVID–19 pandemic effects

Due to social restrictions during the COVID–19 pandemic, many professional sports leagues around the world experienced varying levels of spectator reductions during 2020–2021. Several studies (for example, [2, 6, 8, 14, 18]) have since investigated the effects of these reductions on home performance and referee decisions. The studies reach a general consensus that lower spectator numbers affected both home team performance and referee decisions. In particular, as pointed out by [6], home teams appeared to be treated less favorably during this period due to the reduction in crowd pressure on referees. Additionally, [2] identify a (statistically significant) reduction in the number of bookings for away teams, thus reducing the 'home advantage' further.

These findings have the potential to adversely affect our current work modelling bookings. To assess this potential more formally we investigated goal and booking counts in the UK Premier League during the COVID period using a set of varying coefficient models. These models, introduced by [9] and developed more recently by [7], allow us to estimate smoothly time-varying covariate effects. Preliminary analyses (illustrated in Figures 1 and 2) indicate distinct and atypical changes in base-rates for goal and booking counts as well as relationships between the counts and match locations, and the counts and bookmaker's odds. These observations align broadly with the aforementioned studies and motivate the exclusion of the COVID-affected seasons from our main modelling exercises.

2.2 Team interaction effects

The title of this subsection euphemistically refers to the antagonistic relationships and rivalries between specific pairs of teams, often, but not exclusively, based on geographical proximity. In Figure 3 we present means of counts of bookings for matches for different team pairings. Immediate conclusions are not obvious

from the plot. In principle, very large values in a row or column with mostly small values would indicate that an interaction effect which, if not accounted for, could skew team-specific parameters inappropriately. We suggest that further analysis of the significance and persistence over time of the apparent interaction effects is required before we attempt to incorporate them into models. For now, however, we choose not to do so.

2.3 Team strength disparity effects

Next, we take a closer look at the relationships between a team's strength relative to its opponent in a match and the number of goals scored and bookings incurred in that match. We quantify a team's relative strength as the logarithm of the ratio of its win and loss probabilities as implied by bookmaker's odds. In Figure 4 we illustrate the relationships found in the data. Obviously, we see a strong positive correlation between a team's relative strength and the number of goals it scores. We also see a less strong, but still statistically significant, negative correlation between relative strength and the number of bookings incurred. This finding is in line with the idea that weaker teams resort to foul play more often when faced with a more technically able opponent.

2.4 Temporal effects

Our analyses appear to show no significant systematic variation in marginal goal counts over time. We do, however, detect temporal effects on the booking count. To be more precise, these effects have been quantified by fitting to the counts Poisson regression models whose covariates consist of a linear trend and a pair of sinusoids with a period of one year. While these models are clearly only able to capture very simple dynamics, they generally support arguments against including temporal or seasonal effects in models for goals but for including them in models for bookings. The latter argument is further strengthened by anecdotal reports of teams being more cautious about incurring bookings as they accumulate over a season.

2.5 Home/away effects

The main factors impacting home advantage are believed to be: crowd support (through *both* the encouragement of home players' performance and biasing referee's decisions (e.g. penalties and bookings) in favor of the home team); familiarity with the stadium; and travel fatigue. In-depth discussion of these topics can be found in [17, 18].

In the current work, the home advantage manifests itself numerically in significant coefficient estimates in the models of Section 3 and graphically in Figure 6 where we see goal counts and booking counts skewed in favour of the home team. The significance of both effects is also detectable from simple paired t-tests, for example, which show the average home team scoring approximately 0.35 more goals and incurring 0.24 fewer bookings than the away team.

2.6 Gender effects

Quantifying differences between goal and booking counts for women's and men's football is undermined by the relative scarcity of data relating to the latter. Nevertheless, data collected from the RSSSF `http://www.rsssf.com/intland-women.html`, for example, appear to show qualitatively similar phenomena for both genders. The home advantage effect, for instance, appears as a statistically significant difference in

home and away goal counts of approximately 0.68 for international women's matches. We note that data to inform additional inferences is often available but is generally collated less systematically and distributed less widely than the data for the men's game. We anticipate that this is likely to change as the women's game increases in popularity.

3 Predictive modelling

We test the ability of three varieties of model to predict bookings counts and, for comparison purposes, goals counts. The first two models are Dixon-Coles Poisson and logistic regression models with the characteristic adjustments for intra-match dependence that effectively inflate simultaneous zero counts for both teams. The third model is a 'hurdle model' that combines the first two. More specifically, it involves modelling the non-zero status of the counts, then modelling the counts given that they are greater than zero. All the models use the same set of covariates in order to accommodate:

1. team-specific offensive and defensive effects as in the original Maher and Dixon-Coles models,
2. a home advantage effect,
3. referee-specific effects,
4. a team-disparity effect informed by pre-match bookmaker's odds,
5. 3 parametric temporal effects (linear trend and a pair of seasonal sinusoids).

All models lead to predictive distributions for goal and booking counts. A specific cumulative probability from these distributions will eventually be used to assess them in Section 3.4.

3.1 Poisson regression

The response variables for these models are the goal and booking counts. Regression coefficients are subject to L_1 penalization whose strength is calculated to minimize a cross-validated error estimate.

3.2 Logistic regression

The response variables for these models are values in $\{0, 1\}$, indicating whether the counts are less than or equal to the relevant population median. As will become clear in Section 3.4, these models directly target the probability according to which they will be assessed. The model coefficients are penalized in the same way as the Poisson regression models.

3.3 Logistic/Poisson hurdle models

These models, introduced by [4] and recently applied to football goals by [16], provide an alternative method for accommodating an over- or under-abundance of zero counts in the data. They do not include the intra-match dependence adjustment that characterizes the Dixon-Coles models, which partly serves to accommodate the same phenomenon. The hurdle models essentially describe the distribution of a count via a Bernoulli distribution for a its non-zero status and a truncated Poisson distribution for its value given that it is non-zero.

For fitting these models we employ the *pscl* package for R, which contains code developed by [19]. This code currently does not penalize fitted coefficients.

3.4 Model comparison

Our comparisons for predictive performance are based on whether we can predict if a team's goal or booking count in a given match will fall above or below a specific threshold value, k. The idea is informed by the practical importance of such predictions for under/over-type gambles in betting markets. The specific threshold values in question are chosen to be the marginal median counts across all matches in the data set.

Our model-based predictions are informed by a training subset of the data consisting of matches during the first five of the seven seasons under consideration. The remaining n_{test} matches are held back for evaluation. Specifically, the probability assigned by our models to the outcome (a count being less than or equal to, or greater than the threshold) that did occur is computed for matches in the test set. Geometric averages of these probabilities, corresponding to exponentiated scaled log-likelihoods, are presented in Table 1. Indexing the probabilities for each match in the test set by i and denoting them $\hat{p}_i$, the scores are computed according to the formula

$$GAPS = \text{Geometric Average Probability Score} \tag{2}$$

$$= \left(\prod_{i=1}^{n_{\text{test}}} \hat{p}_i^{1(\text{count } i \text{ is} \leq k)} (1-\hat{p}_i)^{1(\text{count } i \text{ is} > k)} \right)^{1/n_{\text{test}}} . \tag{3}$$

This quantity is an exponentiated average log-score for the predictions $\hat{p}_i$. The benchmark against which we suggest measuring the regression models is the GAPS achieved by a forecaster who specifies for every match the same probability, which is computed as the marginal proportion of all goals or bookings in the training set that are less than or equal to the relevant threshold k.

To reiterate, the GAPSs are (geometric) averages of probabilities assigned by models to outcomes that do occur. Good models will therefore produce high GAPSs. The nature of the under/over-outcomes has been selected so that it is relatively easy for a naive model, which allocates the same outcome probabilities for each event, to achieve a score of around 0.5.

Our results are distinctly underwhelming. For the goal counts, both Poisson and logistic regression models allocate to outcomes that do occur probabilities that are significantly, but only modestly, greater than the naive marginal method. For the bookings, the Poisson and logistic regression models allocate probabilities that are, on average, not significantly different from those of the naive marginal method. The hurdle models perform significantly worse than the naive marginal method for both goals and bookings. In each of the preceding three sentences the word 'significantly' is used in its technical sense and is informed by paired Wilcoxon signed-rank tests comparing pairs of logged probabilities from different models for the same under/over events.

4 Remarks

Summary statistics from the fitted Dixon-Coles models for bookings suggest that they are picking up on and quantifying real effects. For example, the causal effects discussed in Section 2 are all reflected in

	Marginal	Poisson reg.	Logistic reg.	Hurdle model
Goals	0.5115	0.5527	0.5519	0.4447
Bookings	0.4998	0.4995	0.4920	0.4844

Table 1: Geometric averages of probabilities assigned to outcomes in the test data.

correspondingly large fitted model coefficients. Despite this, the predictive skill of the models on test booking data is not significantly different from those of the most naive methods. We tentatively conclude that the model is also (over-)fitting to illusory trends in the training data that are not present in the test data. Shrinking all coefficients towards zero by a degree calibrated to minimize cross-validated prediction errors improves the models' out-of-sample predictive skill but the gain over the naive methods remains very small. We conjecture that further improvements, if possible, will require more nuanced, contextually motivated selection or shrinkage of covariate effects or a reformulation of the prediction problem that targets more predictable quantities.

A particularly promising direction for further investigation, also identified by [11], involves the effects of individual players on a match. Since individuals are arguably more capable of skewing booking counts than goal counts, we ought to expect them to play a greater role in predictive modelling of bookings. However, as discussed above, without very careful treatment, adding player effects is likely to exacerbate over-fitting problems.

We also anticipate opportunities for model improvement in the form of 'match importance' indices to be used as extra covariates in predictive models. These might, for example, be based on relative league position(s) or on a 'derby' indicator that acts as a proxy for reputational standing. The relevance of these indices, as anticipated by [3] for example, is based on the premise that teams perform differently when the consequences of their performance are higher or lower. Computing such measures of importance is likely to benefit from a combination of expert background knowledge and numerical experimentation.

Acknowledgements

The work above is motivated and informed by an undergraduate student project in collaboration with Sky Betting & Gaming. Accordingly, the authors would like to acknowledge valuable contributions from Thomas Hemery to the original project. They would also like to thank the team at Sky Betting & Gaming (including Donough Regan, Jon Carter, Mitch Bond, Will Cook and Fredrik Bjorkeroth) for helpful conversations and insights.

5 Appendix A: Figures

References

[1] Boshnakov, G., Kharrat, T., and McHale, I. G. (2017). A bivariate weibull count model for forecasting association football scores. *International Journal of Forecasting*, 33(2):458–466.

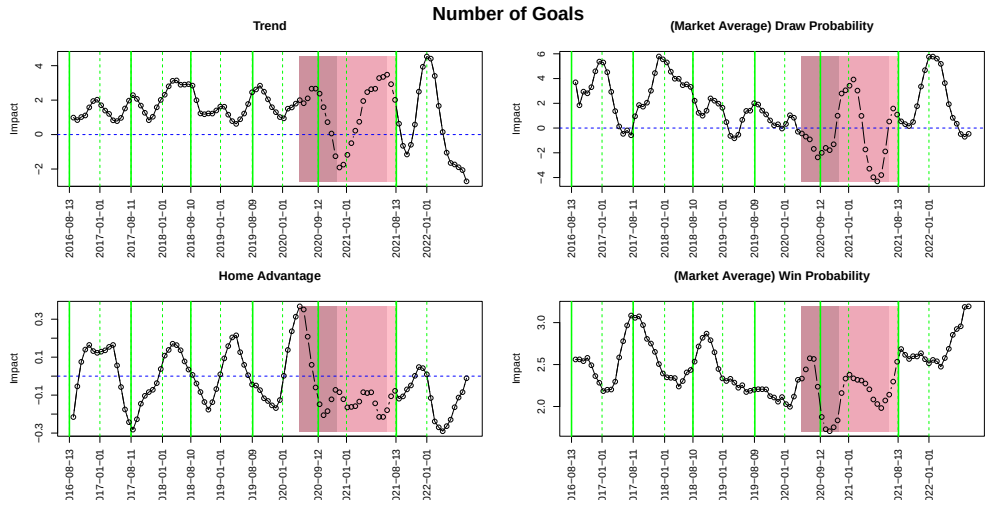


Figure 1: The estimated coefficient functions of the varying coefficient models (Section 2.1): Goals. Vertical green lines indicate: the start of a season (solid); New Year's Day (dashed). Red background indicates COVID restrictions in place (dark red: zero spectators permitted; dark pink: "tiered" restrictions in England (i.e. some grounds permitted restricted spectator numbers); light pink: all grounds permitted restricted numbers of spectators).

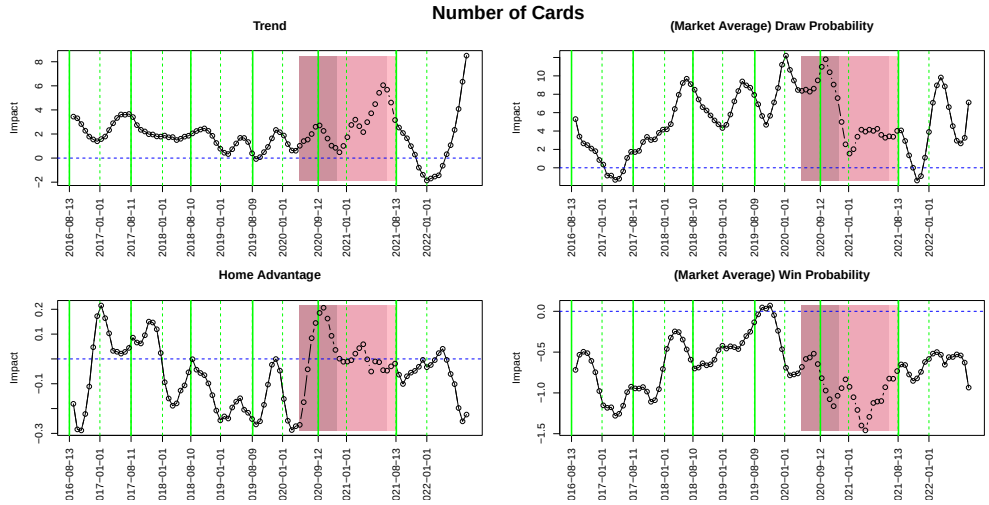


Figure 2: The estimated coefficient functions of the varying coefficient models (Section 2.1): Cards. Vertical green lines indicate: the start of a season (solid); New Year's Day (dashed). Red background indicates COVID restrictions in place (dark red: zero spectators permitted; dark pink: "tiered" restrictions in England (i.e. some grounds permitted restricted spectator numbers); light pink: all grounds permitted restricted numbers of spectators).

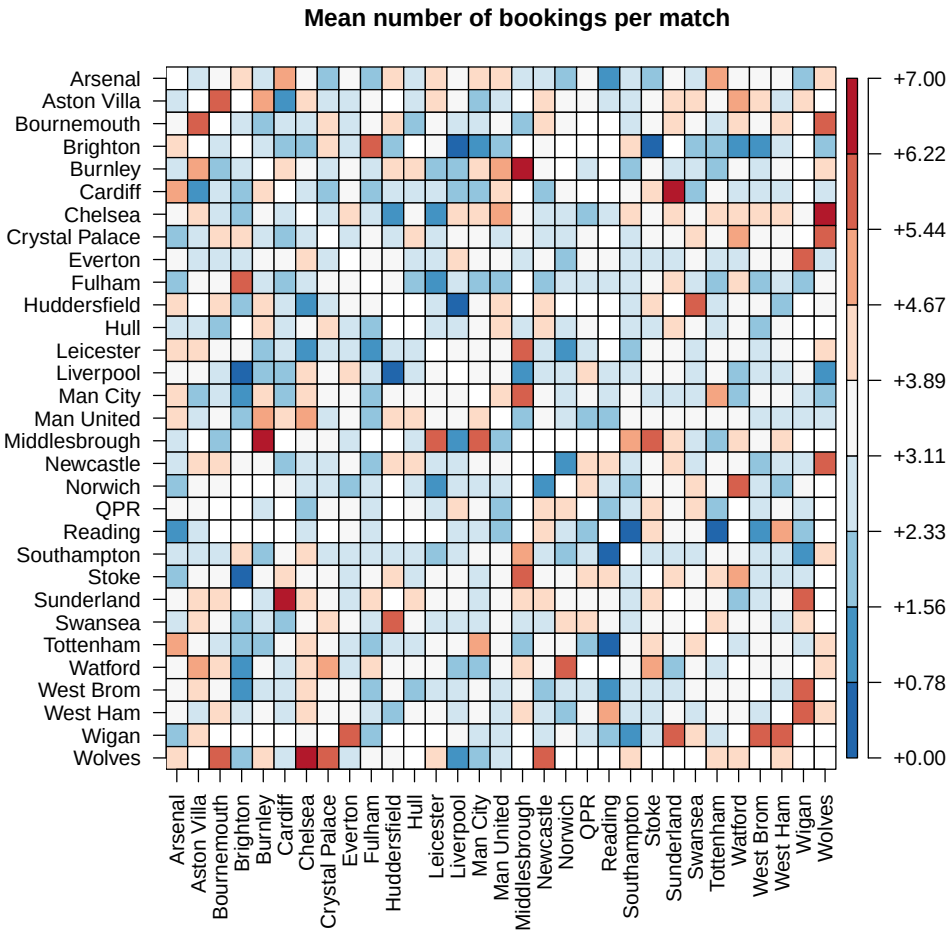
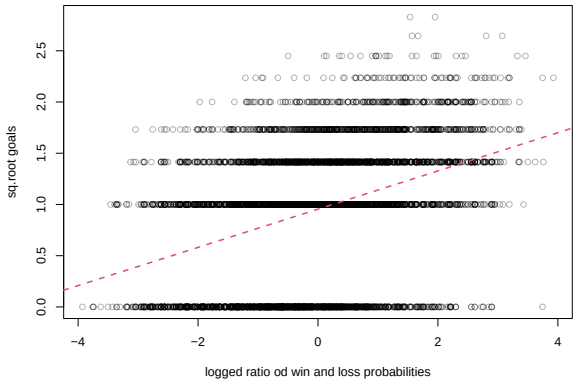
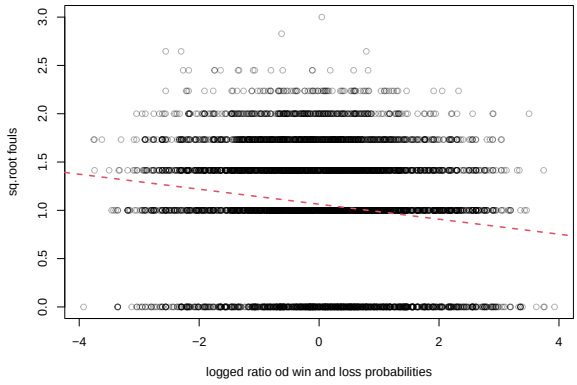


Figure 3: Average booking counts for matches between specific teams. The few pairings that do not occur in the data set are imputed with the mean booking count.

Figure 4: Per-match goal and booking counts for teams with varying win probabilities.

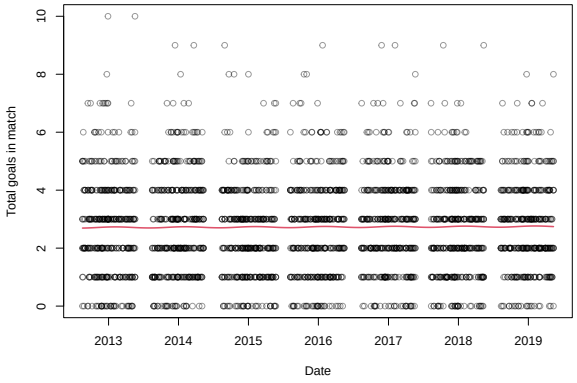


(a) Goals.

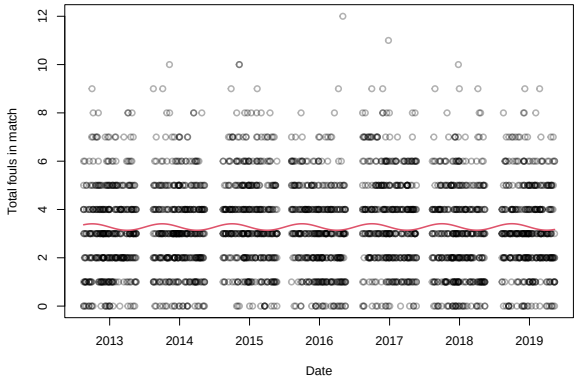


(b) Cards.

Figure 5: Per-match goal and booking counts for both teams over time.

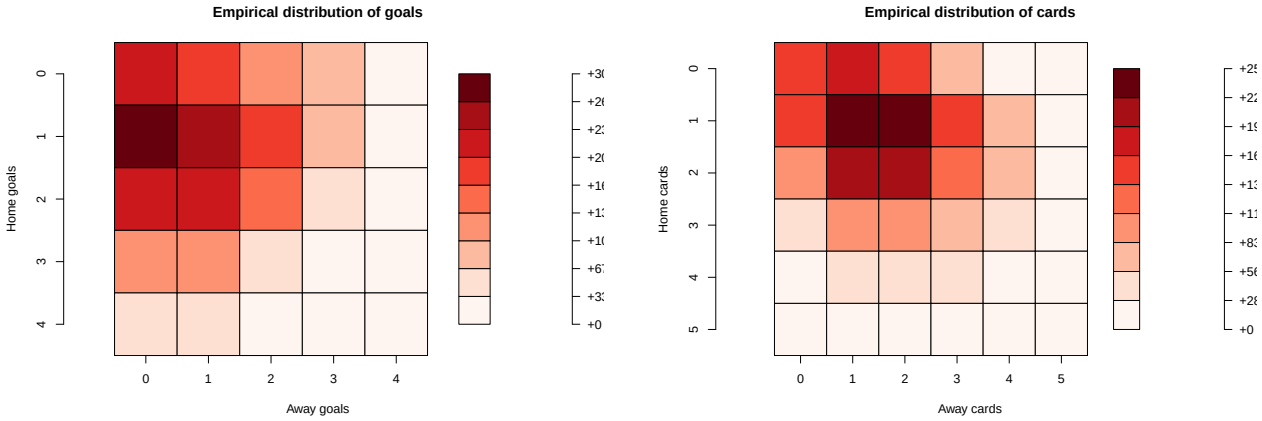


(a) Goals.



(b) Cards.

Figure 6: Counts for matches in which home/away goal and booking counts take different values.



(a) Goals.

(b) Cards.

[2] Bryson, A., Dolton, P., Reade, J. J., Schreyer, D., and Singleton, C. (2021). Causal effects of an absent crowd on performances and refereeing decisions during covid-19. *Economics Letters*, 198:109664.

[3] Buraimo, B., Forrest, D., and Simmons, R. (2010). The 12th man?: refereeing bias in english and german soccer. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(2):431–449.

[4] Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica: Journal of the Econometric Society*, pages 829–844.

[5] Dixon, M. J. and Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):265–280.

[6] Endrich, M. and Gesche, T. (2020). Home-bias in referee decisions: Evidence from "ghost matches" during the covid19-pandemic. *Economics Letters*, 197:109621.

[7] Fan, J. and Gijbels, I. (2018). *Local polynomial modelling and its applications*. Routledge.

[8] Fischer, K. and Haucap, J. (2021). Does crowd support drive the home advantage in professional football? evidence from german ghost games during the covid-19 pandemic. *Journal of Sports Economics*, 22(8):982–1008.

[9] Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(4):757–779.

[10] Karlis, D. and Ntzoufras, I. (2003). Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):381–393.

[11] Kharrat, T., McHale, I. G., and Peña, J. L. (2020). Plus–minus player ratings for soccer. *European Journal of Operational Research*, 283(2):726–736.

[12] Koopman, S. J. and Lit, R. (2015). A dynamic bivariate poisson model for analysing and forecasting match results in the english premier league. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(1):167–186.

[13] Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118.

[14] McCarrick, D., Bilalic, M., Neave, N., and Wolfson, S. (2021). Home advantage during the covid-19 pandemic: Analyses of european football leagues. *Psychology of sport and exercise*, 56:102013.

[15] Owen, A. (2011). Dynamic bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. *IMA Journal of Management Mathematics*, 22(2):99–113.

[16] Owen, A. (2017). The application of hurdle models to accurately model 0-0 draws in predictive models of football match outcomes. In *Proceedings of MathSport International 2017 Conference*, page 295.

[17] Ponzo, M. and Scoppa, V. (2018). Does the home advantage depend on crowd support? evidence from same-stadium derbies. *Journal of Sports Economics*, 19(4):562–582.

[18] Tilp, M. and Thaller, S. (2020). Covid-19 has turned home advantage into home disadvantage in the german soccer bundesliga. *Frontiers in sports and active living*, 2:593499.
[19] Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression models for count data in R. *Journal of Statistical Software*, 27(8).

A Matheuristic for Scheduling Double Round-Robin Sports Tournaments

C. Lamas-Fernandez*, A. Martinez-Sykora**, C.N. Potts***

*Southampton Business School, University of Southampton, UK + email address: C.Lamas-Fernandez@soton.ac.uk
** Southampton Business School, University of Southampton, UK + email address: A.Martinez-Sykora@soton.ac.uk
*** School of Mathematical Sciences, University of Southampton, UK + email address: C.N.Potts@soton.ac.uk

Abstract

Scheduling a double round-robin sports tournament requires assigning matches to time slots. In professional football/soccer, there are various stakeholders such as the football clubs that are concerned about revenue, the league who aim to ensure that the competition is fair, the police who are responsible for safety outside the ground, and TV companies who invest heavily to gain broadcasting rights. The interests of these stakeholders are met by imposing numerous hard and soft constraints on the schedule of matches. An integer linear programming (ILP) model is developed which includes a mixture of hard and soft constraints of the types described above and uses binary variables that assign matches to slots. The solution methodology used is a matheuristic that fixes a large number of variables in the ILP model at each iteration to enable a new solution to be generated relatively quickly. In this fix-and-relax approach, different methods are used to determine which variables are to be fixed. Computational results are provided for the 45 instances that formed the international competition on sports timetabling (ITC2021). These instances have 16, 18 or 20 teams, with 30, 34 or 38 time slots, respectively. The main findings are that the proposed matheuristic finds solutions for most ITC2021 instances relatively quickly with all hard constraints satisfied and generates many best-known solutions for these instances.

1 Introduction

Research on designing algorithms for scheduling sports competitions has taken place for almost fifty years, although interest in this research topic has increased during the last twenty years. For an overview of the major contributions, we refer to Knust [5]. In this paper, we address the scheduling of double round-robin tournaments of the type used in many European and South American football/soccer leagues where each team in the competition plays one home game and one away game against every other team. Goossens and Spieksma [4] provide an overview of the structure of the main football leagues in Europe. As pointed out by Van Bulck et al. [7], most publications focus on designing an algorithm that is suited to the constraints imposed by a particular league. As a consequence, computational work assessing the relative performance of such algorithms is scarce. This has motivated a unified data format for round-robin sports timetabling by Van Bulck et al. [8] that facilitates the comparison of algorithmic approaches. It has also motivated the

International Timetabling Competition on Sports Timetabling, named ITC2021 [9], that "aims to stimulate the development of solvers for the construction of round-robin timetables". This paper reports on the solver developed by the authors for the double round-robin sports scheduling problem (DRRSSP), and provides computational results in the form of objective function values for the 45 instances on which the result of ITC2021 is based.

