



MATHSPORT INTERNATIONAL 2019 CONFERENCE

— Proceedings —

Crowne Plaza Hotel
Athens, 1-3 July 2019

**ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ**



ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

Organized by AUEB Sports Analytics Group

Conference Sponsors



Original title: Proceedings of MathSport International 2019 Conference (e-book)

Editors: Dimitris Karlis, Ioannis Ntzoufras, Sotiris Drikos

©Propobos Publications

53, Patision Street, 10433, Athens, Greece

<https://www.propobos.gr/>

ISBN 978-618-5036-53-9

All rights reserved.

Preface

We are very glad to welcome you at Athens for the 7th Mathematics in Sports International conference. It is our pleasure to host the conference this year after the successful previous ones in Manchester, Groningen, Salford, Leuven, Loughborough and Padova. Since the first MathsSports conference, a lot of research has been made on topics related to Mathematics and Sports and the interest has increased considerably. MathsSports conference is a gathering of Academics and practitioners covering a wide range of sports and methodologies and we hope that this year the meeting will cover even wider variety of sports. We have attempted to arrange a wide range of talks during the 3 days of the conference.

A complete list of all the abstracts of the papers to be presented in the conference can be found in this book. A detailed index of all authors can be found at the end to facilitate easy search.

We hope that you will enjoy the meeting.

Dimitris Karlis
Ioannis Ntzoufras
on behalf of the LOC and SC.

Scientific Committee

Phil Scarf	University of Salford, Manchester (UK)
Anthony Bedford	RMIT University, Melbourne (AU)
Marco Ferrante	University of Padova (IT)
Dries Goossens	Gent University (BE)
Ruud Koning	University of Groningen (NL)
Stephanie Kovalchik	Victoria University (AU)
Alun Owen	De Montfort University - DMU Leicester (UK)
Frits Spijksma	Eindhoven University of Technology (NL)
Ray Stefani	CSULB University (USA)

Local Organizing Committee

Chair

Dimitris Karlis	Athens University of Economics and Business
Ioannis Ntzoufras	Athens University of Economics and Business

Members

Sotiris Drikos	AUEB Sports Analytics Group (GR)
Panagiotis Repousis	Athens University of Economics and Business (GR)
Evgenia Tzoumaka	American College of Greece (GR)
Konstantinos Matzorakis	AUEB Sports Analytics Group (GR)

Contents

Alexander, Jeremy; Spencer, Bartholomew; Sweeting, Alice; Mara Jocelyn and Robertson, Sam	
The influence of match phase and field position on collective team behaviour in Australian Rules football	1
Atan, Tankut and Cavdaroglu, Burak	
Rest Differences among Teams in European Football Leagues	10
Barbiero, Alessandro	
Alternative count regression models for modeling football outcomes	16
Cavdaroglu Burak and Atan, Tankut	
Integrated Break and Carryover Minimization Problem in Round Robin Tournaments	25
Csato, Laszlo	
Overcoming the incentive incompatibility of tournaments with multiple group stages	33
Cueva, Valentina; Rodríguez-Avi, José and Olmo-Jiménez, María José	
Comparison between some European football leagues through related count data variables	57
Curley, Brian; Hopkirk, Gretchen; Lokhorst Ryan and Pilkington, Annette	
Randomness of Play Calling in College Football	68
Drikos Sotiris	
Complex 1 in Male Volleyball as a Markov Chain	80
Egidi, Leonardo and Ntzoufras, Ioannis	
Modelling volleyball data using a Bayesian approach	86
Fonseca, Giovanni and Giummolè, Federica	
Extreme value prediction: an application to sport records	96
Goes, Floris; Kempe Matthias and Lemmink, Koen	
Predicting match outcome in professional Dutch football using tactical performance metrics computed from position tracking data	105

Goossens, Dries; Wang Chang and Vandebroek, Martina	
Champions League or domestic league: a coach's choice	116
Grassetti, Luca; Bellio, Ruggero; Fonseca Giovanni and Vidoni, Paolo	
Play-by-play data analysis for team managing in basketball	129
Guyon, Julien	
Will Groups of 3 Ruin the World Cup?	140
Hirotsu Nobuyoshi and Komine, Ayako	
Analysing the effect of a change of transition probabilities related to possession on scoring a goal in a football match	156
Hubáček, Ondřej; Šourek Gustav and Železný, Filip	
Score-based soccer match outcome modeling - an experimental review	164
Kee, Lyn; Huynh, Minh; Meyer Denny and Marshall, Kelly	
The explosive growth of eSports and the potential for research opportunities	173
Kempe Matthias and Goes, Floris	
Move it or lose it: Exploring the relation of defensive disruptiveness and team success	184
Koevoets, Wim	
An evaluation of the three-point rule in football	192
Konaka, Eiji	
A quantitative method for evaluating the skills of national volleyball teams: Prediction accuracy comparisons of the official ranking system in the worldwide tournaments of 2010s	202
Koning, Ruud and Jan Going, Hidde	
ELO or Coca Cola, which ranking is better?	217
Kouřim, Tomáš	
Random Walks with Memory Applied to Grand Slam Tennis Matches Modeling	220
Lawrence, Steve; Jonker Laura and Verbeek, Jan	
The Age Advantage in Youth Football	228
Leriu, Ilias; Ntzoufras Ioannis and Karlis, Dimitris	
Survival Modelling of Goal Arrival Times in Champions League . . .	234
Meyer, Denny; Huynh, Minh; Marshall Kelly and Pollard, Geoff	
Fame and Fortune in Elite Tennis Revisited	241
Nurmi, Kimmo; Kyngäs Jari and Kyngäs, Nico	
Lessons Learned in Scheduling the Finnish Major Ice Hockey League	251
Owen, Alun	
Statistical Models of Horse Racing Outcomes Using R	264

Polyashuk, Marina	
Multi-Criteria Solutions for Optimizing Lineup in Baseball	277
Sans Fuentes, Carles; Carlsson, Niklas and Lambrix, Patrick	
Player impact measures for scoring in ice hockey	307
Stefani, Raymond	
The Ancient Olympics: Events, Technology, Superstars, Women, Lessons for Them and for Us	318
Twersky, Georgia; Lyman George and Pilkington, Anne	
Combining the Four Factors with the Generalized PageRank (GeM) model for NBA Basketball	328
Uhrín, Matej; Šourek, Gustav; Hubáček, Ondřej and Železný, Filip	
Sports betting strategies: an experimental review	337

The impact of a team numerical advantage on match play in Australian Rules football

Jeremy. P. Alexander ^{1*}, Bartholomew Spencer¹, Alice J. Sweeting ^{1,3}, Jocelyn. K. Mara ², Sam Robertson ^{1,3}

¹*Institute for Health and Sport (IHES), Victoria University,*

²*Research Institute for Sport and Exercise, University of Canberra*

³*Western Bulldogs Football Club*

¹ Corresponding author: jeremyalexander60@hotmail.com

ABSTRACT

The primary aim of this study was to provide a proof of concept that determines the relationship between a team numerical advantage and match play in a continuous manner. The secondary aim was to quantify how players occupy different sub-areas of play, while accounting for match phase and position of the ball in Australian Rules football. Data from Australian football athletes (years 23.9 ± 4.3 ; cm 188.0 ± 7.9 ; kg 86.0 ± 9.4), were collected via 10 Hz global positioning system (GPS) during match simulation. The total number of players, team numerical advantage, and Approximate Entropy (ApEn) were analysed during match phase (offensive, defensive, and contested) and field position (defensive 50, defensive midfield, forward midfield, and forward 50). Results revealed that a team numerical advantage was associated with advantageous match play outcomes. Specifically, the likelihood of gaining possession of the ball increased when teams obtained a numerical advantage. The total number of players increased based on where the ball was positioned, especially if located in the D50. Teams were largely outnumbered when the ball was in their F50 but maintained a numerical advantage when defending in the D50. Variability in ApEn values was greater in team numerical advantage and total players during the middle segments of the ground compared to the F50 and D50. A method that continuously represents how players occupy sub-areas of play may provide coaches and sport science practitioners with a more precise account of how tactical team behaviour influences ensuing match play.

Keywords: Performance analysis, invasion sports, game style, tactical behaviour

1. INTRODUCTION

The advent of player tracking technologies has supported a more detailed approach to the match analysis of invasion sports (Rein and Memmert 2016). Interactions between teammates and opponents can now be captured in a continuous manner that more accurately reflects the constantly changing nature of match play (Travassos, Davids et al. 2013). Analysis of this information can provide an assessment of the collective organisation of players across a field of play (Clemente, Couceiro et al. 2013), which has been used to describe team tactical behaviour and performance outcomes (Vilar, Araújo et al. 2013, Silva, Travassos et al. 2014).

Recently, studies in football have attempted to assess the tactical behaviour of teams by examining how players occupy different sub-areas on a playing field at different timescales (Vilar, Araújo et al. 2013, Silva, Travassos et al. 2014, Clemente, Couceiro et al. 2015). Teams may regulate player positioning to increase offensive effectiveness or instill disorder in opposition defensive structures (Vilar, Araújo et al. 2013). This may be achieved by generating a numerical advantage or dominance at different sub-areas on a field of play by outnumbering the opposing team (Vilar, Araújo et al. 2013, Silva, Travassos et al. 2014). Researchers in football have proposed that match success is associated with a team's ability to generate a numerical advantage during offensive sequences of play (Vilar, Araújo et al. 2013) and to preserve defensive stability by allocating a greater number of players closer to their goal when compared to the opposition (Vilar, Araújo et al. 2013, Clemente, Couceiro et al. 2015).

However, studies that reduce performance to a single aspect of match play may not fully

appreciate the complex, multifaceted, and unpredictable nature of invasion sports (Duarte, Araujo et al. 2012). Performance can be influenced by various contextual variables, such as, match phase, ball location, and quality of opposition (Lago 2009, Duarte, Araujo et al. 2012, Alexander, Spencer et al. 2019). Notwithstanding, studies to date that have investigated tactical team behaviour by measuring how players occupy different sub-areas on a playing field in football have inferred performance by assessing a team's capacity to generate a numerical advantage over a specific area (Vilar, Araújo et al. 2013). As such, a limited understanding exists between a team's numerical advantage and the impact on match play in a continuous manner. In addition, contextual variables such as ball position and match phase are yet to be reported when assessing how players occupy different sub-areas on a playing field (Vilar, Araújo et al. 2013, Clemente, Couceiro et al. 2015). Investigations into how players occupy different sub-areas of play in Australian Football also remain largely absent. Australian Rules football (AF) is a sport where teams compete on an oval shaped field (length = ~160 m, width = ~130 m) with 22 players in total, with 18 on the field and 4 on an interchange (Gray and Jenkins 2010). A goal is scored when a player kicks the ball through the two large goalposts and equates to 6 points (Woods 2016). If a player misses the large goalposts but the ball passes through the small goalposts on either side, a single point is registered (Woods 2016).

Thus, a specific method that can determine a team's numerical advantage in a continuous format could be useful in determining the immediate impact on ensuing match play. Research analysing how players occupy different sub-areas on a playing field that accounts for ball position and match phase also remains absent. Therefore, the primary aim of this study was to provide a proof of concept that determines the relationship between a team numerical advantage and match play in a continuous manner. The secondary aim was to determine how ball position and match phase influence how players occupy different sub-areas of play in AF.

2. METHODS

Data were collected from one training session with 30 male professional AF players (years 23.9 ± 4.3 ; cm 188.0 ± 7.9 ; kg 86.0 ± 9.4) recruited from a single team in the Australian Football League (AFL) competition. Participants took part in a match simulation drill as part of preseason training. All participants received information about the requirements of the study via verbal and written communication, and provided their written consent to participate. The University Ethics Committee approved the study.

Participants were separated into two teams of 15 each at the coach's discretion to ensure a relatively even competition and were labeled Home team and Away team for analysis purposes. The match simulation took place on an oval shaped ground using dimensions 163.7 m x 129.8 m (length x width) with two 20-min halves and a 10-min break between periods. Data for all participants were collected using 10 Hz GPS devices (Catapult Optimeye S5, Catapult Innovations, Melbourne, Australia). The devices were housed in a sewn pocket in the jersey that is located on the upper back. The number of GPS satellites was greater than 8 per second, which ensured adequate signal quality (Corbett, Sweeting et al. 2017).

Spatiotemporal data were exported in raw 10 Hz format. Each file contained a global time stamp and calibrated location (x- and y- location). Match phase was determined via which team had possession of the ball (offensive, defensive or contest). The offensive phase was recorded when a team first gained possession of the ball and maintained it for at least a second and ended when the opposing team gained possession of the ball for at least a second or there was a stoppage in play. For example, the team scored or the ball went out of bounds (Yue, Broich et al. 2008). Using the same conditions, the defensive phase was recorded when the opposing team had possession of the ball (Yue, Broich et al. 2008). If neither team had possession of the ball, for example, when the officiating umpire returned the ball to play, the phase was considered to be in contest until a team gained possession of the ball for at least one second. All periods where the ball was out of play, for example, when there was a break between periods of play, celebration after goals, were excluded from the investigation.

Field position was separated into four zones (defensive 50; D50, defensive mid; DMID, forward mid; FMID, forward 50; F50) by the two 50 m arcs and the center of the ground (see Figure 1), which is conventional for AF research and statistical providers (Jackson 2016). Match event data notated the action of the player who had possession of the ball and was recorded to the nearest second. This information provided an assessment of the relationship between a team numerical advantage and match play. Specifically, if a team numerical advantage was associated with gaining possession of the ball. Total scoring opportunities were also recorded. Teams have the capacity to gain possession of the ball through three different methods including; turnover (TO), which is possession gained from the opposition, a clearance (CL), which is possession gained from a contested situation, and via a kick in (KI), which is when a team gains possession if the opponents scores a behind (Woods 2016). Previous investigations have assessed the validity and reliability of similar match events (Robertson, Gupta et al. 2016). Positional data was synchronised with match event data using the respective global timestamps. This was established using the initial point when the two widest players on the field converged from a stationary position prior to start of each quarter.

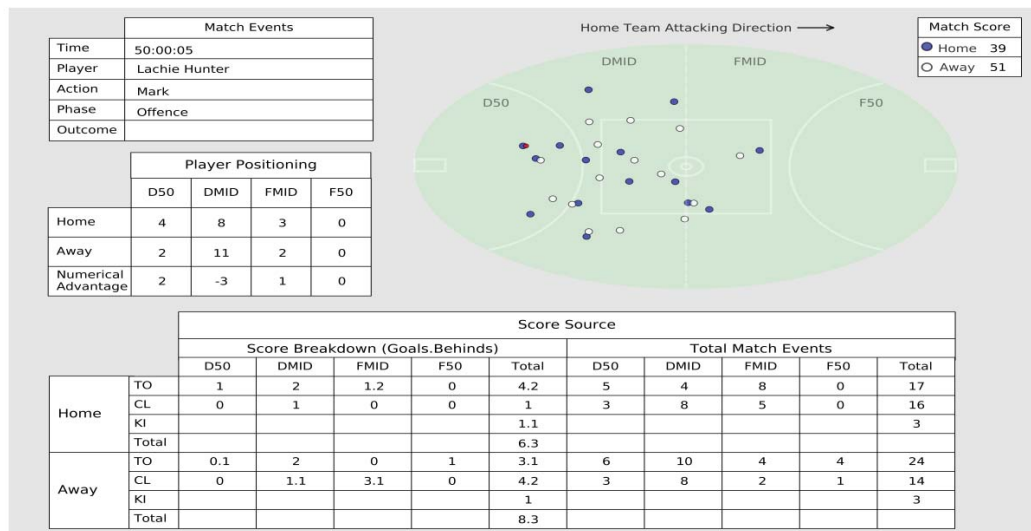


Figure 1: Match events, total player positioning, and team numerical advantage of both teams at the 50-minute mark of the match.

TO, Turnover; CL, Clearance, KI; Kick In; F50, Forward 50; FMID, Forward Midfield; DMID, Defensive Midfield; D50, Defensive 50;

Data Analysis

The total number of players inside the four field positions for the Home team and the Away team was assessed for each point in time. This was used to also determine the team numerical advantage and disadvantage for each team (N_s) following the rule: $N_s = N_s^H - N_s^A$. The total number of players and numerical advantage for each match phase and field position were visualised via frequency histograms. Match play outcomes were assessed via analysing the relationship between the total amount of turnovers and clearances and the respective team numerical advantage. Analysis was processed using the computational package Python version 3.2 with *Spyder*, which is part of the Anaconda software suite (www.python.org).

Statistical Analyses

The variability of total players within the four field positions and the team numerical advantage or disadvantage was calculated using the Approximate Entropy (ApEn) (Pincus, Gladstone et al. 1991). Provided with a given time series of N points (x_1, x_2, \dots, x_N), $ApEn(m, r, N)$ can be used to measure the logarithmic probability that lengths of patterns with m points that are close, continue to be close within a tolerance factor r for the subsequent assessments (Pincus, Gladstone et al. 1991). Put simply, a sequence of data points is more regular if the following data points expand in a similar manner. To calculate $ApEn(m, r, N)$, the parameters m , the length of compared runs, and r , the tolerance factor, need to be consistent for all assessments to ensure reliable analysis (Pincus and Goldberger 1994).

$$ApEn(m, r, N) = \phi^m(r) - \phi^{m+1}(r)$$

ApEn values vary between 0 and 2, with values closer to 2 indicating time series with less regular or more variable patterns. Values closer to 0 imply a more regular or less variable time series (Fonseca, Milho et al. 2013). These calculations were completed using the computational package Python version 3.2 with *Spyder*, which is part of the Anaconda software suite (www.python.org).

3. RESULTS

Distribution of the total number of players and team numerical advantage during each match phase and field position for the Home team and the Away team are displayed in Figure 2 and Figure 3 respectively. Variability in the total number of players and team numerical advantage as expressed by ApEn values during each match phase and field position for the Home team and the Away team are presented in Figure 4 and Figure 5 respectively. The relationship between match play and a team's numerical advantage is displayed in Figure 6. The Away team won the match 51 – 39.

The total number of players increased when the ball was located in either the F50 or D50 when compared to the DMID and FMID. This finding was more pronounced during defence when the ball was in the D50. The Away team was more effective at preserving more players in the D50 during defence when compared to the Home team. Both teams maintained a team numerical advantage when the ball was in the D50 and faced a numerical disadvantage when the ball was in their F50 during all match phases. Contrastingly, both teams endured a numerical disadvantage during the DMID during offence and defence but obtained a numerical advantage during FMID during offence.

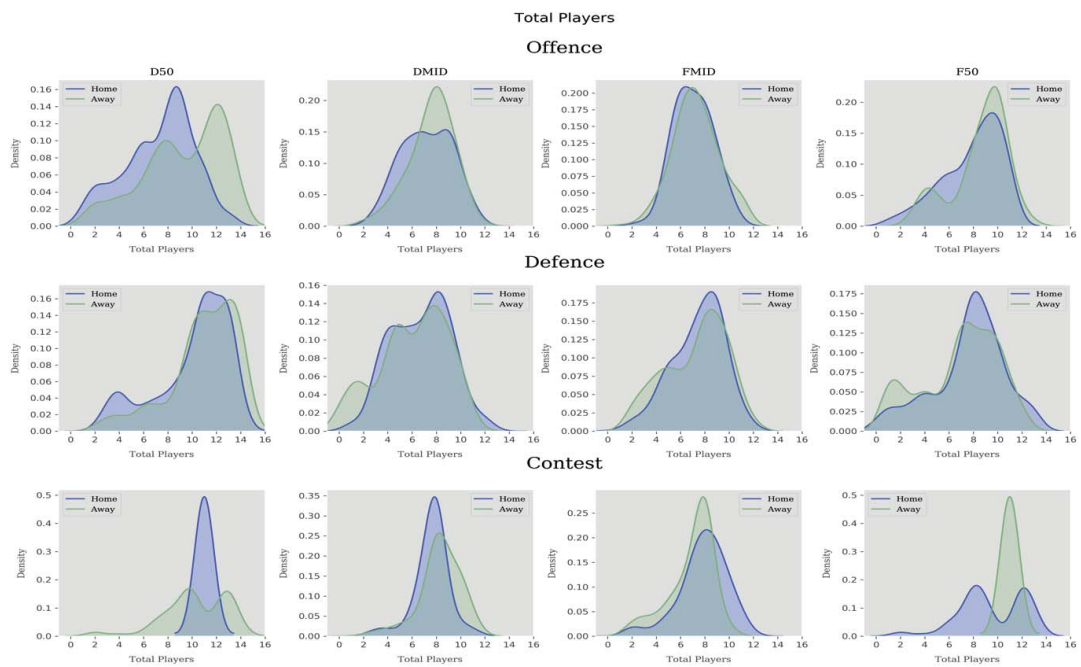


Figure 2: Total number of players for the Home team and Away team in each field position for each phase of play

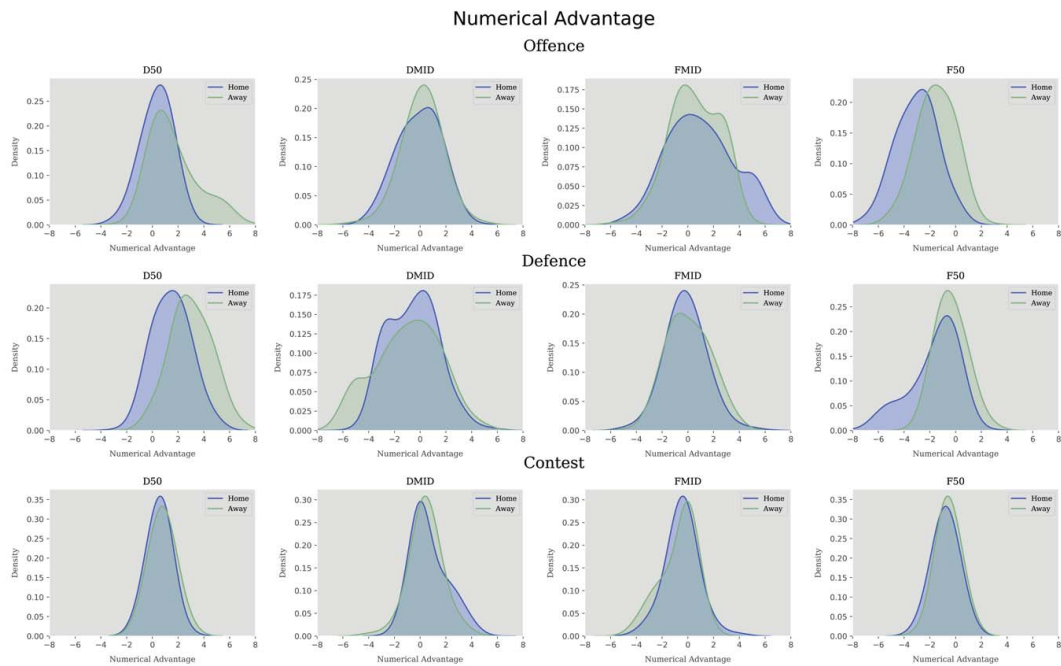


Figure 3: Team numerical advantage for the Home team and Away team in each field position for each phase of play

ApEn values in total players and team numerical were greater during both the FMID and DMID during all phases of play. ApEn values were reduced during contest compared to offence and defence in both the total number of the total number of players and team numerical advantage.