The double round-robin tournaments of ITC2021 are *compact*, meaning that in each time slot every team has exactly one match. We distinguish between *phased* and *unphased* double round-robin tournaments. In a phased tournament, the matches played in the first half of the slots comprise a single round-robin tournament, and similarly for the matches in the second half of the slots, whereas an unphased tournament has no such constraints on the matches. A *break* occurs if a team plays at home in some slot having played at home in its previous slot (*home break*), or if a team plays away in some slot having played away in its previous slot (*away break*).

There are different classes of constraints on the DRRSSP. These arise due to the interests of multiple stakeholders. The football clubs competing in a tournament are concerned about maximizing revenue by having their home matches being played in slots when more fans are likely to be able to watch the game. They also prefer schedules in which home and away matches alternate, thereby reducing the number of breaks. Teams that use the same stadium clearly create constraints, and other events occurring within the vicinity of the stadium in certain time slots may prevent a home match being assigned to these slots. The police are responsible for safety outside of the ground, which may impose constraints on the number of teams playing in the same city in the same slot. TV companies invest significantly in gain broadcasting rights to the matches, and may prefer schedules having the most high-profile matches spread throughout the season.

This aim of our study is to propose a new algorithm for creating schedules for the DRRSSP, and to evaluate this approach on the 45 instances of the ITC2021 competition. In Section 2, we provide a formal description of the DRRSSP that we are addressing, and Section 3 ourlines our integer programming formulation of the problem. Section 4 describes our proposed matheuristic, which takes the form of a fix-and-relax procedure. Computational results obtained by applying our matheuristic to the 45 instances provided within ITC2021 are presented and discussed in Section 5. Lastly, Section 6 contains some concluding remarks.

2 Problem Description

For the DRRSSP, it is required to design a round-robin tournament that allocates matches to time slots. However, Van Bulck et al. [7] observe from various studies reported in the literature that various constraints can affect how the matches are scheduled.

Let n denote the number of teams competing in the double round-robin tournament, where n is even. Further, let $T = \{1,\ldots,n\}$ be the set of all teams. For each pair of teams $i, j \in T$, where $i < j$, there is match (i, j) for which team i plays a home game against team j and a match (j, i) for which team i plays an away match against team j. Thus, each team plays $n-1$ home games at its own venue and $n-1$ away games at its opponent's venue. The tournament has a set S of time slots for the matches. We assume that the minimum number of time slots is used so that $S = \{1,\ldots,2n-2\}$, which produces a compact tournament. If the matches that are played in slots $1,\ldots,n-1$ define a single round-robin tournament, then the matches in slots $n,\ldots,2n-2$ also define a single round-robin tournament, and such a structure is *phased* tournament. For a phased tournament, we define $S' = \{1,\ldots,n-1\}$ to be the set of slots used for the first phase.

There are structural constraints that ensure that the matches assigned to time slots satisfy the conditions of a double round-robin tournament, with additional constraints added if the tournament is phased. However, there are many other types of constraints within ITC2021, as listed below, which can either be hard or soft.

Capacity constraints: within a given set of time slots, a team is forced to play at home or away, and the total number of matches played by a team or by a set of teams has an upper limit.

Game constraints: given a set of time slots and a set of matches, the number of matches assigned to these time slots has a lower limit and an upper limit.

Break constraints: within a given set of time slots, there is an upper limit on the total number of breaks for a given set of teams.

Fairness constraints: given pairs of teams and a set of slots, there is an upper limit on the largest difference in the number of home games played between each pair of teams at the end of each slot in the set.

Separation constraints: for a given set of teams, there is a lower limit on the gap between home and away matches between each pair of teams in this set.

A solution of the DRRSSP has, for each soft constraint, a non-negative deviation that specifies the number of units of violation of the constraint. This deviation is multiplied by a given weight to provide a penalty for a constraint violation. The objective of the problem is to design a double round-robin tournament in which all hard constraints are satisfied so that the sum of weighted deviations for all soft constraints is minimized.

3 Integer Linear Programming Model

Our integer linear programming (ILP) model has variables and structural constraints that are widely used in round-robin sports scheduling, such as in the study of Durán et al. [3]. The key parameters used in our ILP are a set of teams T and a set of slots S, as defined in Section 2, with S' representing a set containing the first half of the slots.

The variables in our ILP that define the structure of the tournament are

$$x_{ijs} = \begin{cases} 1 & \text{if match } (i,j) \text{ is played in slot } s, \\ 0 & \text{otherwise.} \end{cases} \qquad \forall i,j \in T,\ \forall s \in S,\ i \neq j$$

However, some tournament specifications impose constraints on numbers of breaks that are allowed. Thus, we introduce additional variables

$$b_{is}^{\mathrm{HA}} = \begin{cases} 1 & \text{if team } i \text{ has a home or away break in slot } s, \\ 0 & \text{otherwise.} \end{cases} \qquad \forall i \in T,\ \forall s \in S \setminus \{1\}$$

In some cases, it is necessary to differentiate between a home break and an away break. Thus, if $b_{is}^{\mathrm{HA}} = 1$, then $b_{is}^{\mathrm{H}} = 1$ or $b_{is}^{\mathrm{A}} = 1$ depending on whether a home break or an away break occurs for team i in slot s. Also, there is a relationship $b_{is}^{\mathrm{HA}} = b_{is}^{\mathrm{H}} + b_{is}^{\mathrm{A}}$ for all $i \in T$ and $s \in S \setminus \{1\}$. When there are constraints on the separation in terms of the number of slots between the two matches played by a pair of teams i and j, it is useful to introduce the variables

$$y_{ij} = \begin{cases} 1 & \text{if match } (i,j) \text{ occurs before match } (j,i), \\ 0 & \text{otherwise.} \end{cases} \qquad \forall i,j \in T$$

The following subsections provide the classes of constraints that are included in the model. There are some structural constraints given below in Section 3.1 that must be satisfied, while the remaining constraints are non-structural.

Each non-structural constraint has an index by which it is identified. Associated with most constraints c is a threshold value t_c, which is the maximum value that some linear combination of the x_{ijs}, b^{H}_{is}, b^{A}_{is}, b^{HA}_{is} and y_{ij} variables can achieve without incurring a penalty. In such cases, the right-hand side of the constraint is set to $t_c + d_c$, where d_c is an integer deviation variable for constraint c. The remaining constraints c have a threshold value t_c, which is the minimum value that some linear combination of the x_{ijs}, b^{H}_{is}, b^{A}_{is}, b^{HA}_{is} and t_{ij} variables can achieve without incurring a penalty. For these constraints, the right-hand side of the constraint is set to $t_c - d_c$, where d_c is an integer deviation variable for constraint c. If a constraint c of either type is hard, we introduce $d_c = 0$ as a further constraint. Associated with each soft constraint c is a weight w_c that represents the penalty per unit violation of the constraint. The objective function is to minimize $\sum_c w_c d_c$ where the summation is over all soft constraints c.

3.1 Structural constraints

The structural constraints on a double round-robin tournament ensure that the variables are assigned values that create a valid tournament. The structural constraints are all hard. Constraints (4) below ensure that within each slot s team i either has a home game or an away game against some other team j. Constraints (5) impose the condition that a match (i, j) for each pair of teams i and j appears in exactly one slot. Constraints (6), which are applied only when the DRRSSP is phased, force all pairs of teams i and j to play exactly one match in the first half of the time slots S', and consequently exactly one match in the second half of the time slots $S \setminus S'$.

$$\sum_{j \in T \setminus \{i\}} (x_{ijs} + x_{jis}) = 1 \qquad \forall i \in T,\ \forall s \in S \tag{4}$$

$$\sum_{s \in S} x_{ijs} = 1 \qquad \forall i, j \in T, i \neq j \tag{5}$$

$$\sum_{s \in S'} (x_{ijs} + x_{jis}) = 1 \qquad \forall i, j \in T, i \neq j \tag{6}$$

The constraints linking the b^{H}_{is} and b^{A}_{is} variables with the x_{ijs} variables are

$$\sum_{j \in T} (x_{ijs} + x_{i,j,s-1)}) \leq b^{\mathrm{H}}_{is} + 1 \quad \sum_{j \in T} (x_{jis} + x_{j,i,s-1}) \leq b^{\mathrm{A}}_{is} + 1 \quad \forall i \in T, s \in S \setminus \{1\} \tag{7}$$

Further, the constraints linking y_{ij} with the x_{ijs} variables are

$$\sum_{s \in S} s(x_{jis} - x_{ijs}) \leq M y_{ij} \qquad \sum_{s \in S} s(x_{ijs} - x_{jis}) \leq M(1 - y_{ij}) \quad \forall i, j \in T, i \neq j \tag{8}$$

where M is a constant that satisfies the condition $M \geq |S| - 1$.

3.2 Other constraints

The non-structural constraints are typically a combination of hard and soft constraints. Each such constraint has an index by which it is identified. Associated with most constraints c is a threshold value t_c, which

is the maximum value that some linear combination of the x_{ijs}, b^{H}_{is}, b^{A}_{is}, b^{HA}_{is} and y_{ij} variables can achieve without incurring a penalty. In such cases, the right-hand side of the constraint is set to $t_c + d_c$, where d_c is a non-negative integer deviation variable for constraint c that measures the number of units by which the threshold is violated. The remaining constraints c have a threshold value t_c, which is the minimum value that some linear combination of the x_{ijs}, b^{H}_{is}, b^{A}_{is}, b^{HA}_{is} and t_{ij} variables can achieve without incurring a penalty. For these constraints, the right-hand side of the constraint is set to $t_c - d_c$, where d_c is an non-negative integer deviation variable for constraint c. A solution is feasible if $d_c = 0$ for all the hard constraints.

For capacity constraints and game constraints , the left-hand side of each constraint c is the sum of selected x_{ijs} variables. Break constraints are of three types, namely home, away and home-away. The left-hand side of these constraints is the sum of selected b^{H}_{is}, b^{A}_{is} and b^{HA}_{is} variables, respectively. Fairness constraints, which are designed to ensure that pairs of teams have played approximately the same number of home games at the end of selected slots, have a left-hand side comprising the sum of $x_{ijs} - x_{i'js}$ for selected pairs of teams teams i and i'. Lastly, separation constraints, which are imposed to ensure that the two matches (i, j) and (j, i) between selected pairs of teams i and j are not too close together, have a left-hand side of $\sum_s s(x_{ijs} - x_{jis}) + My_{ij}$. We refer to Lamas-Fernandez et al. [6] for an explicit statement of these constraints.

4 Matheuristic

Our matheuristic is based on the ILP model for the DRRSSP outlined in Section 3. The instances created for the ITC2021 competition are sufficiently challenging that, with only the hard constraints considered, our ILP does not find a feasible solution to any instances when using the Gurobi solver (version 9.0) and allocating reasonable computational resources. Thus, our matheuristic provides a framework for producing solutions for ILPs by solving a series of smaller problems. Each of these smaller problems has many variables fixed, thus leaving a relatively small number of decision variables for the ILP optimizer. Thus, we adopt a fix-and-relax (also known as relax-and-fix) approach that has been successful for related problems; see de Oliveira et al. [1], for example.

Our matheuristic involves two stages, where the first stage aims at finding a feasible solution that satisfies all hard constraints, while ignoring all soft constraints, and the second stage uses the first-stage solution as input and considers all constraints with a view to minimizing the total penalty attributed to soft constraint violations. The search mechanisms used in the two stages are variable neighbourhood search (VNS) and variable neighbourhood descent (VND), respectively.

4.1 First stage: using VNS to find a feasible solution

Initially, we find a feasible solution to the ILP but with all constraints removed except for the structural constraints. We then include the remaining hard constraints and use an objective function that sums the values of the deviation variables d_c for each violated hard constraint. The remainder of first stage is based on a fix-and-relax search that uses following five neighbourhoods to select the variables to be fixed. We denote by $\hat{x}_{ijs}$ the value of x_{ijs} in the current solution for all $i, j \in T$ with $i \neq j$, and $s \in S$.

N1: *Slots.* The neighbourhood selects a subset $\bar{S} \subseteq S$ of slots, where $|\bar{S}| = n_1$ and n_1 is a parameter, and then fixes $x_{ijs} = \hat{x}_{ijs}$ and $x_{jis} = \hat{x}_{jis}$ for all $i, j \in T$ and $s \in S \setminus \bar{S}$.

N2: *Teams.* The neighbourhood selects a subset $\bar{T} \subseteq T$ of teams, where $|\bar{T}| = n_2$ and n_2 is a parameter, and then fixes $x_{ijs} = \hat{x}_{ijs}$ and $x_{jis} = \hat{x}_{jis}$ for all $i \in T$, $j \in T \setminus \bar{T}$ and $s \in S$.

N3: *Slots & Teams.* The neighbourhood selects a subset $\bar{S} \subseteq S$ of slots, where $|\bar{S}| = n_1/2$, and a subset $\bar{T} \subseteq T$ of teams, where $|\bar{T}| = 2$, and then fixes $x_{ijs} = \hat{x}_{ijs}$, $\forall i, j \in T \setminus \bar{T},\ s \in S \setminus \bar{S}$.

N4: *Phased* (only for phased tournaments). We fix the x_{ijs} variables one half of the competition and then optimise the other half.

N5: *Home and away*. In this neighbourhood, we allow home and away matches between the same pair of teams to be swapped.

In N1, N2 and N3, slots to include in $\bar{S}$ and teams to include in $\bar{T}$ are chosen either randomly or by considering the constraints most heavily violated.

We apply the five neighbourhoods N1-N5 until no improvement is found. If there is no improvement and the solution is still infeasible, then we increase the weight of the coefficient in the objective function for each hard constraints that is currently violated.

4.2 Second stage: optimizing with soft constraints

In this stage, we consider the full ILP model and the input is the final solution obtained in the first stage. Then we reset all the weights on the objective function as follows. The weight on the d_c variables of the soft constraints is given by the penalty of the constraints, and the weight of all the other d_c variables corresponding to hard constraints is set to 100. We identified that this value is sufficiently large to converge to better solutions without violating too many hard constraints, which may lead to difficulties in recovering feasibility. Next we apply the N1-N5 neighbourhoods until we obtain a local optimum, in which case we increase the two parameters of the algorithm, n_1 and n_2 (see Section 4.1) by 1, until Gurobi cannot solve the model efficiently. In some instances where there are not many constraints, the resulting ILP models are easier to solve and, therefore, higher values for n_1 and n_2 are explored (up to $n_1 = 30$ and $n_2 = 14$ in the best cases). For the larger or more complex instances resulting in more challenging ILPs, the model becomes intractable with $n_1 = 18$ and $n_2 = 8$. We execute multiple runs (n_s) with the same initial solution (by introducing randomness in N1-N3) or by different initial solutions obtained by the VNS algorithm of the first stage.

5 Computational experiments

For the computational experiments we have used 2.6GHz Intel Sandybridge processors, and each run was performed by 4 CPUs with 16GB of memory. We have used Gurobi (version 9.0) to solve the ILP models and we run the algorithm three times with 30, 60 and 600 second per model. In the first stage of the algorithm (VNS), a run continues until a feasible solution is found. The instances created for ITC2021 comprise 15 "Early", 15 "Middle" and 15 "Late" instances, which were released at different times during the competition.

The computation time needed to find a feasible solution with the first stage of the algorithm (Section 4.1) is less than 1 hour for 37 instances, between 1 and 24 hours for 6 instances and between 24 and 100 hours for 2 instances (Early 10 and Middle 2).

In the second stage of the algorithm (Section 4.2), for each run of the algorithm we set $n_s = 60$ (multi-start runs), and we initially set $n_1 = 8$ and $n_2 = 6$. The computation time of one single run of the algorithm strongly depends on the instance that we are solving, but the largest amount of time on a single run was up to 6 days. Our best solutions obtained are reported in Table 1. The VNS/VND columns list the objective function vales for our proposed matheurstic and the ITC2021 columns list the best objective functions found within the ITC2021 competition.

At the end of the competition, the proposed algorithm obtained the best known solution for 22 of the 45 instances, and for another 7 instances the best known solution is within 6% of the solution obtained by the proposed algorithm.

Table 1: Computional results of the ITC2021 instances

	Early		Middle		Late	
Instance	*VNS/VND*	*ITC2021*	*VNS/VND*	*ITC2021*	*VNS/VND*	*ITC2021*
1	362	362	5177	5177	1969	1969
2	222	160	7381	7381	5400	5400
3	1052	1012	9800	9701	2369	2369
4	536	512	7	7	0	0
5	3127	3127	494	413	2218	1939
6	3714	3352	1275	1125	923	923
7	4763	4763	2049	1784	1652	1558
8	1114	1114	129	129	934	934
9	108	108	450	450	563	563
10	3400	3400	1250	1250	2031	1988
11	4436	4436	2608	2511	226	207
12	510	380	923	911	3912	3689
13	121	121	282	253	2110	1820
14	47	4	1323	1172	1363	1206
15	3368	3368	965	495	40	20

The proposed algorithm obtained the best results in 22 out of 45 instances by the end of competition, and in further 7 instances the best-known result is within 6% of the solution obtained by the proposed algorithm.

6 Concluding remarks

In this study, we propose a novel matheuristic algorithm having two stages to solve the DRRSSP. In the first stage, we address the feasibility problem by using a VNS framework, while a second stage optimizes the soft constraint violations using a multi-start algorithm within a VND framework. Both the VNS and the VND both use the same combination of five different neighbourhoods. The results obtained show that both stages of the algorithm perform well, being able to prove optimality for one instance (Late 4, with an objective value of 0), finding a feasible solution for all 45 instances of the ITC2021 competition and producing best-known solutions for 22 of the 45 instances.

Acknowledgements

The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, during this study.

References

[1] de Oliveira L, de Souza CC, Yunes T: Improved bounds for the traveling umpire problem: A stronger formulation and a relax-and-fix heuristic. *European Journal of Operational Research* **236** 592–60 (2014).

[2] Dillenberger C, Escudero LF, Wollensak A, Zhang W: On practical resource allocation for production planning and scheduling with period overlapping setups. *European Journal of Operational Research* **75** 275-286 (1994).

[3] Durán G, Guajardo M, Miranda J, Sauré D, Souyris S, Weintraub A, Wolf R: Scheduling the Chilean soccer league by integer programming. *Interfaces* **37**(6) 539–552 (2007).

[4] Goossens DR, Spieksma FCR: Soccer schedules in Europe: An overview. *Journal of Scheduling* **15** 641-651 (2011).

[5] Knust, S: Classification of sports scheduling literature. http://www2.informatik.uni-osnabrueck.de/knust/sportssched/sportlit_class/.

[6] Lamas-Fernandez C, Martinez-Sykora, Potts CN: Scheduling double round-robin sports tournaments. *Processings of the 13th International Conference on the Practice and Theory of Automated Timetabling - PATAT 2021, Volume II* pp. 435-448 (2021).

[7] Van Bulck D, Goossens D, Schönberger J, Guajardo M: RobinX: A three-field classification and unified data format for round-robin sports timetabling. *European Journal of Operational Research* **280**(2), 568-580 (2020).

[8] Van Bulck D, Goossens D, Beliën, J, Davari, M: ITC—Sports Timetabling Problem Description and File Format. https://www.sportscheduling.ugent.be/ITC2021/ (2021).

[9] Van Bulck D, Goossens D, Beliën, J, Davari, M: The Fifth International Timetabling Competition (ITC 2021): Sports Timetabling. *Proceedings of MathSport International 2021 Conference,* pp. 117-122 (2021).

Break minimisation in sports timetabling using modern SAT solvers

Martin Mariusz Lester

Department of Computer Science, University of Reading, United Kingdom: m.lester@reading.ac.uk

Abstract

We investigate break minimisation in the context of scheduling time-constrained double round robin (2RR) sports timetables using SAT-based solvers. While SAT-based approaches to sports timetabling have been considered in the past, most previous work predates the modern era of SAT solving. This raises several questions. Are modern solvers are more effective at handling break minimisation problems? Does this depend significantly on details of the encoding used? How do they compare with more popular integer programming-based techniques? We answer these questions empirically by benchmarking a range of solvers on a range of problem instances with different sizes and encodings.

1 Introduction

The *time-constrained double round robin* (2RR) is a common format of sports league or tournament. In this format, each of n teams plays every other team exactly twice: once *at home* and once *away*. Furthermore, the timetable is split into time *slots*, with each team playing one game in every slot. Thus there are necessarily $2(n-1)$ slots.

There are many constraints that a league organiser may wish a timetable to satisfy. Of these, one of the most common is to minimise the number of *breaks*. A break is when the same team plays two consecutive games at home (a *home break*) or two consecutive games away (an *away break*). Breaks are considered undesirable because they affect attendance by spectators and because long periods playing away may disadvantage a team.

The minimum number of breaks in a 2RR with n teams is $2(n-2) = 2n-4$. Although there are designs that achieve this minimum number of breaks in the absence of other constraints, real-world sports timetabling problems often include a variety of heterogeneous constraints. There are several popular techniques for constructing timetables satisfying these constraints, including local search and mixed integer programming (MIP). Some of these approaches attempt to solve the problem *monolithically*, all in one go; others decompose the problem into several stages. Rasmussen and Trick [4] give a detailed survey of approaches to break minimisation specifically.

Extending a partial timetable while minimising breaks (or equivalently accommodating various other constraints) is NP-complete [3]. Practically relevant NP-complete problems are often solvable quickly by translating them into SAT and using SAT solvers. It would make sense to apply the same approach here. Our own prior work [2] shows that, while SAT-based approaches are often effective at finding feasible solutions to timetabling problems quickly, they struggle with break minimisation. This motivates our investigation into whether choice of SAT solver and details of the problem encoding affect break minimisation.

When reviewing the literature, apart from our own work, we found only two significant studies investigating use of SAT for sports timetabling. Firstly, Zhang et al. [5] describe using a SAT solver extended to handle cardinality constraints. From a modern perspective, this is actually an early pseudo-Boolean (PB) solver, but predates the first PB solver competition in 2005. (PB extends SAT with integer equality and inequality constraints over Boolean variables, interpreted as 0 when false and 1 when true.) Secondly, Horbach et al. [1] describe using a custom incremental SAT solving process to solve an optimisation problem. (The API for incremental SAT solving was not standardised and included in the SAT Competition until 2016.) Both these works propose adding extra constraints to improve break minimisation. As SAT solvers have advanced massively over the past 20 years, so we thought it worthwhile to review both of these variations with modern solvers.

2 Problem encoding

2.1 Baseline

Before we can investigate how changes in encoding or solver affect break minimisation, we need a baseline against which to compare. For our baseline encoding, we pick the default PB encoding from our existing timetabling tool *Reprobate* Lester [2].

Let us consider a timetable for n teams (numbered 0 to $n-1$) and $2n-2$ slots (numbered 0 to $2n-3$). We use t, t_1 and t_2 to refer to teams and s to refer to slots. These indices are implicitly quantified over these ranges, unless otherwise specified. We introduce variables $M_{t_1,t_2,s}$, true just if team t_1 plays home against team t_2 in slot s. This scheme is also used by both Zhang et al. [5] and Horbach et al. [1].

We introduce feasibility constraints on the M variables, which enforce that a team must play exactly once in every slot:

$$\forall s, t_1 . \sum_{t_2} (M_{t_1,t_2,s} + M_{t_2,t_1,s}) = 1 \tag{9}$$

and that any pair of teams t_1 and t_2 must play twice, once with t_1 at home and once with t_2 at home:

$$\forall t_1, t_2 . \sum_{s} M_{t_1,t_2,s} = 1 \tag{10}$$

To track the number of breaks, we introduce variables $H_{t,s}$ and $B_{t,s}$. $H_{t,s}$ is true just if team t plays at home in slot s:

$$\begin{aligned} &\forall s, t_1, t_2 . -M_{t_1,t_2,s} + H_{t_1,s} \geq 0 \\ &\forall s, t_1, t_2 . -M_{t_1,t_2,s} + -H_{t_2,s} \geq -1 \end{aligned} \tag{11}$$

Then, $B_{t,s}$ is true, indicating a team has a break, if it plays either two consecutive home games or two consecutive away games:

$$\begin{aligned} &\forall s > 0, t . B_{t,s} + -H_{t,s-1} + -H_{t,s} \geq -1 \\ &\forall s > 0, t . B_{t,s} + H_{t,s-1} + H_{t,s} \geq 1 \end{aligned} \tag{12}$$

Our objective is to minimise the total number of breaks, $\sum_{t,s} B_{t,s}$. This is fine if we are working with a solver that supports optimisation problems, but SAT solvers and some PB solvers only handle decision problems. In this case, we need to pick a target bound k, which the number of breaks must not exceed:

$$\sum_{t,s} -B_{t,s} \geq -k \tag{13}$$

2.2 No double break

In addition to our baseline encoding, we consider two sets of constraints proposed specifically for minimising breaks in previous work. Adding extra constraints to a SAT encoding of a problem can help a solver by forcing a conflict more quickly. Zhang et al. [5] suggest what we call a "no double break" restriction, forbidding *each* team from playing three consecutive games at home, or three consecutive games away. As their encoding does not have H variables, they encode the constraints directly on the M variables. Our encoding is much more succinct:

$$\begin{aligned} &\forall s \leq 2n-3, t. -H_{t,s} + -H_{t,s+1} + -H_{t,s+2} \geq -2 \\ &\forall s \leq 2n-3, t. H_{t,s} + H_{t,s+1} + H_{t,s+2} \geq 1 \end{aligned} \tag{14}$$

Note also that, expressed in terms of H, these are "at least one" constraints, which are expressible in SAT using a single clause.

2.3 No triple break period

Next, we implement a constraint from Horbach et al. [1], which we call "no triple break period". This introduces new variables P_s to track whether slot s is a break period (meaning that *any* team has a break in s) and requires that there cannot be three consecutive break periods.

$$\begin{aligned} &\forall s > 0, t. -B_{t,s} + P_s \geq 0 \\ &\forall s > 0, -P_s + \textstyle\sum_t B_{t,s} \geq 0 \\ &\forall s \in [1, 2n-5]. -P_s + -P_{s+1} + -P_{s+2} \geq -2 \end{aligned} \tag{15}$$

Clearly, this constraint is much tighter than "no double break". Both constraints are sound for the abstract problems we consider: they do not prevent the minimum number of breaks from being achieved. In a real-world application, there are other constraints to consider, which may conflict with these heuristic constraints. Nonetheless, they may still prove helpful in reducing the number of breaks.

3 SAT benchmarking

We now investigate the degree to which choice of encoding and solver affect break minimisation when used with a pure SAT solver. Because our baseline encoding uses PB constraints, not SAT constraints, we use *pbencoder* from *PBLib* to convert PB constraints into SAT. For our target bound on the number of breaks, we pick $k = c(n-2)$ with $c \in [2,5]$. We consider problems with up to $n = 20$ teams.

For our baseline solver, we pick the ubiquitous solver *MiniSat*, which won several prizes in the 2005 and 2007 SAT Competitions. For our more modern solvers, we use recent versions of *Kissat*, *CaDiCaL*, *CryptoMiniSat*, *lstech_maple* and *clasp*, which won SAT Competition prizes between 2009 and 2021.

Tables 1–2 summarise some of our findings. All results were generated on a machine running Debian Linux 10 with a 3.4 GHz Intel Core i5-7500 CPU and 64 GB of RAM. Each solver was run on a single core with a timeout of 3600 s (1 hour). Table 1 shows the best solver/encoding combination for different

Teams	Breaks	c	Solver	Encoding	Time (s)
12	20	2	clasp	base	19
14	24	2	cadical	base	21
16	28	2	crypto	base	104
18	32	2	crypto	double	485
20	54	3	crypto	double	1920

Table 1: Best results with SAT solvers. For a fixed number of teams, the table lists the solver/encoding pair with the tightest bound on number of breaks (c, with $c = 2$ optimal); ties were broken by time taken.