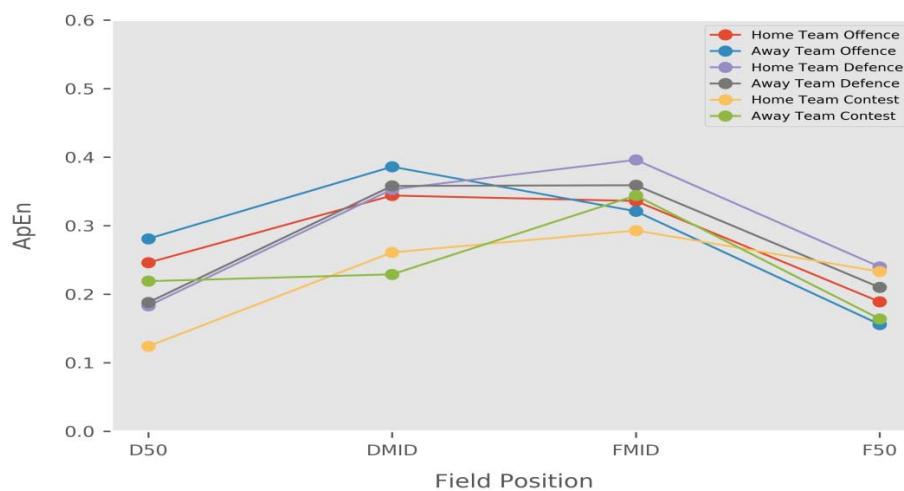


Figure 4: ApEn values in total players in each field position for each phase of match play

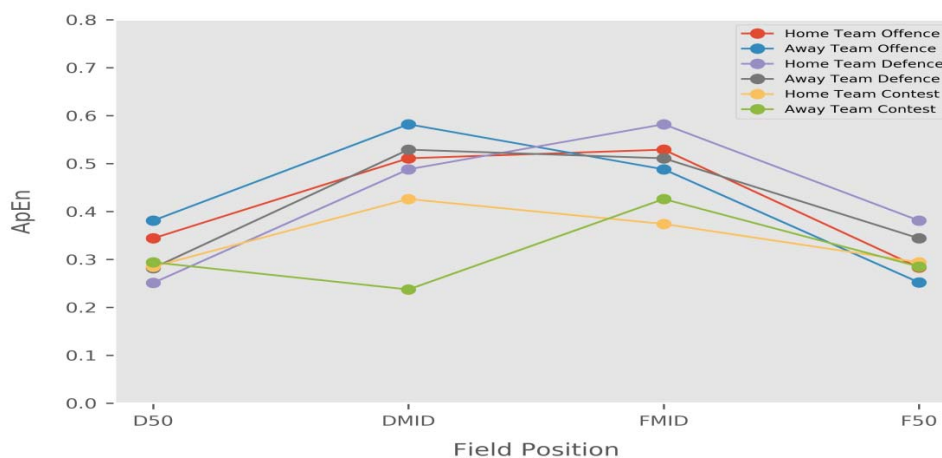


Figure 5: ApEn values in team numerical advantage in each field position for each phase of match play

A total of 43 turnovers were observed throughout the match with the Home team generating 20, while the Away team gathered 23. The Home team obtained 15 clearances, while the Away team gathered 14. Both teams obtained a team numerical advantage when generating a turnover, although the Away team recorded a greater advantage with an average of 1.35 additional players compared to the Home team who had an average 0.75 players extra players. Both the Home team and Away team had a numerical advantage when gaining possession of the ball during clearances of 0.4 and 0.43 extra players respectively.

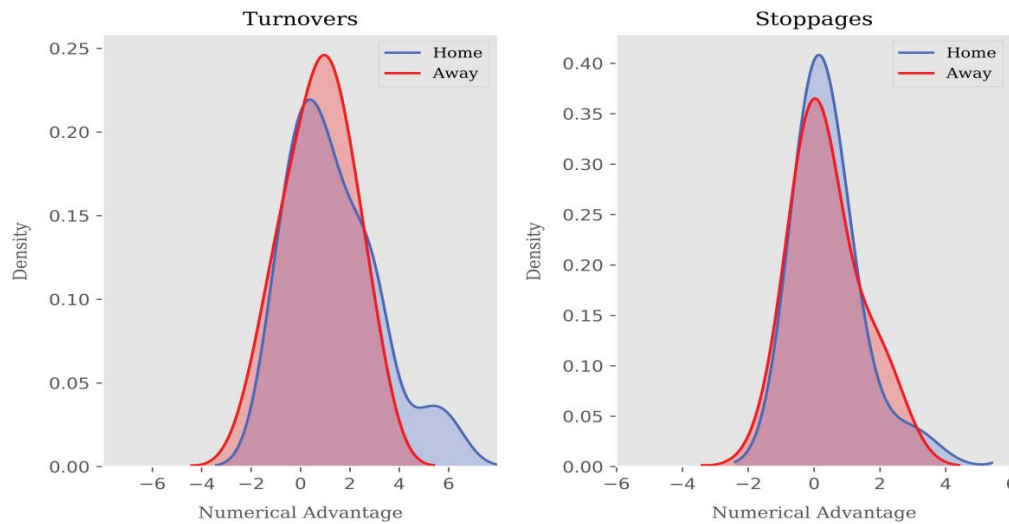


Figure 6: Distribution of team numerical advantage during clearances and turnovers for the Home team and Away team

4. DISCUSSION

This proof of concept study is the first in invasion sports to establish the association between team numerical advantage and match play in a continuous manner. This investigation also provides an enhanced understanding of tactical team behaviour by determining how ball position and match phase influence how players occupy different sub-areas of play in AF.

This study revealed that an increased team numerical advantage was associated with advantageous match play outcomes. Specifically, a team that obtained a numerical advantage displayed an increased likelihood of gaining possession of the ball from turnovers and during contested situations. Other findings included that the total number of players increased based on where the ball was positioned. Increasing the total amount of players within a certain area may constrain opposition movement (Alexander, Spencer et al. 2019). This is supported by research that indicates that increasing the number of defensive players surrounding an attacking team taking a shot at goal is associated with a concomitant decrease in successful scoring attempts (Ensum, Pollard et al. 2004, Wright, Atkins et al. 2011). Teams were largely outnumbered when the ball was in their F50 but maintained a numerical advantage when defending in the D50. This is similar to findings in football, which prescribe that teams generally employ conservative team behaviour by maintaining a numerical advantage in their defensive half (Vilar, Araújo et al. 2013, Clemente, Couceiro et al. 2015). Variability in total players and team numerical advantage measured through ApEn was greater during the middle segments of the ground compared to the F50 and D50. Similarly, other research in football found greater variation in central sectors of the ground (Vilar, Araújo et al. 2013, Clemente, Couceiro et al. 2015). This may be explained by the increased interaction of players in these regions (Clemente, Couceiro et al. 2015). For instance, at any point in time, players in these regions must be willing to create attacking opportunities for their teammates during offensive phases of play and prepared to maintain defensive support when the opposition gains possession of the ball.

Continuous interactions between teammates and opponents transpire that revolve around promoting offensive opportunities and preserving defensive stability (Vilar, Araújo et al. 2013). As such, players are required to alter their movement behaviour during a match due to the emerging nature of match play (Duarte, Araujo et al. 2012). Nonetheless, how teams manage player positioning during

various contextual variables is fundamentally linked to tactical team behaviour. Specifically, teams may strategically position players across a field of play in an attempt to gain a competitive advantage in certain circumstances. For instance, findings from this study indicate that teams who obtained a numerical advantage during contested ball situations had an increased likelihood of gaining possession of the ball. If opposing teams are more successful in gaining a greater amount of possession in these situations, teams may look to allocate more players to limit this impact. However, this may create a numerical imbalance elsewhere on the field that has the potential to influence other aspects of match play. For example, if a player is taken from the forward half of the field to assist in contested situations, the opposition may have a numerical advantage in their defensive half, which may provide the opportunity to create more turnovers in this area of the field. If both teams were to employ a numerical advantage in their defensive half with an aim to increase defensive stability, resulting match play could observe a potential increase in turnovers but a decrease in scoring. This 'positional trade-off' is a constant evolution that coaches, players and sport science practitioners are challenged with when determining their team's tactical behaviour and assessing that of the opposition.

Some limitations relating to sample size and amount of teams included in this study should be recognised. The present study analysed player positioning of one club during a single pre-season match simulation. Thus, additional research should include multiple clubs throughout several matches to construct a more accurate representation of how players occupy sub-areas of play and if any variations exist between various contextual variables. Future investigations may also provide a statistical significance between a team numerical advantage and match play in real time in AF. Future work may also incorporate a more fluid approach to the concept of team spatial dominance. Specifically, dominance should be aligned with how much space a player can theoretically cover, rather to attribute greater dominance to a team that obtains an extra player within a large sub-area of play that may not have a direct influence of match play.

5. CONCLUSIONS

This study investigated the relationship between a team numerical advantage and match play in a continuous manner in AF, along with providing a greater understanding of how players occupy different sub-areas during various contextual variables. Teams that obtained a numerical advantage displayed an increased likelihood of gaining possession of the ball. A method that continuously represents how players occupy sub-areas of play may provide coaches and sport science practitioners with a more precise account of how a team numerical advantage influences ensuing match play.

ACKNOWLEDGMENTS

The authors wish to thank the athletes and support staff of the Western Bulldogs for their participation in this study.

REFERENCES

- Alexander, J. P., B. Spencer, A. J. Sweeting, J. K. Mara and S. Robertson (2019). "The influence of match phase and field position on collective team behaviour in Australian Rules football." Journal of sports sciences: 1-9.
- Clemente, F., M. Couceiro, F. Martins, R. Mendes and A. Figueiredo (2013). "Measuring tactical behaviour using technological metrics: Case study of a football game." International Journal of Sports Science & Coaching 8(4): 723-739.
- Clemente, F. M., M. S. Couceiro, F. M. L. Martins, R. S. Mendes and A. J. Figueiredo (2015). "Soccer team's tactical behaviour: Measuring territorial domain." Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology 229(1): 58-66.

- Corbett, D. M., A. J. Sweeting and S. Robertson (2017). "Weak relationships between stint duration, physical and skilled match performance in Australian Football." Frontiers in physiology **8**.
- Duarte, R., D. Araujo, V. Correia and K. Davids (2012). "Sports teams as superorganisms: implications of sociobiological models of behaviour for research and practice in team sports performance analysis." Sports medicine (Auckland, N Z) **42**(8): 633-642.
- Ensum, J., R. Pollard and S. Taylor (2004). "Applications of logistic regression to shots at goal at association football: Calculation of shot probabilities, quantification of factors and player/team." Journal of Sports Sciences **22**(6): 500-520.
- Fonseca, S., J. Milho, B. Travassos, D. Araújo and A. Lopes (2013). "Measuring spatial interaction behavior in team sports using superimposed Voronoi diagrams." International journal of performance analysis in sport **13**(1): 179-189.
- Gray, A. J. and D. G. Jenkins (2010). "Match analysis and the physiological demands of Australian football." Sports medicine (Auckland, N Z) **40**(4): 347-360.
- Jackson, K. (2016). Assessing Player Performance in Australian Football Using Spatial Data, PhD Thesis, Swinburne University of Technology.
- Lago, C. (2009). "The influence of match location, quality of opposition, and match status on possession strategies in professional association football." Journal of sports sciences **27**(13): 1463-1469.
- Pincus, S. M., I. M. Gladstone and R. A. Ehrenkranz (1991). "A regularity statistic for medical data analysis." Journal of clinical monitoring **7**(4): 335-345.
- Pincus, S. M. and A. L. Goldberger (1994). "Physiological time-series analysis: what does regularity quantify?" American Journal of Physiology-Heart and Circulatory Physiology **266**(4): H1643-H1656.
- Rein, R. and D. Memmert (2016). "Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science." Springerplus **5**(1): 1410.
- Robertson, S., R. Gupta and S. McIntosh (2016). "A method to assess the influence of individual player performance distribution on match outcome in team sports." J Sports Sci **34**(19): 1893-1900.
- Silva, P., B. Travassos, L. Vilar, P. Aguiar, K. Davids, D. Araujo and J. Garganta (2014). "Numerical relations and skill level constrain co-adaptive behaviors of agents in sports teams." PLoS One **9**(9): e107112.
- Travassos, B., K. Davids, D. Araújo and P. T. Esteves (2013). "Performance analysis in team sports: Advances from an Ecological Dynamics approach." International Journal of Performance Analysis in Sport **13**(1): 83-95.
- Vilar, L., D. Araújo, K. Davids and Y. Bar-Yam (2013). "Science of winning soccer: Emergent pattern-forming dynamics in association football." Journal of systems science and complexity **26**(1): 73-84.
- Woods, T. C. (2016). "The use of team performance indicator characteristics to explain ladder position at the conclusion of the Australian Football League home and away season." International Journal of Performance Analysis in Sport **16**(3): 837-847.
- Wright, C., S. Atkins, R. Polman, B. Jones and L. Sargeson (2011). "Factors associated with goals and goal scoring opportunities in professional soccer." International Journal of Performance Analysis in Sport **11**(3): 438-449.
- Yue, Z., H. Broich, F. Seifriz and J. Mester (2008). "Mathematical analysis of a soccer game. Part I: Individual and collective behaviors." Studies in applied mathematics **121**(3): 223-243.

Rest Differences among Teams in European Football Leagues

Tankut Atan¹ and Burak Cavdaroglu²

¹ Department of Industrial Engineering, Bahçeşehir University, Beşiktaş, İstanbul, Turkey, 34353
`sabritankut.atan@eng.bau.edu.tr`

² Department of Industrial Engineering, Kadir Has University, Cibali, İstanbul, Turkey, 34083
`burak.cavdaroglu@khas.edu.tr`

Abstract

When two opponents in a sports game have not rested an equal amount after their most recent game, the more rested team has an unfair advantage over the less rested team. Tournament organizers typically do not pay attention to this fairness criterion when they determine the timings of the games. We provide a general linear integer programming formulation for a given round robin tournament schedule that finds the periods of the games minimizing the total rest difference among the teams. Then, we compare how different European national first division football leagues perform in terms of the rest differences.

1 Introduction

League scheduling is one of the popular applications of operations research in sports. In most relevant research, the focus is on determining the games in the rounds of the competition. Several fairness criteria may be considered in this timetable construction. So-called break and carryover minimizations are widely used. A break occurs when a team has to play two consecutive home or away games. A carryover effect, on the other hand, occurs when a weaker team has to play against two strong opponents in a row. Yet another well known criterion is the minimization of the travel distances when teams have to play several consecutive away games without returning home which is of concern in leagues that cover a large geographic area.

Once an acceptable timetable, i.e. the games in each round, is constructed, it is announced ahead of the competition season. Since each round may actually consist of several days, tournament organizers determine the days of the individual games of each round as the season progresses. Often, league schedules are criticized in popular media when two opposing teams have not rested the same amount of time after their most recent game; especially losing teams become more critical about this when they had less number of days to rest. A rest difference between opposing teams is not desired. Thus, a fair schedule should have as little rest difference as possible. Fairness criteria regarding the team rest durations between the rounds can only be considered at the “day” level of detail and they have not received much attention by researchers. In this study, we concentrate on the problem of determining the day (period) of each competition given a tournament schedule so that total rest difference among teams is minimized. The organization of this paper is as follows. After reviewing relevant literature in Section 2, we give an integer linear formulation for minimizing the sum of rest differences of teams from each game in Section 3. Our computational experiments and their results are given in Section 4 followed by a conclusion.

2 Previous Work

A round robin tournament is a common format for organizing sports events. In a round robin tournament, every team plays every other team a fixed number of times. [3] provide integer programming models for round robin tournaments with several externally given and also fairness constraints such as forbidden matches, observing regional capacities, having few breaks, respecting team's preferences for matchdays, and balancing opponents' strengths. For example, when teams are partitioned into several strength groups, it may not be desirable to have teams to play opponents from the same strength group consecutively. It may also be preferable to have each team to play an opponent of each strength group at most once within a time window. [2] discusses combinatorial properties of strength groups in round robin tournaments. [9] provide a survey on round robin scheduling whereas [6] summarize competition formats and schedules used in 25 European soccer competitions for the season 2008-2009. They also compare these leagues based on several design criteria. While much work regarding league scheduling is theoretical, there are also some reported applications of scheduling theory for finding the official schedules of leagues. Some recent articles include [10] on the Ecuadorian football league, [5] reporting their experience with the Belgian football league, [12] with the German and [4] with the Argentina basketball leagues.

[1] were first to investigate how to construct a league schedule that considers rest imbalances of opposing teams in games. In particular, they look into devising a round robin tournament that minimizes the number of rest mismatches. A rest mismatch is defined as the occurrence of a difference between the rest durations of two opposing teams in a game. Observe that, a rest mismatch does not consider the magnitude of the difference in the rest durations of opposing teams. Both rest differences and rest mismatches can work against a team which did not have a chance to rest as much as its opponent. Team managers frequently complain about having had less rest than their opponents in popular media. [8] studies a single round robin tournament with only a single venue to play the games. In each round, all teams travel to this single venue and play two games each. Since large waiting times between the games of a team are not preferred, [8] constructs schedules that minimize the number of long waiting times and the total waiting time simultaneously for any odd number of teams. [11] investigates asynchronous round robin tournaments where all games are played at different consecutive times with respect to three different fairness criteria: guaranteed rest time, games-played difference index, and rest difference index. Rest difference index is equal to the maximum difference in rest durations of two opponents in a schedule whereas we focus on minimizing the sum of rest differences in a given schedule.

It is also worth mentioning that there is not much reported research on determining the game days for a given schedule. [4] use a phased approach in their basketball league scheduling in Argentina and determine the days of games after fixing the order of games. They mention the inclusion of fairness restrictions such as rest day balance between games across all teams as future research.

3 Total Rest Difference Model

Imagine a double round robin tournament with n teams where n is an even number. There are $2 \cdot (n - 1)$ rounds in the tournament with $n/2$ games in each round. All games in each round are to be played in p consecutive periods with a predetermined number of games in each period. Note that, the number of games in the periods of the same rounds from the first and second half of the tournament can be different even when the tournament is mirrored, i.e. round by round

games in the second half of the tournament are played in the same order as in the first half with only venues changing. Therefore, we consider double round robin tournaments instead of only single round robin tournaments.

3.1 Integer Linear Model

Next, we give the definitions of sets, parameters and decision variables used in the integer programming (IP) formulation of the considered problem followed by the mathematical model.

3.1.1 Sets

P : Periods, $k = 1, \dots, p$. As an example, in a weekly tournament, periods may correspond to days in the week, and the number of periods will be equal to 7.

R : Rounds, $r = 1, \dots, 2 \cdot (n - 1)$.

T : Teams, $i = 1, \dots, n$.

3.1.2 Parameters

λ_r : The number of periods between the first periods of round r and $r - 1$.

$nGames_{r,k}$: The number of games in period k of round r where $nGames_{r,k} \geq 0$.

$play_{i,j,r} = 1$ if teams i and j play against each other in round r ; 0 otherwise.

3.1.3 Decision Variables

$p_{i,r}$: The number of periods team i rested less than its competitor in round r .

$x_{i,r,k} = 1$ if team i has its game in period k of round r ; 0 otherwise.

z : Sum of rest differences of teams.

3.1.4 IP Model

$$\min \quad z = \sum_i \sum_r p_{i,r} \quad (1)$$

subject to:

$$\sum_k x_{i,r,k} = 1 \quad \forall i \in T, \forall r \in R \quad (2)$$

$$play_{i,j,r} \cdot x_{i,r,k} = play_{i,j,r} \cdot x_{j,r,k} \quad \forall i, j \in T, \forall r \in R, \forall k \in P \quad (3)$$

$$\sum_i x_{i,r,k} = 2 \cdot nGames_{r,k} \quad \forall r \in R, \forall k \in P \quad (4)$$

$$\begin{aligned} play_{i,j,r} \cdot \left(\sum_k k \cdot x_{i,r,k} - \left(\sum_k k \cdot x_{i,r-1,k} - \lambda_r \right) + p_{i,r} \right) \\ = play_{i,j,r} \cdot \left(\sum_k k \cdot x_{j,r,k} - \left(\sum_k k \cdot x_{j,r-1,k} - \lambda_r \right) + p_{j,r} \right) \\ \forall i, j \in T, \forall r \in R \setminus \{1\} \end{aligned} \quad (5)$$

$$x_{i,r,k} \in \{0, 1\} \quad \forall i \in T, \forall r \in R, \forall k \in P \quad (6)$$

$$p_{i,r} \geq 0 \quad \forall i \in T, \forall r \in R \quad (7)$$

The objective function minimizes the sum of rest differences. Constraint 2 states that each team must play a game in only one period during each round. Constraint 3 makes sure that if two teams play against each other then they are assigned to the same period. Constraint 4 sets the number of games in a period to the predetermined number of games for that period. Observe that since each game involves two teams of which related x variables are being summed over on the left-hand side, the number of games in the period is multiplied by 2. Constraint 5 determines the rest difference between opponent teams in games. The multiplication by $play_{i,j,r}$ parameters delimits constraints in Constraint 5 to only the ones regarding the actual games in the given schedule of a round. The left-hand side calculates the duration (number of periods) that passes from the game of i in round $r - 1$ to i 's game in round r . A similar calculation is conducted for j , the opponent, on the right-hand side. Ideally, the left- and right-hand sides of all games should be equal to each other. If this is not the case then respective p variable for the team that has less rest is set to the positive difference, and hence shows how many periods less that team has rested in round r . Observe that the p variable for the team that has more rest will be set to zero because the objective function minimizes the sum of all p variables. Constraint 6 and Constraint 7 set the types of decision variables.

4 Computational Experiments

Computer runs were executed on an Intel i5-4570 CPU 3.2 GHz computer with a RAM of 8Gb. Exact solutions to the reported problem instances were obtained with the General Algebraic Modeling System (GAMS). GAMS is a high-level modeling system for mathematical programming and optimization. GAMS first compiles mathematical models formulated by the user at a high level, and then feeds them to a high-performance solver such as [7] as done in this study.

For large size problems, the full rest difference problem can be a hard problem to solve for with commercial solvers. A single round robin tournament example with 40 teams for which the number of games were equally distributed to four periods with five games on each period ran out of memory after about seven hours without being able to show the optimality of the found solution. In the above-mentioned problems the schedules were generated using the circle method.

We investigate how the top division professional football leagues in Europe were doing with regard to rest differences. Table 1 compares several leagues for the 2017-18 season. Besides rest difference comparisons, the table also includes the actual and optimal rest mismatch values of the seasons. To obtain the optimal rest mismatch values we added a set of binary variables to the model that were set to 1 whenever the p variable of a team was positive, and then minimized over the sum of these new rest mismatch indicator variables instead of over the sum of p variables. The league instances were solved in a matter of seconds for the rest difference criterion but the problem with the rest mismatch objective proved to be a harder problem to solve for those smaller instances as well sometimes taking about half an hour to solve for the full problem. The decomposed problem was very quick to solve.

The percentages shown under the column for the rest differences (mismatches) indicate how far away the actual schedule's value is from the optimal whereas the percentages under the column for the maximum possible values show how much the worst possible value has been

Table 1: Comparison of different football leagues for the 2017-18 season

Country	Canonical (Teams)	Rest dif- ferences	Opt	Max pos- sible	Rest mis- matches	Opt	Max pos- sible
Belgium	No (16)	162 (56%)	104	232 (30%)	139 (74%)	80	217 (36%)
England	No (20)	234 (86%)	126	306 (24%)	177 (121%)	80	227 (22%)
France	No (20)	230 (117%)	106	294 (22%)	202 (143%)	83	288 (30%)
Germany	No (18)	188 (71%)	110	224 (16%)	158 (86%)	85	214 (26%)
Italy	No (20)	172 (121%)	78	202 (15%)	160 (142%)	66	192 (17%)
Netherlands	No (18)	190 (94%)	98	298 (36%)	167 (109%)	80	274 (39%)
Portugal	Yes (18)	308 (77%)	174	410 (25%)	225 (91%)	118	293 (23%)
Russia	Yes (16)	200 (54%)	130	288 (31%)	156 (68%)	93	224 (30%)
Spain	Yes (20)	314 (64%)	192	484 (35%)	244 (83%)	133	364 (33%)
Turkey	Yes (18)	280 (52%)	184	394 (29%)	210 (68%)	125	297 (29%)

improved by the actual schedule. To understand how far off the actual rest difference numbers were from the worst case scenario, the *maximum possible* rest differences are also reported in the Table. For finding the worst case scenarios, a decomposed integer linear model was solved as a maximization problem. The worst case solutions were found using the decomposed approach because solving the full model was difficult. For the sake of brevity, the mathematical models of the maximization problems were not included here.

While we do not know the details of how many of these leagues decide about their schedules, the Belgian Jupiler League has long been using sophisticated optimization in its scheduling that is well documented (see for example [5]). Furthermore, they do not use a canonical schedule as in the Turkish league. In Belgium, the regular season -a not mirrored double round robin tournament with 16 teams- is followed by the playoff rounds to win the championship. During the regular season in 2017-18 the sum of rest differences were 162 with 139 rest mismatches. The optimal values were 104 and 80 respectively. While the tournament schedule has about 30% improvement from the maximum possible rest difference, there seems to be room for improvement should the organizers decide to pay attention to this criterion as well. The same can be said for all of the compared leagues. Note that the number of days used in each round and games played on each day of each round differ in each league. These differences also impact the totals for rest differences and rest mismatches in the leagues.

5 Conclusion

When opposing teams in matches have had rest durations of different lengths after their games in the previous round, the team with less rest is disadvantaged physically which makes the competition unfair. Thus, it is important to minimize such differences. In many leagues, the specific dates for playing the games in each round are determined after the schedule is announced. While officials have many other concerns such as security, broadcast ratings and shared stadiums when the dates of the games are determined, the authors believe that considered criteria should also include minimizing or equalizing the resting differences of opposing teams.

References

- [1] T. Atan and B. Çavdaroglu. Minimization of rest mismatches in round robin tournaments. *Computers and Operations Research*, 99:78–89, 2018.
- [2] D. Briskorn. Combinatorial properties of strength groups in round robin tournaments. *European Journal of Operational Research*, 192(3):744–754, 2009.
- [3] D. Briskorn and A. Drexl. IP models for round robin tournaments. *Computers and Operations Research*, 36(3):837–852, 2009.
- [4] G. Durán, S. Durán, J. Marenco, F. Mascialino, and P.A. Rey. Scheduling Argentina’s professional basketball leagues: A variation on the relaxed travelling tournament problem. *European Journal of Operational Research*, 2018.
- [5] D. Goossens. Optimization in sports league scheduling: Experiences from the Belgian pro league soccer. In *International Conference on Operations Research and Enterprise Systems*, pages 3–19. Springer, 2017.
- [6] D. Goossens and F.C.R. Spieksma. Soccer schedules in Europe: An overview. *Journal of Scheduling*, 15(5):641–651, 2012.
- [7] Gurobi Optimization Inc. Gurobi Optimizer Reference Manual, 2016.
- [8] S. Knust. Scheduling sports tournaments on a single court minimizing waiting times. *Operations Research Letters*, 36(4):471–476, 2008.
- [9] R. Rasmussen and M. Trick. Round robin scheduling - A survey. *European Journal of Operational Research*, 188(3):617–636, 2008.
- [10] D. Recalde, R. Torres, and P. Vaca. Scheduling the professional Ecuadorian football league by integer programming. *Computers and Operations Research*, 40(10):2478–2484, 2013.
- [11] W. Suksompong. Scheduling asynchronous round-robin tournaments. *Operations Research Letters*, 44(1):96–100, 2016.
- [12] S. Westphal. Scheduling the German basketball league. *Interfaces*, 44(5):498–508, 2014.

Alternative count regression models for modeling football outcomes

Alessandro Barbiero¹

University of Milan, Milan, Italy
alessandro.barbiero@unimi.it

Abstract

In this work, we propose the use of discrete counterparts of the Weibull distribution along with a copula function for modeling football results, as an alternative to existing bivariate Poisson regression models and extensions thereof. We expect that the choice of the marginal distribution and dependence structure, which try to capture known features of the data, can be beneficial in terms of fitting of the developed models; to check this conjecture, an application to the Italian Serie A championship is provided.

1 Introduction

Football is by far the most popular participant and spectator sport in the world. In many countries, especially in Europe, television and internet companies compete strongly to win the rights to broadcast games. Huge sums of money are involved, from players wages to transfer fees and sports betting. The simplicity of football's objectives and rules along with the uncertainty of games are probably responsible for such an inexhaustible attractiveness. The latter feature has captured the attentions of statisticians, who have proposed a multitude of stochastic models for analyzing (and predicting) several events associated with a football game: the first half or final result (expressed as number of goals scored by the two teams or simply as win-draw-loss), the number of shots-for and shots-against, the time to the first goal, the number of yellow or red cards, etc.

In this work, we propose the use of discrete counterparts of the Weibull distribution for modeling football results, as an alternative to existing bivariate Poisson regression models and modifications/extensions thereof, such as diagonally inflated or generalized Poisson models.

The simple bivariate Poisson model, with independent components, was the first used in football data analysis for modeling the outcome of a game (number of goals scored by the two competing teams) due to its ease of use and interpretation. Later, more complex models allowing for non-null correlation were explored, since real data often show a slight but non-negligible positive correlation between the numbers of goals scored by the two teams; or allowing for overdispersion and excess in draws, which usually characterize football outcomes.

The discrete Weibull distributions derived as analogues of the homonym continuous distribution seem to be more flexible than Poisson, since adjusting their two parameters can model a variety of different features. The numbers of goals scored by the two teams can be regarded as a joint observation from a bivariate random vector with discrete Weibull margins, linked through a copula function that accommodates dependence. The parameters of the distribution are assumed to depend on covariates such as the attack and defense abilities of the two teams and the "home effect". Several discrete Weibull regression models are proposed, by varying the type of discretization, the copula function, the choice of covariates, and are then applied to the Italian Serie A championship.

Even if the interpretation of parameters is less immediate than in Poisson models, yet they represent a suitable alternative, as the application demonstrates, and can be employed

as a statistical tool for better understanding the performance of teams in order to improve predictions, from a betting perspective, or to deploy corrective actions, from a managerial point of view.

The next Section briefly recaps the basic ideas underlying bivariate count regression models usually employed when analysing football results. In Section 3 we will draw our attention on alternative marginal distributions derived as discrete counterparts from the continuous Weibull distribution; in Section 4, we will focus on the choice of the copula function; in Section 5, we will discuss an application to the Italian Serie A championship.

2 Modelling the Numbers of Goal in a Football Game through a Count Regression

Focusing on the final result of a football game, many bivariate models have been discussed in statistical literature. Most of them are an extension of the simple bivariate Poisson model with independent components. These proposals, taking the cue from the bivariate Poisson model by Holgate [10] with correlated components, take into account the specific features these data usually exhibit, namely non-negligible correlation, overdispersion and bivariate zero-inflation, and propose count regression models where the two count variables are regressed towards covariates such as team attack and defence potential, home effect, etc. [16, 15, 6, 7, 11, 12, 1, 13]. More recently, some contributions suggested the use of alternative discrete probability distributions, related to the continuous Weibull random variable [5, 3], and dependence structures, by naturally considering copula functions.

In very general terms, the stochastic model can be structured as follows. Let Y_{1i} be the number of goals scored by the home team in game i , and Y_{2i} the number of goals scored by the away team in game i ; $p_1(y; \theta_{1i})$ and $p_2(y; \theta_{2i})$ are the discrete probability distributions modelling Y_{1i} and Y_{2i} , belonging to the same parametric family, with θ_{1i} and θ_{2i} being the distribution parameters (scalars or, more generally, vectors). These latter, or a transformation thereof, are expressed as a linear model, for example

$$g_j(\theta_{1ji}) = \beta'_{1j} \mathbf{x}_{1ji}, \quad g_j(\theta_{2ji}) = \beta'_{2j} \mathbf{x}_{2ji}$$

with $j = 1, \dots, p$, where p is the dimension of the parameter vectors θ_1 and θ_2 ; \mathbf{x}_{1ji} and \mathbf{x}_{2ji} are the two corresponding vectors of covariates, not necessarily the same; β_{1j} and β_{2j} the vector of regression parameters; $i = 1, \dots, n$, being n the sample size. For example, if we consider the Poisson distribution with parameter λ , being $p = 1$, the model can be written as

$$\begin{cases} Y_{1i} \sim \text{Pois}(\lambda_{1i}), & \log(\lambda_{1i}) = \beta'_1 \mathbf{x}_{1i} \\ Y_{2i} \sim \text{Pois}(\lambda_{2i}), & \log(\lambda_{2i}) = \beta'_2 \mathbf{x}_{2i} \end{cases}$$

In order to accommodate possible association between the two count variables, we resort to copulas. The cumulative distribution functions of the two count variables Y_{1i} and Y_{2i} , say F_{1i} and F_{2i} , are linked through a parametric bivariate copula function $C(u_1, u_2; \theta)$:

$$F(y_{1i}, y_{2i}) = C(F_{1i}(y_{1i}), F_{2i}(y_{2i}); \theta),$$

so that the joint probability mass function is derived as

$$P(Y_{1i} = y_{1i}, Y_{2i} = y_{2i}) = F(y_{1i}, y_{2i}) - F(y_{1i} - 1, y_{2i}) - F(y_{1i}, y_{2i} - 1) + F(y_{1i} - 1, y_{2i} - 1).$$

3 Marginal Distribution: Discrete Analogue of the Continuous Weibull Distribution

At least three probability distributions have been derived so far as a discrete counterpart of the continuous Weibull model.

A first discrete Weibull distribution was introduced by [18] and is usually referred to as ‘type I discrete Weibull distribution’, in order to distinguish it from two other models proposed later by [23] (type II discrete Weibull) and [21] (type III discrete Weibull). A continuous Weibull random variable (rv) T has probability density function given by

$$f_t(t; \lambda, \beta) = \lambda \beta t^{\beta-1} e^{-\lambda t^\beta} \quad t > 0, \quad (1)$$

with $\lambda, \beta > 0$, and cumulative distribution function (cdf)

$$F_t(t; \lambda, \beta) = 1 - e^{-\lambda t^\beta}. \quad (2)$$

If we consider the rv $Y = \lfloor T \rfloor$, where $\lfloor T \rfloor$ denotes the largest integer equal to or smaller than T , it can be easily shown that its probability mass function (pmf), defined on the non-negative integers only, is given by

$$p(y; q, \beta) = F_t(y+1) - F_t(y) = e^{-\lambda y^\beta} - e^{-\lambda (y+1)^\beta} = q^{y^\beta} - q^{(y+1)^\beta} \quad y \in \mathbb{N} = \{0, 1, 2, \dots\}, \quad (3)$$

with $q = e^{-\lambda}$, and then $0 < q < 1$. The corresponding cdf is

$$F(y; q, \beta) = 1 - q^{(y+1)^\beta} \quad y \in \mathbb{N}. \quad (4)$$

This distribution retains the expression of the cumulative distribution function of the continuous Weibull model – just compare Eq.(2) to Eq.(4). The first parameter q has a nice interpretation: since $P(X = 0) = 1 - q$, it represents the probability of a positive value. As to the second parameter β , it does not possess an equally immediate meaning. However, if we define the hazard rate function of Y as $r(y) = p(y)/P(Y \geq y)$, it has been shown [18] that $r(y)$ is a constant function if $\beta = 1$ (in this case, (3) reduces to the geometric pmf), an increasing function if $\beta > 1$, a decreasing function if $\beta < 1$.

Figure 1 displays the pmf of the type I discrete Weibull rv for several value combinations of q and β . From here the role of β , for a fixed value of q , clearly emerges: larger values of β lead to less dispersed distributions, with most of the probability mass concentrated on the first integer values; smaller values of β lead to more dispersed distributions. The expected value of the type I discrete Weibull rv cannot be generally computed in a closed form; it is equal to the infinite sum $\mathbb{E}(Y) = \sum_{y=1}^{\infty} y q^{y^\beta}$, which leads to a closed-form expression if and only if $\beta = 1$: $\mathbb{E}(Y) = q/(1-q)$. It is clear $\mathbb{E}(Y)$, fixed q , is a decreasing function of β . Its value can be approximated recalling the result in [14], involving the expected value $\mathbb{E}(T)$ of the corresponding continuous distribution, which ensures that the value $\mathbb{E}(Y)$ falls between $\mathbb{E}(T) - 1$ and $\mathbb{E}(T)$.

The first parameter of the type I discrete Weibull model can be related to explanatory variables \mathbf{x}_i through a complementary log-log link function: $\log(-\log(q_i)) = \boldsymbol{\alpha}'\mathbf{x}_i$. Additionally, even the second parameter β can be related to explanatory variables \mathbf{z}_i , not necessarily the same as for q , through the following natural link function (remember that β takes only positive values): $\log(\beta_i) = \boldsymbol{\gamma}'\mathbf{z}_i$.

Contrary to the Poisson rv, which cannot adequately model count data whose variance differs from the mean, which is a circumstance often occurring in practice, the type I discrete

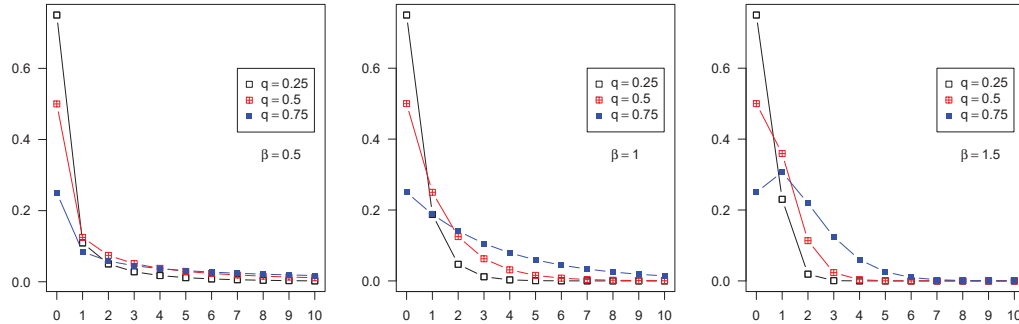


Figure 1: Graphs of the probability mass function of the type I discrete Weibull distribution for some combinations of its parameters q and β

Weibull rv can model both under-dispersed and over-dispersed data [8]. This distribution can also handle count data presenting an excess of zeros, arising in many physical situations (see again [8]); just remember that the probability of 0 is controlled by the q parameter only.

In regard to point and interval estimation of the parameters of the type I discrete Weibull distribution, one can refer to [2] and references therein, where several inferential procedures are considered and discussed and applicability issues are raised. The type I discrete Weibull model is implemented in the R environment [24] through the packages `DiscreteWeibull` [4] and `DWreg` [25].

As for the type II discrete Weibull rv, its distribution is derived by imposing that its hazard function has the same expression as the hazard function of the continuous Weibull rv. The resulting discrete distribution may have a finite or infinite support according to the value taken by the second parameter β of the continuous distribution. Such an odd feature depends on the fact that the hazard rate for a discrete model is bounded between 0 and 1, whereas this restriction is not needed for the hazard rate of a continuous distribution. For more details, we address the reader to the original paper [21].

As for the type III discrete Weibull rv, its pmf can be expressed as

$$P(Y = y; c, \beta) = e^{-c \sum_{j=1}^y j^\beta} [1 - e^{-c(y+1)^\beta}], \quad y \in \mathbb{N}, \quad (5)$$

letting by convention $\sum_{j=1}^y j^\beta = 0$ if $y = 0$; with $c > 0$ and $\beta \geq -1$. Note that $P(Y = 0) = 1 - e^{-c}$. Despite its unequivocal name, the type III discrete Weibull rv is not similar in functional form to any of the functions describing a continuous Weibull distribution, although the negative exponential terms in (5) reminds us of an analogous term in (1).

These latter two discrete models have not attracted much attention so far, due to the complex expression of their pmf, which makes parameter estimation not straightforward. However, their use in a count regression model can be still feasible, although some care has to be devoted to the choice of the link functions for their parameters.

4 Dependence Structure: the Clayton Copula

Lack of independence/incorrelation between the number of goals scored by the two teams in a football match was first claimed by [6]; in [17] the use of copulas for modeling two correlated count distributions related to football games was suggested perhaps for the first time. As

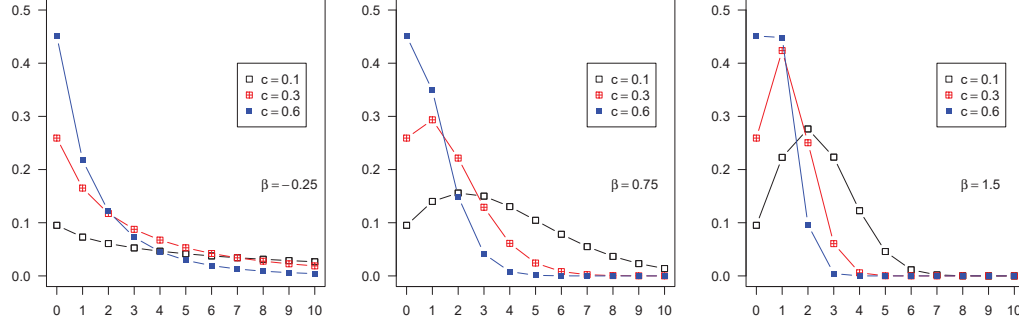


Figure 2: Graphs of the probability mass function of the type III discrete Weibull distribution for some combinations of its parameters c and β

anticipated in Section 2, we assume that the random variables modeling the number of goals scored by home and away teams, Y_{1i} and Y_{2i} , are no longer statistically independent, given the covariates; we model their dependence structure through a specific copula family.

Copulas represent a very flexible tool for modeling dependence among rvs. A bivariate copula is a joint cumulative distribution function in $[0, 1]^2$ with standard uniform margins U_1 and U_2 :

$$C(u_1, u_2) := P(U_1 \leq u_1, U_2 \leq u_2). \quad (6)$$

Sklar's theorem [22] states that if F is a joint distribution function with margins F_1 and F_2 , then there exists a copula $C : [0, 1]^2 \rightarrow [0, 1]$ such that, for all x_1, x_2 in $\mathbb{R} = [-\infty, +\infty]$,

$$F(x_1, x_2) = C(F_1(x_1), F_2(x_2)).$$

If the margins are continuous, then C is unique, otherwise C is uniquely determined on $\text{Ran}(F_1) \times \text{Ran}(F_2)$, with $\text{Ran}(F_j)$ denoting the range of F_j . Conversely, if C is a copula and F_1, F_2 are univariate cdfs, then the function F defined in (6) is a joint distribution function with margins F_1, F_2 . If the margins are continuous, the unique copula C is given by

$$C(u_1, u_2) = F(F_1^{-1}(u_1), F_2^{-1}(u_2)),$$

where F_j^{-1} denotes the generalized inverse of the marginal cdf F_j , i.e., $F_j^{-1}(t) = \inf \{x \in \mathbb{R} : F_j(x) \geq t\}$.

We recall that for any copula C the following constraint holds for any $(u_1, u_2) \in [0, 1]^2$:

$$\max(0, u_1 + u_2 - 1) \leq C(u_1, u_2) \leq \min(u_1, u_2); \quad (7)$$

the left and right members of the inequality are called Fréchet lower bound and Fréchet upper bound, respectively [9]. $M(u_1, u_2) = \min(u_1, u_2)$ is itself a copula, named “comonotonicity copula”, as well as $W(u_1, u_2) = \max(0, u_1 + u_2 - 1)$, the bivariate “countermonotonicity copula”.

From among the multitude of parametric bivariate copulas, we pick Clayton's copula, belonging to the so-called Archimedean family. The expression of the one-parameter Clayton copula is

$$C(u_1, u_2) = \max \left\{ (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}, 0 \right\}, \quad \theta \in (-1, +\infty) \setminus \{0\}. \quad (8)$$

The Clayton copula is interesting as it can model various kinds of dependence, ranging from comonotonicity in the limit as $\theta \rightarrow +\infty$, independence if $\theta \rightarrow 0$, and countermonotonicity if $\theta \rightarrow -1$.

The values of the θ parameter can be better interpreted resorting to the expression of Kendall's correlation ρ_τ for the Clayton copula (valid however for continuous margins only; see [19]):

$$\rho_\tau = \frac{\theta}{2 + \theta}.$$

Moreover, the Clayton copula is also able to capture lower tail dependence. For a bivariate absolutely continuous rv (X_1, X_2) , with marginal cdfs F_1 and F_2 , and generalized inverse functions F_1^{\leftarrow} and F_2^{\leftarrow} , respectively, the coefficient of lower tail dependence is defined as

$$\lambda_L = \lim_{u \rightarrow 0^+} P(X_2 \leq F_2^{-1}(u) | X_1 \leq F_1^{-1}(u)) = \lim_{u \rightarrow 0^+} C(u, u)/u,$$

and for the Clayton copula with $\theta > 0$, we have that

$$\lambda_L = 2^{-1/\theta} > 0.$$

Other well-known one-parameter bivariate copulas, such as the Gauss, the Plackett, and the Frank, do not meet this feature, being all asymptotically lower and upper tail independent. In Figure 3, the bivariate density plot of the Clayton copula is displayed for $\theta = 2$, along with the scatter plot of a bivariate random sample generated from the same copula (size $n = 5,000$). Thus, the Clayton copula may be a suitable candidate for modelling dependence between the numbers of goals scored in football games in a football championship, usually presenting a frequency of 0 – 0 draws higher than that which is caught by standard stochastic models.

[20] considered the Clayton-copula model with negative binomial marginals for modelling simultaneous spike-counts of neural populations, whereas, for computational reason, they are typically modeled by a Gaussian distribution. In [17], the Clayton copula is cited as a possible dependence structure for modelling the numbers of shots-for and shots-against a team in a football game.

5 Empirical Analysis: Italian Serie A Championship

We focus on the main Italian football championship, called “Serie A”, a professional league competition for football clubs located at the top of the Italian football league system. Since 2004-05, there have been 20 clubs playing in Serie A and as in most of the European countries a true round-robin format is used. During the season, each club plays each of the other teams twice; once at home and once away, eventually totaling 38 games. In the first half of the season, called the “andata”, each team plays once against each league opponent, for a total of 19 games. In the second half, called the “ritorno”, the teams play in the same exact order that they did in the first half of the season, the only difference being that home and away situations are switched. Since the 1994-95 season, teams earn three points for a win, one point for a draw and no points for a loss.

Here we are interested in analysing and modeling the final result for all the 380 games played throughout the season. For game i , $1 \leq i \leq 380$, we denote with y_{1i} the number of goals scored by the home team, h_i , and with y_{2i} the number of goals scored by the away team, a_i . Based on these data, one can estimate all the parameters involved in the regression model of Section 2, by using the maximum likelihood method, and for each game construct a theoretical

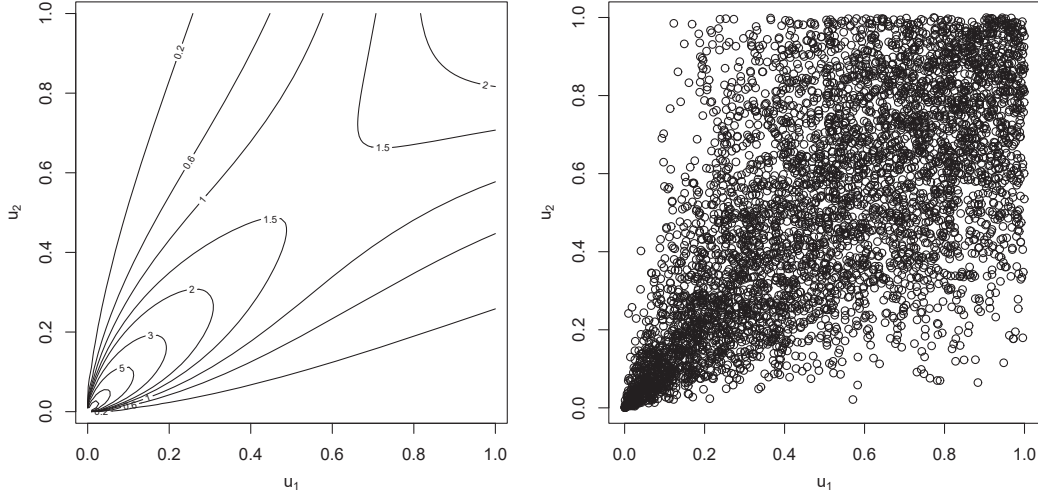


Figure 3: Clayton copula with parameter $\theta = 2$: contour density plot (on the left) and scatter plot of a random sample of size 5,000 (on the right)

joint probability table providing the probability of any possible outcome. As an overall result, by aggregating all the single theoretical outcomes, one can reconstruct the theoretical final scoreboard and compare it with its real counterpart.

We will start from a basic copula-based model, where the margins are assumed to follow the type I discrete Weibull distribution (3) and the dependence structure is induced by the Clayton copula (8). The two q parameters of the Weibull distribution are related to covariates as follows (see [3]):

$$\begin{cases} \log[-\log(q_{1i})] &= \mu^{(q)} + \text{home}^{(q)} + \text{att}_{h_i}^{(q)} + \text{def}_{a_i}^{(q)} \\ \log[-\log(q_{2i})] &= \mu^{(q)} + \text{att}_{a_i}^{(q)} + \text{def}_{h_i}^{(q)} \end{cases}$$

where $\mu^{(q)}$ is a constant term, $\text{home}^{(q)}$ is the “home effect”, $\text{att}_k^{(q)}$ and $\text{def}_k^{(q)}$ are the “attack” and “defence” parameters associated to q for team k . Note that apart from the constant term, the covariates for q are all dummy variables. The parameter β for the marginal distributions and the parameter θ of Clayton copula are assumed to be constant. Estimates for all parameters can be numerically obtained by maximizing the joint log-likelihood function. For the Italian Serie A championship, season 2015/16, the parameter estimates of the model above and their significance are reported in Table 1. Note the value of the estimate of β ($1.866 > 1$), which highlights how the distribution of scored goals is quite concentrated on the first integers; and the value of the estimate of θ (0.142), denoting a very slight correlation between the numbers of scored and conceded goals.