Teams	base	double	triple
12	0	0	0
14	0	1	0
16	7	3	5
18	14	12	19
20	20	17	22

Solver	base	double	triple
cadical	14	11	12
clasp	7	9	8
crypto	13	15	11
kissat	11	11	9
lstech	10	12	14
minisat	6	10	6

Table 2: *Left:* Number of timeouts (out of 24) as encoding and teams vary, for all solvers and $c \in [2,5]$. *Right:* Number of fast (under 600 s) solves (out of 20) for each solver/encoding combination, with $12 \leq t \leq 20$ and $c \in [2,5]$.

numbers of teams, where "best" means tightest bound on number of breaks and least time to solve. The tightest possible bound is achievable up to 18 teams. Although there is significant variation between instances, the combination of *CryptoMiniSat* and "no double break" seems to perform best. Table 2 (left) shows how the number of timeouts varies with number of teams and encoding. The most reliable encoding seems to be "no double break". In general, solvers struggle with 18–20 teams. Figure 1 shows each solver for the baseline encoding with 16 teams, where the number of timeouts is beginning to become significant. As one might expect, relaxing the bound on number of breaks tends to decrease solve time, but this is not always true. Table 2 (right) shows how many problem instances were solvable quickly, in under 600 s (10 minutes), with different solvers/encodings. Again, *CryptoMiniSat* and "no double break" seem to perform best.

4 PB and MIP optimisation benchmarking

Next we investigate the effectiveness of PB solvers for break minimisation, presented as an optimisation problem. There are two changes from the SAT approach here. Firstly, some PB solvers can use cutting planes reasoning on the cardinality constraints in our encoding, which is sometimes more effective than translating the constraint into (for example) an adder circuit or a sorting network. Secondly, the solvers we consider can solve optimisation problems as well as decision problems, which removes the need for us to specify an explicit target bound; this is useful in real-world timetabling problems, where we often want the best solution according to some metric, but do not know what the "best" is.

We benchmark the solvers *clasp*, *UWrMaxSat* (with *CaDiCaL* as the underlying SAT solver) and *Round-*

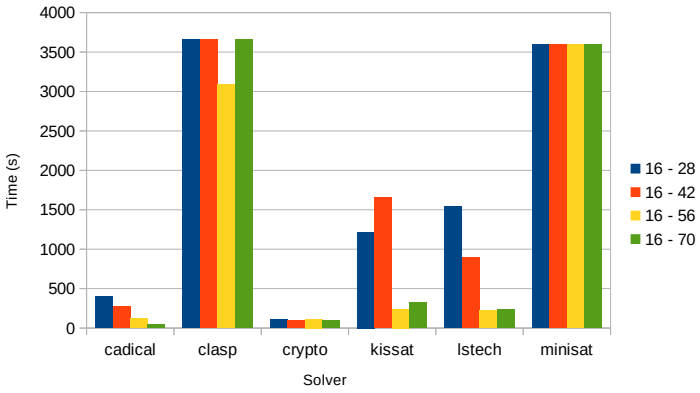


Figure 1: An indicative graph of times for different solvers and bounds, here with base encoding, 16 teams and 28–70 breaks. Note increasing bound usually decreases solving time, but not always.

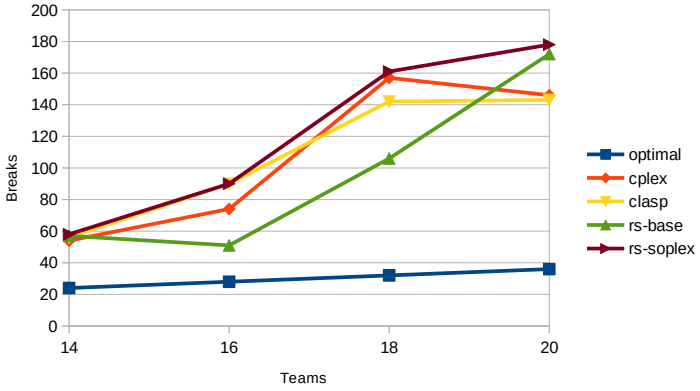


Figure 2: Smallest number of breaks found with baseline encoding for 14–20 teams by different PB solvers and CPLEX. Optimal number of breaks included for comparison.

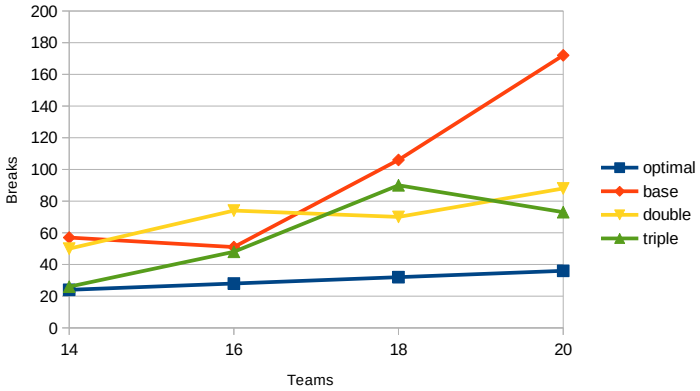


Figure 3: Smallest number of breaks found with different encodings for 14–20 teams by *RoundingSat* (without *SoPlex*).

ingSat (both with and without integration of the *SoPlex* linear programming solver). We also consider the commercial MIP solver CPLEX on an equivalent formulation of our problem. We run each solver with 14–20 teams on our baseline encoding and two variations.

Figure 2 shows the best bound achieved in 1 hour by each solver with our baseline encoding. As *UWrMaxSat* solved only two instances, with poor scores on both, we omit it from our results. In general, the PB solvers performed better with the "no double break" constraint and best with the "no triple break period" constraint. (CPLEX performed much worse with these extra constraints.) Figure 3 shows the variation for *RoundingSat* (without *SoPlex*), which was best overall, but failed to achieve the tight bounds seen with the SAT solvers.

5 Conclusion

The best SAT-based results for break minimisation come from using a SAT solver directly and explicitly specifying a target bound. The "no double break" constraint seems to be an effective heuristic for this purpose, especially when paired with the solver *CryptoMiniSat*. The popular but aging solver *MiniSat* performed badly when compared with more modern solvers. Using a PB solver to treat break minimisation as an optimisation problem gives worse (but still acceptable results). In this case, the more restrictive "no triple break period" seems more effective with most solvers, especially when paired with *RoundingSat* (without integration of the *SoPlex* linear programming solver). The commercial MIP solver CPLEX gave acceptable results with our base encoding, but (for our resource limits) was less effective than *RoundingSat*. It performed much worse with both encoding variations.

The SAT-based approach can achieve an optimal bound for 18 teams, but not 20. For timetabling problems above this size, other approaches, such as techniques based on local search, are more likely to yield good solutions. We have submitted some instances of break minimisation problems to the SAT Competition 2022, which we hope will motivate solver developers to investigate this problem further.

References

[1] Andrei Horbach, Thomas Bartsch, and Dirk Briskorn. Using a SAT-solver to schedule sports leagues. *J. Sched.*, 15(1):117–125, 2012. URL `https://doi.org/10.1007/s10951-010-0194-9`.

[2] Martin Mariusz Lester. Pseudo-boolean optimisation for RobinX sports timetabling. *J. Sched.*, 2022. URL `https://doi.org/10.1007/s10951-022-00737-7`.

[3] Gerhard F. Post and Gerhard J. Woeginger. Sports tournaments, home-away assignments, and the break minimization problem. *Discret. Optim.*, 3(2):165–173, 2006. URL `https://doi.org/10.1016/j.disopt.2005.08.009`.

[4] Rasmus V. Rasmussen and Michael A. Trick. Round robin scheduling - a survey. *Eur. J. Oper. Res.*, 188 (3):617–636, 2008. URL `https://doi.org/10.1016/j.ejor.2007.05.046`.

[5] Hantao Zhang, Dapeng Li, and Haiou Shen. A SAT based scheduler for tournament schedules. In *SAT 2004 - The Seventh International Conference on Theory and Applications of Satisfiability Testing*, 2004. URL `http://www.satisfiability.org/SAT04/programme/74.pdf`.

Rescheduling the NBA regular season via Integer Programming

J.J. Miranda Bront* and N. García Aramouni**

*Universidad Torcuato Di Tella, Buenos Aires, Argentina + email address: jmiranda@utdt.edu

** Consejo Nacional de Investigaciones Científicas y Técnicas,
Argentina + email address: nico.garcia.ara@gmail.com

Abstract

The COVID-19 pandemic generated disruptions across multiple sectors, in particular in the sports industry as many leagues had to re-adapt their competitions as a consequence of lock-downs, travel restrictions and other implemented safety measures. Even when activities started to resume, match suspensions continued throughout the different leagues. Dealing effectively with these unexpected events is extremely relevant regarding both sports and economic aspects of the competitions. In this paper, we propose a framework that builds upon Integer Programming models to systematically reschedule suspended games and generate a contingency fixture that accounts for relevant operational constraints. Using the 2020-21 NBA season as a benchmark, we compare different scheduling approaches under two different objective functions. Computational results show that the approach produces good quality schedules and that the framework has potential to be applied in practice.

1 Introduction

The COVID-19 pandemic has been one of the most challenging events the humanity faced in the last decades. This unexpected event stressed the nationals' health systems across the world, generating millions of cases and deaths, as well as the world economy, that was affected the health and safety protocols implemented. Estimations indicate that global GDP decreased 3.3% in 2020 [1] (as a point of comparison, this number was 1.3% during the global crisis of 2009), and that global unemployment rose from 5.4% to 6.6% in 2020 [2]. The sports industry was not the exception, as most major tournaments undergone an initial indefinite interruption during March-April 2020. Later on, some competitions started to resume their activities with new formats, and generally, without fans on the stadiums. For instance, the uncertainty introduced by the pandemic motivated some tournaments to be played in a "bubble" format to finish the 2019-20 season. where the teams played all games in a central location (including mobility restrictions for players as well) to prevent the teams from traveling among different locations. This is the case of multiple leagues across different sports, such as football (UCL), basketball (NBA) and ice hockey (NHL).

Despite the lack of attendance to the stadiums, the sports leagues faced important challenges throughout this period. Players, coaches and the management could still get sick and sidelined due to the safety protocols. Depending on the number of players affected, some matches could get suspended. This situation generated

new challenges for league organizers, as a large number of suspensions would affect the execution of the schedule. In pre-pandemic contexts, suspensions were rare and unrelated to each other (generally related to a specific reason). On the contrary, during the COVID19 pandemic infected people required several days to recover and clear the safety protocols. In this scenario, it is likely that multiple consecutive matches might get suspended. This is specially problematic for time-relaxed schedules, where generally multiple matches per week are played. For the 2020-21 NBA season, the probability of a team's game being rescheduled given that the previous one was rescheduled was 33.8%. To reduce the impact, different actions were considered. For instance, the NBA allowed teams to sign 10-day contracts with free agents, that acted as temporary replacements for players entering the protocols. If necessary, matches were also suspended and rescheduled to be played later in the season. These contingency plans affect the planning, but also impact the competition.

To our knowledge, the literature on reschedudling strategies in sports competitions is rather scarce. Yi et. al. [3] present both proactive and reactive strategies to reschedule football matches in a pre-pandemic scenario for time-constrained tournaments. However, we are not aware of research involving rescheduling in a time-relaxed setup. Based on this preliminary analysis, in this paper we explore different strategies to systematically adapt time-relaxed schedules when affected by games suspensions due to unexpected or uncontrolled events during its execution. More specifically, we propose two alternative strategies, that build on integer linear programming models to reschedule suspended matches within the original, planned schedule respecting different rules, such as the number of consecutive games, the distance travelled, among others. We consider two different potential objective functions that might be considered by league organizers, and compare the alternative rescheduling strategies using different quality metrics through computational experiments. We concentrate on the 2020-21 NBA season to explore our models, although our approach could fit in other contexts.

2 Problem definition

In this section we provide a high level description of the key rules for our approach. We assume the tournament has an *original schedule* for the season, indicating the planned date and location for each game throughout the planning horizon. During the execution, due to either unexpected or uncontrollable events, some games need to be suspended from their scheduled date and postponed to be played later in the season. We further assume that all suspended games need to be played to conclude the season, eventually after the final date in the original schedule. Indeed, a somehow straightforward alternative is to wait until the end of the season and reschedule all the suspended games afterwards, assuming there is enough time available before the next stages of the competition (e.g., playoffs). Such an approach, however, may affect some specific teams due to fairness issues, specially considering that their final standings would by affected by the fact that a large number of rescheduled games should be played in a short period of time and, potentially, right before decisive stages of the competition.

We formulate the problem in a more general fashion. Assume the schedule has been executed up to some specific time in the original planning horizon (eventually, the end of the season), and that some games have been suspended. Intuitively, the (sports) timetable rescheduling problem involves finding feasible dates for the cancelled games in the rest of the season. In this work, we aim to insert the suspended games into the schedule without modifying the game dates for matches from the original schedule.

Following the definitions introduced in [3] for time-constrained tournaments, also present in other

rescheduling contexts, each game in the executed timetable that is played before or after its scheduled round is considered a *disruption*. In our setup, each suspended game will therefore translate into a disruption that needs to be rescheduled in the remaining of the season schedule. Then, the challenge is to find feasible rounds to reschedule each disruption, obtaining a feasible rescheduled timetable.

Given a disrupted game, a date t (round) is considered a candidate if the following conditions hold:

1. t occurs after the disruption's date in the original schedule;
2. there are no planned games for the teams involved in the disruption on date t;
3. there are no scheduling rules violations if the corresponding match is scheduled on t; and
4. both teams travel a *reasonable* distance if the corresponding match is scheduled on t.

Even in a time-relaxed system, there are several rules that are relevant in order to create a reasonable and fair schedule. Condition 3 goes in that direction. For example, we impose that no team should have games on three consecutive days to make the schedule realistic regarding rest times between games. Therefore, if a team has matches on Match 13^{th} and 14^{th}, no disrupted game should be rescheduled on March 15^{th}. We carry out this analysis, evaluating the number of matches (both home, away and in general) that are being held on a set of consecutive 1 to 7 days. Another relevant aspect involves considering the total distance travelled, one of the most common metrics used to evaluate sports schedules. For example, if a match held in Los Angeles has to be rescheduled (let's say between the Lakers and the Rockets), it could make sense to avoid using a date during a tour in the east coast, for instance after a game in New York and before a game in Boston. Thus, we prioritize dates that generate traveled distances closer to the ones incurred in the case the match was not suspended. Condition 4 incorporates this aspect to obtain a new fixture having reasonable tours.

We note that there are situation where some disruptions may not be rescheduled within the rest of the season even when these sets contain feasible dates. In that case, we determine that match would be played after the last game of each team, respecting that no franchise should play on three consecutive days, and potentially adding new dates to the competition if needed, in order to play these games.

3 Mathematical Model

We first introduce some basic definitions and notation used to model the problem. Let $S = \{1,2,\ldots,m\}$ be the set of teams (in our case, NBA teams), and $T = \{1,2,\ldots,r\}$ the set of original tournament days (rounds). Let (j,k) denote a match between teams $j,k \in S$, and we further represent a *scheduled game* by a pair $\alpha =< (j,k),t >$ indicating that game (j,k) takes place on day $t \in T$. With these definitions, the schedule for the season can be modelled as a set R of games.

Given a team $i \in S$, we define $R_i, R_i^{\text{dis}} \subseteq R$ as the set of scheduled games (i.e., do not need to be rescheduled) and the set of original disrupted matches for team i (i.e., that need to be rescheduled), respectively. Let $R^{\text{dis}} = \cup_{i \in S} R_i^{\text{dis}}$ denote the set of all the disrupted matches, including their original date, to be rescheduled. For every disruption game α, let T_α^{free} as the set of potential candidate dates in Ts to reschedule match α. In order to incorporate the conditions imposed to the schedule, given $t_1, t_2 \in T$, $t_1 < t_2$, let MG_{t_1,t_2} indicate the maximum number of games a team can play within every window of $t_2 - t_1$ days. In our case, we consider sliding windows of $1 \leq t_2 - t_1 \leq 7$ days, each of them with value corresponding value for MG_{t_1,t_2}. Finally, let

$k^i_{t_1,t_2}$ denote the number of games originally scheduled for team $i \in S$ between t_1 and t_2. For each disruption $\alpha \in R^{\text{dis}}$ and $t \in T^{\text{free}}_\alpha$ we define binary variables $x_{\alpha t}$ that are equal to 1 if the new date for match α is t. The ILP mathematical model reads:

$$\max \sum_{\alpha \in R^{\text{dis}}} \sum_{t \in T^{\text{free}}_\alpha} x_{\alpha t} \tag{16}$$

$$\text{s.t.} \sum_{t \in T^{\text{free}}_\alpha} x_{\alpha t} \leq 1 \qquad \forall \alpha \in R^{\text{dis}} \tag{17}$$

$$\sum_{\alpha \in R^{\text{dis}}_i} \sum_{\substack{t_1 \leq t \leq t_2, \\ t \in T^{\text{free}}_\alpha}} x_{\alpha t} + k^i_{t_1,t_2} \leq MG_{d_{t_1,t_2}} \qquad \forall t_1, t_2 \in T, 1 \leq t_2 - t_1 \leq 7, i \in S \tag{18}$$

$$x_{\alpha t} \in \{0,1\} \qquad i \in R^{\text{dis}}, t \in T^{\text{free}}_\alpha \tag{19}$$

The objective function (16) maximizes the number of matches rescheduled within the original set of dates of the season. We refer to this objective function as MAXG. Constraints (17) force each match to be rescheduled at most once during the original schedule dates. Constraints (18) enforce the new schedule satisfies Condition 3, while constraint (19) defines the variables to be binary.

In addition, for $\alpha \in R^{\text{dis}}$ and $t \in T^{\text{free}}_\alpha$, we define $d_{\alpha t}$ that indicates the number of days between the date in the original schedule for α and the potential new date, t. For experimental purposes, we consider the additional objective that minimizes the sum day difference between the original date and the new date. which reads

$$\min \sum_{\alpha \in R^{\text{dis}}} \sum_{t \in T^{\text{free}}_\alpha} d_{\alpha t} x_{\alpha t} \tag{20}$$

We refer to objective function (20) as MIND. We remark that constraints (17) are set as equality, rescheduling every disrupted match. In this case, only matches $\alpha \in R^{\text{dis}}$ with $T^{\text{dis}_\alpha} \neq \emptyset$ are considered.

It is important to mention that these formulations differ from other ILP models, such as the traveling tournament problem (TTP). Since our objective is to modify the original schedule as little as possible, the only we only consider decision variables to reschedule suspended matches. Thus, the constraints aim to impose some desired quality conditions on the new generated, even when some feasible dates may exist for some disrupted games. In all cases, the matches that cannot be rescheduled before the end of the season are scheduled afterwards, but still satisfying conditions 1 - 3.

4 Preliminary experimental results

We conducted computational experiments to analyze the differences among the planned NBA 2020-21 season schedule, the executed NBA 2020-21 season schedule, and scheduled obtained following the ideas presented in sections 2 and 3. During this season, 31 games were rescheduled due to COVID19 protocols. The mathematical models are implemented using Python 3 and CPLEX as an ILP solver.

We consider the following strategies implementing the model. For each strategy, we consider the MAXG and the MIND models.

- NBA exec: the timetable executed by the NBA. Recall that the schedule for the second half (after the All Star game) was defined from scratch including the suspended games form the first half. Thus, in addition to the benchmark with the implemented schedule, it provides a comparison with a proactive strategy with almost all the information known in advance.
- *monthly*: games are rescheduled on a monthly basis. We generate batches with the games suspended in a given month, which are rescheduled at the beginning of the next one throughout the remainder of the season. For the upcoming months, this new schedule is used as input.
- *monthly**: the *monthly* strategy with no distance restrictions. Also, if a game was rescheduled by the NBA within the month, we use that date.
- *Post All-Star*: we generate a unique batch including all the suspended games between the beginning of the season and the start of the All-Star Weekend, which are rescheduled in the remainder of the season. Suspended games after the All Star are rescheduled using the *monthly* strategy.
- *Post All-Star**: the *Post All-Star* adapted accordingly to *monthly**.

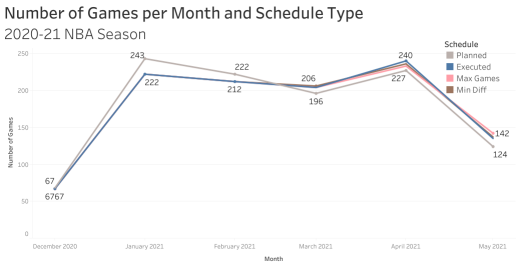
We consider first the instance provided by the scenario faced by the NBA during the 2020 - 2021 regular season, using as inputs the schedule and those games that were indeed suspended due to COVID protocols. Each combination between model and strategy, including the executed NBA season, is evaluated considering the total distance travelled and the number of breaks. For our approaches, since they are reactive strategies, we further report the number of additional rounds and games scheduled after the end of the season according to NBA exec. In Table 1 we report the percentage difference of MAXG and MIND with respect to the planned NBA schedule (i.e., with no disruptions).

metric / method	NBA exec	*monthly*		*monthly**		*Post All-Star*		*Post All-Star**	
		MAXG	MIND	MAXG	MIND	MAXG	MIND	MAXG	MIND
distance	-0.2%	0.9%	1.0%	1.5%	0.8%	1.3%	0.8%	1.3%	0.4%
breaks	0.6%	-0.2%	-0.1%	-0.3%	0.2%	-0.5%	0.2%	-0.4%	0.4%
# dates added	-	11	7	8	4	8	5	6	3
games after	-	14	8	9	3	15	8	10	3

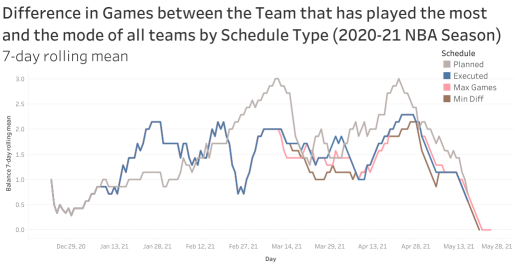
Table 1: Comparison for the different strategies, relative to the NBA planned schedule.

The main message in Table 1 is that there are no major differences over these metrics when compared to the planned NBA schedule. Then, our approach is competitive compared to the approach used by the NBA, with all metrics fluctuating around a difference of 1%. One potential reason for this behavior may be the few potential reschedule options for each suspended match, considering that teams have, on average, two days between consecutive games. We note that, in all cases, MIND reschedules less games after the end of the season than MAXG, requiring also less additional rounds. In addition, we observe in almost all scenarios an increase on the distance travelled and an improvement in the total breaks. Finally, the comparison between each strategy with its relaxed version (i.e., *monthly* vs. *monthly** and *Post All-Star* vs. *Post All-Star**) shows that Condition 4 has a significant impact regarding the number of additional rounds needed.

In Figure 1, we also evaluate the distribution of games per month and the balance level that is generated by each schedule to obtain deeper insights on these results. Figure 1a shows the number of games per months for the NBA original schedule, the NBA executed schedule, and for the *monthly** strategy under the MIND and the MAXG objectives. As expected, our solution yields similar results than the ones that were presented



(a) Number of games per month by schedule.



(b) Balance of the schedule, *monthly** strategy.

Figure 1: Number of matches by month and balance level by schedule type - *monthly**

by the NBA, as basically the same number of games are played in all scenarios. However, it is interesting to observe that the NBA originally planned for a more concentrated schedule on the first half of the season, leaving some flexibility for second half to eventually reschedule suspended games by implicitly using these *buffers* defined at the planning stage. This proactive action is consistent with some recommendations obtained by [3] for time-constrained schedules. It is important to mention that this worked on practice because the amount of suspensions due to COVID cases peaked during December - February and decreased afterwards.

Figure 1b analyzes the *balance* of the different schedules, computed as a 7-day rolling mean between the team that played the greater number of games up to that point and the mode of all teams, for the *monthly** strategy. This visualization shows the differences among the schedules depending on the objective function used. We highlight the high variability in the original NBA planned schedule.

These results are heavily influenced by the particular disruptions that happened during the 2020-21 NBA season. More generally, it is difficult to assess whether these strategies, including the NBA executed schedule, would have worked in a context where the COVID19 cases remain constant over time. Thus, we generated larger instances, using the real season as input, we consider larger instances with more disruptions created as follows: (i) 15 additional disruptions, randomly selected throughout the tournament (50% + disruptions); (ii) 25 additional disruptions, randomly selected throughout the tournament (80% + disruptions); and (iii) 15 additional disruptions, all selected from games in March, modeling a second COVID19 wave.

instance / metric	distance	breaks	# dates added	games after
15 more games	3.3%	-0.3%	9	15
25 more games	5.8%	-0.2%	9	20
15 more games in March	1.8%	0.5%	9	17

Table 2: Difference against the planned NBA schedule, *monthly** strategy and MIND objective.

Table 2 shows the results for the *monthly** strategy using the MIND objective for these three additional scenarios, relative to the executed NBA schedule. Briefly, this experiment suggests that under more stressed scenarios, the impact on the different metrics may not be negligible, affecting eventually the business and the fairness of the competition.

References

[1] Gdp growth (annual %). *World Bank.*, 2022.

[2] Unemployment, total (% of total labor force) (modeled ilo estimate). *World Bank.*, 2022.

[3] X. Yi, D. Goossens, and F. T. Nobibon. Proactive and reactive strategies for football league timetabling. *European Journal of Operational Research*, 282(2):772–785, 2020.

Alcohol and Soccer in Brazil

A. Owen* and T. Bason** and G. Buso*** and A. May****

*Coventry University, UK: aa5845@coventry.ac.uk
**Coventry University, UK ab2135@coventry.ac.uk
***Coventry University, UK. guilhermebuso@gmail.com
****Birmingham City University, UK. Anthony.May@coventry.ac.uk

Abstract

A survey of 623 Santos FC fans in Brazil in 2019 was used to explore their perceptions of the link between alcohol and violence, and investigate how introducing alcohol sales may potentially impact future attendance. Perceptions of the links between alcohol and violence was the main driver behind the likelihood of fans avoiding attending future matches if alcohol was on. However, this relationship was found to have greater complexity, being more about (i) fans' attitudes to alcohol more generally and (ii) their likelihood of avoiding attending if violence increased in the stadium. This suggests that the decision to avoid attending matches if alcohol was on sale is not simply about individuals' perceptions of the links between alcohol and violence. Gender and age did not have any bearing on the decision to avoid attending, nor did the extent to which fans attend matches in the stadium. However, there was evidence that those fans that watch live matches in public places are more likely to avoid attending matches in the stadium if alcohol was sold.

1 Introduction

In 2003, following a decade of fan violence in and around football matches, the Brazilian government enacted 2003 *Estatuto de Defesa do Torcedor* (Fans' Bill of Rights Act). This law prevented fans from carrying objects, drinks or prohibited substances susceptible to generate acts of violence, and led to the Brazilian Football Confederation to prohibit the consumption of alcohol within football stadiums. This decision was made despite there being an unclear link between alcohol and violence at football stadiums, and little evidence that such bans reduce alcohol intake in fans; fans who are not able to consume alcohol during a match may instead drink to excess prior to the event (Pearson & Sale, 2011). Indeed, rather than leading to violence, Giulianotti (1995) found that alcohol consumption can lead to a 'carnival' type atmosphere amongst fans. Work by Nepomuceno et al. (2017) found that the Estatuto de Defesa do Torcedor had little impact on violent incidents within football stadiums, with the violence that is associated with Brazilian football instead being linked to wider societal issues within the country (dos Reis & Lopes, 2016; Nepomuceno et al., 2017; Raspaud & da Cunha Bastos, 2013). The violence, and lack of safety within football stadiums are thought to be a key reason for fans staying away from stadiums (Rocha & Fleury, 2017).