Additional models can be constructed by considering the other two discrete Weibull distribution, alternative copula functions, and different sets of covariates for the distribution parameters.

team	att ^(q)	def ^(q)
Atalanta	0.230	0.089
Bologna	0.322.	0.087
Carpi	0.314.	−0.250
Chievo	0.121	0.056
Empoli	0.184	0.030
Fiorentina	−0.334*	0.217
Frosinone	0.287.	−0.622***
Genoa	0.122	−0.058
Inter	0.072	0.243
Juventus	−0.557***	0.975***
Lazio	−0.138	−0.094
Milan	0.000	0.097
Napoli	−0.712***	0.400*
Palermo	0.334*	−0.378*
Roma	−0.714***	0.153
Sampdoria	−0.022	−0.285.
Sassuolo	−0.020	0.226
Torino	−0.117	−0.134
Udinese	0.283.	−0.387*
other parameters		
$\mu^{(q)}$	−1.037***	
home ^(q)	−0.385***	
β	1.866***	
θ	0.142.	

Table 1: Parameter estimates for the model applied to Italian Serie A championship 2015/2016. Attack and defense parameters satisfy the sum-to-zero constraint; so, for the last team in alphabetical order, Verona, we have $\text{att}^{(q)} = 0.346$ and $\text{def}^{(q)} = -0.364$. Significance codes for p -values: 0 “***” 0.001 “**” 0.01 “*” 0.05 “.” 0.1 “” 1

Acknowledgments

The author acknowledges the financial support to the present research by the Italian Ministry for Education. University and Research (FFABR 2017).

References

- [1] Gianluca Baio and Marta Blangiardo. Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, 37(2):253–264, 2010.
- [2] Alessandro Barbiero. A comparison of methods for estimating parameters of the type I discrete Weibull distribution. *Statistics and its Interface*, 9(2):203–212, 2016.
- [3] Alessandro Barbiero. Discrete Weibull regression for modeling football outcomes. *International Journal of Business Intelligence and Data Mining*, 2018.
- [4] Alessandro Barbiero. DiscreteWeibull: Discrete Weibull Distributions (type 1 and 3), 2018.
- [5] Georgi Boshnakov, Tarak Kharrat, and Ian G. McHale. A bivariate Weibull count model for forecasting association football scores. *International Journal of Forecasting*, 33(2):458–466, 2017.

- [6] Mark J. Dixon and Stuart G. Coles. 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, 1997.
- [7] David Dyte and Steven R. Clarke. A ratings based Poisson model for world cup soccer simulation. *Journal of the Operational Research Society*, 51(8):993–998, 2000.
- [8] James D. Englehardt and Ruochen Li. The discrete Weibull distribution: An alternative for correlated counts with confirmation for microbial counts in water. *Risk Analysis*, 31:2011, 2011.
- [9] Maurice Fréchet. Sur les tableaux de corrélation dont les marges sont données. *Annales de l’Université de Lyon. Section A: Sciences mathématiques et astronomie*, 3(14):53–77, 1951.
- [10] Philip Holgate. Estimation for the bivariate Poisson distribution. *Biometrika*, 51(1-2):241–287, 1964.
- [11] Dimitris Karlis and Ioannis Ntzoufras. Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):381–393, 2003.
- [12] Dimitris Karlis and Ioannis Ntzoufras. Bayesian modelling of football outcomes: using the Skellam’s distribution for the goal difference. *IMA Journal of Management Mathematics*, 20(2):133–145, 2009.
- [13] Dimitris Karlis and Ioannis Ntzoufras. Robust fitting of football prediction models. *IMA Journal of Management Mathematics*, 22(2):171–182, 2011.
- [14] Muhammad S. Ali Khan, Abdul Khaliq, and Abdulrehman M. Abouammoh. On estimating parameters in a discrete Weibull distribution. *IEEE Transactions on Reliability*, 38(3):348–350, 1989.
- [15] Alan J Lee. Modeling scores in the premier league: is Manchester United really the best? *Chance*, 10(1):15–19, 1997.
- [16] Michael J Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 1982.
- [17] Ian McHale and Phil Scarf. Modelling soccer matches using bivariate discrete distributions with general dependence structure. *Statistica Neerlandica*, 61(4):432–445, 2007.
- [18] Toshio Nakagawa and Shunji Osaki. The discrete Weibull distribution. *IEEE Transactions on Reliability*, 24(5):300–301, 1975.
- [19] Roger B. Nelsen. *An Introduction to Copulas*. New York, Springer-Verlag, 1999.
- [20] Arno Onken, Steffen Grünewälder, Matthias HJ Munk, and Klaus Obermayer. Analyzing short-term noise dependencies of spike-counts in macaque prefrontal cortex using copulas and the flash-light transformation. *PLoS computational biology*, 5(11):e1000577, 2009.
- [21] WJ Padgett and John D Spurrier. On discrete failure models. *IEEE Transactions on Reliability*, 34(3):253–256, 1985.
- [22] Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l’Institut de Statistique de l’Université de Paris*, 8:229–231, 1959.
- [23] William E Stein and Ronald Dattero. A new discrete Weibull distribution. *IEEE Transactions on Reliability*, 33(2):196–197, 1984.
- [24] R. Development Core Team. *R: A language and environment for statistical computing*. Vienna, Austria, 2019. version 3.5.3.
- [25] Veronica Vinciotti. *DWreg: Parametric Regression for Discrete Response*, 2016. R package version 2.0.

Integrated Break and Carryover Minimization Problem in Round Robin Tournaments

Burak Cavdaroglu¹ and Tankut Atan²

¹ Department of Industrial Engineering, Kadir Has University, Cibali, İstanbul, Turkey, 34083
burak.cavdaroglu@khas.edu.tr

² Department of Industrial Engineering, Bahçeşehir University, Beşiktaş, İstanbul, Turkey, 34353
sabrutankut.atan@eng.bau.edu.tr

Abstract

League scheduling is a field of operations research that has attracted scientists for many years. Break minimization and carryover effect minimization are considered to be two important criteria of fairness in league scheduling. There have been recent studies that integrate both criteria in a computationally hard problem. Some of these studies try to minimize the carryover effect in tournaments so that the number of breaks does not exceed a specific level, while some others apply schedule-then-break approach which first schedule the teams ignoring home-away requirements, then determine the home-away pattern for each team. In this work, we develop a solution method for this integrated problem which produces comparable results with that of a recent study. We show that our method drastically improves carryover effects value at the expense of an occasional increase in the number of breaks.

1 Introduction

League scheduling with respect to various fairness criteria is one of the popular research areas of sports scheduling literature. [7] is an annotated bibliography which provides a broad discussion of most popular fairness criteria considered in sports scheduling such as minimizing breaks, carryover effects, and travel distances or balancing time periods and venues. The minimization of breaks and the minimization of carryover effects are two of these criteria, especially focused on while timetabling the round robin tournaments.

If a team plays two home or away matches in two successive rounds, the alternating home-away pattern for the team is said to be *broken* and the team has a "break". Break minimization problem tries to minimize the total number of breaks in a round robin tournament. An example of a single round robin tournament with 12 teams (ranging from A to L) is given in Table 1. The first team of each match represents the Home team. Table 2, on the other hand, shows the home-away pattern (HAP) of each team. A highlighted "H" or "A" value designates the occurrence of a break. The last column shows how many breaks each team experiences through the season. The total number of breaks in this instance happens to be 34.

[2] shows that a single round robin tournament must have at least $n - 2$ breaks, where $\{1 \dots n\}$ is the set of teams and n is even. [2] also proves that this lower bound can be attained with a schedule (the so-called *canonical schedule*), in which the games of each round $R_i, i \in \{1 \dots n - 1\}$ is given by

$$R_i = \{[n, i]\} \cup \{[i + k, i - k]; k = 1, 2, \dots, n/2 - 1\} \quad (1)$$

where the numbers $i + k$ and $i - k$ are expressed as $i + k \bmod (n - 1)$ and $i - k \bmod (n - 1)$, respectively. Then, he identifies a HAP for each team that enables the canonical schedule to have $n - 2$ breaks with the following rule:

		Rounds											# Break
		R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	
		D-C	K-C	B-D	I-L	K-H	H-I	D-J	G-J	F-H	D-I	C-E	
Games		K-L	I-A	H-J	J-B	I-J	B-C	K-G	L-C	K-B	F-K	I-K	
		F-J	B-F	F-L	G-A	A-C	E-F	A-F	B-A	C-G	G-L	J-A	
		E-B	L-D	C-I	C-F	L-B	L-A	I-B	D-K	L-J	B-H	G-B	
		I-G	G-H	A-K	H-D	F-G	J-K	E-L	F-I	A-D	E-A	D-F	
		A-H	J-E	E-G	K-E	D-E	G-D	C-H	H-E	E-I	J-C	H-L	
Teams	A	H	A	H	A	H	A	H	A	H	A	A	1
	B	A	H	H	A	A	H	A	H	A	H	A	2
	C	A	A	H	H	A	A	H	A	H	A	H	3
	D	H	A	A	A	H	A	H	H	A	H	H	4
	E	H	A	H	A	A	H	H	A	H	H	A	3
	F	H	A	H	A	H	A	A	H	H	H	A	3
	G	A	H	A	H	A	H	A	H	A	H	H	1
	H	A	A	H	H	A	H	A	H	A	A	H	3
	I	H	H	A	H	H	A	H	A	A	A	H	4
	J	A	H	A	H	A	H	A	A	A	H	H	3
	K	H	H	A	H	H	A	H	A	H	A	A	3
	L	A	H	A	A	H	H	A	H	H	A	A	4

Table 1: Games of a round robin tournament

Table 2: Break occurrences

- (a) For each game $[n, i]$, team i plays at home if i is even, and away if i is odd.
- (b) For each game $[i + k, i - k]$, team $i + k$ plays at home if k is even, and away if k is odd.

Carryover effect, first introduced by [11], is the effect of a team on its opponent, which is transferred to the next game of that opponent. For instance, if Team k plays against Teams i and j in two consecutive rounds, Team k 's performance against Team j may be affected from the game between Teams k and i in the previous round. In this instance, Team i (causing team) is said to have a carryover effect on Team j (receiving team) through Team k (transferring team). The number of such carryovers Team j receives from Team i in a tournament is defined as c_{ij} . The carryover effect minimization problem aims to distribute these carryover effects among the team pairs as evenly as possible. In a balanced tournament, all c_{ij} values (where $i \neq j$) should be equal to one. In any tournament, the degree of how balanced carryover effects are distributed among team pairs can be measured by the *carryover effects (coe) value*, which is given by $\sum_{ij} c_{ij}^2$. In a balanced round robin tournament with n teams, *coe* value is equal to $n^2 - n$. For the tournament schedule of Table 1, we can specify the opponent of each team in each round as given in Table 3. For example, one can easily verify that, three times during the tournament, Team I plays with Team B 's opponent in the following round (i.e. $C_{BI} = 3$). Table 4 gives the c_{ij} value for each team pair. The carryover effects value can be computed as $\sum_{ij} c_{ij}^2 = 308$.

There are not many studies concerned in minimizing the total break and carryover effects value simultaneously. This paper considers the "Integrated Break and Carryover Minimization Problem", which aims to find a round robin schedule with at most one break per team and a reasonably small *coe* value.

The organization of this paper is as follows. After reviewing relevant literature in Section 2, we describe a solution method for solving the integrated problem of break and carryover minimization in Section 3. Our computational experiments and results are given in Section 4.

2 Previous Work

Minimizing break and minimizing carryover effects value are two criteria addressed separately in plenty of sports scheduling studies. [2] formulates a method for finding schedules with $n - 2$ breaks when n is even and no break when n is odd. Given a feasible tournament schedule without home-away assignment, [3] conjectures that the problem of finding a HAP for each team that

		Rounds										
		R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11
Teams	A	H	I	K	G	C	L	F	B	D	E	J
	B	E	F	D	J	L	C	I	A	K	H	G
	C	D	K	I	F	A	B	H	L	G	J	E
	D	C	L	B	H	E	G	J	K	A	I	F
	E	B	J	G	K	D	F	L	H	I	A	C
	F	J	B	L	C	G	E	A	I	H	K	D
	G	I	H	E	A	F	D	K	J	C	L	B
	H	A	G	J	D	K	I	C	E	F	B	L
	I	G	A	C	L	J	H	B	F	E	D	K
	J	F	E	H	B	I	K	D	G	L	C	A
	K	L	C	A	E	H	J	G	D	B	F	I
	L	K	D	F	I	B	A	E	C	J	G	H

Table 3: Opponent of each team in each round

		Teams												$\sum c_{ij}$
		A	B	C	D	E	F	G	H	I	J	K	L	
Teams	A	0	1	2	0	2	3	1	0	2	0	1	0	24
	B	1	0	0	1	0	2	0	2	3	1	0	2	24
	C	2	2	0	0	1	0	1	0	1	1	0	4	28
	D	0	1	0	0	1	2	1	0	0	3	4	0	32
	E	2	0	1	3	0	2	1	2	0	1	0	0	24
	F	1	2	2	2	2	0	0	0	2	0	0	1	22
	G	1	0	1	1	3	0	0	1	0	3	1	1	24
	H	0	2	0	0	2	0	1	0	2	1	3	1	24
	I	2	1	1	0	0	2	0	2	0	0	2	2	22
	J	0	1	1	1	1	0	3	3	0	0	1	1	24
	K	1	0	0	4	0	0	3	1	2	1	0	0	32
	L	2	2	4	0	0	1	1	1	0	1	0	0	28

Table 4: Carryover effects

minimizes the number of breaks is NP-hard. [9] proposes a polynomial-time algorithm which finds a home-away assignment for a given feasible tournament schedule if a solution with $n - 2$ breaks exists for the schedule, else returns ‘infeasible’.

[11] introduces the concept of carryover effect in round robin tournaments and identifies the problem of minimizing the carryover effects value. It is shown that when the number of teams is a power of two, a balanced schedule can always be achieved. The paper also conjectures that balanced schedules do not exist for other numbers of teams, proposes a method for unbalanced cases, and reports *coe* values for the team numbers 6, 10, 12, 14, 18, and 20. [1] improves *coe* values previously found by [11] for several team numbers by introducing the algebraic concept of starters in the group \mathbb{Z}_{n-1} . The study finds balanced schedules for $n = 20$ and 22 as well and thus disproves the conjecture of [11]. The procedure proposed by [1] gives the best solutions known to date, except for $n = 12$, which is obtained by another study [4].

However, despite this abundant amount of literature on each individual problem, studies focusing on solving both problems simultaneously are scarce. [12] proposes a two-phase method in which the first phase schedules the teams without any HAP and then the second phase assigns a feasible HAP for each team. In this approach, the first phase only focuses on some particular requirements (such as fixing the rounds of some games a priori or setting an upper bound for the carryover effects value) that do not involve HAPs. The second phase reduces to the break minimization problem for a given tournament schedule. Even though, they do not attempt to minimize carryover effects value in the first phase, they claim their approach can be extended to implement the carryover restrictions. To the best of our knowledge, the only study so far that attempts to solve the “Integrated Break and Carryover Minimization Problem” is [5], which first formulates an IP model, next discusses why the problem is computationally expensive, and finally provides a heuristic approach to solve the problem with at most one break for each team. They also discuss how the carryover effects can be further reduced at the expense of larger number of breaks per team.

3 Solution Method

It is possible to obtain good results for reducing *coe* value of a tournament by random permutations of weeks of the canonical schedule [10]. Encouraged by the success of such a simple approach, we are curious on how such an algorithm would work on the integrated problem of

break and carryover minimization. We apply a similar round swapping procedure for single round robin tournaments. The first step of the procedure is to construct a canonical schedule with Equation 1 and apply the rule of home-away assignment for each game. For instance, for $n = 12$, the games of the first three rounds in the canonical schedule would be as given in Table 5. The first team in each game shows the home team.

Round	Games					
1	12-1	11-2	3-10	9-4	5-8	7-6
2	2-12	1-3	4-11	10-5	6-9	8-7
3	12-3	2-4	5-1	11-6	7-10	9-8

Table 5: First three rounds of a canonical schedule for $n = 12$

The canonical schedule is known to have maximum carryover effects value [8] and minimum number of breaks [2]. Therefore, any schedule obtained by changing the order of rounds in the canonical schedule (without disturbing the home-away assignments) is likely to have a smaller *coe* value and larger number of breaks as against the canonical schedule. We run a round swapping algorithm which randomizes the order of rounds to construct schedules isomorphic [6] to the canonical schedule, and calculates the *coe* and total break values for each of them. After generating a number of isomorphic random schedules, the schedules with the best *coe* and the best break values are identified. The results for single round robin tournaments with several number of teams are shown in Table 6. The third (fourth) row denotes the *coe* value (the number of breaks) of the best schedule among one million schedules and the number of breaks (the *coe* value) corresponding to this schedule. The Pareto frontier for one million schedules of $n = 18$ is also provided in Figure 1. Even though *coe* values we found are relatively better than the values reported in [5], the break values are far worse. This is mainly because randomizing the rounds without disturbing the home-away assignments drastically increases the number of breaks. In order to overcome this issue, we have decided to utilize the round swapping algorithm only for identifying the schedule with the smallest *coe* value. Next, we have solved a break minimization problem for this particular schedule.

# of teams (n)	10		12		14		16		18	
	Coe	Break	Coe	Break	Coe	Break	Coe	Break	Coe	Break
Best coe	136	26	192	60	254	68	330	118	406	126
Best break	468	8	392	16	634	22	660	40	884	44

Table 6: Best *coe* and best break values

The break minimization problem we need to solve for a given schedule can be defined as follows. Assume T and R are the sets of teams and rounds, respectively. Let $x_{i,j,r}$ be the parameter denoting the games of each round in a schedule. $x_{i,j,k}$ equals to 1 if team i plays against team j in round r , and 0 otherwise. $h_{i,r}$ is a decision variable, which is equal to 1 if team i plays at home in round r , and 0 otherwise. $b_{i,r}$ is the other binary variable which shows whether there is an occurrence of break for team i in round r . The integer programming formulation of the problem is:

$$\min \quad z = \sum_{i \in T} \sum_{r \in R} b_{i,r} \quad (2)$$

subject to:

$$h_{i,r} + h_{j,r} \geq x_{i,j,r} \quad \forall i, j \in T \mid i \neq j, \forall r \in R \quad (3)$$

$$2 - h_{i,r} + h_{j,r} \geq x_{i,j,r} \quad \forall i, j \in T \mid i \neq j, \forall r \in R \quad (4)$$

$$h_{i,r} + h_{i,r+1} \leq 1 + b_{i,r+1} \quad \forall i \in T, \forall r \in R \quad (5)$$

$$2 - h_{i,r} - h_{i,r+1} \leq 1 + b_{i,r+1} \quad \forall i \in T, \forall r \in R \quad (6)$$

$$h_{i,r}, b_{i,r} \in \{0, 1\} \quad \forall i \in T, \forall r \in R \quad (7)$$

The objective function minimizes the sum of breaks. Constraint 3 (Constraint 4) states that if there is a game between team i and team j in round r , at least one team should play at home (away). Constraint 5 (Constraint 6) determines the occurrence of a break in round $r + 1$ if team i plays at home (away) in the successive rounds r and $r + 1$. Constraint 7 sets the types of decision variables.

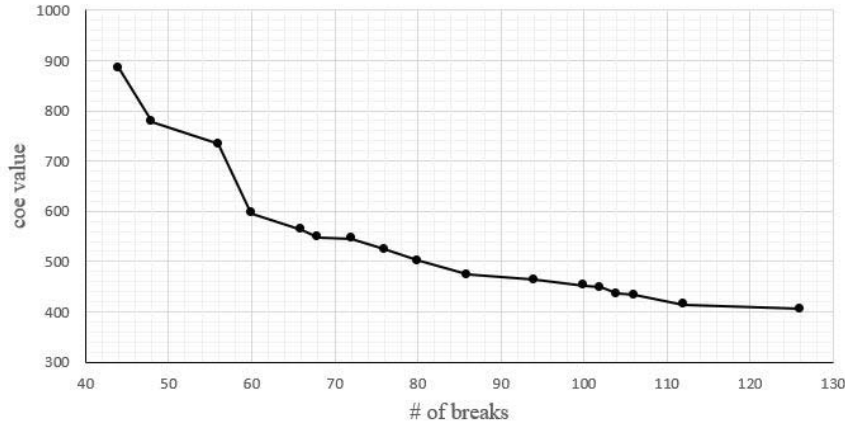


Figure 1: The Pareto frontier for $n = 18$

4 Computational Experiments

Round swapping procedure is coded in VBA and executed on an Intel Core Xeon, 2.4 GHz computer with a RAM of 12Gb and 8 processors. We run the code until one million random schedules are generated, which takes less than half an hour in each experiment. Break minimization problems are solved in the same computing system using GAMS with GUROBI solver. Each problem takes less than 5 minutes to be solved to optimality.

Table 7 reports the best known results to date for the problem of *coe* value minimization. The second row shows the lower bound of the problem for each number of teams n . The third row gives the best *coe* value (and the first study attaining it) without considering the number

of breaks. For $n = 16$, the *coe* value is equal to the lower bound since a balanced schedule can be found. For the remaining n values, the optimality has not been proven yet. The last three rows provide the best *coe* values achieved so far given that the number of breaks for each team is less than or equal to one, two and three respectively. The values in parenthesis give the number of breaks in the corresponding schedule. These values indicate a significant increase in *coe* compared to the values of the third row since the number of breaks is restricted as well in these instances.

# of teams (n)	10	12	14	16	18
$n \cdot (n - 1)$	90	132	182	240	306
any # of breaks	108 [1]	160 [4]	234 [1]	240 [11]	340 [1]
$b_i \leq 1$	192 (8) [5]	318 (10) [5]	446 (12) [5]	626 (14) [5]	944 (16) [5]
$b_i \leq 2$	144 (12) [5]	212 (18) [5]	344 (26) [5]	472 (30) [5]	646 (30) [5]
$b_i \leq 3$	144 (12) [5]	212 (16) [5]	302 (24) [5]	396 (34) [5]	556 (40) [5]

Table 7: Best *coe* value for each number of teams

In Table 8, we summarize the experimental results of our solution method for several n values. The last three rows provide the solution of the break minimization problem with a “break per team” restriction (the number of breaks per team $\leq 2, 3$, or ∞ respectively) for the schedule with the best *coe* value. When our results are compared with the results found by [5], it can be noted that our method gains a significant improvement in *coe* value at the expense of a slight increase in the number of breaks. For example, with the restriction of $b_i \leq 3$, we improve the *coe* value by 6%, 9%, 16%, 17%, and 27% for $n = 10, 12, 14, 16$, and 18 respectively. Meanwhile, the number of breaks has increased by 0%, 13%, 8%, 0%, and 15% respectively. This comparison indicates that the level of improvement in *coe* value becomes better with the increasing number of teams while the derogation in the number of breaks has an irregular pattern. For reference purposes, in Table 9 of Appendix A we provide a template for the single round robin tournament with 18 teams where the number of breaks per team is not restricted.

# of teams (n)	10		12		14		16		18	
	Coe	Break	Coe	Break	Coe	Break	Coe	Break	Coe	Break
$b_i \leq 2$		12		18		infsbl		infsbl		infsbl
$b_i \leq 3$	136	12	192	18	254	26	330	34	406	46
b_i unbounded		12		18		26		32		42

Table 8: Computational results

References

- [1] Ian Anderson. Combinatorial designs and their applications. In K. Quinn, B. Webb, C. Rowley, and F.C. Holroyd, editors, *Combinatorial Designs and their Applications*, pages 1–16. CRC Press, 1997.
- [2] Dominique de Werra. Scheduling in sports. In P. Hansen, editor, *Studies on Graphs and Discrete Programming*, pages 381–395. North-Holland, 1981.

- [3] Matthias Elf, Michael Jünger, and Giovanni Rinaldi. Minimizing breaks by maximizing cuts. *Operations Research Letters*, 31(5):343–349, 2003.
- [4] Allison C.B. Guedes and Celso C. Ribeiro. A heuristic for minimizing weighted carry-over effects in round robin tournaments. *Journal of Scheduling*, 14(6):655–667, 2011.
- [5] Dilek Günneç and Ezgi Demir. Fair-fixtue: Minimizing carry-over effects in football leagues. *Journal of Industrial and Management Optimization*, 2018. doi: 10.3934/jimo.2018110.
- [6] Tiago Januario, Sebastián Urrutia, Celso C. Ribeiro, and Dominique De Werra. Edge coloring: A natural model for sports scheduling. *European Journal of Operational Research*, 254(1):1–8, 2016.
- [7] Graham Kendall, Sigrid Knust, Celso C. Ribeiro, and Sebastian Urrutia. Scheduling in sports: An annotated bibliography. *Computers and Operations Research*, 37:1–19, 2010.
- [8] Erik Lambrechts, Annette M.C. Ficker, Dries Goossens, and Frits C.R. Spieksma. Round-robin tournaments generated by the circle method have maximum carry-over. *Mathematical Programming*, 172(1-2):277–302, 2018.
- [9] Ryuhei Miyashiro and Tomomi Matsui. A polynomial-time algorithm to find an equitable home-away assignment. *Operations Research Letters*, 33(3):235–241, 2005.
- [10] Ryuhei Miyashiro and Tomomi Matsui. Minimizing the carry-over effects value in a round-robin tournament. In *Proceedings of the 6th International Conference on the Practice and Theory of Automated Timetabling*, pages 460–463. PATAT, 2006.
- [11] Kenneth G. Russell. Balancing carry-over effects in round robin tournaments. *Biometrika*, 67(1):127–131, 1980.
- [12] Michael A. Trick. A schedule-then-break approach to sports timetabling. In E. Burke and W. Erben, editors, *International Conference on the Practice and Theory of Automated Timetabling III*, volume 2239, pages 242–253. Springer, 2001.

A Schedule Template for 18 Teams

Table 9 provides a schedule template for the single round robin tournament with 18 teams. The first team is the one who plays at home. The *coe* value is 406 and the number of breaks is 42 (2 teams with no break, 2 teams with one break, 6 teams with two breaks, 6 teams with three breaks, 2 teams with five breaks). The occurrence of a break for a team is highlighted in **bold**.

Round	Games								
1	11-1	10-12	13-9	8-14	15-7	6-16	17-5	4-18	2-3
2	1-10	9-11	12-8	7-13	14-6	5-15	16-4	3-17	18-2
3	6-1	7-5	4-8	9-3	10-2	11-18	17-12	13-16	15-14
4	1-9	8-10	7-11	12-6	5-13	14-4	3-15	16-2	18-17
5	1-7	6-8	9-5	4-10	11-3	2-12	13-18	17-14	15-16
6	12-1	13-11	10-14	9-15	16-8	7-17	18-6	5-2	3-4
7	1-13	14-12	11-15	10-16	17-9	8-18	2-7	6-3	4-5
8	1-17	18-16	15-2	3-14	13-4	12-5	11-6	7-10	9-8
9	16-1	17-15	14-18	2-13	3-12	4-11	5-10	6-9	8-7
10	1-5	6-4	7-3	8-2	9-18	10-17	11-16	12-15	13-14
11	14-1	15-13	16-12	17-11	18-10	2-9	3-8	4-7	5-6
12	1-4	5-3	6-2	7-18	8-17	9-16	15-10	11-14	12-13
13	18-1	17-2	3-16	4-15	14-5	13-6	7-12	11-8	10-9
14	1-3	2-4	5-18	6-17	16-7	15-8	9-14	10-13	12-11
15	1-15	14-16	13-17	18-12	11-2	3-10	4-9	8-5	7-6
16	8-1	9-7	10-6	5-11	12-4	3-13	2-14	15-18	17-16
17	1-2	18-3	4-17	16-5	6-15	7-14	13-8	9-12	11-10

Table 9: Schedule template for $n = 18$

Overcoming the incentive incompatibility of tournaments with multiple group stages

László Csató*

Institute for Computer Science and Control, Hungarian Academy of Sciences (MTA SZTAKI)
Laboratory on Engineering and Management Intelligence, Research Group of Operations
Research and Decision Systems

Corvinus University of Budapest (BCE)
Department of Operations Research and Actuarial Sciences

Budapest, Hungary

19th January 2019

Es ist dabei selbst die historische Wahrheit eine Nebensache, ein erfundenes Beispiel könnte auch dienen; nur haben historische immer den Vorzug, praktischer zu sein und den Gedanken, welchen sie erläutern, dem praktischen Leben selbst näher zu führen.¹

(Carl von Clausewitz: *Vom Kriege*)

Abstract

The paper discusses the incentive incompatibility of tournaments with multiple group stages. This design divides the competitors into round-robin groups in the preliminary and main rounds. The higher ranked teams from the preliminary round qualify to the next stage such that matches are not repeated in the main round if two qualified teams have already faced each other. It is proved that these tournament systems, widely used in handball and other sports, violate strategy-proofness since the contestants prefer to carry over better results to the main round. Some historical examples are presented where a team was *ex ante* disinterested in winning by a high margin. We suggest two incentive compatible mechanisms and compare them with the original format via simulations. Carrying over half of the points scored in the preliminary round turns out to be a promising policy.

JEL classification number: C44, C63, D71, Z20

MSC class: 62F07, 68U20, 91A80, 91B14

Keywords: OR in sports; tournament design; strategy-proofness; simulation; handball

* E-mail: laszlo.csato@uni-corvinus.hu

¹ “Historical correctness is a secondary consideration; a case invented might also serve the purpose as well, only historical ones are always to be preferred, because they bring the idea which they illustrate nearer to practical life.” (Source: Carl von Clausewitz: *On War*, Book 2, Chapter 6 – On Examples, translated by Colonel James John Graham, London, N. Trübner, 1873. <http://clausewitz.com/readings/OnWar1873/TOC.htm>)

1 Introduction

It is known at least since Arrow's pioneering work (Arrow, 1950) that the real world is full of decision paradoxes. This is true even though Arrow's impossibility theorem neglects the fact that voters are strategic actors. According to the famous Gibbard-Satterthwaite theorem (Gibbard, 1973; Satterthwaite, 1975), all fair voting rules are susceptible to tactical voting in the case of at least three alternatives: there always exists a voter who can achieve a better outcome by being insincere.

This result may suggest that strategy-proofness is difficult to achieve in practice. Nonetheless, there are several cases when incentive compatible designs exist, but a widely used procedure is manipulable. For example, Tasnádi (2008) demonstrated that the Hungarian mixed-member electoral system, applied between 1990 and 2010, suffers from the population paradox as the governing coalition may lose seats either by getting more votes or by the opposition obtaining fewer votes. Similarly, the invariant method (Pinski and Narin, 1976) – characterised by Palacios-Huerta and Volij (2004), and used to quality-rank academic journals – is subject to manipulation because a journal can boost its performance by making additional citations to other journals (Kóczy and Strobel, 2009).

Strategy-proofness is an especially relevant issue in sports where all contestants are familiar with the high-stake decisions involved, and they can obviously behave as strategic actors. Consequently, any tournament design should provide the players with the appropriate incentives to perform (Szymanski, 2003).

Scientific analysis of sports ranking rules from the perspective of incentive compatibility has started recently, although sporting applications of operations research proliferate in the academic world (Wright, 2009, 2014). Kendall and Lenten (2017) provides probably the first comprehensive review of sports regulations resulting in unexpected consequences. On the basis of the examples presented, three possible situations can be identified in which a team might prefer losing a game to winning it: (1) when a team might gain advantages in the next season; (2) when a lower ranked team can still qualify and it might face a preferred competitor in a later stage; (3) when a team is strictly better off by losing in certain situations due to ill-constructed rules.

The classical example for the first situation arises from the reverse order applied in the traditional set-up of player drafts, which aims to increase competitive balance over time: if a team is still certainly eliminated from the play-off, it creates a perverse incentive to tank in the later games (Fornwagner, 2018; Lenten et al., 2018).

The second situation occurred in Badminton at the 2012 Summer Olympics – Women's doubles (Kendall and Lenten, 2017, Section 3.3.1), and has inspired some game-theoretical works addressing the strategic manipulation problem (Pauly, 2014; Vong, 2017).

However, in the first case, the rules are deliberately designed to support underdogs, and in the second case, the team gains only in expected terms. Here, the remaining third situation will be discussed, when tournament rules are constructed such that a team is guaranteed to benefit from performing weaker. Probably the first academic work studying this issue is Dagaev and Sonin (2017), where the authors prove that tournament systems, consisting of multiple round-robin and knockout tournaments with noncumulative prizes, are often incentive incompatible.

While several football tournaments – such as qualifications for FIFA Worlds Cups (Dagaev and Sonin, 2013; Csató, 2017a), UEFA club competitions (Dagaev and Sonin, 2017; Csató, 2018f), and UEFA European Championships (Csató, 2018c,g) – have been shown to be vulnerable to manipulation, it is far from trivial to identify a misaligned rule

in practice since there is a low probability of failure because a scandal usually involves such an enormous cost that the particular design is almost certain to be never used again.

Perhaps the most famous case is a football match, [Barbados vs Grenada \(1994 Caribbean Cup qualification\)](#), when a sudden-death goal scored in extra time counted as double, creating an incentive to concede a goal at the end of the match in order to gain additional time for a necessary two-goal win ([Kendall and Lenten, 2017](#), Section 3.9.4). The Barbadians exploited this perverse rule by scoring an own goal in the 87th minute ([Dagaev and Sonin, 2017](#), Note 1). Nonetheless, this match had not affected any third team, so one can agree with the decision of FIFA not to penalise Barbados as the players were striving for the best outcome conditional upon the prevailing rules. Despite that, the strange regulation has not been applied since then.

We have reviewed two other football matches experiencing similar problems. In the first, a Dutch team, SC Heerenveen was better off by losing than by playing a draw ([Csató, 2019b](#)). In the second, both teams were interested in achieving a draw in order to grab the only chance to qualify ([Csató, 2018d](#)).

A similar situation was prevented by a particular FIBA (International Basketball Federation) rule saying that ‘*if a player deliberately scores in the team’s own basket, it is a violation and the basket does not count*’: in the [men’s tournament of the 2014 Asian Games Basketball Competition](#), a Philippine player shot at his own basket against Kazakhstan in order to force overtime and thus increase the margin of victory ([Carpio, 2014](#)).

The paper will highlight that tournaments with multiple group stages, in which some match results from the preliminary round are carried over to the main round, suffer from incentive incompatibility. First, we present a handball math where a team had an incentive not to win by a high margin. Second, it is proved that this particular tournament design violates strategy-proofness in its current form. Finally, we give a mechanism to guarantee incentive compatibility: to carry over a monotonic transformation of all preliminary round results to the main round, regardless that some matches were played against teams already eliminated from the tournament. It is also shown via simulations that carrying over half of the points scored in the preliminary round does not affect essentially the selective ability and the competitive balance of the tournament, while it provides strategy-proofness and reduces the influence of seeding the teams into pots.

The rest of the paper proceeds as follows. Section 2 brings an example from handball, which may be even more serious than the football match Barbados vs Grenada as a seemingly unfair behaviour of a team led to the elimination of a third team. Section 3 contains the theoretical model and proves that a tournament with multiple group stages usually violates strategy-proofness. Section 4 lists some recent tournaments applying this design. In Section 5, we provide two incentive compatible mechanisms for organising these tournaments and explore their characteristics with respect to selective ability and competitive balance via simulations. Finally, Section 5 summarises our main findings.

2 A real-world example of manipulation

The [European Men’s Handball Championship](#) is the official biannual competition for the senior men’s national handball teams of Europe since 1994, organised by the EHF (European Handball Federation), the umbrella organization for European handball.² The

² This section is mainly based on the official homepage of the [11th Men’s European Handball Championship \(EHF Euro 2014\)](#). We will cite only those documents which concern the ranking of teams.

11th European Men's Handball Championship (EHF Euro 2014) was held in Denmark between 12 and 26 January 2014. Sixteen national teams participated in the tournament. In the preliminary round, they were divided into four groups (A-D) to play in a round-robin format. The top three teams in each group qualified to the main round. Teams from Groups A and B of the preliminary round composed the first main round group X, while teams from Groups C and D of the preliminary round composed the second main round group Y. The main round groups were also organised in a round-robin format, but all matches (consequently, results and points), played in the preliminary round between the teams that were in the same main round group, were kept and remained valid for the ranking of the main round. Figure A.1 in the Appendix gives an overview of this tournament design.

In the groups of the preliminary and main rounds, two points were awarded for a win, one point for a draw and zero points for a defeat. Teams were ranked by adding up their number of points. If two or more teams had an equal number of points, the following tie-breaking criteria were used after the completion of all group matches (EHF, 2014a, Articles 9.12 and 9.24):

- a) *Higher number of points obtained in the group matches played amongst the teams in question;*
- b) *Superior goal difference from the group matches played amongst the teams in question;*
- c) *Higher number of goals scored in the group matches played amongst the teams in question;*
- d) *Superior goal difference from all group matches (achieved by subtraction);*
- e) *Higher number of goals scored in all group matches.*

A strange situation emerged in Group C of the preliminary round, which requires further investigation. On 16 January 2014, each team in the group had one more game to play. Table 1 shows the known results and the preliminary standing of the group.

Consider the possible scenarios from the perspective of Poland. This team is certainly eliminated if it does not win against Russia. Poland carries over 0 points, 46 goals for and 48 goals against to the main round if it wins against Russia and Serbia plays at least a draw against France. On the other hand, if Poland wins by x goals against Russia and Serbia loses, there will be three teams with 2 points, which obtained 2 points in the group matches played among them. Consequently, the further tie-breaking criteria should be applied: Poland, Russia, and Serbia will have head-to-head goal differences of $x - 1$, $2 - x$ and -1 , respectively. As $x - 1 > -1$, Poland will qualify.

Serbia is eliminated as the fourth team if $1 \leq x \leq 2$. Russia and Serbia have the same head-to-head goal difference if $x = 3$, hence higher number of goals scored against the three teams with 2 points breaks the tie. It is 45 for Serbia and at least 27 for Russia, thus Russia qualifies if it scores at least 19 goals against Poland (if Poland vs Russia is 21-18, then the third place will depend on the result of Serbia vs France). If $x \geq 4$, then Serbia has a better head-to-head goal difference than Russia, so Serbia qualifies and Russia is eliminated.

To summarise, if Poland wins, it carries over its result against Russia (2 points) or Serbia (0 points) to the main round, thus Poland has every incentive to qualify together

Table 1: 11th European Men’s Handball Championship (EHF Euro 2014) – Group C

(a) Match results

Date	First team	Second team	Result
13 January 2014, 18:00	Serbia	Poland	20-19
13 January 2014, 20:15	France	Russia	35-28
15 January 2014, 18:00	Russia	Serbia	27-25
15 January 2014, 20:15	Poland	France	27-28
17 January 2014, 18:00	Poland	Russia	to be played
17 January 2014, 20:15	Serbia	France	to be played

(b) Standing after two matchdays

Pos = Position; W = Won; D = Drawn; L = Lost; GF = Goals for; GA = Goals against; GD = Goal difference; Pts = Points. All teams have played 2 matches.

Pos	Team	W	D	L	GF	GA	GD	Pts
1	France	2	0	0	63	55	8	4
2	Serbia	1	0	1	45	46	-1	2
3	Russia	1	0	1	55	60	-5	2
4	Poland	0	0	2	46	48	-2	0

with Russia. Hence, it is unfavourable for Poland to win by more than three goals against Russia as this scenario yields no gain in the main round but may lead to a loss of 2 points if Serbia is defeated by France. Russia is clearly better off by a smaller defeat.

In fact, Poland vs Russia was 24-22 and Serbia vs France was 28-31, so France, Poland, and Russia qualified to the main round with 4, 2 and 0 points, respectively. Naturally, it is not a proof that the Polish team deliberately manipulated, but the circumstances are at least suspicious. For example, the result of Poland vs Russia was 10-14 after 30 minutes (half-time), while the match stood at 21-16 in the 48th, 22-17 in the 50th, and 23-18 in the 52nd minute (EHF, 2014b).³

The potentially unfair behaviour of Poland resulted in the elimination of a third, innocent team, Serbia, which makes the example especially worrying. Furthermore, the situation could not have been improved by playing the last group matches simultaneously because Poland’s dominant strategy was independent of the result of the game played later. It seems to be a persuading argument against the rules of 11th European Men’s Handball Championship (EHF Euro 2014).

3 The model

In this section, we build a model of a tournament consisting of round-robin preliminary and main rounds, where matches played in the preliminary round against teams qualified to the same main round group are carried over. It will be revealed that these systems are incentive incompatible, that is, they are vulnerable to a manipulation such as the one presented in Section 2. Our notations follow Csató (2018c) in certain details since the

³ A video of the match Poland vs Russia is available at <https://www.youtube.com/watch?v=dQvEAzyBgGo>.

qualification system discussed there is also based on round-robin groups.

Definition 3.1. *Round-robin tournament:* Let X be a nonempty finite set of at least two teams, $x, y \in X$ be two teams and $v : X \times X \rightarrow \{(v_1; v_2) : v_1, v_2 \in \mathbb{N}\} \cup \{\star\}$ be a function such that $v(x, y) = \star$ if and only if $x = y$, where \mathbb{N} denotes the set of nonnegative integers. The pair (X, v) is called a *round-robin tournament*.

Function v describes game results with the number of goals scored by the first and second team, respectively.

Definition 3.1 allows for a home-and-away round-robin tournament, any two teams may play each other once at home and once at away. The first team is the one playing at home.

Definition 3.2. *Single round-robin tournament:* Round-robin tournament (X, v) is *single* if $v_1(x, y) = v_2(y, x)$ for all $x, y \in X$.

In a single round-robin tournament, any two teams play each other only once (often at a neutral site), so the order of the teams has no significance.

Definition 3.3. *Incomplete round-robin tournament:* Let X be a nonempty finite set of at least two teams, $x, y \in X$ be two teams and $v : X \times X \rightarrow \{(v_1; v_2) : v_1, v_2 \in \mathbb{N}\} \cup \{\star\}$ be a function such that $v(x, y) = \star$ if $x = y$ and $v(x, y) = \star$ implies $v(y, x) = \star$ if $x \neq y$. The pair (X, v) is called an *incomplete round-robin tournament*.

In an incomplete round-robin tournament, some matches between the teams may remain to be played. Any round-robin tournament is an incomplete round-robin tournament, too.

Definition 3.4. *Ranking in incomplete round-robin tournaments:* Let \mathcal{X} be the set of incomplete round-robin tournaments with a set of teams X . A *ranking method* R maps any function v of \mathcal{X} into a strict order $R(v)$ on the set X .

Let (X, v) be an incomplete round-robin tournament, $R(v)$ be its ranking and $x, y \in X$, $x \neq y$ be two different teams. x is said to be ranked higher (lower) than y if and only if $x \succ_{R(v)} y$ ($x \prec_{R(v)} y$).

Let $x, y \in X$, $x \neq y$ be two different teams and $v(x, y) = (v_1(x, y); v_2(x, y))$. It is said that team x wins over team y if $v_1(x, y) > v_2(x, y)$ (home) or $v_1(y, x) < v_2(y, x)$, team x loses to team y if $v_1(x, y) < v_2(x, y)$ or $v_1(y, x) > v_2(y, x)$ and teams x draws with team y if $v_1(x, y) = v_2(x, y)$.

In some professional team sports (basketball, ice hockey, volleyball, etc.) draws are prohibited. Since we want to keep the model as general as possible, it is assumed that no matches result in a draw. Introducing this constraint will cause no problem because we are searching for situations that are vulnerable to manipulation, which becomes more difficult on a smaller domain.

Assumption 1. *No matches result in a draw: $v_1(x, y) \neq \star$ implies $v_1(x, y) \neq v_2(x, y)$ for any incomplete round-robin tournament (X, v) and teams $x, y \in X$.*

The ranking is usually based on the number of points scored.

Definition 3.5. *Number of points:* Let (X, v) be an incomplete round-robin tournament and $x \in X$ be a team. Denote by $N_v^w(x)$ the number of wins and by $N_v^l(x)$ the number of losses of team x in (X, v) , respectively. The *number of points* of team x is $s_v(x) = \alpha N_v^w(x) + \beta N_v^l(x)$ such that $\alpha > \beta$.

In other words, a win means α points and a loss gives β points.

Number of points does not necessarily induce a strict order on the set of teams, hence some tie-breaking rules are required.

Definition 3.6. *Goal difference:* Let (X, v) be an incomplete round-robin tournament and $x \in X$ be a team. The *goal difference* of team x is

$$gd_v(x) = \sum_{y \in X, v(x,y) \neq \star} (v_1(x, y) - v_2(x, y)) + \sum_{y \in X, v(x,y) \neq \star} (v_2(y, x) - v_1(y, x)).$$

Goal difference is the difference between the number of goals scored for team x and the number of goals conceded by team x .

Definition 3.7. *Head-to-head results:* Let (X, v) be a round-robin tournament and $x \in X$ be a team. Denote by $L \subseteq X \setminus \{x\}$ a set of teams.

The *head-to-head number of points* of team x with respect to L in (X, v) is

$$s_v^L(x) = \alpha (|\{y \in L : v_1(x, y) > v_2(x, y)\}| + |\{y \in L : v_1(y, x) < v_2(y, x)\}|) + \beta (|\{y \in L : v_1(x, y) < v_2(x, y)\}| + |\{y \in L : v_1(y, x) > v_2(y, x)\}|)$$

The *head-to-head goal difference* of team x with respect to L in (X, v) is

$$gd_v^L(x) = \sum_{y \in L} (v_1(x, y) - v_2(x, y)) + \sum_{y \in L} (v_2(y, x) - v_1(y, x)).$$

In accordance with EHF (2014a, Articles 9.12 and 9.24), head-to-head results are calculated only in complete round-robin tournaments, after all group matches were played.

Definition 3.8. *Monotonicity of group ranking:* Let \mathcal{X} be the set of incomplete round-robin tournaments with a set of teams X , and R be a ranking method. R is said to be *monotonic* if for any function v and for any different teams $x, y \in X$:

1. $s_v(x) > s_v(y) \Rightarrow x \succ_{R(v)} y$;
2. $s_v(x) = s_v(y)$, furthermore, $gd_v(x) > gd_v(y)$ and if (X, v) is a round-robin tournament, then $s_v^L(x) > s_v^L(y)$, or $s_v^L(x) = s_v^L(y)$ and $gd_v^L(x) > gd_v^L(y)$ where $z \in L$ if and only if $s_v(x) = s_v(y) = s_v(z) \Rightarrow x \succ_{R(v)} y$.

Monotonicity implies that (a) a team should be ranked higher if it has a greater number of points (criterion 1); and (b) a team should be ranked higher compared to another with the same number of points, an inferior goal difference and worse head-to-head results against all teams with the same number of points (criterion 2). Monotonicity still does not lead to a strict ranking. The complexity of Definition 3.8 is necessary in order to cover the two different tie-breaking concepts, goal difference, and head-to-head results. For example, in association football, FIFA usually uses the former, while UEFA applies the latter.

Definition 3.9. *Preliminary round:* The *preliminary round* \mathcal{G} consists of k groups of round-robin tournaments $(X^1, v^1), (X^2, v^2), \dots, (X^k, v^k)$ such that $X^i \cap X^h = \emptyset$ for any $i \neq h$.

Definition 3.10. *Main round:* The *main round* \mathcal{M} consists of ℓ groups of incomplete round-robin tournaments $(Y^1, w^1), (Y^2, w^2), \dots, (Y^\ell, w^\ell)$ such that $Y^j \cap Y^h = \emptyset$ for any $j \neq h$.

Definition 3.11. *Qualification rule:* Let \mathcal{G} be the preliminary round and \mathcal{M} be the main round. The *qualification rule* is a mapping $\mathcal{Q} : \mathcal{X}^1 \times \mathcal{X}^2 \times \dots \times \mathcal{X}^k \rightarrow \mathcal{Y}^1 \times \mathcal{Y}^2 \times \dots \times \mathcal{Y}^\ell$.

Team $x \in X^i$ is said to be *qualified* to the main round if $x \in \cup_{j=1}^\ell Y^j$.

Definition 3.12. *Tournament with multiple group stages:* A *tournament with multiple group stages* is a triple $(\mathcal{G}, \mathcal{M}, \mathcal{Q})$ consisting of the preliminary round \mathcal{G} , the main round \mathcal{M} , and the qualification rule \mathcal{Q} .

Definition 3.13. *Regularity of the qualification rule:* Let $(\mathcal{G}, \mathcal{M}, \mathcal{Q})$ be a tournament with multiple group stages. Qualification rule \mathcal{Q} is *regular* if:

- a) $\cup_{j=1}^\ell Y^j \subseteq \cup_{i=1}^k X^i$;
- b) there exists a common monotonic ranking R in each group of the preliminary round \mathcal{G} such that $x, y \in X^i$, $1 \leq i \leq k$ and $x \succ_{R(v^i)} y$, $y \in \cup_{j=1}^\ell Y^j$ imply $x \in \cup_{j=1}^\ell Y^j$;
- c) $x, y \in X^i \cap Y^j$ implies $w(x, y) = v(x, y)$;
- d) $x \in X^i$, $y \in X^h$, $i \neq h$ and $x, y \in Y^j$ imply $w(x, y) = \star$;
- e) there exists a common monotonic ranking R in each group of the main round \mathcal{M} .