The ban on alcohol sales lasted just over a decade, until the 2014 FIFA World Cup and 2016 Olympic Games were held in Brazil. Commercial pressure from FIFA and the IOC resulted in the law being relaxed to cover the tournaments (Ireland et al., 2019). Following the Rio Olympic Games, several states in Brazil have reintroduced the sale of alcohol at sport events. One state that, as yet, has resisted is São

Paulo, home to Santos FC. In 2015 Santos City Council legalised alcohol sales at football stadia, but this was overturned 12 months later by the Ministry of Justice. Thus, Santos FC fans were temporarily able to consume alcohol at matches, while the club have attempted to pressure the state of São Paulo to legalise alcohol sales at football

The present study used Santos FC as a case study to (1) explore fan perceptions of the link between alcohol and violence, and (2) investigate how introducing alcohol sales may potentially impact future attendance.

2 Questionnaire and Methodology

An online survey, loosely based on the work of Gee et al. (2016) was distributed to Santos FC fans in 2019. The survey was initially advertised via social media, but the support of Santos FC's marketing department, official fan groups and local sport journalists resulted in 623 respondents, of whom 562 (90.2%) identify as Santos FC fans. The overall gender distribution, with 11.3% Female is consistent with Zuaneti Martins et al. (2022) which reports women represent 10 to 15% of organized supporters in Brazil. The median age range was 26-35 years (mean age based on frequencies = 37.4 years) which is consistent with (Moraes et al., 2020) who found the mean age amongst Brazilian football fans to be 32.8 years. The sample would therefore seem to be a reasonable reflection of football fans in Brazil.

For brevity, we only refer here to the questions included in the questionnaire of relevance to this paper. In particular, participants recorded their level of agreement with the following three statements (with response categories: strongly disagree, disagree, neither agree nor disagree, agree and strongly agree), all of which refer to links between alcohol, violence and attendance:

> Q11.3: "If there was alcohol on sale at the stadium I would avoid attending games";
>
> Q12.1: "I think alcohol on sale at the stadium would lead to increase violence among fans";
>
> Q12.2: "If the violence increases in the stadium, I would not attend football games".

For ease of exposition, in what follows, we refer to these three statements respectively as follows: Avoid Attendance (Q11.3), Alcohol Increases Violence (Q12.1) and Violence Not Attend (Q12.2). These three variables form the primary focus of an ordinal logistic regression model, with **Avoid Attendance** as an ordinal dependent variable (shown in bold to highlight this as the DV). The aim being to examine factors related to fans decisions not to attend matches if alcohol was on sale. For simplicity and in order to allow for more robust estimation in the model, **Avoid Attendance** was re-coded using a scoring system of 1 for a response of strongly disagree or disagree, a score of 2 for neither agree nor disagree and a score of 3 for strongly agree or agree. Responses to the two remaining questions *Alcohol Increases Violence* and *Violence Not Attend*, being treated as two key explanatory variables (factors) of interest in this model (names of explanatory factors are highlighted in italics).

Further explanatory factors were derived using an Exploratory Factor Analysis on the following 14 questions: Six questions related to participants' alcohol consumption and their overall attitude to alcohol more widely (Q8.1, Q9.1, Q9.2, Q10.1, Q10.4 and Q11.1), along with eight additional questions relating to participants' behaviours with regard to watching Santos FC matches (Q5.1, Q5.2, Q6.1, Q6.2, Q7.1, Q7.2, Q10.2 and Q10.3). A list of these questions is provided later in Section 3.

The ordinal logistic regression model is specified by letting y_i represent the response to **Avoid Attendance** for participant i ($y_i = 1, 2, 3$; $i=1,\ldots,623$). The probability of participant i responding with

a category score of j or lower (j = 1, 2) is then defined as p_j = prob($y_i \leq j$). The usual (ordinal) odds can then be defined as $\theta_j = p_j / (1 - p_j)$, which represents the odds of participant i responding to **Avoid Attendance** with a score of j or lower (j = 1, 2). Note that since $p_3 = 1$, θ_j is only defined for $j \leq 2$. The ordinal logistic regression model can then be specified by relating the log-odds to a linear combination of the explanatory variables as follows:

$$\text{Log}_e(\theta_j) = \alpha_j - \boldsymbol{X\beta}. \quad (1)$$

In (1), the matrix $\boldsymbol{X}$ contains the observed data from the explanatory factors, whilst the vector $\boldsymbol{\beta}$ contains the model parameters to be estimated which describe the effect of the explanatory factors on the log-odds. Finally, the α_j are "threshold" parameters which are not of particular interest here as they simply serve a similar role as the intercept in a linear regression model. The inclusion of a negative sign in (1), allows increasing values for explanatory factors being associated with a greater likelihood of agreeing with **Avoid Attendance**.

3 Results

3.1 Overall responses to Questions on Links between Alcohol, Violence and Attendance

Figure 1 summarises the distributions of agreement with **Avoid Attendance**, *Alcohol Increases Violence* and *Violence Not Attend*. The vast majority (80%) disagreed with **Avoid Attendance**, suggesting that the sale of alcohol would not impact on most fans decision to attend if alcohol was sold in stadia. However, we wish to better understand what might be behind those fans who disagreed with this statement, or either remained neutral on this question or indeed agreed with it.

In terms of perceived links between alcohol and violence, the majority also disagreed with *Alcohol Increases Violence* with just 22% sharing the view that alcohol on sale at the stadium would lead to increase violence among fans. There was also a slight majority agreeing with *Violence Not Attend*, where 56% saying that if violence increases in the stadium, I would not attend football games. Hence, whilst this suggests violence in stadia would lead to lower attendance, only a minority believe that alcohol on sale in stadia would actually lead to more violence.

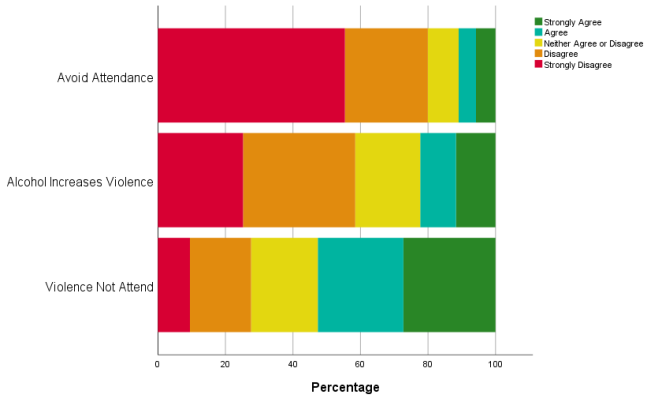


Figure 1. Distribution of responses to three statements.

3.2 Exploratory Factor Analysis

An Exploratory Factor Analysis was undertaken, which included six questions related to participants' alcohol consumption and their overall attitude to alcohol more widely, along with eight additional questions relating to participants' behaviours with regard to watching Santos matches. Table 1 lists the questions asked and the resulting factor loadings (smaller factor loadings below 0.4 are not shown to assist with interpretation of the results). The results suggest that there were five different factors with eigenvalues above one, which together explained 76% of the variance in the data. These questions loaded onto five factors, which we define as *Attitude to Alcohol*, *Watched Public*, *Watched TV*, *Attended* and *Atmosphere*, respectively. The method of principal components with Varimax rotation was used and all model assessments supported this being a good model for these factors. Alternative methods of rotation were examined but all gave the same conclusions. For the first factor, *Attitude to Alcohol*, a Cronbach's alpha value of 0.87 provided further support for combining those six questions into one factor.

Table 1: Factor Loadings (showing cumulative variance explained)

Proposed Factor Name	Question	PC1 (32%)	PC2 (46%)	PC3 (58%)	PC4 (68%)	PC5 (76%)
Attended	5.1. How many State Championship Santos FC games have you attended at the stadium last season?				0.93	
	5.2. How many Brazilian Championship Santos FC games have you attended at the stadium last season?				0.93	
Watched Public	6.1. How many State Championship Santos FC games have you watched in a public place (e.g. pub, bar, restaurant)?		0.93			
	6.2. How many Brazilian Championship Santos FC games have you watched in a public place?		0.91			
Watched TV	7.1. How many State Championship Santos FC games have you watched live on TV last season?			0.96		
	7.2. How many Brazilian Championship Santos FC games have you watched live on TV last season?			0.96		
Attitude to Alcohol	8.1. How often do you consume alcohol?	0.80				
	9.1. How many alcoholic drinks do you consume watching Santos FC game on TV?	0.70				
	9.2 How many alcoholic drinks do you consume watching Santos FC games in a public place?	0.73				
	10.1. "I think drinking alcohol is part of the football atmosphere""	0.71				
	10.4. "The atmosphere of football events makes me feel like drinking alcohol"	0.72				
	11.1. "If there was alcohol sale at the stadium, I would definitely drink"	0.88				
Atmosphere	10.2. "I attend football matches because of the atmosphere					0.68
	10.3. "I attend football matches because of the sport"					0.75

Simple aggregated mean scores were derived for each of these five groups of questions for each participant. Others might instead choose to utilise the factor scores from the Factor Analysis, but our approach provides for easier interpretability of the results and accessibility in understanding the data, as well as easier replication of the results by others. Furthermore, the loadings on each factor as shown in Table 1 are all broadly very similar, which would also support our more straightforward approach.

3.3 Modelling Avoid Attendance

The first three rows of Table 2 summarise Models 1 (a), (b) and (c), all with **Avoid Attendance** as the ordinal response, where each has only one of the three explanatory variables *Alcohol Increases Violence*, *Violence Not Attend* or *Attitudes to Alcohol* included in the model. However, note that when included as explanatory variables, *Alcohol Increases Violence* and *Violence Not Attend* are both treated as scale variables (1=strongly disagree, 2=disagree, 3=neither agree nor disagree, 4=agree and 5= strongly agree). Although the data for these two explanatory factors are inherently ordinal, we actually treat them as scale variables for two reasons. Firstly, this approach is supported by comparison with the results, which follow below, with those from models where both are treated as 'categorical' explanatory variables (not shown for brevity). We argue that our approach is reasonable, since the parameter estimates from those 'categorical' models suggested that their impact on **Avoid Attendance** was broadly the same when either *Alcohol Increases Violence* and *Violence Not Attend* move from one ordinal category to the next (i.e. from a score of 1 to 2, 2 to 3, etc.). Secondly, this approach facilitates more robust parameter estimation and allows for easier interpretation and of the results.

The results for Models 1 (a), (b) and (c) in Table 2, suggest that all three explanatory variables are individually statistically significant ($p<0.001$). The positive parameter estimate for *Alcohol Increases Violence* (2.24) in Model 1(a), suggests that an increased perception that alcohol leads to violence is associated with a greater likelihood of avoiding attending matches if alcohol was on sale. Similarly, the positive parameter estimate for *Violence not Attend* (1.75) in Model 1(b), indicates that a greater predisposition to not attend if there was violence is also associated with a greater likelihood of avoiding attending if alcohol was on sale in stadia. The negative estimate for *Attitudes to Alcohol* (-1.04) in Model 1(c), suggests that those with a more 'positive' attitude to alcohol and/or greater consumption, is associated with being less likely to avoid attending if alcohol was on sale in stadia.

These are of course all intuitively sensible outcomes. However, Table 2 also shows that the single most important factor amongst these three, in terms of their impact on **Avoid Attendance**, is in fact *Alcohol Increases Violence,* since Model 1(a) has the largest Nagelkerke R^2 value (0.47). This suggests that one of the most important factors that influences whether people elect not to attend if alcohol is on sale, is whether they perceive the sale of alcohol leading to increased violence.

Table 2: Initial Models for **Avoid Attendance** as the Ordinal Response

Model	**Explanatory Variable**						**Nagelkerke R^2**
	Alcohol Increases Violence		*Violence Not Attend*		*Attitudes to Alcohol*		
	Estimate (sd)	p	Estimate (sd)	p	Estimate (sd)	p	
1(a)	2.24 (0.18)	<0.001					0.47
1(b)			1.75 (0.24)	<0.001			0.22
1(c)					-1.04 (0.12)	<0.001	0.17
2(a)	2.00 (0.18)	<0.001	1.13 (0.25)	<0.001			0.51
2(b)	2.09 (0.18)	<0.001			-0.44 (0.14)	0.002	0.49
3	1.88 (0.19)	<0.001	1.10 (0.25)	<0.001	-0.39 (0.14)	0.006	0.52

However, we begin to see a more complex set of relationships when we examine Models 2(a) and (b) in Table 2 that combine *Alcohol Increases Violence* with either *Violence Not Attend* or *Attitudes to Alcohol*, respectively. These suggest that both *Violence Not Attend* and *Attitudes to Alcohol* have an additional impact on the likelihood of not attending if alcohol was on sale ($p<0.001$), after having

accounted for fans' perceptions of whether alcohol leads to increased violence. Table 2 shows that the original estimated effect of *Alcohol Increases Violence* of 2.24 in Model 1(a), is actually changed very little by the inclusion of either of these additional explanatory variables, reducing only slightly to 2.00 or 2.09 in Models 2(a) and 2(b), respectively.

What is most interesting to note in Table 2 however, is that the size of the estimates for *Violence Not Attend* and *Attitudes to Alcohol* are much reduced from 1.75 and -1.04 in Models 1(b) and (c) respectively, to 1.13 and -0.44 in Models 2(a) and (b). This actually suggests that *Violence Not Attend* and *Attitudes to Alcohol* both potentially have not only a direct effect on **Avoid Attendance**, but also an indirect effect, mediated through *Alcohol Increases Violence*. These mediation effects are indeed confirmed by models AV1 and AV2 in Table 3, where *Alcohol Increases Violence* is taken to be an ordinal response in an ordinal logistic regression model with either *Violence Not Attend* or *Attitudes to Alcohol* as the single explanatory variable. Mediation is evidenced in both cases by the fact these relationships are significant ($p<0.001$). See Figure 2 later for a visual representation of these links, which illustrates both the direct effects of *Violence Not Attend* and *Attitudes to Alcohol* on **Avoid Attendance** (solid lines) and their indirect effects (dotted lines) mediated through *Alcohol Increases Violence*. We discuss the implications of these results in the discussion.

The effects of all three explanatory variables on **Avoid Attendance** are little changed and all remain statistically significant when all are included in the same model in Model 3 (Table 2). This suggests that the mediation relationships summarised above remain valid in Model 3, which also has an increased Nagelkerke R^2 value of 0.52.

Table 3: Models for ***Alcohol Increases Violence*** as the Ordinal Response*

Model	Explanatory Variable	Estimate (sd)	p	Nagelkerke R^2
AV1	*Violence Not Attend*	1.11 (0.12)	<0.001	0.20
AV2	*Attitudes to Alcohol*	-1.02 (0.10)	<0.001	0.20

*For model robustness, ***Alcohol Increases Violence*** was also re-defined as 1 = strongly disagree or disagree, 2 = neither agree nor disagree and 3 = strongly agree or agree.

The next step was to explore whether any of the additional factors from the Factor Analysis, not yet considered, provide any additional insights into the reasons why people might avoid attendance if alcohol was on sale. Table 4 shows the results of fitting Model 4 which includes the same explanatory variables as above in Model 3, but also all the additional explanatory variables.

Table 4: Final Model 4 for **Avoid Attendance** as the Ordinal Response (Nagelkerke R^2=0.54)

Explanatory Variable	Estimate (sd)	p
Alcohol Increases Violence	1.91 (0.19)	<0.001
Violence Not Attend	1.06 (0.26)	<0.001
Attitude to Alcohol	-0.54 (0.16)	0.001
Attended	0.03 (0.15)	0.85
Watch TV	-0.09 (0.11)	0.39
Watch Public	0.35 (0.14)	0.012
Atmosphere	-0.05 (0.18)	0.78
Gender (Males)	0.05 (0.34)	0.88
Age (18-35)	-0.50 (0.27)	0.069

The results suggest that the decision to avoid attending if alcohol was on sale is not influenced by the extent to which people watch football in stadia (p=0.85) or on TV (p=0.39), but it is influenced by the extent to which people watch football in public places (p=0.012). The positive estimate of 0.35 for *Watch Public* suggested that those who watch football more so in public places are also more likely to avoid attending if alcohol was on sale. Finally, there was no evidence that gender (p=0.88) or age (p=0.069) have a direct influence on whether individuals are likely to avoid attending if alcohol was on sale. Again, the value of Nagelkerke R^2 was very high at 0.54 suggesting the model is actually very good at explaining what might be behind a person's decision to avoid attending if alcohol was on sale. In addition a test of the assumption of parallel lines (assumed in all models) suggested there was no evidence to doubt this assumption (p=0.32).

Figure 2 provides a visual representation for this final model, which includes the effect of *Watch Public*, as well as illustrating both the direct effects (solid lines) and indirect effects (dotted lines) of *Violence Not Attend* and *Attitudes to Alcohol* (mediated through *Alcohol Increases Violence*) on **Avoid Attendance**.

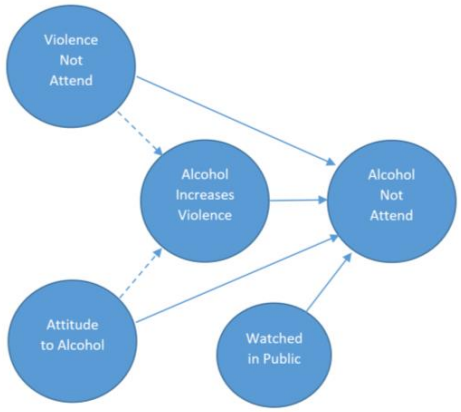


Figure 2: Visual Representation of the Final Model for **Avoid Attendance** as the Ordinal Response

4 Discussion

The key driver behind the decision to avoid attending matches if alcohol was on sale initially appeared to be fans' perceptions of the links between alcohol and violence. However, this relationship is actually being driven by two other factors: fans' attitudes to alcohol more generally and their likelihood of avoiding attending if violence increased in the stadium. These two factors also have an additional direct impact on the decision to avoid attending matches if alcohol was on sale, which suggests that the decision to avoid attending matches if alcohol was on sale is not simply about perceptions of the links between alcohol and violence. The results show that fans' attitudes to alcohol and the degree to which violence in the stadium would put them off attending, would also influence their decision not to attend if alcohol was on sale. Interestingly, gender and age did not to have any bearing on this decision, nor did the extent to which fans attend matches in the stadium. However, there was evidence that those fans that watch their matches live in public places would be more likely to avoid attending matches in the stadium if alcohol was sold.

References

dos Reis, H. H. B., & Lopes, F. T. P. (2016). O Torcedor Por Detrás Do Rótulo: Caracterização E Percepção Da Violência De Jovens Torcedores Organizados. *Movimento (ESEFID/UFRGS)* 22(3):693 https://doi.org/10.22456/1982-8918.57150

Gee, S., Jackson, S. J., & Sam, M. (2016). Carnivalesque culture and alcohol promotion and consumption at an annual international sports event in New Zealand. *International Review for the Sociology of Sport* 51(3):265–283 https://doi.org/10.1177/1012690214522461

Giulianotti, R. (1995). Football and the Politics of Carnival: An Ethnographic Study of Scottish Fans in Sweden. *International Review for the Sociology of Sport* 30(2):191–220 https://doi.org/10.1177/101269029503000205

Ireland, R., Bunn, C., Reith, G., Philpott, M., Capewell, S., Boyland, E., & Chambers, S. (2019). Commercial determinants of health: Advertising of alcohol and unhealthy foods during sporting events. *Bulletin of the World Health Organization* 97(4):290–295 https://doi.org/10.2471/BLT.18.220087

Moraes, I. F., De Oliveira Cruz Carlassara, E., Mazzei, L. C., & Rocco Jr., A. J. (2020). Football in Brazil: what brings fans/consumers to stadiums and arenas in the city of São Paulo. *International Journal of Sport Management and Marketing*, 20(3/4):193 https://doi.org/10.1504/ijsmm.2020.10032919

Nepomuceno, T. C. C., de Moura, J. A., e Silva, L. C., & Cabral Seixas Costa, A. P. (2017). Alcohol and violent behavior among football spectators: An empirical assessment of Brazilian's criminalization. *International Journal of Law, Crime and Justice*, 51, 34–44 https://doi.org/10.1016/j.ijlcj.2017.05.001

Pearson, G., & Sale, A. (2011). "On the Lash" - revisiting the effectiveness of alcohol controls at football matches. *Policing and Society*, 21(2):150–166 https://doi.org/10.1080/10439463.2010.540660

Raspaud, M., & da Cunha Bastos, F. (2013). Torcedores de futebol: Violence and public policies in Brazil before the 2014 FIFA World Cup. *Sport in Society*, 16(2):192–204 https://doi.org/10.1080/17430437.2013.776251

Rocha, C. M., & Fleury, F. A. (2017). Attendance of Brazilian soccer games: the role of constraints and team identification. *European Sport Management Quarterly*, 17(4):485–505 https://doi.org/10.1080/16184742.2017.1306871

Zuaneti Martins, M., Santos Silva, K. R., & Borel Delarmelina, G. (2022). Tensions between fan culture and the feminist identities of female football fans in Brazil. *Soccer and Society*, 23(3):285–297 https://doi.org/10.1080/14660970.2022.2037212

Predicting Cross-country Skiing FIS Points with Taper Load Sequences and Neural Networks

Joonas Pääkkönen*

*Department of Data and Information Management, Dalarna University + email address: jpa@du.se

Abstract

We investigate the training diary dataset of a former elite female cross-country skier who won, *e.g.*, a gold medal in the 2014 Olympic relay. The dataset spans from spring 2012 through spring 2018 with a yearly average of 708.95 training hours. These 6 years cover 119 competitions averaging 69.29 FIS points. We use a multilayer perceptron (MLP) with rectified linear units and 28-day taper load sequence (28-TLS) inputs to predict FIS points. Plots demonstrate that averaged MLP outputs yield promisingly accurate predictions. We also present taper load sequences that correspond to select realised and MLP-predicted FIS point minima. The timing error between the predicted minimum and the realised interval start minimum is only 3 days over a time span of more than 4.5 years. According to our results, the problem of FIS point prediction based on taper load sequences appears to be complex, yet tractable through a neural network.

1 Introduction

The planning of professional cross-country skiing training is heavily results-oriented. It is thus of extreme importance to understand the relationship between training stimuli and competition outcomes. Training program planning largely boils down to understanding the timing of training session modes, loads, intensities and frequencies such that the expected performance level is maximised at competition time. This especially applies to the sensitive time periods close to important competitions, commonly referred to as *taper periods*. Such periods typically involve training load reductions.

One approach to training program planning is through mathematical fitness-fatigue models (see, *e.g.*, [1, 2, 3, 4] and the references therein). However, such models have traditionally relied on explicit formulae, such as simple differential equations, which may lead to overly simplified models. One might argue that artificial neural networks offer a more intriguing approach due to their ability to capture complicated, and non-linear, relationships between inputs and outputs. We argue that there is a non-linear relationship between competition results and training loads, and that this is due to the fact that competition results do not indefinitely improve or deteriorate with training load – once the training load exceeds or drops below certain tipping point threshold values, competition performance is expected to decline drastically. One might hence argue that for any given competition, there exists a non-trivial sequence of daily taper loads that yields the best expected performance outcome, *i.e.*, the lowest FIS point value.

In this paper, we use an artificial neural network model with rectified linear unit activation functions to predict cross-country skiing competition performance with taper load sequences. Here a taper load sequence

is defined as an ordered tuple of daily training time-durations (*loads*, in hours). Our work largely builds around the ideas discussed in [5], where a neural network was successfully applied to a swimming result prediction problem, which inspires further efforts in building performance prediction models for training program planning.

While the training characteristics of highly successful cross-country skiers have been extensively reported [6, 7, 8], to the best of our knowledge this is the first work that applies artificial neural networks to predict cross-country skiing FIS points.

2 Training Diary Dataset

We study the training diary of a former elite female cross-country skier. The diary spans over 6 years from April 23, 2012 to April 22, 2018 (ages 26–32) covering 2,686 training sessions and 4,256 training hours with an average weekly load (AWL) of 13.60 hours of which 90.33% is low intensity training (LIT, training zones A1–A2, 60–84% of maximum heart rate), while the rest, 9.67%, is high intensity training (HIT, training zones A3–A3+, $\geq$85% of maximum heart rate). 119 FIS point competitions (68 interval starts, 23 mass starts, 17 pursuits, 11 skiathlons) were attended producing an average of 69.29 FIS points with a minimum (maximum) of 6.90 (207.95) FIS points. We especially focus on daily *loads*, *i.e.*, the total daily time durations for which the athlete trained during a day.

Figure 1 presents all 119 FIS points (blue circles), 28-day trailing average weekly loads (28-TAWL, red line), over all 6 competition season spans (grey). In total, the dataset contains data over 2,191 days.

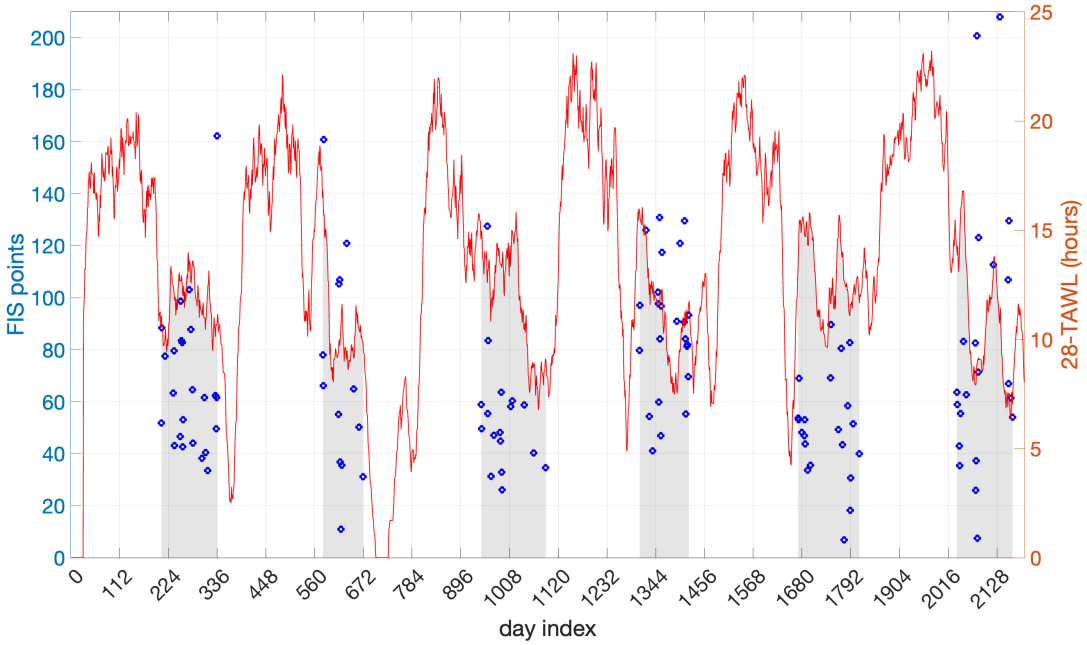


Figure 1: All 119 FIS point realisations (blue circles), 28-day trailing average weekly loads (28-TAWLs, red line), and all 6 competition season spans (grey).

For example, we see that on day 1680, the skier was carrying a training load of approximately 15 hours of weekly training from the past 4 weeks (28 days).