The idea behind a regular qualification rule is straightforward. Some top teams of the preliminary round groups qualify to the main round (conditions *a*) and *b*)), where they are divided into new groups such that matches already played against other qualified teams are carried over to the main round (conditions *c*) and *d*)). Furthermore, rankings in the preliminary and main round groups are required to be monotonic (conditions *b*) and *e*)).

Perhaps these ideas have inspired the decision-makers of EHF.

Definition 3.14. *Manipulation:* Consider a tournament with multiple group stages $(\mathcal{G}, \mathcal{M}, \mathcal{Q})$ and a set of preliminary round results $V = \{v^1, v^2, \dots, v^i, \dots, v^k\}$. A team $x \in X^i$ can *manipulate* $(\mathcal{G}, \mathcal{M}, \mathcal{Q})$ if there exists $\bar{V} = \{v^1, v^2, \dots, \bar{v}^i, \dots, v^k\}$ with $\bar{v}_2^i(x, y) \geq v_2^i(x, y)$ and $\bar{v}_1^i(y, x) \geq v_1^i(y, x)$ for all $y \in X^i$, furthermore, $x \in \cup_{j=1}^\ell Y^j$ according to both $\mathcal{Q}(V)$ and $\mathcal{Q}(\bar{V})$ such that $s_w(x) < s_{\bar{w}}(x)$, or $s_w(x) = s_{\bar{w}}(x)$ and $gd_w(x) < gd_{\bar{w}}(x)$.

Manipulation means that team x can increase its number of points, or at least improve its goal difference with preserving its number of points in the main round by conceding more goals in a match of the preliminary round.

Definition 3.15. *Strategy-proofness:* A tournament with multiple group stages $(\mathcal{G}, \mathcal{M}, \mathcal{Q})$ is called *strategy-proof* if there exists no set of group results $V = \{v^1, v^2, \dots, v^k\}$ under which a team can manipulate it.

The main contributions concern the strategy-proofness of tournaments with multiple group stages and a regular qualification rule. Note that manipulation certainly worsens a team's goal difference (and sometimes its number of points, too) in its preliminary round group as the ranking rule applied here is monotonic, but – provided that the team still qualifies – it may pay off in the main round when some matches of the preliminary round are discarded.

Theorem 3.1. Let $(\mathcal{G}, \mathcal{M}, \mathcal{Q})$ be a tournament with multiple group stages such that \mathcal{Q} is a regular qualification rule and the following conditions hold:

- there exists $x, y \in X^i \cap Y^j$ for some $1 \leq i \leq k$ and $1 \leq j \leq \ell$;
- for at least one $1 \leq i \leq k$, there exists $u, v \in X^i$ with $u \in Y^j$ implying $v \notin Y^j$.

Then the tournament with multiple group stages $(\mathcal{G}, \mathcal{M}, \mathcal{Q})$ does not satisfy strategy-proofness.

According to the conditions of Theorem 3.1, the result of at least one match played in the preliminary round (between the teams x and y) is carried over to main round, and the result of at least one such match (between the teams u and v) is ignored.

Proof. An example is presented where a team can manipulate a tournament with multiple group stages that satisfies all criteria of Theorem 3.1.

Table 2: Group 1 of Example 3.1

GF = Goals for; GA = Goals against; GD = Goal difference; Pts = Points.

The last but one row contains the group winner's benchmark results that are carried over to the main round.

The last row contains the group winner's alternative results that are carried over to the main round after it manipulates.

Position	Team	a	b	c	GF	GA	GD	Pts
1	a	*	0-1	4-0	4	1	3	$\alpha + \beta$
2	b	1-0	*	0-2	1	2	-1	$\alpha + \beta$
3	c	0-4	2-0	*	2	4	-2	$\alpha + \beta$
1	a	*	0-1	*	0	1	-1	β
1*	a^*	*	*	2-0*	2*	0*	2*	α^*

Example 3.1. Let $X^1 = \{a, b, c\}$ be a single round-robin group.

Consider the regular qualification rule \mathcal{Q} with $\ell = 1$ group in the main round and $x \in Y^1$ if and only if $\{z \in X^i : x \succ_{R(v^i)} z\} \neq \emptyset$. \mathcal{Q} says that the group winner and the runner-up qualify for the main round.

A possible set of results in Group 1 is shown in Table 2. Team a is the group-winner since it has the best (head-to-head) goal difference (see criterion 2 of a monotonic group ranking method), and it is considered with $s_w(a) = \beta$ points in the main round, after discarding its match against team c , the last in Group 1 due to criterion 2 of a monotonic group ranking method (see the last but one row of Table 2).

However, examine what happens if $\bar{v}^1(a, c) = (2; 0)$, thus $\bar{v}^1(c, a) = (0; 2)$. Then teams a , b , and c remain with $\alpha + \beta$ points, but they have head-to-head goal differences of $+1$, -1 and 0 , respectively, therefore a is the first and c is the second according to criterion 2 of a monotonic group ranking method. Consequently, team a is considered with $s_{\bar{w}}(a) = \alpha > \beta = s_w(a)$ points in the main round (see the last row of Table 2).

To conclude, team a has an opportunity to manipulate this simple tournament with multiple group stages under the set of group results V , so it violates strategy-proofness.

Example 3.1 contains only three teams, which is minimal under the conditions of Theorem 3.1. It is clear that the number of groups and the number of teams in them can

be increased without changing the essence of the counterexample. Groups can be double round-robin tournaments instead of single ones, too. \square

Theorem 3.1 also remains valid if draws are allowed in a tournament with multiple group stages.

Remark 3.1. The 11th European Men's Handball Championship (EHF Euro 2014), discussed in Section 2, fits into the model presented above. The number of groups in the preliminary round is $k = 4$, the number of groups in the main round is $\ell = 2$, and the qualification rule is regular (EHF, 2014a):

- a) $Y^1 \subset X^1 \cup X^2$ and $Y^2 \subset X^3 \cup X^4$;
- b) Ranking in the preliminary round groups is monotonic as it is based on the number of points with tie-breaking through head-to-head results, and the first three teams qualify for the main round;
- c) Matches played during the preliminary round against opponents which qualified to the main round are kept and remain valid for the ranking of the main round;
- d) Matches of the main round are played in groups with each team facing three opponents which did not participate in its preliminary round group;
- e) Ranking in the main round groups is monotonic as it is based on the number of points with tie-breaking through head-to-head results.

Proposition 3.1. *The 11th European Men's Handball Championship (EHF Euro 2014) is not strategy-proof.*

Proof. The scenario presented in Section 2 shows that team Poland = $x \in X^3$ can manipulate since there exist sets of group results $V = \{v^1, v^2, v^3, v^4\}$ and $\bar{V} = \{v^1, v^2, \bar{v}^3, v^4\}$ such that $\bar{v}^3 = v^3$, $\bar{v}_2^3(x, y) = v_2^3(x, y) = 22$ except for $\bar{v}_1^3(x, y) = 26 > 24 = v_1^3(x, y)$, where team Russia = $y \in X^3$ and Poland qualifies according to $\mathcal{Q}(V)$ and $\mathcal{Q}(\bar{V})$, but $s_w(x) = 0 < 2 = s_{\bar{w}}(x)$.

Theorem 3.1 can also be applied due to Remark 3.1. \square

Now we state a positive result, a 'pair' of Theorem 3.1.

Theorem 3.2. *Let $(\mathcal{G}, \mathcal{M}, \mathcal{Q})$ be a tournament with multiple group stages such that \mathcal{Q} is a regular qualification rule and at least one of the following conditions hold:*

- *there does not exist $x, y \in X^i \cap Y^j$ for any $1 \leq i \leq k$ and $1 \leq j \leq \ell$;*
- *$u, v \in X^i$ and $u \in Y^j$ implies $v \in Y^j$ for all $1 \leq i \leq k$.*

Then the tournament with multiple group stages $(\mathcal{G}, \mathcal{M}, \mathcal{Q})$ is strategy-proof.

Proof. If all preliminary round results obtained against other qualified teams are ignored (first condition), or carried over to the main round (second condition), then it makes no sense to perform weaker in the preliminary round due to the monotonicity of rankings in all groups. \square

Theorem 3.2 essentially says that teams qualifying from the same preliminary round group should be drawn into different main round groups (it is guaranteed if only one team qualifies from each preliminary round group), or all teams from a given preliminary round group should qualify for the same main round group.

It is also clear from the match discussed in Section 2 that head-to-head results cannot be used to break a tie in the main round between two teams from the same preliminary round group, otherwise there exists some incentives to influence the set of qualified teams.

Our main result is somewhat related to – but entirely independent of – the finding of [Vong \(2017\)](#) that in general multi-stage tournaments, the necessary and sufficient condition of strategy-proofness is to allow only the top-ranked player to qualify from each group. The difference is that in the model of [Vong \(2017\)](#), teams deliberately lose matches in order to meet preferred opponents in the next round, so they only gain in expected value. Contrarily, we have discussed the possibility that a team can be strictly better off by a weaker performance.

4 Tournaments with multiple group stages

The [European Men’s Handball Championship](#) between 1994 and 2000 consisted of a group stage followed by a knockout stage, hence they were incentive compatible. Since 2002, its format is the same as outlined in Section 2 and presented in Figure A.1 in the Appendix: a preliminary round with four groups of four teams each such that the first three teams qualify for the main round with two groups of six teams each, and they carry over the matches played against the two teams in their preliminary round group. The winners and runners-up of the main round groups qualify to the semifinals.

During the [10th Men’s European Handball Championship \(EHF Euro 2012\)](#), a situation analogous to the one presented in Section 2 emerged. Slovenia played its last match in Group D against Iceland when Croatia had 4 points after it won against Iceland and Slovenia, Norway had 2 points because of its win against Slovenia by 28-27, and Iceland had also 2 points due to its win against Norway by 34-32. Consequently, Slovenia should have won against Iceland for qualification to the main round, but it would be better not to win by more than 3 goals in order to carry over its result against Iceland. The actual results were Iceland vs Slovenia 32-34, and Croatia vs Norway 26-22, so the manipulation of Slovenia turned out to be successful (with Iceland vs Slovenia 31-34 or 32-35, Iceland still would have qualified, but 30-34, 31-35, or 32-36 would be unfavourable for Slovenia).

The [Women’s European Handball Championship](#) is the official competition for senior women’s national handball teams of Europe. It takes place in the same years as Men’s European Handball Championship and is organised according to same design, so it was also strategy-proof until 2000, but it is incentive incompatible from 2002.

The [Women’s EHF Champions League](#) is an annual official competition for women’s handball clubs of Europe since the season of 1993/94. It is the most competitive and prestigious tournament for the top clubs of the continent’s leading national leagues. The tournament is organised with multiple group stages since 2013/14. The preliminary round consists of four groups of four teams each, playing each other twice in home and away matches such that the best three teams qualify. In the main round, two groups of six teams are formed, and teams play twice, in home and away matches against those three teams they have not already faced. The top four teams from each group advance to the quarter-finals.

The [World Men's Handball Championship](#), organised by the IHF (International Handball Federation), takes place in every second year since 1993. From 1995, the number of competing teams has increased to twenty-four, and four different tournament formats have been used ([Csató, 2019a](#)). Among them, three designs violate incentive compatibility because of Theorem 3.1.

Since 1995, the [World Women's Handball Championship](#) is played in the same years as the men's tournament, and it is designed in the same format with the exception of 1995, 2003, and 2011.

Table 3: Handball tournaments with multiple group stages

Notes: S = single round-robin (in groups); D = double round-robin (in groups); Gr. = Number of groups in the preliminary and main round, respectively, which are denoted by k and ℓ in the theoretical model of Section 3; Teams = Number of teams in each group of the preliminary and main round, respectively; Q = Number of teams qualified from each group of the preliminary and main round, respectively
Abbreviations: EHF Euro Men (Women) = European Men's (Women's) Handball Championship; EHF Women's CL = Women's EHF Champions League; IHF World Men (Women) = IHF World Men's (Women's) Handball Championship

Tournament	Year(s)	Type	Preliminary round			Main round		
			Gr.(k)	Teams	Q	Gr.(ℓ)	Teams	Q
EHF Euro Men	2002–	S	4	4	3	2	6	2
EHF Euro Women	2002–	S	4	4	3	2	6	2
EHF Women's CL	2013/14–	D	4	4	3	2	6	4
IHF World Men	2003	S	4	6	4	4	4	1
IHF World Men	2005, 2009-2011, 2019–	S	4	6	3	2	6	2
IHF World Men	2007	S	6	4	2	2	6	4
IHF World Women	2003-2005, 2009	S	4	6	3	2	6	2
IHF World Women	2007	S	6	4	2	2	6	4

Table 3 summarises the incentive incompatible handball tournaments discussed above. They all contain two multiple group stages, and the number of qualified teams in the main round (see the last column) is the number of teams which have a chance to win the tournament.

Tournaments with multiple group stages are also used in other sports, for instance, in basketball ([EuroBasket 2013](#)), cricket ([2007 Cricket World Cup](#)) ([Scarf et al., 2009](#)), and volleyball ([2014 FIVB Volleyball Men's World Championship](#)). There was a match played by Australia and West Indies in the [1999 Cricket World Cup](#), in which Australia probably attempted a manipulation similar to the one presented in Section 2 ([Kendall and Lenten, 2017](#), Section 3.7.2). However, this plan – if there was one – did not work out entirely.

The [1999-2000 UEFA Champions League](#), as well as the following three seasons of this tournament, also included two group stages: from the first group stage of eight groups with four teams each, eight winners and eight runners-up were drawn into four groups of four teams each, containing two group winners and two runners-up such that teams from the same country or from the same first round group could not be drawn together. Consequently, no results were carried over to the second group stage, guaranteeing the incentive compatibility of the design according to Theorem 3.2.

5 Two ways of overcoming incentive incompatibility

It is clear from our theoretical results, presented in Section 3, that there is no straightforward way to guarantee the strategy-proofness of tournaments with multiple group stages, in contrast to tournament systems consisting of multiple round-robin and knockout tournaments (Dagaev and Sonin, 2017), or group-based qualifiers with a repechage (Csató, 2018c).

Theorem 3.2 shows that incentive compatibility is met if either all points scored in the preliminary round are considered in the main round (directly or after an arbitrary monotonic transformation), or all of them are discarded, which is against the essence of these tournaments. Consequently, if the administrators want to organise a strategy-proof tournament with multiple group stages, the only solution is to carry over *all* preliminary round results to the main round, perhaps after a monotonic transformation, regardless that some matches were played against teams already eliminated from the tournament.

However, it seems that if all results are carried through, then the subsequent stage loses a bit of excitement because there will be greater variation in points at the commencement of that stage, and teams entering bottom will find it much harder to catch up to the teams entering the stage on top.

This effect can be mitigated by carrying over only half of the points from the preliminary round. The idea comes from the Belgian First Division A, the top league competition for association football clubs in Belgium, where the sixteen competitors play a double round-robin tournament in the regular season, followed by a championship play-off for the first six teams such that the points obtained during the regular season are halved.

For tie-breaking purposes, we suggest retaining the number of goals scored and conceded in the preliminary round. Theoretically, they can be ignored, too, but it seems to be unfair when there was a match played in the preliminary round against a team from the same main round group. In the case of Belgian First Division A, goal difference is not among the tie-breaking criteria in the championship playoffs.

Therefore, two strategy-proof versions of each tournament design with multiple group stages can be defined. In the following, the consequences of these modifications will be explored as a kind of cost-benefit analysis via simulations.

Our starting point is a comparison of tournament formats for the World Men's Handball Championships (Csató, 2019a). Section 4 has revealed that this tournament has applied recently three formats containing multiple group stages (see Table 3). We investigate two of them since the third suffers from various problems and seems to be inefficient (Csató, 2019a). They are the following:

- Format *G66*: This design, presented in Figure A.2, has been used first in the 2005 World Men's Handball Championship and has been applied in 2009 and 2011. The 2019 championship is also organised in this format (IHF, 2018). The preliminary round (see Figure A.2a) consists of four groups of six teams each such that the top three teams qualify for the main round. The main round consists of two groups of six teams, each created from two preliminary round groups. The top two teams of every main round group advance to the semifinals in the knockout stage (see Figure A.2b).
- Format *G46*: This design, presented in Figure A.3, has been used in the 2007 World Men's Handball Championship, hosted by Germany. Teams are drawn into six groups of four teams each in the preliminary round

(see Figure A.3a) such that the top two teams proceed to the main round. The main round consists of two groups, each created from three preliminary round groups. Four teams of every main round group advance to the quarterfinals in the knockout stage (see Figure A.3b).

While the knockout stage of both tournament formats is immediately determined by the preceding group stage, the competing teams should be drawn into groups before the start of the tournament, so the seeding policy may affect the outcome, too (Guyon, 2015; Dagaev and Rudyak, 2016; Guyon, 2018a; Laliena and López, 2018).

Hence – similarly to Csató (2019a) – two variants of each tournament design, called *seeded* and *unseeded*, are considered. In the seeded version, the preliminary round groups are drawn such that in the case of groups with k teams ($k = 4$ for $G46$ and $k = 6$ for $G66$), the strongest k teams are placed in Pot 1, the next strongest k teams in Pot 2, and so on. The unseeded version uses fully random seeding. Consequently, some strong teams, allocated in a harsh group will have more difficulties in qualifying than certain weaker teams, allocated in an easier group.

Table 4: Tournament designs

Notation	Format	Seeding policy	Description
$G66/S$	G66	seeded	original incentive incompatible
$G66/R$	G66	unseeded	original incentive incompatible
$G66\Diamond/S$	G66	seeded	all points are carried over
$G66\Diamond/R$	G66	unseeded	all points are carried over
$G66\star/S$	G66	seeded	half of all points are carried over
$G66\star/R$	G66	unseeded	half of all points are carried over
$G46/S$	G46	seeded	original incentive incompatible
$G46/R$	G46	unseeded	original incentive incompatible
$G46\Diamond/S$	G46	seeded	all points are carried over
$G46\Diamond/R$	G46	unseeded	all points are carried over
$G46\star/S$	G46	seeded	half of all points are carried over
$G46\star/R$	G46	unseeded	half of all points are carried over

Table 4 shows the twelve tournament designs to be analysed.

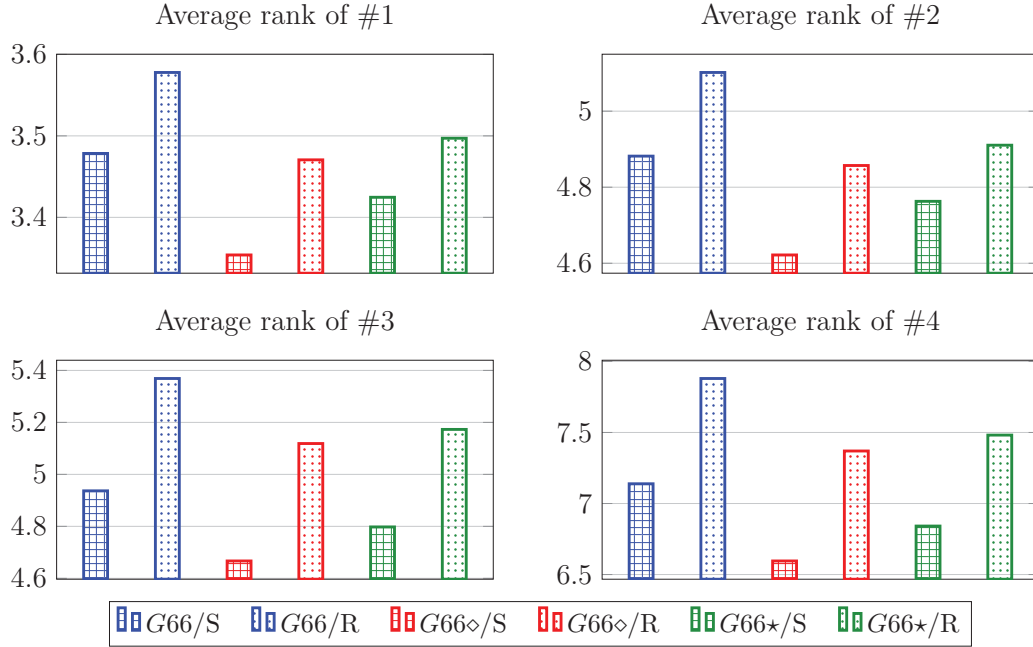
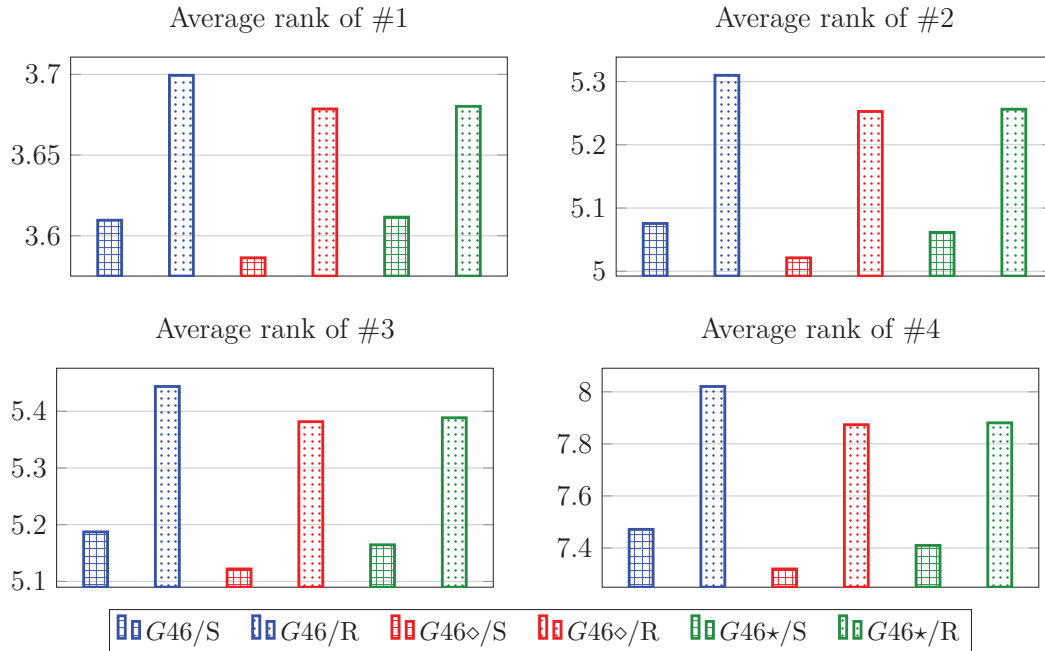
The tournament metrics applied are as follows:

- the average pre-tournament ranking of the winner, the second-, the third- and the fourth-placed teams;
- the expected quality of the final (the sum of the finalists' pre-tournament ranking);
- the expected competitive balance of the final (the difference between the finalists' pre-tournament ranking).

The simulation procedure is detailed in Csató (2019a). According to the arguments presented there, all simulations have been implemented with one million runs.

Figure 1 shows the average pre-tournament ranking of the first four teams. If all points are carried over from the preliminary round, then the result of the tournament becomes more predetermined as the expected ranking slightly decreases. Preserving only half of these points significantly mitigates the loss of excitement, except in the unseeded variant of

Figure 1: Expected pre-tournament ranking of the first four teams

(a) Tournament format $G66$

(b) Tournament format $G46$


format $G46$. On the other hand, the average rank of the winner is even higher in the case of seeded $G46$ according to this solution than under the original incentive incompatible design. Furthermore, carrying over half of all points minimises the effect of the seeding policy, which seems to be desirable because it is a factor not influenced by the competitors.