Fig. 1 presents no predictions, only realisations. There are essentially two figures in one figure: the scatterplot of all FIS points and the line plot of TAWLs. A competition season span (grey) is here defined as the day index span from the first FIS competition of a season to the last one. The numbers of competitions per season were: season 1: 24, season 2: 13, season 3: 17, season 4: 23, season 5: 21, season 6: 21.

Fig. 1 shows that for each of the 6 training seasons, the highest 28-TAWL was over 20 hours, peaking at over 23 hours, while the 28-TAWLs dropped to as low as 7–10 hours in the competition seasons with an AWL of 11.88 hours over all 28-day taper periods. Over all taper periods, 78.89% of all training was *specific* training (on-snow skiing and roller skiing, *i.e.*, not running, cycling, or ski-walking). Of the specific training in taper periods, 96.30% was on-snow skiing and 3.70% was roller skiing, while 78.86% of non-specific training was running. The LIT/(LIT+HIT) load ratio was 89.49% over all taper period days.

One interesting observation we make is that, *e.g.*, in competition season 6, the lowest FIS points were produced when there was a clear drop in the 28-TAWL. However, at the end of competition season 6 there was an even deeper drop, yet the lowest FIS points were not produced during this deepest drop. These observations provide us with concrete evidence that while training load reductions are needed to reach peak performance, FIS points do not simply seems to be directly proportional to 28-TAWLs.

3 Prediction Models

FIS point value P scored by a skier in a competition is loosely speaking inversely proportional to the skier's time-result T as $P(T) \coloneqq (T/T_0 - 1)\,F$, where T_0 is the winner's time-result and F is a constant that depends on competition type [9]. The models that we build can predict FIS points for any day $i > t$ whether or not there was a competition on day i, where t is the taper period length in days. If the m-day trailing moving average for the final prediction is desired, we may predict for any index $i > (t+m)$.

In terms of random variables, we study the relative effect of load predictor X on FIS point response variable P, and thus build regression models to predict $Y = \log P$, where $\log(\cdot)$ is the natural logarithm. The log-transformation dampens the effect of large training errors on large target variable values, including outliers. Also, we are especially interested in models that can predict low FIS point values, and we do not, therefore, punish models for prediction errors on exceptionally large values of P.

As will be discussed in what follows, we train two models, the LS model and the MLP model, both with 28-day taper period load data, so that the mean squared logarithmic error (MSLE) between training set target values (FIS points) and model predictions is minimised. The training set consists of all the 24 FIS point observations of season 1 competitions (13 interval starts, 5 mass starts, 3 pursuits, 3 skiathlons).

3.1 LS model

We define the *load scaler* (LS) model as $\hat{y}_i(x_i) \coloneqq cx_i$, where scalar $c = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x})^{-1}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}$ is derived from the ordinary least squares estimate (OLS) with training set t-day TAWL vector $\mathbf{x}$ and training set vector $\mathbf{y}$ of the corresponding FIS points, t-day TAWL scalar $x_i = \frac{1}{t}\sum_{j=1}^{t} d_{i-j}$, where scalar d_i is the training load of day i, and scalar $\hat{y}_i$ is the logarithmic FIS point prediction at average load x_i over t days prior to a competition on day i. In other words, we fit a logarithmic model $\log p = y = cx$ of FIS point values p.

3.2 MLP model

We use MATLAB function `fitrnet` [10] to fit a (28,14,2,1) fully connected multilayer perceptron (MLP) regression artificial neural network (ANN) with rectified linear unit (ReLU) activation functions $f(x) = \max(0,x)$ with the Standardize flag set to *true*. The quadruple (28,14,2,1) refers to 28-day taper load sequences (28-TLS) compressed into 14 neural network inputs, 1 hidden layer with 2 neurons and 1 regression output. We choose the ReLU activation function due to the non-linear nature of athlete training stimuli and responses as discussed in, *e.g.*, [5], where, however, a hyperbolic tangent activation was used. We argue that ReLU activation functions suit the problem best due to their sharp transitions. These sharp transitions correspond to the arguably steep degradation of competition performance once the training load becomes too high or too low. These phenomena are known as *overtraining* and *undertraining*, respectively.

MLP inputs before compression (dimensionality reduction) consist of $t = 28$ daily loads prior to a prediction (competition). The final predictions are trailing moving averages of $m = 28$ predictions. The input dimension is reduced from $t = 28$ to $l = 14$ by choosing the principal component analysis (PCA) scores that correspond to the 14 highest eigenvalues of the sample covariance matrix of the 28-TLSs of all possible 2,163 days. These 14 PCA components explain approximately 69.26% of the total load variance.

Training the MLP model gives the following biases and weights: hidden layer biases: (1.0274, -0.1055), output layer bias: 2.7072, first layer weights: ((0.3670, -0.4695), (-0.1483, -0.1039), (-0.4165, 0.3117), (-1.0270, 0.5839), (0.1063, 0.6387), (-0.4568, 0.4378), (0.1145, -1.1717), (0.5305, -0.7801), (0.1250, -0.1198), (-0.4564, 0.9249), (1.4783, -1.7030), (-0.7047, -0.0617), (0.1671, 0.3221), (-0.6352, 0.0607)), second layer weights: (0.5148, 0.5496).

4 Results

Figs. 2 and 3 present the prediction results for the LS model and the final predictions (28-day trailing moving averages) of the MLP model, respectively. Season 1 is used for training the models, while seasons 2–6 are used for visual testing. The MLP predictions in Fig. 3 are visibly better than the LS predictions in Fig. 2.

4.1 MLP predictions

We make several interesting observations regarding the MLP predictions in Fig. 3, where arrows indicate select critical points, such as the MLP-predicted global FIS point minimum on day 2076 and the realised global interval start FIS point minimum on day 2079. While the scaler model of Fig. 2 incorrectly predicts that the best results of season 6 will be produced at the end of the season, the MLP model correctly predicts that it will occur before the halfway mark of the competition season. The MLP model seems to understand that the 28-TAWL is too low (see Fig. 1) at the end of season 6 to yield low FIS points.

Four of the lowest FIS points were produced either in a pursuit or a mass start competition and all these four competitions produced clearly better results than the fifth best competition. The fifth best competition was the best interval start competition, a 10 km classic, on day 2079. This is only 3 days away from the predicted best day on day 2076 despite the prediction being made 1,741 days (more than 4.5 years) after the last training set observation.

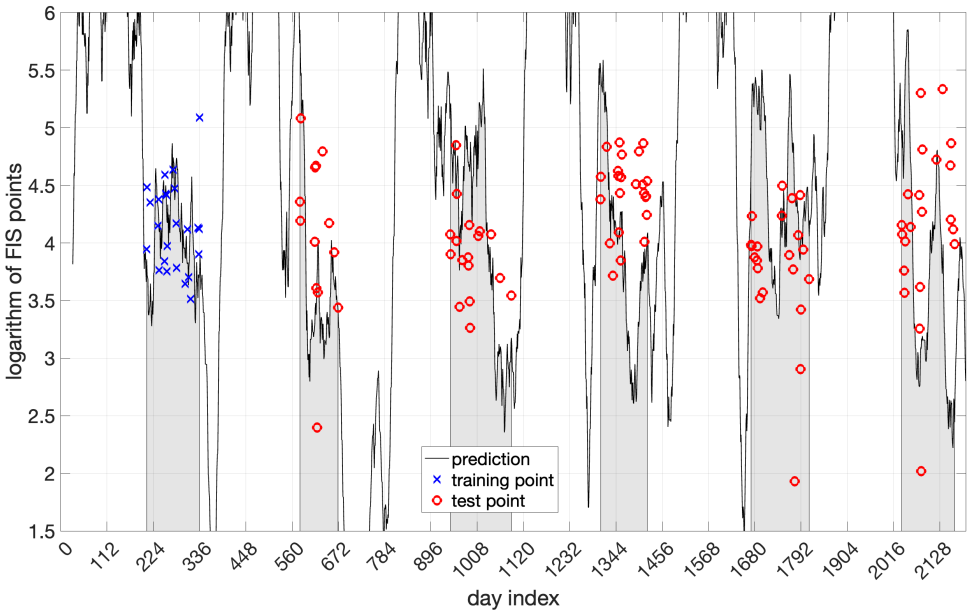


Figure 2: Load scaler (LS) predictions (black line), training set points (blue crosses), test set points (red circles), and competition season spans (grey). The black line corresponds to the red line in Fig. 1.

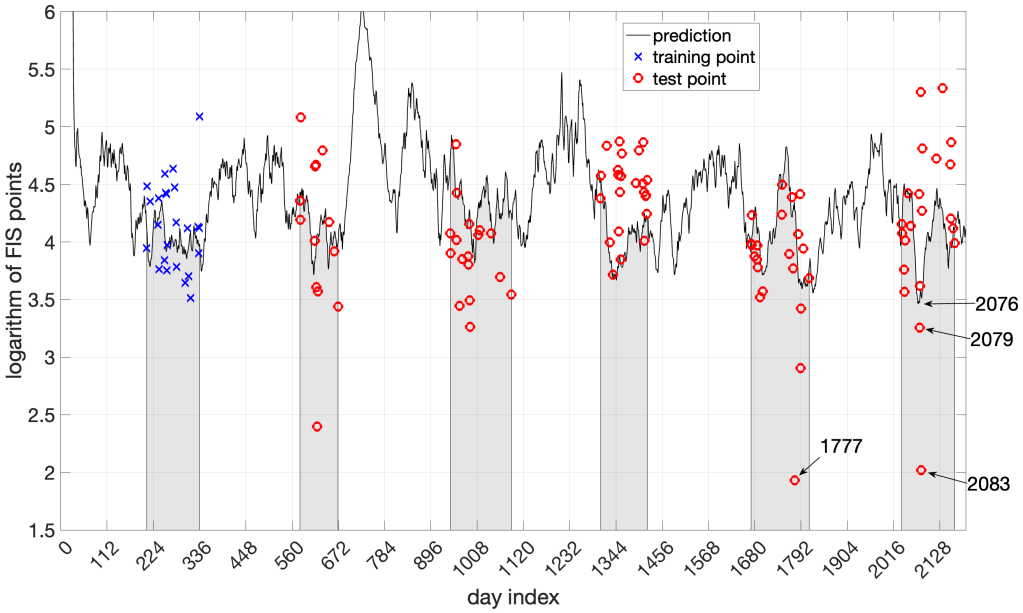


Figure 3: Multilayer perceptron (MLP) predictions (black line), training set points (blue crosses), test set points (red circles), and competition season spans (grey).

The global FIS point minimum realisation occured on day 1777 (a 30 km freestyle mass start), the MLP-predicted global minimum on day 2076, the realised global interval start minimum on day 2079, and the season-specific minimum realisation of season 6 on day 2083 (a 10 km freestyle mass start). Fig. 3 shows that practically for every test season, the lowest FIS point MLP-prediction coincides with the lowest realisation.

4.2 Best taper load sequences

Fig. 4 shows select 56-TLSs that produced or predict low FIS points. All bars are daily loads in training hours. Table 1 presents the 56-TAWLs and the 28-TAWLs for the sequences of Fig. 4 pointed out in Fig. 3.

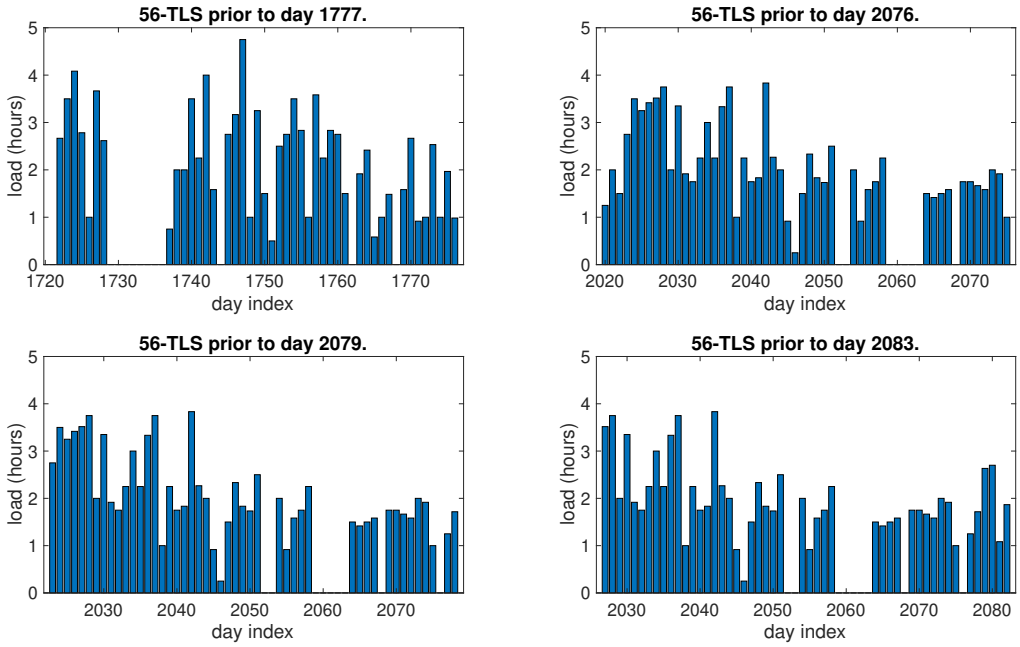


Figure 4: 56-day taper load sequences (56-TLSs) for competitions in Table 1 and pointed out in Fig. 3. Global FIS point minimum realisation on day 1777, MLP-predicted global minimum on day 2076, global interval start minimum realisation on day 2079, the minimum realisation of season 6 on day 2083.

Table 1: TAWLs of the critical points pointed out in Fig. 3.

Index i	Season	Minimum of	56-TAWL at i	28-TAWL at i	FIS points
1777	5	realised global FIS point	12.36 hours	12.70 hours	6.90
2076	6	MLP-predicted global FIS point	12.59 hours	8.64 hours	31.85
2079	6	realised interval start FIS point	12.46 hours	7.91 hours	25.98
2083	6	realised season-specific FIS point	11.79 hours	8.85 hours	7.52

The 56-TAWLs are similar for each sequence in Table 1, while most of the 28-TAWLs are significantly lower than the 56-TAWLs. Roughly speaking, the TAWL decreases from 16 hours to 8 hours across the two consecutive 28-day periods as the mean over the 56-TAWLS is typically approximately 12 hours. Note that on day 2076 there was no competition in actuality, but if there had been, it would have yielded the lowest FIS point realisation according to the MLP model. This lowest prediction value was 31.85 FIS points, so the skier outdid the predictions on several occasions, such as on days 1777, 2079 and 2083, where the realisations were 6.90, 25.98 and 7.52 FIS points, respectively.

5 Conclusions

We have studied the training data of a former elite female cross-country skier, and shown that it is possible to build a multilayer perceptron model with taper load sequence inputs that appears to produce accurate predictions several years into the future. The model seems to be especially adept at predicting the timing of the best relative performances.

We have further observed that 28-day trailing average weekly loads typically at least halve from approximately 20 to 10 hours when transitioning from training season into competition season. We have further presented taper load sequences that correspond to low FIS points, and noted that both the predicted and the realised 28-day trailing average weekly loads usually halve from approximately 16 to 8 hours over two consecutive 28-day periods prior to low FIS point competitions.

Our findings underline the importance of taper period training program planning. Moreover, we have found that it is not only taper loads but rather taper load sequences that appear to be strong predictors for cross-country skiing competition performance.

Acknowledgements

The author is deeply indebted to Anna Jönsson Haag for all her input, including her enthusiastic support. The author would also like to thank Martina Höök, Øyvind Karlsson, and Mikael Svarén for inspiring discussions and helpful assistance.

References

[1] Calvert, T. W., Banister, E. W., Savage, M. V., & Bach, T. M., "A Systems Model of the Effects of Training on Physical Performance," *IEEE Trans. Syst. Man. Cybern.*, 6, 94–102, 1976.

[2] Fitz-Clarke, J. R., Morton, R. H., & Banister, E. W., "Optimizing Athletic Performance by Influence Curves," *J. App. Physiol.*, 71 3, 1151–8, 1991.

[3] Stephens Hemingway, B., Greig, L., Jovanovic, M., Ogorek, B., & Swinton, P. (2021), "Traditional and Contemporary Approaches to Mathematical Fitness-Fatigue Models in Exercise Science: A Practical Guide with Resources. Part I.," *SportRxiv (Preprint)*, https://doi.org/10.31236/osf.io/ap75j

[4] Swinton, P., Stephens Hemingway, B., Rasche, C., Pfeiffer, M., & Ogorek, B. (2021), "Traditional and Contemporary Approaches to Mathematical Fitness-Fatigue Models in Exercise Science: A Practical Guide with Resources. Part II.," *SportRxiv (Preprint)*, https://doi.org/10.31236/osf.io/5qgc2

[5] Edelmann-nusser, J., Hohmann A., & Henneberg, B., "Modeling and Prediction of Competitive Performance in Swimming upon Neural Networks," *European J. Sport Sci.*, 2:2, 1-10, 2002. DOI: 10.1080/17461390200072201

[6] Tønnessen E., Sylta Ø., Haugen T. A., Hem E., Svendsen I. S., & Seiler S. (2014), "The Road to Gold: Training and Peaking Characteristics in the Year Prior to a Gold Medal Endurance Performance," *PLoS ONE*, 9(7): e101796. doi:10.1371/journal.pone.0101796

[7] Solli, G. S., Tønnessen, E., & Sandbakk, Ø. (2017), "The Training Characteristics of the World's Most Successful Female Cross-Country Skier," *Front. Physiol.*, 8:1069. doi: 10.3389/fphys.2017.01069

[8] Torvik, P.-Ø., Solli, G. S., & Sandbakk, Ø. (2021), "The Training Characteristics of World-class Male Long-Distance Cross-Country Skiers," *Front. Sports Act. Living* 3:641389. doi: 10.3389/fspor.2021.641389

[9] International Ski Federation (French: Fédération Internationale de Ski, FIS), "Rules for FIS Cross-country Points 2021–2022," https://assets.fis-ski.com/image/upload/v1624282947/fis-prod/assets/FIS_points_rules_2021-2022_clean.pdf

[10] MATLAB 9.12.0 (R2022a) Statistics and Machine Learning Toolbox, The MathWorks, Inc., Natick, Massachusetts, USA.

Are women 'more emotional'? Evaluating collective decision making by gender using international football

J.J. Reade* and C. Singleton**

*Department of Economics, University of Reading + email address: j.j.reade@reading.ac.uk
**Department of Economics, University of Reading + email address: c.a.singleton@reading.ac.uk

Abstract

In this note we evaluate a claim made that women are more likely to react badly to adverse events in competitive environments. We construct a minute-by-minute dataset on international football matches played by men and women, and consider whether in the aftermath of conceding a goal women's teams are more likely than men's teams to further concede goals.

We find evidence to support this claim.

1 Introduction

On April 13 2022, after watching his Northern Ireland Women's National Team lose a crucial World Cup Qualifying clash with England 5-0, head coach Kenny Shiels said 'Women are more emotional than men. So, they take a goal going in, they don't take that very well.'[1] This comment naturally provoked strong condemnation from within and outside of sport. It is, however, an empirical question, as indeed Sheils acknowledged: "I'm sure you will have noticed if you go through the patterns – when a team concedes a goal, they concede a second one in a very, very short space of time". Could it actually be the case that there is some substance in Shiels's comments?

In this note we use a large sample of international football matches by both women and men, and we test whether, in the immediate aftermath of conceding a goal women's teams are more likely than men's teams to concede further goals.

In Section 2 the relevant previous literature is reviewed, in Section 4 the modelling methodology adopted is set out, in Section 3 our dataset and sources are introduced, in Section 5 results from the econometric estimations are presented, and Section 6 concludes.

[1] The full quote was: "I felt [England] were struggling a wee bit at times to open us up until the psychology of going 2-0 up in the women's game," said Shiels on Tuesday night.

"I'm sure you will have noticed if you go through the patterns – when a team concedes a goal, they concede a second one in a very, very short space of time.

"Right through the whole spectrum of the women's game, because girls and women are more emotional than men. So, they take a goal going in not very well.

"When we went 1-0 down we tried to slow it down to give them time to get that emotional imbalance out of their heads.

"That's an issue we have. Not just in Northern Ireland but all of the countries in the world."

2 Literature

Our study has a natural setting in sport, and football in particular. This has significant advantages, as sport is very well measured, with highly transparent actions being taken, both off the field and on it. As such, it has very often proven to be an effective setting for empirical studies in a range of academic fields [1, 4].

In particular, in the context of gender studies, increasing the abundant data on both men's and women's football is being used to draw distinctions between the genders, and help understand some of the variation in outcomes, and in particular aspects of gender discrimination [3].

3 Data

Our data is collected from the website www.worldfootball.net. We collect the timings of all goals in international matches involving both women and men.

The dataset consists of 15,861 international matches, of which 1,503 are women's matches, and 14,358 are men's. Of the 1,503 women's matches, 316 are youth matches and 1,187 are full matches. Of the 14,358 men's matches, 5,019 are youth matches, and 9,339 are full matches.

In those matches we have 50,357 goals, hence an average of 3.2 goals per game. We construct a minute-by-minute dataset, which is 1,553,188 observations in length. We then consider each team in the match, and hence the dataset is 3,106,376 observations in total.

In Figure 1 we provide a graphical representation of the incidence of goals per minute across a football match. From the summary statistics table in Table 1 there is, on average, 0.017 goals per minute in matches across our entire sample. In Figure 1 this average varies throughout the match from around 0.008 in the very first minute, to 0.038 in the 45th minute and 0.074 in the 90th minute. These two enlarged figures reflect the fact that the 90th minute is an 'elastic' minute: at the end of each half of a football match, 'injury time' is added at the discretion of the referee for stoppages during play.

In Table 1 on page 84 we present summary statistics, including on the frequency of goals. It is plausible that the likelihood of a team scoring a goal is a function of the scoreline at the time: a team slightly behind might chase the game harder than a team multiple goals ahead, for example. We include this variable which, on average, is zero, and ranges between -24 and +24, but from the 11th quartile through to the 89th ranges only between -1 and 1.

The vast literature on the home advantage in sport suggest we ought to include a variable for this. Because our data is on international football, often matches take place on neutral territory — World Cups or regional championships. In our dataset, on average, 34.7% of the time a team is at home.

4 Methodology

We employ an event study methodology. We ask whether, in and around a goal being conceded by a team, they become more likely to concede a further goal. We then consider whether the magnitude of the effect for women's teams differs from that of men's teams. It makes most sense to model this on a minute-by-minute basis, and hence we construct a variable for a goal scored (conceded) by team i against team j in minute m of

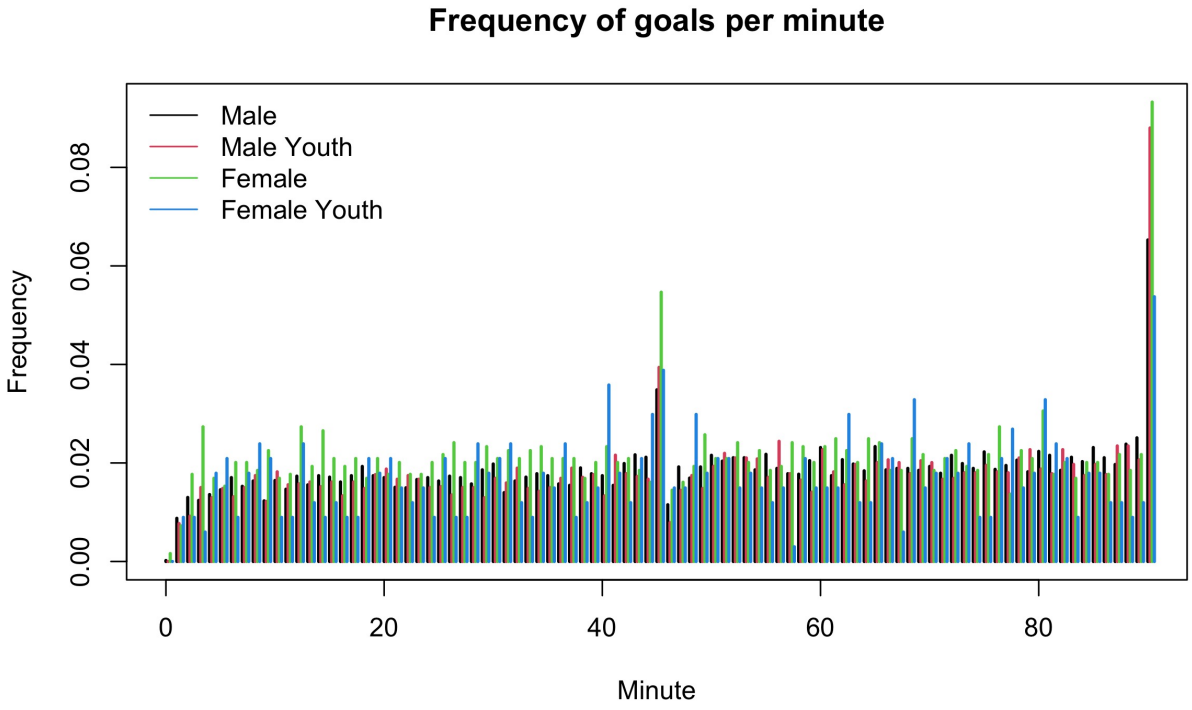


Figure 1: Frequency of goals per minute in international football matches.

Table 1: Summary statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Goal	1,935,042	0.017	0.129	0	0	0	1
Elostrength	1,935,042	1,042.028	156.350	431	967.5	1,138.3	1,463
Score Difference	1,935,042	0.000	1.546	−24	−1	1	24
Score Difference=-1	1,935,042	0.197	0.398	0	0	0	1
Score Difference=-2	1,935,042	0.072	0.258	0	0	0	1
Home Advantage	1,935,042	0.347	0.476	0	0	1	1

a match at time t, $goal_{ijmt}$, as:

$$goal_{ijmt} = \begin{cases} 1 & \text{if team } i \text{ scores (concedes) a goal,} \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

Our regression model is thus:

$$goal_{ijmt} = \alpha_0 + \sum_{k=-15,k\neq 0}^{k=15} \alpha_k goal_{ijt,m+k} + \beta X_{ijmt} + e_{ijmt}. \tag{22}$$

We consider two types of goal: goals scored, and goals conceded. We allow a window of up to 15 minutes either side of a goal, and as is conventional in event study analyses like these, we model both before and after the event in question. Note that this does mean we treat each team in a match separately, and as we model minute by minute, it means we have 180 observations per match (90 for each team). In (22) we include fixed effects for the team in question, and also for the minute of a match. We also include a set of control variables for the teams involved, and the time in the match. These are:

- The Elo rating (strength) for each team.
- The score difference at the time, $hgoals_{mt} - agoals_{mt}$.
- A dummy for if the score difference at the time is -1 or -2.
- Home advantage: a dummy that is one if the team is at home.

5 Results

In Figure 2 we present graphically our results. In this figure we present four cases:

- Likelihood of conceding having recently conceded (top left).
- Likelihood of conceding having recently scored (top right).
- Likelihood of scoring having recently scored (bottom left).
- Likelihood of scoring having recently conceded (bottom right).

In each plot, the 15 points to the left of the vertical line are the 15 minutes before a goal occurs, and the 15 points to the right are the 15 minutes after a goal occurs.[2] We plot the coefficients for four types of matches:

1. Men's matches (solid blue line and circles).
2. Women's matches (solid pink line and circles).
3. Men's Youth matches (dotted blue line and squares).
4. Women's Youth matches (dotted pink line and squares).

[2]We varied the size of the estimation window from 15 minutes either side, to 10 and 5, and found no difference in our results.

Each is plotted with standard error bars, which give a sense of the significance of the difference between the different types of matches.