Figure 2: Characteristics of the tournament final

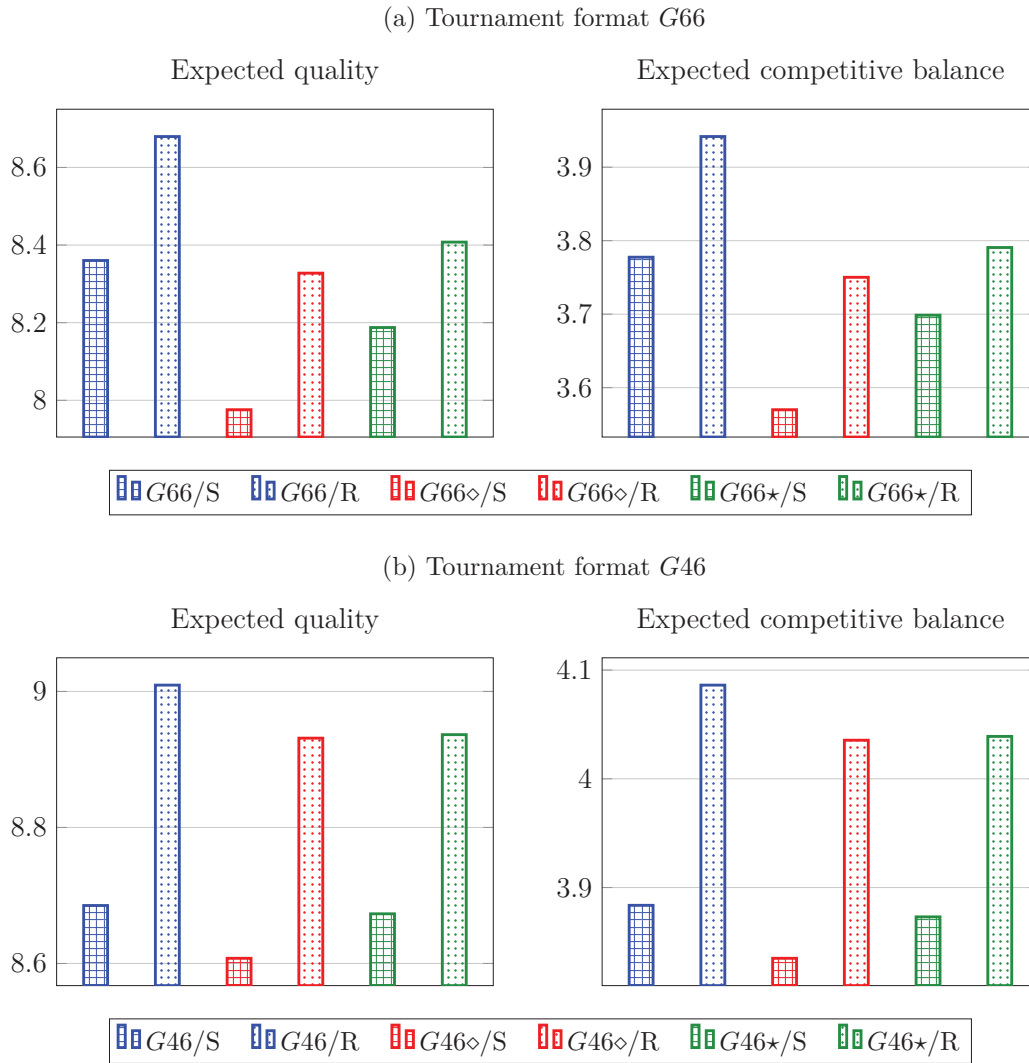


Figure 2 reinforces these findings by focusing on the final of the tournament: if the half of all points scored in the preliminary round are carried over instead of only the results against the teams qualified to the main round, then the final may become a bit more boring but usually involves stronger teams. It also decreases the influence of seeding, especially in the format $G66$.

Following Scarf et al. (2009), we have also made a robustness check by calculating the metrics for more and less competitive tournaments than the baseline version, in the same way as Csató (2019a). All of our qualitative results are insensitive to the distribution of teams' strength.

The comparison of Figures 1.a and 1.b, as well as Figures 2.a and 2.b, reveals that the choice of the tournament format is more important than the effect of how points are carried over to the main round (see the scales on the vertical axis). Since there is no consensus in the former, at least for the Men's (Women's) World Handball Championships, it makes not much sense to risk the possible problems caused by the original incentive

incompatible designs.

Thus the price of guaranteeing incentive compatibility seems to be negligible – at least compared to other features of the design like the particular tournament format or the seeding policy –, and we suggest to carry over half of the points scored in the preliminary round. Applying this solution has another, unexpected advantage by minimizing the effects of the preliminary seeding of the teams into pots.

6 Conclusions

Optimal design of sports ranking rules is an important topic of economics and operations research. We have revealed that administrators should not miss analysing strategy-proofness: a simple shortcoming in the design of a tournament can lead to perverse incentives in a sporting contest that is supposed to be genuine, and as such is sold to the public as having full integrity. While the actual probability of manipulation can be relatively small, and the audience does not necessarily recognise the problem, it makes no sense to risk a potential scandal which has enormous financial and reputational costs. Our simulation model has also proved that the price of guaranteeing incentive compatibility can be marginal, and the use of a fair mechanism does not affect essentially the selective ability and the competitive balance of a tournament with multiple group stages.

It is somewhat surprising that we have not found any controversy about the particular match presented in Section 2. We think it is because of its non-trivial detection as compared to the football and basketball matches discussed in Section 1, it was enough to make some mistakes in defence or attack, without the need to score own goals. One can understand that EHF remained silent on this issue, and the audience obviously did not study the tie-breaking rules carefully. On the other hand, it is almost sure that the coaches and players knew that they should not make great efforts to win by a higher margin. Hopefully, our discussion will contribute to placing this match in the category of the notorious ‘*Nichtangriffspakt (Schande) von Gijón*’⁴ (Kendall and Lenten, 2017, Section 3.9.1) in the history of sports.

There are some directions for future scientific research. First, by the quantification of team strengths and the modelling of match outcomes, the probability of manipulation can be estimated. Second, fairness of a tournament is a more general notion than incentive compatibility as the 2016 UEFA European Championship (Guyon, 2018a), the 2026 FIFA World Cup (Guyon, 2018b), the scheduling of round-robin tournaments (Krumer and Lechner, 2017; Krumer et al., 2017a,b; Sahm, 2018), or the problem of penalty shootouts (Palacios-Huerta, 2012, 2014; Brams and Ismail, 2018; Csató, 2018a) show. The final aim may be an extensive axiomatic discussion and comparison of sports ranking rules, which has started recently (Arlegi and Dimitrov, 2018; Berker, 2014; Csató, 2017b, 2018b,e; Vaziri et al., 2018).

⁴ Kendall and Lenten (2017) use the term ‘Shame of Gijón’, and Wikipedia calls it ‘Disgrace of Gijón’. The name is given to a 1982 FIFA World Cup football match played between West Germany and Austria at Gijón, Spain, on 25 June 1982. A win by one or two goals for West Germany would result in both them and Austria qualifying at the expense of Algeria. West Germany took the lead after 10 minutes, and the remaining 80 minutes were characterised by few serious attempts by either side to score. Both teams were accused of match-fixing although FIFA ruled that they did not break any rules.

Acknowledgements

This paper could not be written without my father, who have coded the simulations in Python.

We are grateful to *Liam Lenten* and *Tamás Halm* for useful advice.

Three anonymous reviewers provided valuable comments and suggestions on an earlier draft.

We are indebted to the [Wikipedia community](#) for contributing to our research by collecting valuable information.

The research was supported by OTKA grant K 111797 and by the MTA Premium Post Doctorate Research Program.

References

- Arlegi, R. and Dimitrov, D. (2018). Fair competition design. Manuscript. <http://www.gtcenter.org/Downloads/Conf/Dimitrov2839.pdf>.
- Arrow, K. J. (1950). A difficulty in the concept of social welfare. *Journal of Political Economy*, 58(4):328–346.
- Berker, Y. (2014). Tie-breaking in round-robin soccer tournaments and its influence on the autonomy of relative rankings: UEFA vs. FIFA regulations. *European Sport Management Quarterly*, 14(2):194–210.
- Brams, S. J. and Ismail, M. S. (2018). Making the rules of sports fairer. *SIAM Review*, 60(1):181–202.
- Carpio, G. (2014). Gilas beats Kazakhstan but misses semis. *Philippine Star*. 29 September 2014. <http://www.philstar.com/sports/2014/09/29/1374376/gilas-beats-kazakhstan-misses-semis>.
- Csató, L. (2017a). 2018 FIFA World Cup qualification can be manipulated. Manuscript. <http://unipub.lib.uni-corvinus.hu/3053/>.
- Csató, L. (2017b). On the ranking of a Swiss system chess team tournament. *Annals of Operations Research*, 254(1-2):17–36.
- Csató, L. (2018a). A fairer penalty shootout design in soccer. Manuscript. [arXiv: 1806.01114](#).
- Csató, L. (2018b). An impossibility theorem for paired comparisons. *Central European Journal of Operations Research*, in press. DOI: [10.1007/s10100-018-0572-5](#).
- Csató, L. (2018c). Incentive compatible designs for tournament qualifiers with round-robin groups and repechage. Manuscript. [arXiv: 1804.04422](#).
- Csató, L. (2018d). It may happen that no team wants to win: a flaw of recent UEFA qualification rules. Manuscript. [arXiv: 1806.08578](#).
- Csató, L. (2018e). Some impossibilities of ranking in generalized tournaments. Manuscript. [arXiv: 1701.06539](#).

- Csató, L. (2018f). UEFA Champions League entry has not satisfied strategy-proofness in three seasons. Manuscript. [arXiv: 1801.06644](#).
- Csató, L. (2018g). Was Zidane honest or well-informed? How UEFA barely avoided a serious scandal. *Economics Bulletin*, 38(1):152–158.
- Csató, L. (2019a). A simulation comparison of tournament designs for the World Men's Handball Championships. Manuscript. [arXiv: 1803.10975](#).
- Csató, L. (2019b). When UEFA rules had inspired unfair behaviour on the field. Manuscript. [arXiv: 1806.03978](#).
- Dagaev, D. and Rudyak, V. (2016). Seeding the UEFA Champions League participants: Evaluation of the reform. Manuscript. DOI: [10.2139/ssrn.2754127](#).
- Dagaev, D. and Sonin, K. (2013). Game theory works for football tournaments. Manuscript. <http://voxeu.org/article/world-cup-football-and-game-theory>.
- Dagaev, D. and Sonin, K. (2017). Winning by losing: Incentive incompatibility in multiple qualifiers. *Journal of Sports Economics*, in press. DOI: [10.1177/1527002517704022](#).
- EHF (2014a). *EHF Euro Regulations*. Applied on the 11th Men's European Handball Championship (EHF Euro 2014). http://den2014.ehf-euro.com/fileadmin/Content/DEN2014M/Files/Other_pdf/EUR0reg_Final_131212.pdf.
- EHF (2014b). Report: 11th Men's European Handball Championship (EHF Euro 2014), Preliminary round – Group C, Match No. 17, Poland against Russia. 17 January 2014. <http://handball.sportresult.com/hbem14m/PDF/17012014/M17/M17.pdf>.
- Fornwagner, H. (2018). Incentives to lose revisited: The NHL and its tournament incentives. *Journal of Economic Psychology*, in press. DOI: [10.1007/10.1016/j.joep.2018.07.004](#).
- Gibbard, A. (1973). Manipulation of voting schemes: A general result. *Econometrica*, 41(4):587–601.
- Guyon, J. (2015). Rethinking the FIFA World CupTM final draw. *Journal of Quantitative Analysis in Sports*, 11(3):169–182.
- Guyon, J. (2018a). What a fairer 24 team UEFA Euro could look like. *Journal of Sports Analytics*, 4(4):297–317.
- Guyon, J. (2018b). Will groups of 3 ruin the World Cup? Manuscript. DOI: [10.2139/ssrn.3190779](#).
- IHF (2018). *Regulations for IHF Competitions*. International Handball Federation. Edition: 14 January 2018. http://ihf.info/files/Uploads/NewsAttachments/0_Regulations%20for%20IHF%20Competitions_GB.pdf.
- Kendall, G. and Lenten, L. J. A. (2017). When sports rules go awry. *European Journal of Operational Research*, 257(2):377–394.
- Kóczy, L. Á. and Strobel, M. (2009). The invariant method can be manipulated. *Scientometrics*, 81(1):291–293.

- Krumer, A. and Lechner, M. (2017). First in first win: Evidence on schedule effects in round-robin tournaments in mega-events. *European Economic Review*, 100:412–427.
- Krumer, A., Megidish, R., and Sela, A. (2017a). First-mover advantage in round-robin tournaments. *Social Choice and Welfare*, 48(3):633–658.
- Krumer, A., Megidish, R., and Sela, A. (2017b). Round-robin tournaments with a dominant player. *The Scandinavian Journal of Economics*, 119(4):1167–1200.
- Laliena, P. and López, F. J. (2018). Fair draws for group rounds in sport tournaments. *International Transactions in Operational Research*, in press. DOI: [10.1111/itor.12565](https://doi.org/10.1111/itor.12565).
- Lenten, L. J. A., Smith, A. C. T., and Boys, N. (2018). Evaluating an alternative draft pick allocation policy to reduce ‘tanking’ in the Australian Football League. *European Journal of Operational Research*, 267(1):315–320.
- Palacios-Huerta, I. (2012). Tournaments, fairness and the Prouhet-Thue-Morse sequence. *Economic Inquiry*, 50(3):848–849.
- Palacios-Huerta, I. (2014). *Beautiful game theory: How soccer can help economics*. Princeton University Press, Princeton, New York.
- Palacios-Huerta, I. and Volij, O. (2004). The measurement of intellectual influence. *Econometrica*, 72(3):963–977.
- Pauly, M. (2014). Can strategizing in round-robin subtournaments be avoided? *Social Choice and Welfare*, 43(1):29–46.
- Pinski, G. and Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management*, 12(5):297–312.
- Sahm, M. (2018). Are sequential round-robin tournaments discriminatory? *Journal of Public Economic Theory*, in press. DOI: [10.1111/jpet.12308](https://doi.org/10.1111/jpet.12308).
- Satterthwaite, M. A. (1975). Strategy-proofness and Arrow’s conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10(2):187–217.
- Scarf, P., Yusof, M. M., and Bilbao, M. (2009). A numerical study of designs for sporting contests. *European Journal of Operational Research*, 198(1):190–198.
- Szymanski, S. (2003). The economic design of sporting contests. *Journal of Economic Literature*, 41(4):1137–1187.
- Tasnádi, A. (2008). The extent of the population paradox in the Hungarian electoral system. *Public Choice*, 134(3-4):293–305.
- Vaziri, B., Dabadghao, S., Yih, Y., and Morin, T. L. (2018). Properties of sports ranking methods. *Journal of the Operational Research Society*, 69(5):776–787.
- Vong, A. I. K. (2017). Strategic manipulation in tournament games. *Games and Economic Behavior*, 102:562–567.

Wright, M. (2009). 50 years of OR in sport. *Journal of the Operational Research Society*, 60(Supplement 1):S161–S168.

Wright, M. (2014). OR analysis of sporting rules – A survey. *European Journal of Operational Research*, 232(1):1–8.

Appendix

Figure A.1: The tournament format which was used in the 2014 European Men's Handball Championship

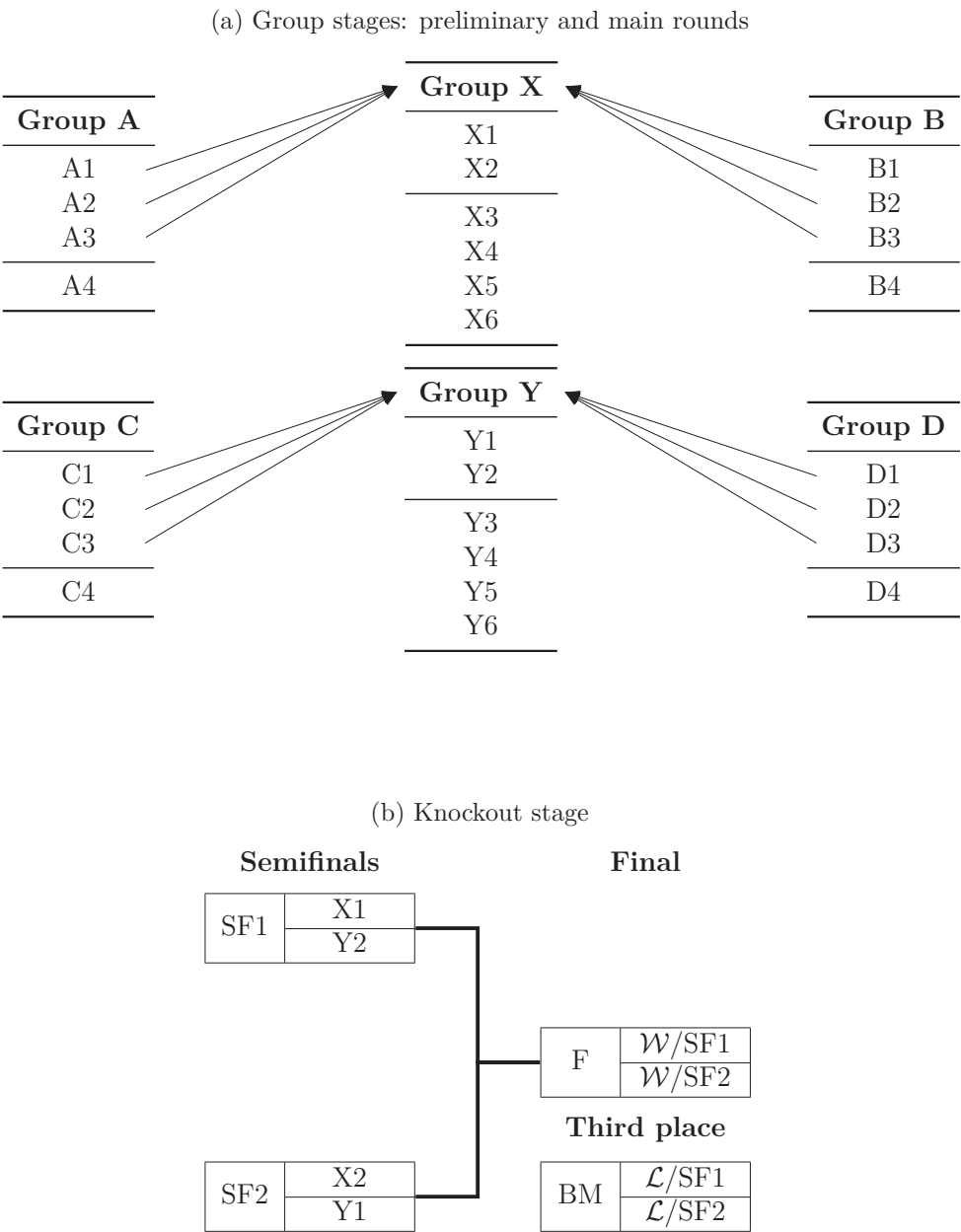


Figure A.2: Tournament format *G66*, which was used in the 2011 World Men’s Handball Championship, and again in the 2019 World Men’s Handball Championship

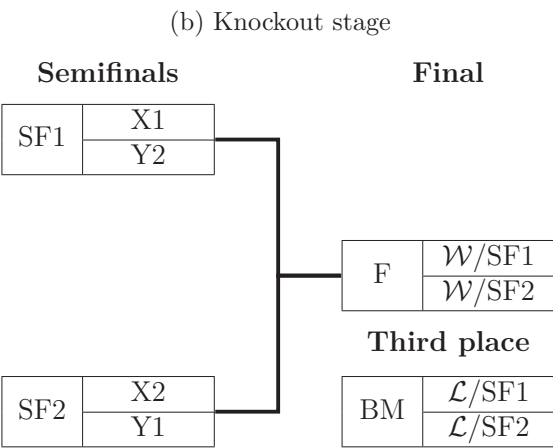
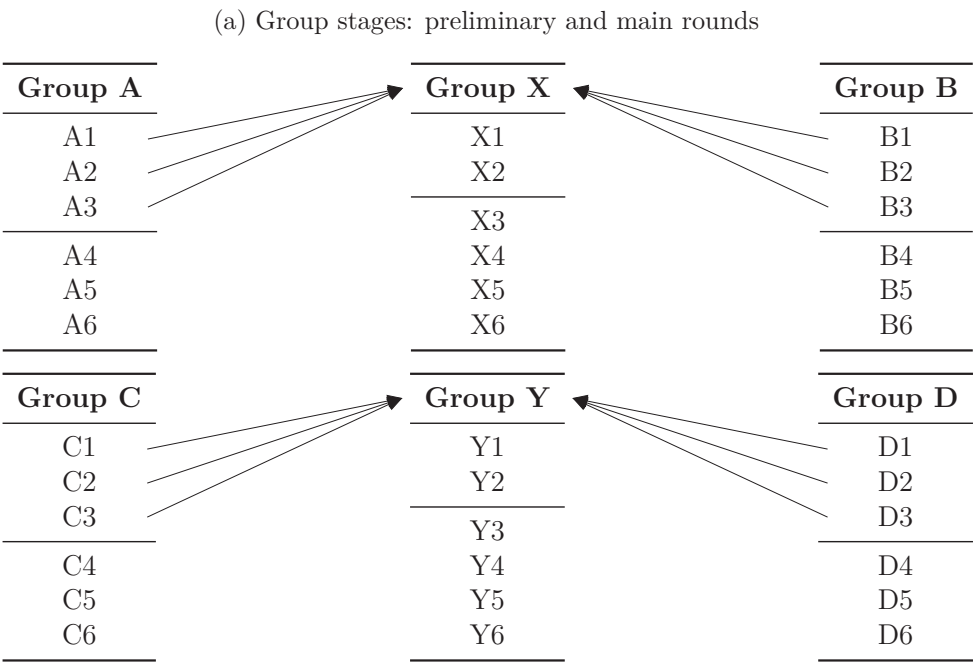
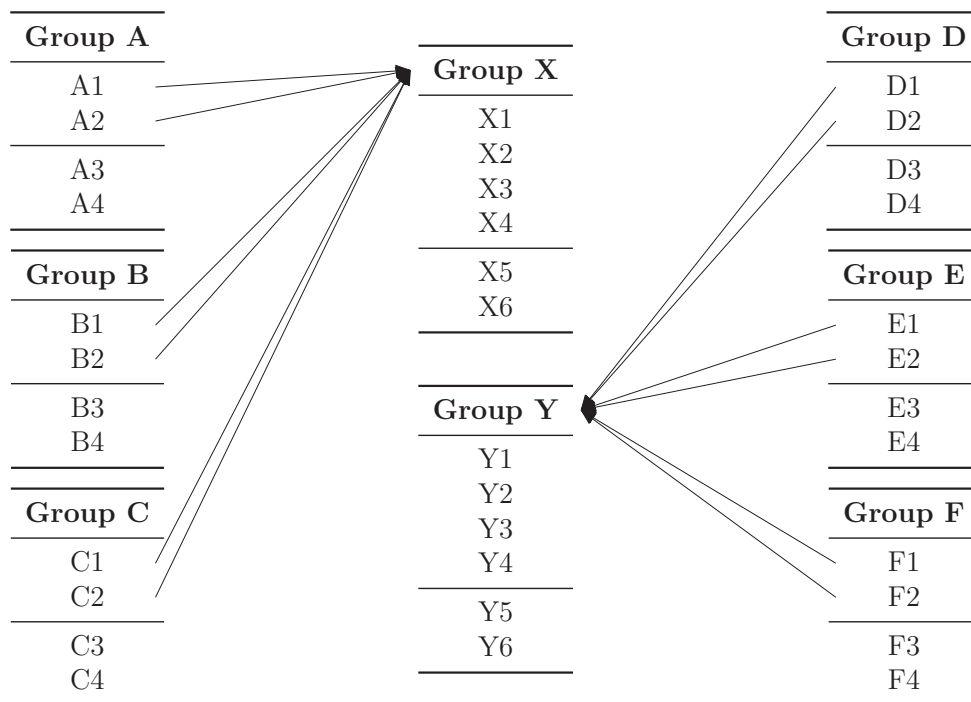
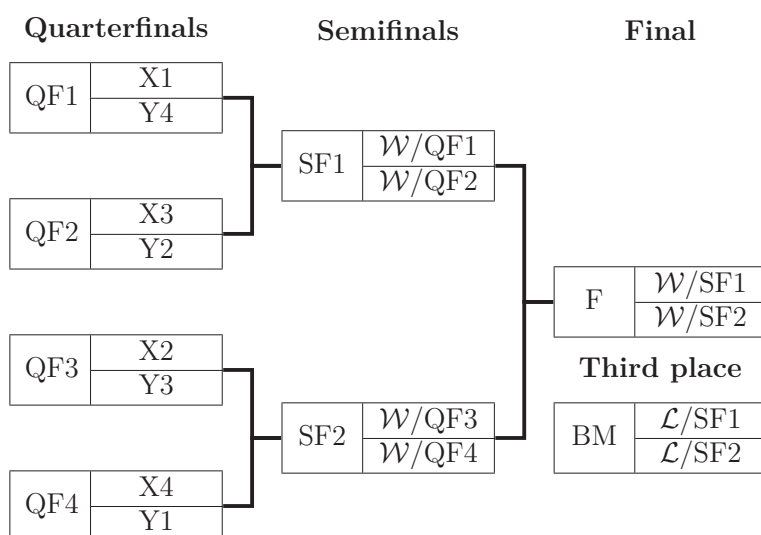


Figure A.3:
Tournament format *G46*, which was used in the 2007 World Men's Handball Championship

(a) Group stages: preliminary and main rounds



(b) Knockout stage



Comparison among some European football leagues through related count data variables

Valentina Cueva-López¹, José Rodríguez-Avi¹, and María José Olmo-Jiménez¹

University of Jaén, Jaén, Spain

vcueva@ujaen.es, jravi@ujaen.es, mjolmo@ujaen.es

Abstract

Football is, probably, the most popular sport in Europe, in terms of supporters, news generated, public interest and movement of funds, among others. Weekly, hundreds of thousands of people go to the stadiums and millions watch the football match on TV, through the Internet or listen to it on the radio. Even, the number of betting offices increases and all the aspects related with football are specially important.