A significant coefficient to the left of the vertical bar would indicate that *before* conceding a goal a team was more likely to concede. A significant coefficient to the right of the vertical bar would indicate that that many minutes *after* conceding a goal a team was more likely to concede.

We note from Figure 2 that *before* scoring or conceding, teams are not more likely to score or concede. Goals are, as has been pointed out in other studies [2], essentially unforecastable 'news' events.

The Shiels Hypothesis is that after conceding a goal, women's teams are more likely than men's teams to subsequently concede. That is, there should be a significant difference between the men's and women's lines in the top left plot in Figure 2. Visually, from the top left plot, teams conceding having recently conceded, the points for women's matches, in pink, are different from the points for men's matches. It appears that, in any given minute after recently conceding, a women's team is more likely than a men's team to concede. Both teams are *less* likely to concede, but a women's team is more likely relative to a men's team.

The regression results from the top left plot, the case closest to the Sheils Hypothesis, are tabulated in Table 3 (on page 90) where each category of match (men's, women's, adult and youth) has a separate column. That teams are not more likely to conceded before conceding seems plausible. The negative coefficient slightly before a goal, in minute -1, is potentially the effect of the discrete timing of goals: they must be placed in a descrete minute, rather than timed precisely. The magnitude of those coefficients compared to all other minutes prior to the event is similar to the magnitude of the minute 1 coefficient to the coefficients on minutes 2 to 15 after conceding, which further supports this notion.

From minutes 2 to 15 after conceding a goal, a team is *less* likely to further concede a goal. The magnitude of this effect is generally between 1 and 2 percentage points. The effect appears different, since the pink points are generally above the blue ones, but the significance of this difference needs to be established.

We summarise the results across the different types of goals, and also test for the significance of the difference between men's and women's matches, in Table 2 (on page 89). Here we present all the coefficients for the four types of goal scoring plotted in Figure 2 in a separate column. We here combine the dummy variables for a goal conceded or scored in the previous 15 minutes into a single dummy variable, in order to interact that with a dummy variable for whether the match was a women's match.

We find that in the case of teams further conceding having conceded in the previous 15 minutes (top left plot in Figure 2 and column (1) in Table 2 on page 89), that a male team is 1.4 percentage points less likely to concede, but that a female team is only 0.6 percentage points less likely to concede, and that the difference, 0.8 percentage points, is statistically significant with $p < 0.01$.[3] This is evidence in favour of the Sheils Hypothesis.

Conversely, the difference between men's and women's football is insignificant when it comes to the likelihood of teams responding to conceding goals. If a team has scored in the last 15 minutes, they are 1.2 percentage points more likely to concede. This is indistinguishably different between men's and women's matches.

[3]This effect is symmetric, as would be anticipated, with teams that have scored in the last 15 minutes being significantly less likely to score, and that for women's teams, that effect is significantly smaller.

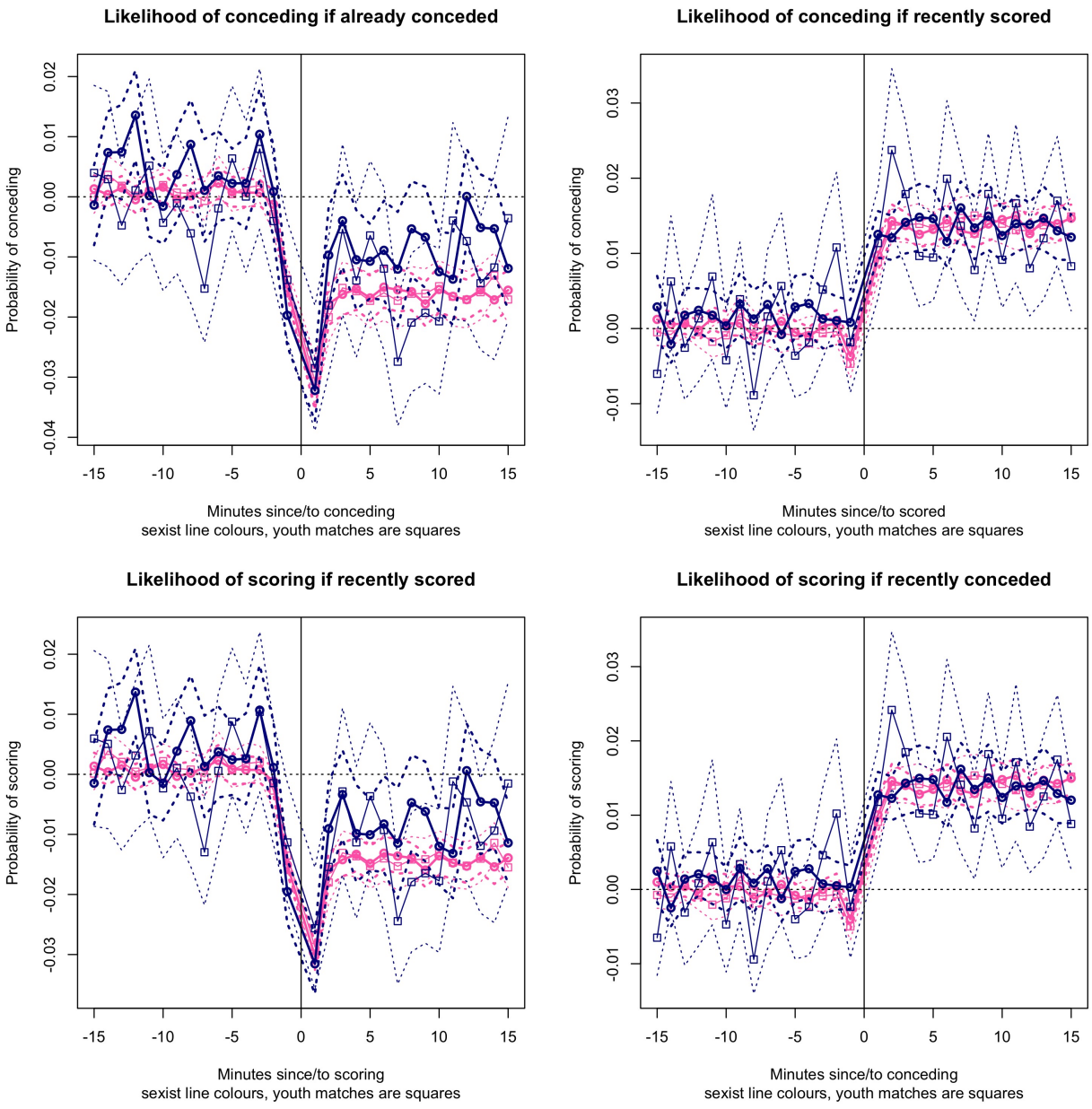


Figure 2: Coefficients for scoring after conceding.

6 Conclusions

In this note we evaluate a claim made that women are more likely to react badly to adverse events. We construct a minute-by-minute dataset on international football matches played by men and women, and consider whether in the aftermath of conceding a goal women's teams are more likely than men's teams to further concede goals.

We find evidence to support this claim.

References

[1] M. Bar-Eli, A. Krumer, and E. Morgulev. Ask not what economics can do for sports-ask what sports can do for economics, 2020.

[2] K. Croxson and J.J. Reade. Information and Efficiency: Goal Arrival in Soccer Betting. *Economic Journal*, 124:62–91, March 2014.

[3] U. Rinne and H. Sonnabend. Female workers, male managers: Gender, leadership, and risk-taking. *Southern Economic Journal*, 88(3):906–930, 2022.

[4] S. Szymanski. The assessment: the economics of sport. *Oxford Review of Economic Policy*, 19(4): 467–477, 2003.

Table 2: Regression results for various goal types

	Dependent variable:			
Event 2	Concede	Concede	Score	Score
Event 1	Concede	Score	Score	Concede
	(1)	(2)	(3)	(4)
female	0.020***	0.017***	0.018***	0.014***
	(0.003)	(0.003)	(0.003)	(0.003)
Elo Rating	0.00000	0.00000	0.00000	0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
female:Elo Rating	−0.00002***	−0.00001***	−0.00001**	−0.00000
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Opponent Elo Rating	−0.00000**	−0.00000	−0.00000***	−0.00000***
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
female:Opponent Elo Rating	−0.00000	0.00000	−0.00001**	−0.00001**
	(0.00000)	(0.00000)	(0.00000)	(0.00000)
Score Difference	−0.010***	−0.010***	0.012***	0.012***
	(0.001)	(0.001)	(0.001)	(0.001)
female:Score Difference	0.001**	0.0002	−0.003***	−0.002***
	(0.0003)	(0.0004)	(0.0004)	(0.0004)
Score Difference=-1	0.008***	0.007***	−0.004***	−0.005***
	(0.002)	(0.002)	(0.001)	(0.0005)
female:score.diff.m1	−0.006***	−0.003***	−0.005***	−0.005***
	(0.001)	(0.001)	(0.001)	(0.001)
Score Difference=-2	0.019***	0.015***	0.006***	0.003***
	(0.003)	(0.002)	(0.001)	(0.001)
female:score.diff.m2	−0.008***	−0.004**	−0.008***	−0.006***
	(0.002)	(0.002)	(0.001)	(0.001)
Home Advantage	−0.0003	−0.0002	−0.0003	−0.0002
	(0.0002)	(0.0002)	(0.0003)	(0.0003)
female:home.adv	0.002***	0.001**	0.002***	0.002**
	(0.001)	(0.001)	(0.001)	(0.001)
Goal conceded in last 15 minutes	−0.014***			0.012***
	(0.001)			(0.001)
female:Goal conceded in last 15 minutes	0.008***			0.001
	(0.001)			(0.001)
Goal scored in last 15 minutes		0.012***	−0.013***	
		(0.001)	(0.001)	
female:Goal scored in last 15 minutes		0.001	0.006***	
		(0.001)	(0.001)	
Observations	2,410,872	2,410,872	2,410,872	2,410,872
R^2	0.024	0.024	0.023	0.023
Adjusted R^2	0.024	0.023	0.023	0.023
Residual Std. Error (df = 2410491)	0.132	0.132	0.132	0.132
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 3: Event study regressions considering the impact of conceding a goal on further conceding by type of match.

	Dependent variable:			
	opp.goal			
	Male	Female	Male Youth	Female Youth
	(1)	(2)	(3)	(4)
Elo strength	−0.00001***	−0.00000	0.00000	0.00002
	(0.00000)	(0.00001)	(0.00000)	(0.00002)
Opponent Elo strength	−0.00000**	0.00000	−0.00000	−0.00001
	(0.00000)	(0.00000)	(0.00000)	(0.00001)
Score difference	−0.014***	−0.011***	−0.014***	−0.012***
	(0.001)	(0.001)	(0.001)	(0.001)
Home Advantage	−0.001***	0.001	−0.00003	0.002
	(0.0003)	(0.001)	(0.0003)	(0.003)
Goal (-15)	0.001	−0.002	−0.0003	0.004
	(0.001)	(0.003)	(0.001)	(0.007)
Goal (-14)	−0.0001	0.007**	0.003**	0.003
	(0.001)	(0.003)	(0.002)	(0.007)
Goal (-13)	0.001	0.007*	0.002	−0.005
	(0.001)	(0.004)	(0.002)	(0.005)
Goal (-12)	−0.001	0.013***	−0.00004	0.001
	(0.001)	(0.004)	(0.001)	(0.006)
Goal (-11)	0.0004	0.00005	0.001	0.005
	(0.001)	(0.003)	(0.001)	(0.007)
Goal (-10)	0.001	−0.002	0.002	−0.005
	(0.001)	(0.003)	(0.001)	(0.006)
Goal (-9)	−0.001	0.004	0.0005	−0.001
	(0.001)	(0.004)	(0.002)	(0.006)
Goal (-8)	−0.0003	0.009**	0.0003	−0.006
	(0.001)	(0.004)	(0.001)	(0.006)
Goal (-7)	0.001	0.001	−0.001	−0.015***
	(0.001)	(0.004)	(0.002)	(0.004)
Goal (-6)	0.002	0.003	0.003*	−0.002
	(0.001)	(0.004)	(0.002)	(0.007)
Goal (-5)	0.0004	0.002	0.0002	0.006
	(0.001)	(0.003)	(0.001)	(0.006)
Goal (-4)	0.0002	0.002	0.001	0.0002
	(0.001)	(0.004)	(0.002)	(0.006)
Goal (-3)	0.0002	0.010***	0.002	0.008
	(0.001)	(0.004)	(0.001)	(0.007)
Goal (-2)	−0.002	0.001	−0.003*	−0.004
	(0.001)	(0.004)	(0.002)	(0.005)
Goal (-1)	−0.016***	−0.020***	−0.014***	−0.014***
	(0.001)	(0.003)	(0.001)	(0.005)
Goal (+1)	−0.030***	−0.031***	−0.028***	−0.027***
	(0.001)	(0.002)	(0.001)	(0.005)
Goal (+2)	−0.015***	−0.008**	−0.017***	−0.016***
	(0.001)	(0.004)	(0.002)	(0.006)
Goal (+3)	−0.014***	−0.003	−0.013***	−0.004
	(0.001)	(0.004)	(0.002)	(0.007)
Goal (+4)	−0.013***	−0.009**	−0.013***	−0.012*
	(0.001)	(0.004)	(0.002)	(0.006)
Goal (+5)	−0.014***	−0.009***	−0.015***	−0.005
	(0.001)	(0.003)	(0.002)	(0.006)
Goal (+6)	−0.013***	−0.008*	−0.013***	−0.010
	(0.001)	(0.004)	(0.002)	(0.007)
Goal (+7)	−0.013***	−0.011**	−0.015***	−0.025***
	(0.001)	(0.005)	(0.002)	(0.005)
Goal (+8)	−0.013***	−0.004	−0.014***	−0.019***
	(0.001)	(0.004)	(0.002)	(0.006)
Goal (+9)	−0.015***	−0.006	−0.014***	−0.017***
	(0.001)	(0.004)	(0.002)	(0.006)
Goal (+10)	−0.013***	−0.011***	−0.012***	−0.019***
	(0.002)	(0.003)	(0.002)	(0.006)
Goal (+11)	−0.014***	−0.013***	−0.014***	−0.002
	(0.001)	(0.004)	(0.002)	(0.008)
Goal (+12)	−0.015***	0.001	−0.015***	−0.005
	(0.001)	(0.004)	(0.002)	(0.007)
Goal (+13)	−0.013***	−0.004	−0.013***	−0.012**
	(0.001)	(0.004)	(0.002)	(0.006)
Goal (+14)	−0.015***	−0.004	−0.011***	−0.010
	(0.001)	(0.004)	(0.002)	(0.008)
Goal (+15)	−0.013***	−0.011***	−0.015***	−0.002
	(0.001)	(0.003)	(0.002)	(0.008)
Observations	1,139,358	144,814	612,318	38,552
R^2	0.023	0.030	0.022	0.029
Adjusted R^2	0.022	0.029	0.021	0.025
Residual Std. Error	0.127 (df = 1138974)	0.138 (df = 144642)	0.128 (df = 612091)	0.134 (df = 38410)
Note:				*p<0.1; **p<0.05; ***p<0.01

The Path to GOAT-ness: Classifying Tennis Strokes

Padmanaba Srinivasan* and Ashish Agrawal** and William J. Knottenbelt*

*Department of Computing, Imperial College London, London, SW7 2AZ, UK, Email: {ps3416, wjk}@ic.ac.uk
**Infosys, Mumbai, Maharashtra, India, Email: ashish.agrawal2123@gmail.com

Abstract

Stroke classification in tennis is a fine-grained, nuanced activity where a model must understand a player's subtle movements while selectively making use of a complex and dynamic background. We present a method to perform robust and fine-grained classification on strokes hit by professional tennis players participating in the 2019 and 2020 French Open tournaments. Much progress has been made in the field of action recognition, yet the application to sports has been limited to performing recognition to determine which sport is being played; fine-grained classification of intra-sport strokes is a more subtle problem that is in its relative infancy. Existing approaches for stroke classification perform coarse classification on broadcast video, or granular classification on content filmed in near-ideal conditions, such as the videos from the THETIS dataset. We propose a model to perform granular classification on broadcast tennis video and evaluate our results to show that our model is able to learn informative and interpretable spatiotemporal features and achieves an 85.7% classification rate.

Data in the sport of tennis is becoming increasingly granular and most importantly, *available*. For years, sports data consisted of manually curated, coarse statistics such as percentage of first serve points won, in tennis. With ever higher resolutions of broadcast video, there is an increasing richness to the data, further supplemented by spatiotemporal information from HawkEye [27] such as locations of players, and location, speed and acceleration of the ball. This paper focuses on classifying shots played in tennis using broadcast RGB video.

Popular sports-focused video datasets include the UCF-Sports [30, 35] and Sports 1-M dataset [20] which contain videos from many different sports with the objective to classify which sport is being played in a video. In this work we are interested in classifying which fine-grained shot is being played within the game of tennis – a problem that is far more subtle in the level of understanding required as we solve an intra-sport classification task [24].

Previous work in tennis video classification has focused on either coarse-grained classification using broadcast or broadcast-like video, or fine grained classification using specially curated video datasets. [47, 48] use broadcast tennis video, computing Histograms of Optical Flow followed by a Support Vector Machine (SVM) classifier to classify shots as *left-swing* or *right-swing*. [9] compute the Histogram of Oriented Gradients in 3D (HOG3D) from broadcast video, along with Linear Discriminant Analysis (LDA) to classify motion into *hit*, *non-hit* or *serve* categories. [10] train a model to indicate whether a *hit*, *serve* or *other* action has occurred in one of four regions of a broadcast video stream before continuing on to caption generation. [32] use frame-differencing and obtain skeleton structures, encoding these using Histogram of Oriented Gradients (HOG) and an SVM classifier to classify a shot as either a *forehand*, *backhand* or *no stroke*.

Outside of broadcast video, a popular dataset is THETIS [13] which contains RGB video as well as (optionally) depth and skeleton information. THETIS contains 55 players (both professionals and amateurs) shown hitting twelve tennis shots, repeated three times, in a well-framed setting with actors facing the camera, in a non-tennis-court setting and without using a ball. [40] and [4] use the RGB video only to perform action recognition across the twelve classes using Recurrent Neural Networks (RNNs) to model time-dependencies. [2] employ a similar approach, using OpenPose [6] to estimate 3D pose and tracking player positions and combining to predict one-of-three shot directions. Related also are several approaches to stroke classification in table tennis [25, 23, 22].

In our work, we perform fine-grained classification directly on broadcast tennis video using convolutional neural networks. The main contributions of this paper are as follows:

- The presentation of a backbone 3D convolutional architecture along with extensions to improve performance on the stroke classification task.
- Application of the proposed architectures to the stroke classification task on broadcast tennis video.
- Quantitative evaluation, comparing the proposed models and their performance on our dataset.
- Qualitative evaluation, investigating why a model makes a decision and identify sources of errors.

1 Related Work

1.1 Action Recognition

Action recognition is a diverse field which deals with data of many modalities as well as varied situations under which actions occur. Different data modalities can yield specific insight into a task, for example, 3D skeleton or RGB-D data would provide a high amount of detail for human action recognition; however, in this work, we use monocular RGB video data as concurrent multi-view video is not available. Some variations in modalities and tasks are summarised below:

- Video properties: high or low resolution, monocular or multi-view, RGB or grayscale.
- Background effects: whether the video is filmed with a controlled background or real-world setting.
- Background importance: whether the background provided information that can help in classifying the action, being near the net when performing a stroke can mean it is far more likely that a volley will be played.
- Foreground interaction: whether the interaction of the subject with other objects/people is of importance.
- Action granularity: coarse or fine-grained classification, e.g. forehand-side hit vs forehand slice or forehand volley.

Deep learning is a subset of machine learning that used Artificial Neural Networks (ANNs) to learn an internal representation of some input data which is then used to inform a regression/classification task. In computer vision, this is achieved through the use of convolutions layers, and networks that use these are called Convolution Neural Networks (CNNs). The primary advantage of deep learning is the ability to learn which features to extract and how to extract these. Self-learning of features also means that any detected structures may be more informative and generalizable. For 2D image classification problems, CNNs have

consistently outperformed other state of the art methods [36]. More recently, CNNs in action recognition have shown improved performance on a variety of datasets [49, 17].

Video understanding is more complex than merely understanding static images as actions can have different levels of coarseness that necessitate temporal understanding, therefore, a model must also learn about feature correspondence between images. Correspondence between images can be weakly enforced by stacking images channel-wise, moderately enforced by fusion techniques, or more strongly enforced through the use of RNNs [31] or 3D convolution [19]. Previous convolution-based methods used in video understanding can be placed in one of three categories:

- 2D CNN with channel-wise stacking of images, or temporal fusion techniques
- 2D CNN followed by a RNN
- 3D CNN

2D CNN two-stream approaches came in 2014 with the work of [34], and [20]. [34] explicitly separate spatial understanding from temporal-dynamics understanding by providing a single RGB image as input to one stream and stacked optical flow frames calculated from several RGB frames to the other stream. [20] allow temporal understanding in both streams with fusion of sequences of frames, but design one stream to receive a centre-cropped version of frames to foster *near* understanding and the other stream to receive uncropped frames for *far* spatial understanding.

Using 2D CNNs for spatial understanding with Long-Short Term Memory (LSTM) [15] for temporal context was pioneered by [8] and [45]. [45] explore using raw RGB frames and optical flow frames with temporal understanding using pooling operations or LSTMs and show that LSTMs outperform pooling for fusing spatial features. [8] use their model for image captioning and achieve impressive performance.

3D CNNs were applied to action recognition by [19] which outperformed existing 2D CNN models and has spawned further development using 3D convolution such as SlowFast networks [12] and X3D networks [11]. 3D convolution is computationally expensive and can lead to overfitting. Attempts have been made to reduce this using separable convolution, where spatial convolutions are performed separately from temporal convolutions [28, 29, 38].

1.2 Attention

When processing visual input, humans have the ability to focus on certain parts of an image instead of the entire scene. In computer vision, attention is achieved through saliency which are scalar matrices that represent the relative importance of pixels in a scene [41]; a higher response indicates a more important pixel.

Attention in CNNs can be classified in one of two ways: as post-training network analysis, or, as trainable attention mechanisms (self-attention). Post-training analysis has been conducted by [33] and [5] who analyse saliency maps and class gradients from input images and perform post-facto computation with [5] introducing attention using binary activations between layers. Self-attention can be classified either soft (deterministic) or hard (stochastic). Soft attention has been used for query based tasks [1, 43] and computer vision tasks [18, 42, 16] and have been shown to improve performance over standard CNNs.

Stroke	Number of samples
Forehand Groundstroke	3856
Forehand Slice	771
Forehand Volley	235
Backhand Groundstroke	3096
Backhand Slice	2068
Backhand Volley	207
Overhead (serve+smash)	941

Table 1: Number of samples of each class in our dataset.

2 Methodology

2.1 Dataset

We use a video dataset consisting of 941 professional singles points played in the 2019 and 2020 French Open tournaments. It contains 11 174 shots from both the men's and women's draw (7 557 men's, 3 617 women's, 22 unique players overall) as monocular RGB videos. Each shot is classified as one of seven strokes with the number of samples for each shown in 1. The videos captured are from two different courts and are drawn from 20 matches. Temporal demarcation of shots is obtained from HawkEye data that accompanies each video. Although the video data is likely publicly available, the corresponding HawkEye is proprietary and so this dataset cannot be made public.

When collating the dataset, we observed that volleys and slices are infrequent, with an imbalance of volleys likely further exacerbated by clay court characteristics. In order to overcome the severe class imbalance, we undersample majority classes and oversample minority ones, resampling at the beginning of every epoch. Smashes on their own would form another minority classes, hence, we combine serves and smashes as one class based on the observation that the overhead strokes share a high level of similarity.

When training or evaluating CNNs, images are often resized to a smaller dimension for computational efficiency. Our raw videos are either at a resolution of 1920×1080 pixels or 1280×720 pixels. Downsizing these videos directly would result in a loss of detailed information about the player while retaining unnecessary information about the wider scene, such as the umpire, line judges and spectators/stands.

To provide a high-detail image to the model, we track players using a siamese network-based tracker [2] and crop a square region around the player when they hit the ball. Different size of square-crop are used for the near and far players, with the crop then resized to 224×224. We split the dataset by assigning 63% to the training set, 30% to the test set, and 7% to the validation set. When performing the train-test split, we ensure that players found in the test set are not present in the training set.

2.2 Backbone Model

Our action recognition model uses a backbone ResNet (2+1)D (R(2+1)D) [38] network based on ResNet-18 [14]. As [38], [28] and [29] have shown, a standard 3D convolution can be decomposed into two distinct convolution operations; a spatial convolution followed by a temporal convolution. Approximating 3D con-

volution with two separable convolution operations can result in fewer parameters, reducing computational requirements and potentially reducing overfitting.

Given a standard ResNet-18, let N_i denote the number of convolutional filters in the i-th block. The size of the filter in layer i is $N_{i-1} \times d \times d$ where d denotes the spatial height and width. When a 2D ResNet is expanded to (2+1)D, we aim to approximate a full 3D convolution using a 2D spatial filter and a 1D temporal filter. The spatial filter in layer i is of size $N_i \times 1 \times d \times d$ and the temporal filter is of size $M_i \times t \times 1 \times 1$ where t is the temporal extent of the filter. M_i is the number of temporal filters which, following [38], we choose to be:

$$M_i = \left\lfloor \frac{td^2 N_{i-1} N_i}{d^2 N_i - t N_i} \right\rfloor. \tag{1}$$

Which results in the (2+1)D block having approximately the same number of parameters as an equivalent, full 3D convolutional block. This means the model has the same number of parameters as a 3D convolutional model, but with twice the number of nonlinearities (with the extra ReLU applied between the spatial and temporal filters) which should increase the complexity of the function we can model.

2.3 Non-Local Attention

We also experiment with the use of non-local (NL) blocks [42]. A non-local mean operation, originally used for image denoising [3], is defined as:

$$\mathbf{y}_i = \frac{1}{C(\mathbf{x}_i)} \sum_{\forall j} w(\mathbf{x}_i, \mathbf{x}_j) v(\mathbf{x}_j). \tag{2}$$

Where i is an output position, j enumerates all possible positions and $\mathbf{x}$ is the input image or feature map. $w(\cdot,\cdot)$ is a pairwise function that calculates an affinity-relationship or weighting between two positions and $v(\cdot)$ is the value of the image at location j. $C(i)$ is a normalisation factor computed as a sum of weights over all positions in $\mathbf{x}$ as $C(\mathbf{x}_i) = \sum_{\forall j} w(\mathbf{x}_i, \mathbf{x}_j)$. Following the implementation in [42], $v(\cdot)$ is formulated as a linear function, $v(\mathbf{x}_j) = W_v \mathbf{x}_j$, performed as a $1 \times 1 \times 1$ convolution. The weighting function, $w(\cdot,\cdot)$, has multiple forms [37, 3, 42, 39], we use the embedded Gaussian formulation $w(\mathbf{x}_i, \mathbf{x}_j) = \exp((W_\theta \mathbf{x}_i)^T W_\phi \mathbf{x}_j)$ which becomes a *softmax* along dimension j and can be written as $\mathbf{y} = softmax((W_\theta \mathbf{x})^T W_\phi \mathbf{x}) g(\mathbf{x})$.

Non-local attention has been shown to perform best when multiple NL blocks are inserted deeper into the network [42, 44] where they can attend to higher level features. However, [42] also found that placing NL blocks after layers with excessively small spatial size can yield less improvement due to the lack of precise spatial information. Following this, we augment our R(2+1)D network with two NL blocks after the res_3 layer.

2.4 Auxiliary Tasks

We recognise that, in our dataset, there are issues with class imbalance; slices on the forehand and backhand side are hugely underrepresented, together making up approximately 4% of all shots. Using a resampling technique [7], imbalance can be reduced, however, training on oversampled minority classes can still lead to overfitting. We theorise that most errors in mis-classifying minority classes arise due to the similarity of stroke between slices and volleys – the distinction between the two is whether or not the ball has bounced

Model name	NL	Auxiliary task
R(2+1)D	×	×
R(2+1)D + NL	√	×
R(2+1)D + NL + aux	√	√

Table 2: The model variations we consider.

before the shot is hit. When using our cropping technique, in many cases the ball is not in frame, is occluded or is too small to see. Generally, when the player is close to the net, they are more likely to be hitting a volley than when positioned behind the service line. Therefore, knowing *where* a shot is hit can inform the classification of the shot and so, we encourage the model to learn spatial awareness. This is achieved through an auxiliary task where the model predicts which section of the court the player is at the time of contact between the racket and the ball.

We separate the court into three regions: region 1 between the service line and the net, region 2 between the service line and baseline and region 3 behind the baseline. This auxiliary task is likely to encourage the model to look for location-informing features such as the net, or intersections of lines to inform to inform region classification. This auxiliary task is a multi-class classification problem where the loss function we optimise is:

$$\mathscr{L}_{combined} = \mathscr{L}_{primary} + \lambda_{aux}\mathscr{L}_{aux}. \tag{3}$$

where $\mathscr{L}_{primary}$ and $\mathscr{L}_{aux}$ are cross-entropy losses for the primary and auxiliary tasks, respectively, and λ_{aux} is a weight term applied to the auxiliary loss. We use $\lambda_{aux} = 0.5$ after experimenting with different values of λ_{aux} and evaluation on the validation set.

We have described the R(2+1)D backbone model which is used along with additional modules and auxiliary tasks to improve learning. We consider three variants of R(2+1)D, summarised in 2.

2.5 Clip Generation and Augmentation

In our dataset, we know the times at which a shot is hit and for a shot hit during a rally, we consider half the time between the previous shot and the current and half the time between the current shot and the next to be part of the stroke of the current shot. For example, if a player hits a shot at time $t-2$, a shot at t and a shot at time $t+3$, for the shot hit at time t we define the temporal extent of the shot to run from $t-1$ to $t+1.5$. For shots with no previous or next shot, we consider at most two seconds of frames before or after the shot as part of the temporal extent.

We use a model clip size of 32 frames in our experiments and so in most cases will need to uniformly sample every n-th frame in a shot's temporal extent. In most cases, there is at most two seconds between consecutive shots, leading to a temporal extent of $t \pm 2$ and with a video frame rate of 25 frames per second (fps) we have an extent covering 100 frames. To generate a 32 frame clip, we must sample roughly every third frame from the frames in the temporal extent. As the temporal extent of each sample shot is different, we dynamically calculate the stride needed to sample 32 frames when generating each sample clip.

Following [21], we perform augmentation during training to reduce the effects of overfitting. We jitter the temporal extent of samples in time by replacing a static ratio of 0.5 (indicating that the temporal extent

Model Name	Top-1	Top-2
R(2+1)D	82.8	96.5
R(2+1)D + NL	85.5	96.8
R(2+1)D + NL + aux	**85.7**	**97.3**

Table 3: Top-1 and Top-2 rate for each model.

includes exactly half the time between the hit at t and the previous and next hits) with a random ratio, uniformly sampled between 0.3 and 0.8. In the case of our previous example, a ratio of 0.4 for pre-hit extent and a ratio of 0.8 for post-hit extent, the temporal extent of the shot would be between $t-0.8$ and $t+2.4$. This also has the effect of decentering the hit frame from the centre of the clip, thereby encouraging the model to search for hit in the clip.

During training, we also vary the crop around the player by selecting a random (square) crop dimension, performing the crop then resizing it to 224×224, which varies the scale of the player relative to the frame. Frames in clips are also randomly flipped horizontally, a random brightness and contrast are applied and RGB channels are shifted via the addition of a randomly generated value; this can produce different colours of court including the commonly used colours of green and blue, although, we do not explore the applicability of our model on other court types. We also randomly shuffle channels with the aim of changing the court colour to other commonly encountered court colours (blue and green) while also minimising the change in appearance of players. All sampling and augmentations performed with probability, where applicable, are performed consistently on all frames that are part of the same clip. When generating clips for oversampled classes, different augmentations are applied at random.

2.6 Optimisation

We use Stochastic Gradient Descent (SGD) to optimise our models with a mini-batch size of four clips, momentum of 0.9 and weight decay of 0.0005. We use an initial learning rate of 0.01 with learning rate adjusted by hand when validation error ceases to improve.

3 Results

We first look at performance metrics for each model on the dataset before analysing the network predictions and looking at the results for the auxiliary task in the R(2+1)D + NL + aux model.

3.1 Stroke Classification: Quantitative Results

3 shows the top-1 and top-2 rates for all models. We see that the R(2+1)D + NL + aux outperforms both other models. The R(2+1)D + NL model improves on the standard R(2+1)D model's top-1 rate by 2.7%, showing the improvement added by the inclusion of NL blocks while only increasing the number of parameters by 0.8%. The R(2+1)D + NL + aux improves on the top-1 rate by 0.2% and the top-2 rate by 0.5% compared to the R(2+1)D + NL model.

Class	Precision	Recall	F1
Forehand Groundstroke	0.96	0.83	0.89
Forehand Slice	0.52	0.85	0.65
Forehand Volley	0.55	0.46	0.50
Backhand Groundstroke	0.91	0.88	0.90
Backhand Slice	0.85	0.89	0.87
Backhand Volley	0.48	0.44	0.46
Overhead	0.88	0.97	0.92

Table 4: Per-class precision, recall and F1 rates for the R(2+1)D + NL + aux model.

To better analyse performance on imbalanced classes in the dataset, we look at per-class precision, recall and F1 rates for the R(2+1)D model with NL and the auxiliary task in 4. Performance is high across all classes except for volleys; volley classification tends to be the worst performing category with many volleys being misclassified as slices. At the professional level, on both the backhand and forehand sides, these two strokes tend to have similar arcs and slices are often used as set up shots or precursors to volleys. As a result, we cannot rely solely on identifying location information and must also know whether or not the ball has bounced prior to the hit.

We also find that there are a significant number of groundstrokes classified as slices on the forehand side. The strokes in their classical form are not very similar; the topspin or flat forehand makes a long, looping backswing and the slice makes a more compact downward motion. However, in practice there can be several examples of forehand shots hit while running or stretching where the player has no time for a full backswing and may and up slapping or chipping at the ball resulting in a more slice-like backswing and imparting little to no topspin on the ball.

3.2 Stroke Classification: Qualitative Analysis

We now begin a qualitative analysis of our results: we look first at the Uniform Manifold Approximation and Projection (UMAP) [26] dimension reduction of feature vectors used for classification to visualise the topological structure of sample-classes; and we look at Class Activation Maps (CAMs) [46] to interpret classification decisions made by the model. Unless otherwise stated, the UMAP and saliency maps we view are for the best performing R(2+1)D + NL + aux model. CAMs are displayed across eight frames, uniformly sampled from the original 32 clip frame input to the model.

Strokes can be classified into one of three types: forehand-side strokes, backhand-side strokes and overhead strokes. We expect these to be distinct clusters with overheads clearly separated from forehands and backhands. As we include both left-handed and right-handed players in our dataset, and perform horizontal flipping as part of augmentation, we expect forehand-stroke clusters to be closer to backhand-stroke clusters. With a broad idea of what to expect, we can look at the UMAP projection in 1. As expected, forehand-strokes, backhand-strokes and overheads show a great degree of separation. Backhand slices show clear separation from backhand groundstrokes, which we attribute to the prevalence of the two-handed backhand groundstroke and single-handed backhand slice. By comparison, forehand slices tend to be more closely associated with forehand groundstrokes due to their propensity to be hit with a single hand. Forehand and backhand

volleys form their own clusters, however, with several samples of overheads and slices interspersed.

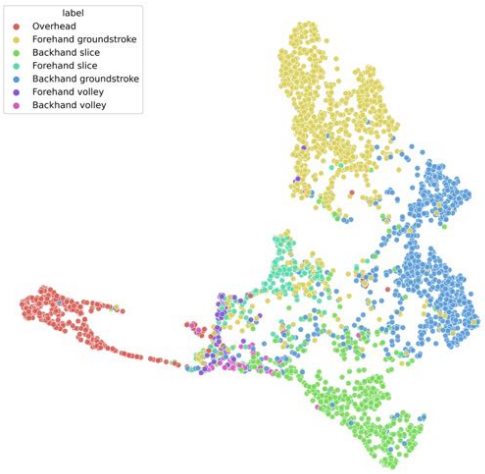


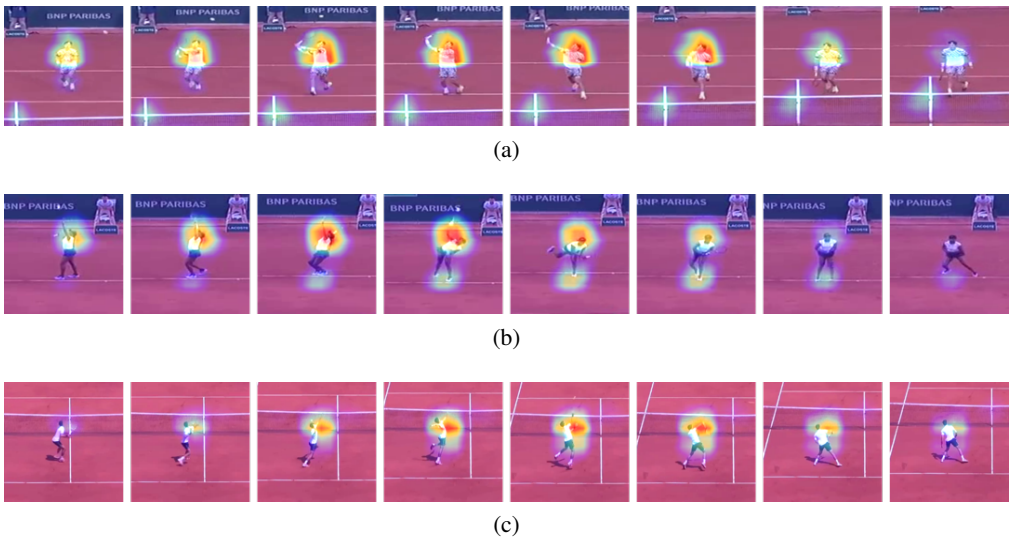
Figure 1: UMAP projection of feature vectors from the R(2+1)D + NL + aux model. Each colour corresponds to a different class, indicated in the key.

The presence of overhead samples so close to volleys is interesting and we look at the CAMs in 2a where the model classifies this stroke as a forehand volley, when in fact it is an overhead. We attribute this incorrect classification to the dubious nature of the stroke – although technically hit overhead, the stroke pattern is more akin to that of a volley. In fact, this shot sits in a *gray zone* where the stroke has aspects of both a forehand volley and an overhead shot. In 2a, we also observe that the model focuses on the area involving the centre service line and the net as an informative region – this region tends to play an important role in the classification of almost all shots predicted as volleys. Through the clip, we see the discriminative region changing in both magnitude and position, with maximum focus in the middle of the clip when around the time of impact and with significantly less attention paid to the final few frames.

From the UMAP projection in 1, we can see that overhead shots are quite different from other shot types and so we expect them to have very different CAMs. 2b shows the CAMs for an overhead (serve) shot where we see the model using, primarily, the stroke features, but in addition, exploiting the consistent positioning of the service stroke just behind the baseline. The model focuses on the feet region of the server near the baseline, suggesting incorporation of wider spatial knowledge of the court. We verify this in 2c where a mid-rally overhead smash is hit and the model focuses only on the stroke, in this case the positioning of the feet is less informative for classification.

1 shows that there is overlap between volley and slice samples. We explore samples in this region by looking at CAMs of backhand slices and backhand slices in 3. 3a shows a volley misclassified as a slice, which is likely due to the lack of bounce information and because the player is in the more ambiguous-shot region between the service line and baseline. 3b shows a correctly classified backhand slice where we note the similarities between the strokes and see that the model correctly focuses on the stroke arcs.

Qualitative evaluation of predictions and their associated CAMs provide a valuable insight into *why* the model classifies an action as it does. We have verified that the model is able to focus on the region of interest spatially and temporally, as well as pick up on wider court cues to inform predictions.



(a)

(b)

(c)

Figure 2: Informative regions for classification of overheads and volleys. (a) and (c) are overhead smashes and (b) is a serve. Analysing the CAMs in these shows correct identification of the stroke and motion cues around the frame of impact, as well as utilisation of wider court features.

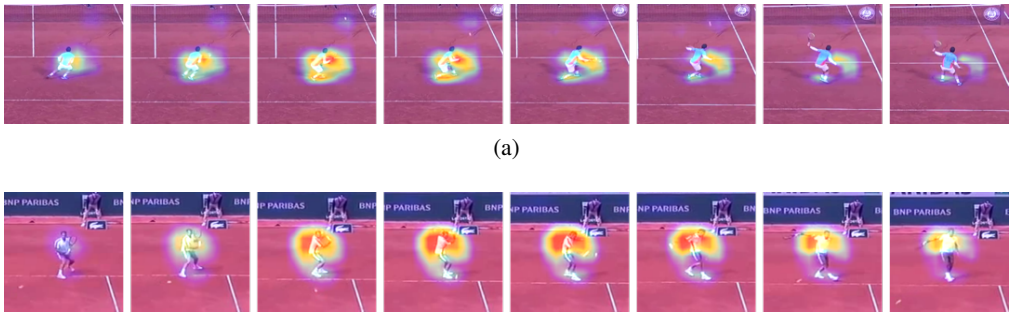
3.3 Auxiliary Task

We evaluate the performance of the R(2+1) + NL + aux model on the auxiliary task. The model achieves an overall 0.93 classification rate with a mean Average Precision (mAP) of 0.84. We visualise the CAMs for region classification (4) to gain insight into the court-features the model uses to classify region. The model, in addition to recognising prominent court-features such as the net or service box (4c), also uses environmental features such as the overlay scorecard (4b) and back of the court (4a). The environmental features, although not part of the tennis court, are consistently found in broadcast video matches and are acceptable features to use. In addition to recognising these features, the model is also aware that the location of the feet around the time of contact between the ball and the racket is most informative for this task and we see focus on the feet and hit region at their maximum in the frames around contact.

4 Conclusion

In this work we have presented a ResNet (2+1)D model based on ResNet-18 to classify seven tennis shots played broadcast video tennis matches.

We have shown that our 3D CNN architecture was able to learn subtle motions that characterise tennis strokes from broadcast video which is characterised by complex, moving backgrounds along with occlusions and non-uniform quality of video. We have performed quantitative and qualitative analysis which showed that classification errors are interpretable with common sources of error stemming from lack of knowledge of ball dynamics or court awareness. The model was trained on video data containing both left-handed and right-handed players, and manages to achieve impressive results, with stroke arcs correctly recognised for both players. Moreover, the model focuses correctly on player stroke arcs, with distinct activation maps



(a)

(b)

Figure 3: Comparing backhand volleys and slices. Both shots share a similar stroke arc and knowledge of whether the ball has bounced beforehand is key to classifying these correctly.

generated for different strokes and for sub-categories of strokes, such as different activation maps for serves and smashes, both of which are contained in the overhead category.

Analysing activation maps also yielded valuable insight into how the model makes a decision when faced with similar stroke patterns – the model was shown to pay attention to wider frame features that indicate location such as the net and scorecard. However, despite the model achieving some spatial awareness, we found performance on volleys to be reduced due to the lack of ball awareness. Following this, we suggest future research to explore strategies that explicitly incorporate ball dynamics, either as quantitative tracking data or multi-scale, multi-stream networks that can capture player and ball interactions.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016.

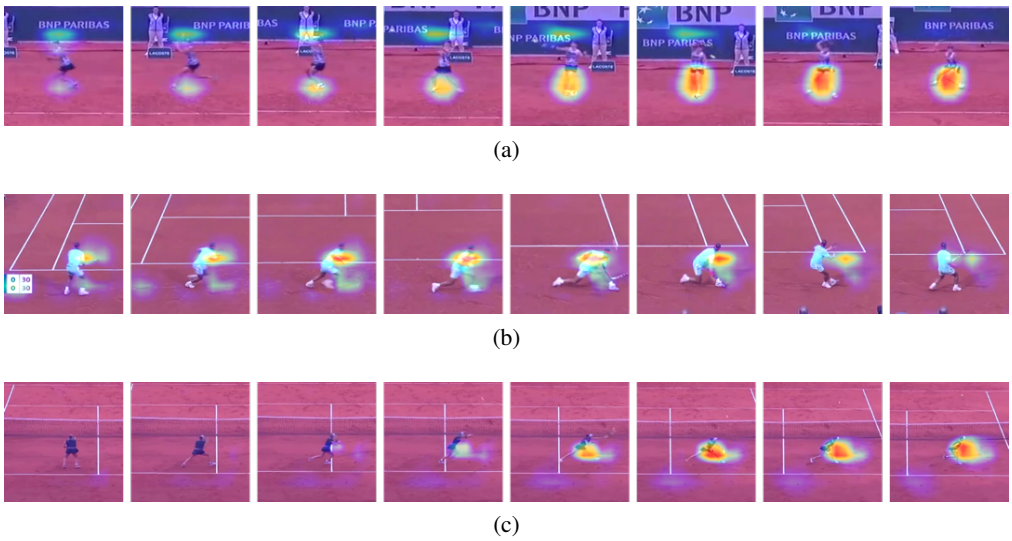
[3] A. Buades, B. Coll, and J. Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65 vol. 2, 2005.

[4] Jiaxin Cai and Xin Tang. Rgb video based tennis action recognition using a deep historical long short-term memory. *Journal*, page arXiv:1808.00845, 2018.

[5] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, and Wei Xu. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2956–2964, 2015.

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[7] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.



(a)

(b)

(c)

Figure 4: Class Activation Maps for location prediction. A range of features are used to make this classification, combining both the stroke style, the hit frame in the clip and location cues.

[8] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *Journal*, page arXiv:1411.4389, 2014.

[9] Nazli FarajiDavar, Teófilo De Campos, Josef Kittler, and Fei Yan. Transductive transfer learning for action recognition in tennis games. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1548–1553. IEEE, 2011.

[10] H. Faulkner and A. Dick. Tenniset: A dataset for dense fine-grained event recognition, localisation and description. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8, 2017.

[11] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. *Journal*, page arXiv:2004.04730, 2020.

[12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *Journal*, page arXiv:1812.03982, 2018.

[13] S. Gourgari, G. Goudelis, K. Karpouzis, and S. Kollias. Thetis: Three dimensional tennis shots a human action dataset. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 676–681, 2013.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[17] Matthew Hutchinson and Vijay Gadepally. Video action understanding: A tutorial. *Journal*, page arXiv:2010.06647, 2020.

[18] Max Jaderberg, Karen Simonyan, and Andrew Zisserman. Spatial transformer networks. *Advances in neural information processing systems*, 28:2017–2025, 2015.

[19] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2013.

[20] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, 2017.

[22] Kaustubh Milind Kulkarni and Sucheth Shenoy. Table tennis stroke recognition using two-dimensional human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4576–4584, 2021.

[23] Pierre-Etienne Martin, Jenny Benois-Pineau, Boris Mansencal, Renaud Péteri, Laurent Mascarilla, Jordan Calandre, and Julien Morlier. Sports video classification: Classification of strokes in table tennis for mediaeval 2020. In *MediaEval 2020 Workshop*, 2021.

[24] Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, and Julien Morlier. Fine grained sport action recognition with twin spatio-temporal convolutional neural networks. *Multimedia Tools and Applications*, 79(27):20429–20447, 2020.

[25] Pierre-Etienne Martin, Jenny Benois-Pineau, Renaud Péteri, and Julien Morlier. Three-stream 3d/1d cnn for fine-grained action classification and segmentation in table tennis. In *Proceedings of the 4th International Workshop on Multimedia Content Analysis in Sports*, pages 35–41, 2021.

[26] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[27] N. Owens, C. Harris, and C. Stennett. Hawk-eye tennis system. In *2003 International Conference on Visual Information Engineering VIE 2003*, pages 182–185, 2003.

[28] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. *Med Image Comput Comput Assist Interv*, 16(Pt 2):246–53, 2013.

[29] R. Rigamonti, A. Sironi, V. Lepetit, and P. Fua. Learning separable filters. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2754–2761, 2013.

[30] Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.

[31] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[32] Hitesh Shah, Prakash Chokalingam, Balamanohar Paluri, Nalin Pradeep, and Balasubramanian Raman. Automated stroke classification in tennis. Image Analysis and Recognition, pages 1128–1137. Springer Berlin Heidelberg, 2007.

[33] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[34] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014.

[35] Khurram Soomro and Amir R Zamir. *Action recognition in realistic sports videos*, pages 181–208. Springer, 2014.

[36] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *Journal*, page arXiv:1409.4842, 2014.

[37] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pages 839–846, 1998.

[38] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[40] Silvia Vinyes Mora and William J Knottenbelt. Deep learning for domain-specific action recognition in tennis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 114–122, 2017.

[41] Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5):2368–2378, 2017.

[42] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *Journal*, page arXiv:1711.07971, 2017.

[43] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[44] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. *Journal*, page arXiv:2006.06668, 2020.

[45] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. *Journal*, page arXiv:1503.08909, 2015.

[46] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[47] Guangyu Zhu, Changsheng Xu, Wen Gao, and Qingming Huang. Action recognition in broadcast tennis video using optical flow and support vector machine. In Thomas S. Huang, Nicu Sebe, Michael S. Lew, Vladimir Pavlović, Mathias Kölsch, Aphrodite Galata, and Branislav Kisačanin, editors, *Computer Vision in Human-Computer Interaction*, pages 89–98. Springer Berlin Heidelberg, 2006.

[48] Guangyu Zhu, Changsheng Xu, Qingming Huang, Wen Gao, and Liyuan Xing. Player action recognition in broadcast tennis video with applications to semantic analysis of sports game. In *Proceedings of the 14th ACM international conference on Multimedia*, page 431–440. Association for Computing Machinery, 2006.

[49] Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R. Manmatha, and Mu Li. A comprehensive study of deep video action recognition. *Journal*, page arXiv:2012.06567, 2020.

Using geostatistics to model and visualize batting ability in baseball

D. Sylvan*

*Department of Mathematics and Statistics,
Hunter College of the City University of New York,
695 Park Avenue, New York, NY 10065, USA,
+ email address: dsylvan@hunter.cuny.edu

Abstract

Many fields of applied research employ geostatistical methods to analyze spatial patterns in data with complex structures. In recent years, with the advent of big data and the emerging field of data science, extensive statistical literature has been devoted to baseball data, however, comparatively less is known about the utility of geostatistical techniques aiming to produce accurate and comprehensive heat maps. The freely available Sportvision PITCHf/x data provides continuous location coordinates for individual pitches using high-speed cameras. This detailed spatial information can be employed to visualize a batter's ability across regions in and around the strike zone.

In this note we summarize classical geostatistical methodology, show how it can be applied to real data, and present comprehensive heat maps based on Major League Baseball pitches from 2006 to 2018. The stochastic process underlying batting ability is assumed to be a spatial Gaussian field with isotropic covariance that is estimated from the aforementioned data. We then use spatial interpolation to obtain best estimates of heat maps of batting ability for individual players. Uncertainty in these estimates is assessed through conditional simulations, and resulting percentile heat maps are displayed for illustration.

1 Introduction

The recent advent of big data and the overarching research field of data science lead to an increased focus on mathematical modeling and statistical inference for sports data in general, and baseball in particular. As a consequence, there have been large contributions on models and software development coming form many fields of applied research, including but not limited to economics, engineering, environmental science, mathematics and statistics. For an overview of methods and models we refer to the textbooks Miller (2016) for general sports applications, and to Marchi, Albert and Baumer (2019) for baseball applications. Baseball is considered America's national pastime, and the abundance and variety of freely available baseball data makes the sport a favorite for teachers and students alike. Baseball data contains spatial information connected to location coordinates for individual pitches captured by high-speed cameras. This information can be used to produce accurate estimates of hot zones (areas around the strike zone in which batters are likely to hit the

ball well) for individual players. In this note we show how geostatistics methodology involving modeling of spatial dependence can be used to accurately estimate the distribution of batting ability for individual players. Percentile estimates of the heat maps that can accurately identify "hot" and "cold" zones for each batter are shown for illustration. These maps can be useful to batting coaches to help batters identify weaknesses and correct their swings. More importantly, pitching coaches may use these visuals to develop strategies to minimize the performance of opposing batters.

2 Geostatistics and baseball

Spatial processes have been increasingly analyzed in recent years due in part to the explosive growth and affordability of computing capabilities. Many fields of applied research employ geostatistical methods to analyze spatial patterns in data with complex structures. In baseball, a hot zone is defined as the area in which batters are likely to hit the ball well. A visual frequently used in the past consists of 3×3 or 5×5 grids of the strike zones displaying the batting averages of individual players in the respective cells. Baumer and Draghicescu (2010) identify shortcomings of such visuals and propose new estimates of hot zones obtained via bivariate kernel smoothing and spatial interpolation. Over the last decade, a wealth of information, commentary and resources can be freely accessed online, see for example `https://www.statsperform.com`, as well as numerous blogs and podcasts. Concerning geostatistcal applications in baseball, we refer to Sylvan and Cross (2019).

2.1 Open source baseball data

Sportvision has tracked Major League Baseball pitches using a system named PITCHf/x that uses two cameras, one behind home plate and the other behind first base capturing roughly 20 images of each pitch on its path to the plate. From these images the entire path of the ball is reconstructed. From 2008 to 2016, Major League Baseball Advanced Media (MLBAM) recorded the results of each pitch in real time, classified each pitch based on its velocity and movement. From 2015 through the present, MLBAM has used Statcast to track pitches. Statcast uses "a combination of two different tracking systems – a Trackman Doppler radar and high definition Chyron Hego cameras. The radar, installed in each ballpark in an elevated position behind home plate, is responsible for tracking everything related to the baseball at 20,000 frames per second. This radar captures pitch speed, spin rate, pitch movement, exit velocity, launch angle, batted ball distance, arm strength, and more.(http://m.mlb.com/glossary/statcast)." Our analysis is based on PITCHf/x, Statcast and MLBAM data provided by Fangraphs.com.

2.2 Framework

Let $Z(s)$ be the random field of batting ability (fixed player) defined for $s \in D \subset \mathbb{R}^2$. In order to map the random field Z we need to determine $Z(s_0)$ for any point $s_0 \in D$. Assume that $\{s_1, s_2, \ldots, s_n\}$ is the set of points where the process is observed. The data is thus the collection $\{Z(s_1), Z(s_2), \ldots, Z(s_n)\}$. To keep notation simple and without confusion we will refer to the sampled values as $Z_i = Z(s_i), i = 1, 2, \ldots, n$. We assume that the Z_i's are a realization of a second-order isotropic random field, meaning that the underlying process has constant mean and the covariance between any two locations depends only on the distance between them: $E(Z(s)) = \mu$ for all $s \in D$, and $\text{Cov}(Z_i, Z_j) = C(||s_i - s_j||)$ for all $s_i, s_j \in D$. Let us denote the $n \times n$ covariance

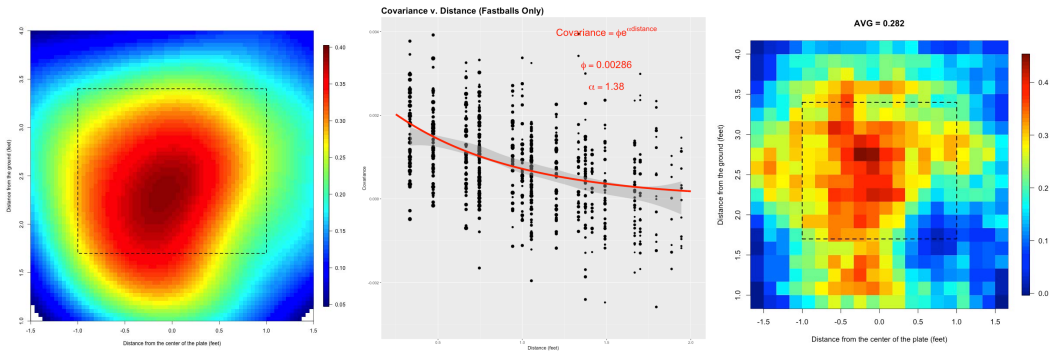


Figure 1: Estimated league batting average (left), estimated covariance function (middle), and one realization of a simulated batter (right).

matrix associated with these observations by C, where $C_{ij} = \text{Cov}(Z_i, Z_j)$. The best linear unbiased predictor (BLUP) of $Z(s_0)$ is then obtained as a linear combination of the observations $Z^*(s_0) = \sum_{i=1}^n \lambda_i Z_i$, such that $\sum_{i=1}^n \lambda_i = 1$. The weights λ_i are completely specified by the covariance matrix C, assumed to be known (since it is built by using the deterministic covariance function). If $c_0 = C(s_i, s_0)$, then the vector of weights is computed as $\Lambda = (\lambda_1, \ldots, \lambda_n)^T = C^{-1}c_0$, with corresponding prediction variance $\sigma(s_0) = C(0)\lambda(s_0)c(0)$. For computational details connected to this procedure (known as *ordinary kriging*) we refer to Chilès and Delfiner (1999). In real life, the mean and covariance of the underlying process are unknown and need to be estimated from the same data, making the BLUP an estimated or empirical BLUP, or EBLUP. This added error is hard if not impossible to derive analytically. Resampling and conditional simulations are typically used to correct errors of EBLUPs, see next subsection for an example.

2.3 Heat maps of batting ability

We assume that the stochastic process underlying batting ability is Gaussian having the overall league batting average as mean function and a deterministic, two-parameter covariance function $C(i, j) = \phi e^{-\alpha \|s_i - s_j\|}$. Sylvan and Cross (2019) provide a thorough analysis and interpretation of this exponential decay model. In practice, we use maximum likelihood estimates $\hat{\phi}$, $\hat{\alpha}$, and $\hat{\sigma}_0$ instead of the unknown true parameters, and therefore the resulting kriging errors need to be adjusted to correct for these additional sources of uncertainty. For this reason we employ a conditional simulation scheme, see Chapter 6 in Stein (1999) for a comprehensive discussion on kriging with estimated parameters.

We simulate batters as realizations of the aforementioned multivariate normal distribution with batting ability generated on a 20×20 grid around the strike zone. Figure 1 shows the estimated league average, the estimated spatial covariance, and a simulated batter. Pitches were generated as Bernoulli random variables with probability of a hit being determined as the simulated batter's ability at each grid location. For a full description of this procedure we refer to Sylvan and Cross (2019). For visualization we show "spatial boxplots" for individual players. This visual tool that can be construed as a spatial analog of the boxplot by displaying the 10's 25th, 50th, 75th and 90th percentile heat maps as an analog to five-number summary, see Figure 2. More examples can be found in Cross and Sylvan (2015) and Sylvan and Cross (2019).

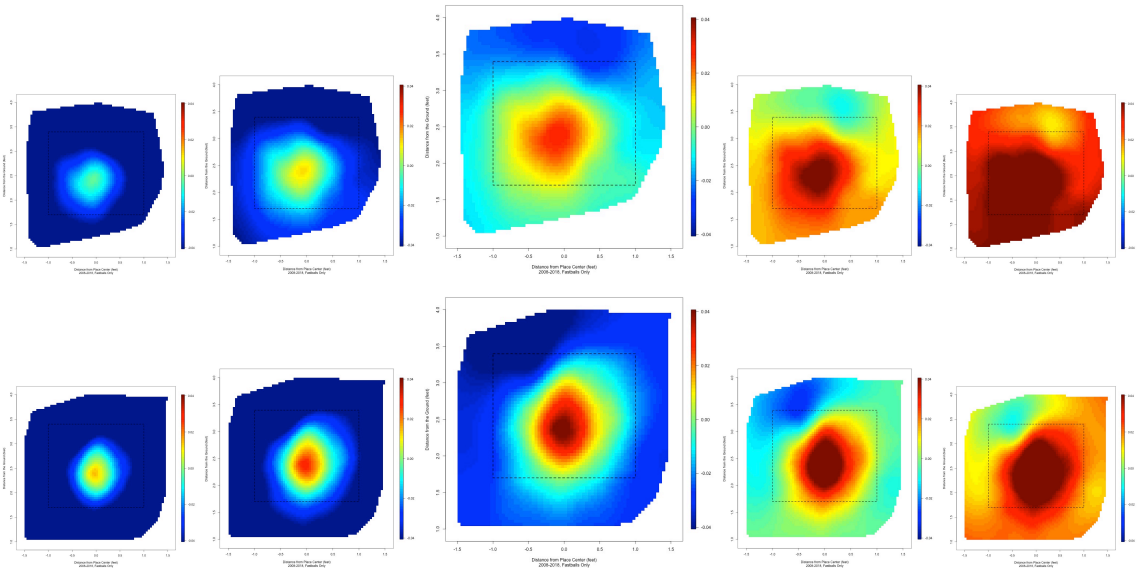


Figure 2: "Spatial boxplots" based on five-heat map summaries – 10's, 25's, 50's, 75's, 90's percentiles, respectively: Kris Bryant of the Colorado Rockies (top), Ian Kinsler of the Texas Rangers (bottom).

3 Discussion

In this note we summarized an approach aimed to estimate the spatial distribution of batting ability for individual players. An analog of Tukey's five-number summary displaying five percentile heat maps is given for illustration. The framework and methodology came from geostatistics, the stochastic process underlying batting ability assumed to be a spatial Gaussian random field with isotropic covariance. The resulting "spatial boxplots" are useful for understanding each batter's strengths and weaknesses and may be decisive in winning contests between batters and pitchers which are at the core of any baseball game. The abundance of freely available baseball data and software made possible to implement and run data-driven algorithms and statistical simulations to improve these spatial predictions. This comprehensive approach provides a fast, accurate, and informative exploratory tool for detection of spatial patterns. Moreover, the visual tool presented here may be used in other fields under the data science umbrella, including but not limited to education, environmental sciences, epidemiology, finance, public health.

Acknowledgement: I would like to thank my former student, Jared Cross, for implementing the statistical methodology and for running the simulations and applications.

References

B. Baumer and D. Draghicescu (2010). Mapping batter ability in baseball using spatial statistics techniques. JSM, American Statistical Association, 3811–3822.

J. Chilès and P. Delfiner (1999). Geostatistics. Modeling Spatial Uncertainty. Hoboken, New Jersey: John

Wiley & Sons, Inc.

J. Cross and D. Sylvan (2015). Modeling spatial batting ability using a known covariance matrix. Journal for Quantitative Analysis of Sports, Volume 11 (3), 155–167. DOI 10.1515/jqas-2014-0089

M. Marchi, J. Albert and B. Baumer (2019). Analyzing baseball data with R. Chapman & Hall.

T. W. Miller (2016). Sports analytics and data science. Winning the game with methods and models. Pearson Education, Inc.

M. L. Stein (1999). Interpolation of spatial data: some theory of kriging. Springer.

D. Sylvan and J. Cross (2019). Using geostatistical techniques to improve heat maps of batting ability. Chapter 5 in Essential Topics in Baseball: From Performance Analysis to Injury Prevention, editor Erik Welch, Nova Publishers, 173–183.

Forecasting Football Match Result with GAP Rating and Player Rating

Calvin C. K. Yeung*

*The Chinese University of Hong Kong, calvinckyeung@link.cuhk.edu.hk

Abstract

Tree Boosting Models appears to be the best performance model at this time in the development of football match results forecast. Nonetheless, there is still room for improvement in feature engineering. Most studies have used historical match statistics as one of the most important features. It can, however, be replaced with the Generalised Attacking Performance (GAP) ratings, which predict non-rare match statistics and improves the model's performance. We attempted to explore the performance of combining the Boosting Tree Model and GAP rating in this study, as well as resolve the limitation of the GAP rating. In addition, we attempted to propose an alternative method for predicting match statistics using common machine learning models and player ratings. Using 5 years of data, over 1700 matches from the premier league, the result shows that in match statistics prediction, the GAP rating performs the best followed by the historical average then the player rating and machine learning approaches. As in match result forecast, the purposed two-approach outperformed the betting odds and each of them has various pros and cons that allow a football team to adopt for coach's tactical decision making.

4 Introduction

Forecasting football match results is a popular topic in football research. However, many approaches have ignored the practical usage for a team, whether to view match result forecast as a statistical modeling challenge or to profit and arbitrage from betting. With the advancement of machine learning and open-source data collection in football, developing a model that forecasts match results and provides practical utility to a team is now more feasible than ever.

There are two main approaches to forecasting football match results. The first approach is statistical modeling, which can be traced back to 1997 [18]. Recent methods include fitting independent Poisson [2], Negative Binomial, or Generalized Extreme Value distributions [4, 5, 10, 15]. The Bradley-Terry-Elo model [12] is also worth mentioning. The second approach is machine learning models. Common methods include Naive Bayes, Tree Model, Regression and SVM. When the above models are compared, it is possible to conclude that tree boosting methods (e.g., XGBoost) have the best performance at the current time [1, 9, 11, 16, 17, 19]. As a result, data engineering is a more feasible research direction for improving performance further.

The majority of studies have relied heavily on the historical match statistics feature. Wheatcroft [20] has proposed the Generalised Attacking Performance (GAP) ratings to predict nonrare match statistics and proven

success. However, the GAP rating performance other than with linear regression is unknown and the GAP rating excluded several matches (e.g., the beginning and ending stages) in forecasting match results, where teams' quality and motivation may differ in those matches.

In this study, by extending the idea of predicting match statistics, we propose the player rating with machine learning models (linear regression and artificial neural network) to predict match statistics and verified the performance of various methods. Moreover, since shot on target and shot off target are rare in most matches, the combination of them was also predicted and separated as two sets of features for forecasting the match results.

In forecasting the match results, we used all matches and investigated the effect of change in the first 6 and last 6 stages on forecasting performance. We provide the model with more information about the teams' quality (such that it gives desirable forecasting performance even in the first 6 and last 6 stages) by incorporating the use of GAP rating, player rating, player coordinate, and stage into XGBoost. In addition to the approach described above, we stack other regressors for match statistics prediction and XGBoost. The performance of the two approaches was then compared using betting odds. Finally, we will discuss how a team can utilize the proposed approaches.

The following are the main contributions of this work: (i) the proposed method allows for a higher degree of model interpretation and allows teams to utilize it, which was not emphasized in previous studies; (ii) We extend the concept of using predicted non-rare match statistics in the forecast of match results. Methodologically, we proposed using player ratings to predict match statistics with common machine learning model. We validate the proposed methods and show that they outperform the betting odds. In addition, we examined how the rare match statistics could be combined to improve the model's performance. We also show how different stages of the seasons affect the model's performance.

5 Materials and Methods

Dataset. Various open-source databases have been evaluated in this study. European Soccer Database, Betting Sites, engsoccerdata, and The 2017 Soccer Prediction Challenge dataset are all included. We examined the dataset based on the volume of data and the features that were available. Finally, the European Soccer Database [13] from Kaggle is the best database for this study. Despite the fact that it only contains data from 2008 to 2016, it offers unique features such as player rating and player coordinate. Nonetheless, due to a formatting error, we have replaced the detailed match statistics with the European Soccer Database Supplementary Database [22]. For the sake of simplicity, we only consider data from the English Premier League (EPL) seasons 2011/2012 to 2015/2016.

Data Prepossessing. Starting with FIFA rating and player coordinate data, we apply the Carpita, Ciavolino, and Pasca methods [7] to handle such data with a twist. Instead of taking the average, we use aggregation to account for player numbers advantage. As a result, we will have a set of ratings for each role (i.e., attacker, midfielder, defender, and goalkeeper), as shown in Table 1. Furthermore, outfield players rarely serve as goalkeepers and vice versa. Goalkeeping rating is dropped for defenders, midfielders, and attackers, while Skill, Attacking, and Defending ratings are dropped for goalkeepers. From now on, we will refer to these ratings as the team's ratings.

Moreover, as Wheatcroft [20] has pointed out, a rare event does not provide much information to the models in match statistics. As a result, red cards, yellow cards, and fouls will be ignored, while corners and throw-ins will be combined into crosses as in the previous study. We further investigated how the rarity of

Ratings	Features from "Player_Attributes" table
Power (POW)	shot_power, jumping, stamina, strength, long_shots
Mentality (MEN)	aggression, interceptions, positioning, vision, penalties
Skill (SKI)	dribbling, curve, free_kick_accuracy, long_passing, ball_control
Movement (MOV)	acceleration, sprint_speed, agility, reactions, balance
Attacking (ATT)	crossing, finishing, heading_accuracy, short_passing, volleys
Defending (DEF)	marking, standing_tackle, sliding_tackle
Goalkeeping (GOL)	gk_diving, gk_handling, gk_kicking, gk_positioning, gk_reflexes

Table 1: Team's Ratings Construction from FIFA Rating Features [7]

shot-on target and shot-off target affect models' performance in this study because these two statistics are uncommon in our sampled data set. We created two datasets because we are unsure whether combining the two as shot is a better option. Besides that, because of missing data, feature possession and matches with missing player ratings have been dropped. There are 1784 matches left at the end (793 wins and 991 draws or loses).

Finally, Wheatcroft [20] has excluded several matches (e.g., the beginning and ending stages) in forecasting match results, where teams' quality and motivation may differ. In this study, we attempted to overcome this limitation by providing more information to the model, specifically the stage of the match.

To summarize, we have two datasets: **Dataset 1**: team's rating (groups of FIFA rating for each of the four roles), match statistic (shot-on, shot-off, and cross), and stage of the season from 1 to 38. **Dataset 2**: replace shot-on, shot-off in Dataset 1 with shot. Where the binary target variable is set to 1 if the home team wins and 0 if the home team draws or loses.

Proposed Approach. The match result forecast is divided into two parts. First, we predict the match statistics, and then we forecast the match result using the predicted match statistics and unused features. In the first part, we introduced two additional models in addition to the two models used by Wheatcroft [20], rolling historical average (AVG) and GAP rating [21] (GAP). The models are built on the premise that a team with a high attacker rating versus a team with a low defender rating is more likely to have more shooting opportunities, resulting in higher match statistics. As a result, we attempted to model the linear or nonlinear relationship between team rating and match statistics using linear regression (LR), linear regression with elastic net (LRE), and deep artificial neural network (ANN). We feed the ANN with the entire dataset, unlike AVG, GAP, and LR, which model each match statistic independently. Hence, we name the ANN that trains with Dataset 1 as ANN1 and the ANN that trains with Dataset 2 as ANN2.

As previously stated, the best model for part two forecasting match result is the XG Boosting Tree Model; the remaining question is what to feed into the model. It is obvious that the predicted match statistics and game stage will be the highlights. This is true for the LR and ANN methods; however, the AVG and GAP methods do not use the team's ratings when predicting match statistics. At this part, the team's ratings will be used for the latter two models (fit into the XG Boosting Tree).

Models Training Method. The GAP rating will be trained using Excel Solver, while the rest of the models will be trained using Python 3.7.10 with the sklearn package, functions LinearRegression(), ElasticNet(), and MinMaxScaler() with MLPRegressor(solver="adam", max_iter=500) (as the default maximum number of iterations is set to 200 and does not converge well) for LR,LRE and ANN respectively. The package xgboost

with the function xgb.XGBClassifier(objective="binary:logistic") is used for XG Boosting Tree.

To calibrate the models, random search and grid search are commonly used. However, when random search is used, Baboota and Kaur [3] have pointed out that the global minima will be skipped in similar task, hence, grid search will be used; in sklearn, the function GridSearchCV() will be used. The function consists of cross-validation, however, it has been disabled to avoid look-ahead bias.

In the grid search of LRE, "alpha" and "l1_ratio" are tuned. In ANN "hidden_layer_sizes"and "activation" ("tanh", "relu","logistic",) are tuned, for simplicity, maximum number of hidden layers is set to 3. In XG Boosting Tree, "gamma" the minimum gain, "learning_rate", "max_depth" and "n_estimators" are tuned, for simplicity, maximum depth is set to 5 and max number of estimators is set to 300.

In Training Set and Validation Set, we have to avoid the look-ahead bias. Seasons 2011/2012 and 2012/2013 are used as the first training set, season 2013/2014 as the first validation set, season 2012/2013 and 2013/2014 as the second training set, and season 2014/2015 as the second validation set to predict future match statistics. The average performance across both validation sets will be reported. In match result forecasting, the prediction on the two validation sets is used as the training set, with the season 2015/2016 serving as the validation set.

Models Evaluation Method. To assess the performance of match statistics prediction, we use MAE (as in [20]) and RMSE (to penalize predictions that deviate far from the target value). In this part, the AVG model serves as a benchmark model.

Brier Score, Log Loss, F1 Score, and AUC-ROC will be reported for match result forecasting. By reporting these common proper and improper scoring rules, we can compare the results to other studies if necessary. In this part, betting odds (ODDS) will be used to benchmark. Bet365 was chosen as a betting odds provider because it has been used in previous studies. The betting odds are converted to the probability of the home team winning. Furthermore, we hope to validate the need for the GAP rating and how much information the team's rating can solely provide. Therefore, in addition, the performance of the XG Boosting Tree model using only the team's rating (TR) will be reported.

6 Results

Models Calibration Result In predicting match statistics, we applied the models multiple times for each match statistic, each with a different set of optimal parameters. In the following, we will only provide a broad range for the optimal parameter. GAP with Lambda of 0.5-0.7 and Phi 1 and 2 of 0.5-0.6. LRE is equivalent to LR where "alpha" and "l1_ratio" are both 0. ANN with ReLu activation function, 1 hidden layer with 10 nodes; however, many other architectures provides comparable performance. Furthermore, ANN predicts with a constant value for all predictions, even when both the many to one and many to many approaches are tested, hence, it is ignored in the next part. Finally, match result forecast with the XG Boosting Tree, minimum gamma 0, learning rate 0.1/0.2, maximum depth 3/4, and number of boosting rounds 200/300 gives the best result.

Match Statistics Prediction Because MAE and RMSE produce similar results, it has no bearing on interpretation, and due to space constraints, only the RMSE will be reported in this study, as shown in Table 2. When the RMSE of different approaches is compared, GAP has the best overall performance and all other approaches outperform AVG. Demonstrating that AVG is the worst choice in most cases. Meanwhile, LR and ANN are positioned in the middle and produce similar results. Additionally, it is difficult to determine whether the use of Shot (Dataset 1) or Shoton and Shotoff (Dataset 2) would be more ideal based on the above

results; in such cases, both methods will be retained for the next part.

Match Statistics	AVG	GAP	LR	ANN1	ANN2
Home_Shoton	7.2291	**6.7712**	7.0166	7.0913	N/A
Home_Shotoff	5.9626	**5.8791**	6.0547	6.0177	N/A
Home_Shot	10.7825	**9.7472**	10.3016	N/A	10.5842
Home_Cross	18.6025	17.7807	17.8424	17.5620	**17.4969**
Away_Shoton	5.8085	**5.6760**	5.7547	5.8428	N/A
Away_Shotoff	5.5969	5.2847	5.2825	**5.2755**	N/A
Away_Shot	9.3732	**8.4792**	8.7469	N/A	9.0031
Away_Cross	15.6025	15.2452	15.3436	14.9671	**14.9336**

Table 2: Future Match Statistics Prediction Total RMSE

Match Result Forecast From Table 3, comparing the use of shot (Dataset 1) or shoton and shot off (Dataset 2), GAP1 has the highest improper scores, implying that it gives a better prediction in terms of result. Whereas GAP2 has the lowest proper score, implying that it gives a better estimation in terms of distribution. LR1 outperforms LR2 in both proper and improper scoring. As a result, the dataset should be selected based on the motivation and method chosen.

When comparing different models, the ranking for estimating the distribution (proper scoring) are GAP2, ODDS, TR, and LR1. Whereas the ranking for result prediction (improper scoring) are GAP1, LR1, TR, and ODDS. Besides, using Table 4, we can see how the approaches perform when the first and last six stages of the season are removed, as in [20]. GAP's performance has increased significantly, while LR's has increased slightly and even decreased slightly. This demonstrates that differences in team quality and motivation have less of an impact on the LR approach. While the additional features stage does not aid the GAP approach in resolving the issue.

Models	Brier Score	Log Loss	F1 Score	AUC-ROC
GAP1	0.2793	0.7919	**0.4984**	**0.5580**
GAP2	**0.2771**	0.7798	0.4889	0.5475
LR1	0.3066	0.9568	0.4702	0.5446
LR2	0.3181	1.0483	0.3864	0.5218
ODDS	0.2797	**0.7631**	0.3897	0.5141
TR	0.2970	0.8166	0.4438	0.5015

Table 3: Match Results Forecast Performance
(GAP1, LR1 utilize the Dataset 1 and GAP2, LR2 utilize the Dataset 2)

Models Interpretation Unfortunately, the group of ratings for the attacker and midfielder within the team's ratings are highly correlated, causing the multicollinearity problem in predicting match statistics. (For example, in the LR model for home team shoton, the coefficient for home team attacker rating is -0.0943. This implies that a higher attacker rating has a significant negative effect on the number of shots on the team, which is illogical.) The F-score of the GAP 2 (GAP 1) approach, on the other hand, indicates that the home

Models	Brier Score	Log Loss	F1 Score	AUC-ROC
GAP1	0.2707	0.7678	0.5268	0.5819
GAP2	0.2655	0.7534	0.5268	0.5819
LR1	0.3080	0.9473	0.4821	0.5419
LR2	0.3276	1.0280	0.3627	0.4943
ODDS	0.2761	0.7538	0.3834	0.5102
TRO	0.2912	0.8014	0.4766	0.5210

Table 4: Match Results Forecast Performance without Stage 1-6 and 33-38

team and away team shot (shot-off) are the most important features in forecasting match results. Explaining that the team that creates more opportunities is more likely to win.

7 Discussion and Conclusion

In this study, we proposed two methods for forecasting football match results using GAP ratings and player ratings. The findings indicate that the GAP and LR approaches are useful models that outperform the betting odds. Both approaches allow the team to test out their formation prior to the game in order to optimize their game strategy. Even though GAP has the best performance, LR's property, which is independent of the team's match result history in forecasting, allows the team to test out hypothetical formation and consistent performance throughout the season makes it appealing.

At the same time, we validated that Historical average is a suboptimal match statistics prediction and that the merge of shot on target and shot off target is depends on the method applied. Even though using FIFA ratings caused the multicollinearity problem, it needs to be verified whether this is due to the nature of player ratings or simply an isolated case.

According to Table 3, the AUC-ROC of all models is between 0.56 and 0.5, which is close to a random classifier. There are two possible explanations: the nature of the sport consists of a large number of random events, and the approaches require more features and information. As a result, in future research, more features (e.g., player's form [14], chemistry [6], and alternative rating [8]) can be considered to improve performance. Alternatively, using the more complex version of the model, XGBoosting Forest.

Acknowledgments

The author would like to express thanks and gratitude to Prof. Keisuke Fujii for his patient guidance, encouragement, and constructive criticism of this research work. This work was supported by JSPS KAKENHI (Grant Number 20H04075).

References

[1] Y. F. Alfredo and S. M. Isa. Football match prediction with tree based model classification. *I.J. Intelligent Systems and Applications*, 7, 2019.

[2] H. R. Azhari1, Y. Widyaningsih, and D. Lestari. Predicting final result of football match using poisson regression model. *Journal of Physics: Conference Series*, 1108, 2018.

[3] R. Baboota and H. Kaur. Predictive analysis and modelling football results using machine learning approach for english premier league. *International Journal of Forecasting*, 35:741–755, 2019.
[4] E. Bittner, A. Nußbaumer, W. Janke, and M. Weigel. Football fever: goal distributions and non-gaussian statistics. *arXiv:physics/0606016*, 1, 2006.
[5] E. Bittner, A. Nußbaumer, W. Janke, and M. Weigel. Self-affirmation model for football goal distributions. *arXiv:0705.2724*, 1, 2007.
[6] L. Bransen and J. V. Haaren. Player chemistry: Striving for a perfectly balanced soccer team. *arXiv:2003.01712*, 1, 2020.
[7] M. Carpita, E. Ciavolino, and P. Pasca. Exploring and modelling team performances of the kaggle european soccer database. *Statistical Modelling 2019*, 19:74–101, 2019.
[8] T. Decroos, L. Bransen, J. V. Haaren, and J. Davis. Actions speak louder than goals: Valuing player actions in soccer. *arXiv:1911.08138*, 2, 2019.
[9] E. Eryarsoy and D. Delen. Predicting the outcome of a football game: A comparative analysis of single and ensemble analytics methods. *52nd Hawaii International Conferenceon System Sciences*, 2019.
[10] J. Greenhough, P. C. Birch, S. C. Chapman, and G. Rowlands. Football goal distributions and extremal statistics. *arXiv:cond-mat/0110605*, 2, 2002.
[11] A. Groll, C. Ley, G. Schauberger, and H. V. Eetvelde. Prediction of the fifa world cup 2018 – a random forest approach with an emphasis on estimated team ability parameters. *arXiv:1806.03208*, 3, 2018.
[12] F. J. Király and Z. Qian. Modelling competitive sports: Bradley-terry-Él˝o models for supervised and on-line learning of paired competition outcomes. *arXiv:1701.08055*, 1, 2017.
[13] H. Mathien. European soccer database. `https://www.kaggle.com/hugomathien/soccer`, 2016.
[14] M. Otting and A. Groll. A regularized hidden markov model for analyzing the "hot shoe" in football. *arXiv:1911.08138*, 1, 2019.
[15] M. Petretta, L. Schiavon, and J. Diquigiovanni. Mar-co: a new dependence structure to model match outcomes in football. *arXiv:2103.07272*, 1, 2021.
[16] C. M. F. C. M. Rosli, M. Z. Saringat, N. Razali, and A. Mustapha. A comparative study of data mining techniques on football match prediction. *Journal of Physics: Conference Series*, 1020:971–980, 2018.
[17] T. Sleeuwaert. Evaluation of the current state of football match outcome prediction models. *Ghent University Master dissertation*, 2020.
[18] R. T. Stefani. Football and basketball predictions using least squares. *IEEE Transactions on systems, man, and cybernetics*, 7:117–21, 1977.
[19] J. Stubinger and J. Knoll. Beat the bookmaker – winning football bets with machine learning. *SGAI-AI 2018*, page 219–233, 2018.
[20] E. Wheatcroft. Forecasting football matches by predicting match statistics. *arXiv:2001.09097*, 1, 2020.
[21] E. Wheatcroft. A profitable model for predicting the over/under market in football. *International Journal of Forecasting*, 2020.
[22] willinghorse. European soccer database supplementary. `https://www.kaggle.com/jiezi2004/soccer`, 2017.

ISBN 978-946498883-3