From a statistical point of view, football can also be seen as a generator of statistical variables that may be studied. In this work, we want to focus on some count data variables that reflect many of the most outstanding football aspects. Specifically, we consider the number of goals scored by a footballer and the number of yellow cards that are shown to a footballer. These variables have been collected in some of the most important football leagues in Europe for several years such as the Spanish, English, German and Italian leagues. For each variable we have proposed several count data model, such as the Negative Binomial, the Univariate Generalized Waring, the Extended Biparametric Waring or the Complex Biparametric and Triparametric Pearson distributions, among others. In each case we select the best fit using the Akaike Information Criterion or the chi-square goodness of fit test. We compare the results obtained for the different leagues considered.

1 Introduction

It can be considered that football, as the most popular sport in modern history, spans more than 150 years. It began in 1863 in England, when rugby football and association football branched off on their different courses. Thus, the most ancient football association was founded.

Although both games have the same root, there are at least half a dozen different games, varying to different degrees, and to which the historical development of football has been traced back.

In its origins, controlling the ball with the feet was considered a feat, since it raised the admiration of other citizens. The first reference of the game is a manual of military exercises

that dates back to the China of the Han dynasty in the 2nd and 3rd centuries BC, known as “Ts’uh Kúh”. This game consisted in introducing a leather ball, filled with feathers and hair, into a small net, from an opening of 30 to 40 centimeters. In that same document, another modality of the same game is described, in which, the players would have to overcome the rival while they went to the goal (you could never catch the ball with your hand).

We can also mention the Japanese Kemari which began some 500-600 years later and is still played today. In this game the players do not struggle to possess the ball but they pass it to each other trying not to let it touch the ground.

More lively were the “Epislyros” Greek and the “Harpastum” Roman. The latter was played with a smaller ball and two teams on a rectangular field marked by boundary lines and a centre line. The objective was to get the ball over the opposition’s boundary lines and as players passed it between themselves, trickery was the order of the day. The game remained popular for 700-800 years, but, although the Romans took it to Britain with them, the use of feet was so small as to scarcely be of consequence.

Undoubtedly, football is nowadays the most popular sport in the world. It is important not only on the sport level, as a game and pastime, but also on the social level, since it joins people, social groups, clubs or even countries. Moreover, football is one of the sports which generates more money in Europe, Latin America, Asia and, recently, in United States. Thus the FIFA (Fédération Internationale de Football Association) World Cup, as many other international competitions, entail a significant capital movement by multinationals, societies or individuals.

As occurs with other sports, where statistical studies are essential part of them (baseball, American football, basketball...) almost all aspects of the game are gathered and analysed. These data are referred to results (number of goals, assists, dribbles, ...) but also to personal characteristics of each player. And many of these databases are available for the researchers. In this sense, the aim of our study is to model and compare two important aspects for the footballer performance expressed in form of count data variables: The number of goals scored and the number of yellow cards received. To do this analysis we have taken into account the global history of each player in the same team. We compare these two variables for the two first teams of each one of the leagues in Spain, England, Germany and Italy.

The work is structured as follows. In Section 2 the models employed are briefly described. Section 3 details the variables used and Section 4 shows the fits obtained in each case. Finally some conclusions are stated.

2 Count data models

One of the advantages of the Calculus of Probabilities is to give models that may be employed for the probabilistic approach of data. The characteristics of the theoretical model can be used for the interpretation of the fitted model and so providing a better knowledge of the studied phenomenon. In particular, the analysis of count data variables through the use of discrete probability distributions is very useful in many disciplines. The basis model for these discrete data sets is the Poisson distribution which has the property of equidispersion, that is to say, the variance is always equal to the mean. But count data sets often verify that the variance is greater than the mean, which is called overdispersion. For this situation, multiple models have been developed, in many cases through a Poisson mixture, such as the Negative Binomial (*NB*) or the Univariate Generalized Waring (*UGW*) distributions, as well as through other models such as the Generalized Poisson (*GP*) or the Complex Biparametric Pearson (*CBP*) distributions.

In this work we use these distributions in order to model count data variables related to football. Briefly we describe the models aforementioned.

- Negative Binomial distribution [3]: $X \sim NB(\theta, \mu)$ with $\theta, \mu > 0$ and probability mass function (pmf) given by

$$P(X = x) = \frac{\Gamma(\theta + x)}{\Gamma(\theta)x!} \left(\frac{\theta}{\theta + \mu} \right)^\theta \left(\frac{\mu}{\theta + \mu} \right)^x, \quad x = 0, 1, \dots$$

- Generalized Poisson distribution [1]: $X \sim GP(\lambda, \theta)$, $\lambda > 0$, $\max(-1, -\lambda/m) < \theta < 1$ with pmf given by

$$P(X = x) = \begin{cases} \frac{\lambda(\lambda + \theta x)^{x-1}}{x!} e^{-\lambda - \theta x} & x = 0, 1, \dots \\ 0 & x > m \end{cases}$$

where $m \geq 4$ is the largest positive integer for which $\lambda + m\theta > 0$ when $\theta < 0$. Let us observe that, in that case, the distribution has finite range (from 0 to m). This lower bound on θ is imposed in order to assure at least five points in the sample space with positive probabilities when θ is negative. As a consequence of the definition, when $\theta < 0$, the distribution does not sum to unity in $0, 1, \dots, m$, and must be normalized.

- Univariate Generalized Waring distribution [2, 10, 6, 9]: $X \sim UGW(a, k, \rho)$ with $a, k > 0$, $\rho > 2$ and pmf given by

$$P(X = x) = \frac{\Gamma(a + \rho)\Gamma(k + \rho)}{\Gamma(a)\Gamma(k)\Gamma(\rho)} \frac{\Gamma(a + x)\Gamma(k + x)}{\Gamma(a + k + \rho + x)\Gamma(x + 1)}, \quad x = 0, 1, \dots$$

- Complex Biparametric Pearson distribution [4, 7]: $X \sim CBP(b, \gamma)$ with $b, \gamma > 0$ and pmf given by

$$P(X = x) = \frac{\Gamma(\gamma + bi)\Gamma(\gamma - bi)}{\Gamma(\gamma)^2} \frac{(bi)_x(-bi)_x}{(\gamma)_x} \frac{1}{x!}, \quad x = 0, 1, \dots \quad (1)$$

where i is the imaginary unit and $(\alpha)_r = \Gamma(\alpha + r)/\Gamma(\alpha)$, $\alpha > 0$.

- Complex Triparametric Pearson distribution [5, 8]: $X \sim CTP(a, b, \gamma)$ with $a \in \mathbb{R}, b, \gamma > 0$ and pmf given by

$$P(X = x) = f_0 \frac{(a + ib)_x(a - ib)_x}{(\gamma)_x} \frac{1}{x!}, \quad x = 0, 1, \dots$$

where f_0 is the normalizing constant whose expression is

$$f_0 = \frac{\Gamma(\gamma - a - ib)\Gamma(\gamma - a + ib)}{\Gamma(\gamma)\Gamma(\gamma - 2a)}. \quad (2)$$

3 Description of data

Describing and documenting data is essential in ensuring that those people who may need to use the data can make sense of them and understand the processes that have been followed in the collection, processing, and analysis of the data. So, this section is devoted to this task.

We focus on the most famous European football leagues which are the Spanish, German, Italian and English football leagues. However, we do not study them in the same period. Specifically, we consider the data about the:

- Spanish football league from 1970 to 2018,
- English football league from 1992 to 2014,
- Italian football league from 1988 to 2018, and
- German football league from 1985 to 2014.

The variables selected for the study are:

- Number of yellow cards received by a footballer in the corresponding team (along the span time considered).
- Number of goals scored by a footballer in the corresponding team (along the span time considered).

We have made the analysis for all the teams that have belonged or belong to each one of these football leagues, but in this work we only have included the two teams that more titles of the national championship have achieved throughout the history of the league of their country. Thus, the teams selected within each league are:

- Spanish football league:
 - Real Madrid
 - Barcelona
- English football league:
 - Manchester United
 - Liverpool
- Italian football league:
 - Juventus
 - Milan
- German football league:
 - Bayern Munich
 - Borussia Dortmund

Data have been collected from the web <https://www.bdfutbol.com/>. In a first approach to these data, we make a descriptive summary of them. Table 1 contains the first and third quartiles, the mean, the median, the standard deviation (s.d.) as well as the minimum and the maximum for each one of the variables considered (by football team).

As it can be observed from Table 1, data exhibit overdispersion in all the teams, that is, the variance is greater than the mean.

4 Fit of data

In this section we model the data using the count distributions described in Section 2. For all the fits the maximum likelihood estimates (MLE) of the parameters, the Akaike Information Criterion (AIC) and the Pearson χ^2 –goodness of fit test have been obtained. In addition, graphs with the observed versus the expected frequencies are shown.

Spanish league							
Number of yellow cards							
Statistics	Min.	Q_1	Median	Mean	Q_3	Max.	s.d.
Barcelona	0.00	0.00	2.00	8.20	9.00	81.00	13.92
Real Madrid	0.00	0.00	2.00	8.59	11.00	111.00	15.60
Number of goals scored							
Statistics	Min.	Q_1	Median	Mean	Q_3	Max.	s.d.
Barcelona	0.00	0.00	2.00	10.99	9.75	383.00	27.86
Real Madrid	0.00	0.00	1.50	11.19	8.75	311.00	30.55
English league							
Number of yellow cards							
Statistics	Min.	Q_1	Median	Mean	Q_3	Max.	s.d.
Manchester United	0.00	0.00	1.00	7.56	8.00	96.00	14.26
Liverpool	0.00	0.00	2.00	5.89	7.00	66.00	9.18
Number of goals scored							
Statistics	Min.	Q_1	Median	Mean	Q_3	Max.	s.d.
Manchester United	0.00	0.00	2.00	12.46	10.00	158.00	25.88
Liverpool	0.00	0.00	2.00	8.30	7.25	128.00	18.97
Italian league							
Number of yellow cards							
Statistics	Min.	Q_1	Median	Mean	Q_3	Max.	s.d.
Juventus	0.00	0.00	3.00	7.15	9.00	59.00	10.94
Milan	0.00	0.00	3.00	7.01	8.00	97.00	12.96
Number of goals scored							
Statistics	Min.	Q_1	Median	Mean	Q_3	Max.	s.d.
Juventus	0.00	0.00	2.00	7.27	7.00	188.00	17.65
Milan	0.00	0.00	1.00	6.51	6.00	127.00	14.44
German league							
Number of yellow cards							
Statistics	Min.	Q_1	Median	Mean	Q_3	Max.	s.d.
Bayern Munich	0.00	1.00	4.00	9.36	13.00	59.00	12.50
Borussia Dortmund	0.00	0.00	2.00	7.94	11.75	66.00	12.02
Number of goals scored							
Statistics	Min.	Q_1	Median	Mean	Q_3	Max.	s.d.
Bayern Munich	0.00	0.00	4.00	11.79	12.00	107.00	19.75
Borussia Dortmund	0.00	0.00	2.00	7.89	8.00	116.00	15.78

Table 1: Descriptive summary of data

4.1 Number of yellow cards

Table 2 contains the AIC value and the p -value corresponding to the χ^2 -goodness of fit test for the fits about the number of yellow cards received by a footballer in the teams selected. The best AIC and p -value within each team are highlighted in bold.

The MLEs of the parameters and their standard errors (in brackets) for the best fit according to the AIC are included in Table 3.

As it can be seen from Table 2, not all the studied teams behave in the same way when

Football team				
	<i>Bayern Munich</i>	<i>Borussia Dortmund</i>	<i>Manchester United</i>	<i>Liverpool</i>
Model	AIC			
<i>NB</i>	1099.58	1253.97	819.95	1000.25
<i>GP</i>	1112.53	1261.56	829.51	1000.49
<i>CBP</i>	1164.12	1310.62	859.94	1033.83
<i>UGW</i>	1101.60	1256.04	821.96	1001.91
<i>CTP</i>	1129.08	1277.71	841.56	1009.94
Model	<i>p</i> -value			
<i>NB</i>	0.63	0.04	0.34	0.76
<i>GP</i>	0.06	0.11	0.06	0.50
<i>CBP</i>	0.00	0.00	0.00	0.00
<i>UGW</i>	0.56	0.30	0.26	0.70
<i>CTP</i>	0.00	0.00	0.00	0.13
	<i>Barcelona</i>	<i>Real Madrid</i>	<i>Juventus</i>	<i>Milan</i>
Model	AIC			
<i>NB</i>	2081.58	2041.53	1492.48	1603.37
<i>GP</i>	2093.52	2061.78	1511.09	1600.09
<i>CBP</i>	2170.39	2143.54	1584.37	1656.69
<i>UGW</i>	2083.60	2043.50	1494.50	1604.00
<i>CTP</i>	2118.49	2087.93	1532.02	1612.32
Model	<i>p</i> -value			
<i>NB</i>	0.63	0.92	0.12	0.23
<i>GP</i>	0.06	0.03	0.00	0.36
<i>CBP</i>	0.00	0.00	0.00	0.00
<i>UGW</i>	0.57	0.89	0.09	0.19
<i>CTP</i>	0.00	0.00	0.00	0.00

Table 2: AIC value and χ^2 -goodness of fit test for fits about yellow cards data

<i>Bayern Munich</i>	<i>Borussia Dortmund</i>	<i>Manchester United</i>	<i>Liverpool</i>
<i>NB</i>			
$\hat{\theta} = 0.485(0.056)$	$\hat{\theta} = 0.436(0.047)$	$\hat{\theta} = 0.280(0.037)$	$\hat{\theta} = 0.518(0.062)$
$\hat{\mu} = 9.358(1.047)$	$\hat{\mu} = 7.943(0.852)$	$\hat{\mu} = 7.560(1.189)$	$\hat{\mu} = 5.894(0.637)$
<i>Barcelona</i>	<i>Real Madrid</i>	<i>Juventus</i>	<i>Milan</i>
<i>NB</i>		<i>GP</i>	
$\hat{\theta} = 0.354(0.029)$	$\hat{\theta} = 0.322(0.027)$	$\hat{\theta} = 0.405(0.041)$	$\hat{\lambda} = 1.355(0.091)$
$\hat{\mu} = 8.204(0.745)$	$\hat{\mu} = 8.590(0.823)$	$\hat{\mu} = 7.154(0.716)$	$\hat{\theta} = 0.806(0.022)$

Table 3: MLEs and standard errors (in brackets) for the best fit of yellow cards data

modelling the number of yellow cards. According to the AIC, the best model is the *NB* for all the teams except for the Milan football club which selects the *GP* model. This performance is more irregular according to the χ^2 -goodness of fit test, since there are several cases in which it disagrees with the AIC. Thus, for the Borussia Dortmund team the best fit, according to the test, is that provided by the *UGW* model; however, the fit with lowest AIC is the corresponding to the *NB* model. For the Milan football club both criteria agree on the *GP* model. For the rest

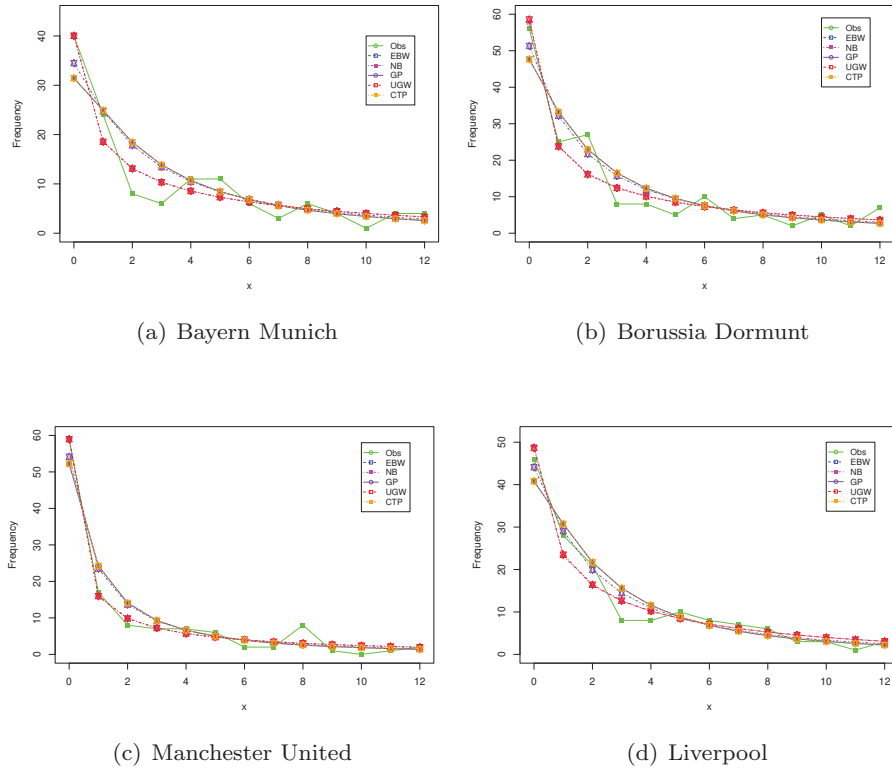


Figure 1: Observed and expected frequencies for data about the number of yellow cards

of the teams the NB distribution is the best model with both criteria. It should be emphasized that in some teams the yellow cards data are adequately modelled by several distributions, such as the Liverpool football club, in which - in addition to the NB distribution - the UGW and CTP distributions are also suitable.

Similar conclusions can be obtained from Figures 1 and 2, where the observed and expected frequencies for each fitted model are shown.

4.2 Number of goals scored

Next we study the number of goals scored by the teams. We have followed the same procedure as in Section 4.1.

If we analyze the AIC value from Table 4, we can observe that the NB model provides the best fit for the German teams and the Manchester United football club, whereas the GP fit is the best for the rest of the teams. Using the χ^2 -goodness of fit test, there is a wider range of appropriate models:

- The case of Manchester United can be considered special, since according to the p -value of the goodness of fit test, both the GP and NB distributions could be adequate models.
- For the goals scored data corresponding to the Bayern Munich team there are two appropriate models, the NB and the UGW , although the NB fit is the best one according to the AIC.

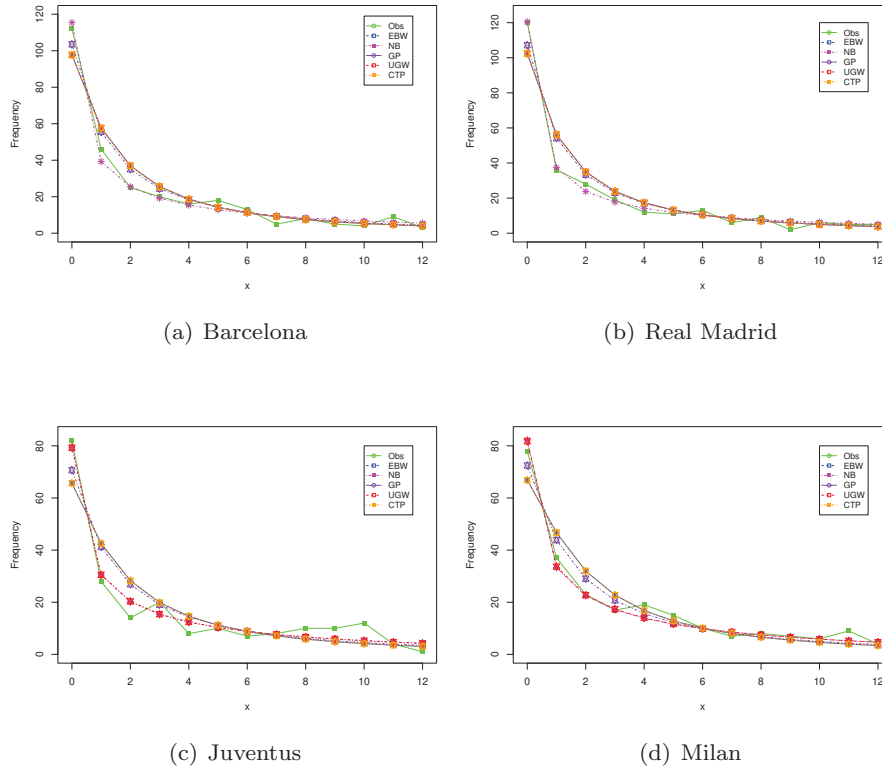


Figure 2: Observed and expected frequencies for data about the number of yellow cards

- Besides these two distributions, the Borussia Dortmund and Real Madrid teams add the *GP* distribution, although the *NB* fit continues being the best in terms of the AIC.
- The Barcelona and Juventus teams also add the *CTP* distribution, although the best fit is that related to the *GP* distribution.
- Goals data for the Liverpool team can be modelled by the majority of the distributions used (the only ones that do not fit appropriately are the *CBP* and *UGW* distributions). However, in this case, the *NB* fit is not the best, as in the previous cases, but the *GP* fit.
- Finally, we have another particular case, the Milan football club, since for this team only the *CBP* distribution does not provide an appropriate fit.

Figures 3 and 4 show the observed and expected frequencies for each fitted model depending on the football team.

5 Conclusions

As it has been observed in the previous sections, the *NB* and *GP* distributions are the best models for the number of yellow cards received and goals scored by the most famous European football teams. The performance of the two variables is rather similar among leagues and also

<i>Bayern Munich</i>	Football team			
	<i>Borussia Dortmund</i>	<i>Manchester United</i>	<i>Liverpool</i>	
Model	AIC			
<i>EBW</i>	1151.47	1189.71	810.80	924.01
<i>NB</i>	1128.00	1170.98	796.76	924.36
<i>GP</i>	1138.86	1177.41	1177.41	917.88
<i>CBP</i>	1191.13	1218.55	828.55	941.92
<i>UGW</i>	1130.00	1172.29	6003.01	7483.96
<i>CTP</i>	1153.47	1191.71	812.80	926.01
Model	<i>p</i> –value			
<i>EBW</i>	0.00	0.04	0.00	0.42
<i>NB</i>	0.45	0.86	0.33	0.44
<i>GP</i>	0.04	0.33	0.33	0.75
<i>CBP</i>	0.00	0.00	0.00	0.00
<i>UGW</i>	0.38	0.82		
<i>CTP</i>	0.00	0.04	0.01	0.27
	<i>Barcelona</i>	<i>Real Madrid</i>	<i>Juventus</i>	<i>Milan</i>
	AIC			
<i>EBW</i>	1956.44	1863.66	1300.40	1355.69
<i>NB</i>	1932.34	1865.35	1289.98	1345.64
<i>GP</i>	917.88	917.88	917.88	1344.33
<i>CBP</i>	941.92	941.92	941.92	1383.34
<i>UGW</i>	1930.64	1866.53	1289.74	1347.59
<i>CTP</i>	926.01	926.01	926.01	1357.69
Model	<i>p</i> –value			
<i>EBW</i>	0.00	0.00	0.04	0.34
<i>NB</i>	0.46	0.63	0.67	0.67
<i>GP</i>	0.75	0.06	0.75	0.87
<i>CBP</i>	0.00	0.00	0.00	0.00
<i>UGW</i>	0.50	0.57	0.66	0.60
<i>CTP</i>	0.27	0.00	0.27	0.28

Table 4: AIC value and p –value of χ^2 –goodness of fit test for fits about goals scored data.

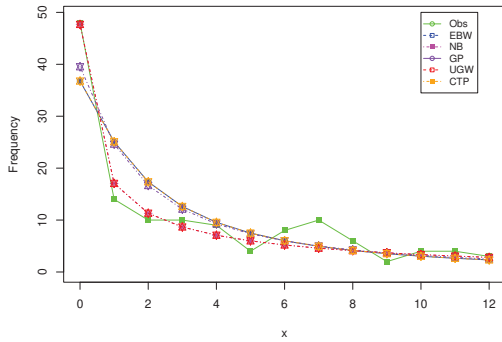
among teams. In general, the number of yellow cards received is modelled by a NB distribution, whereas the number of goals scored is modelled by a GP distribution. Parameter estimates are quite similar among teams in the same country, except for the case of the Liverpool football club, since the variable follows a GP distribution.

References

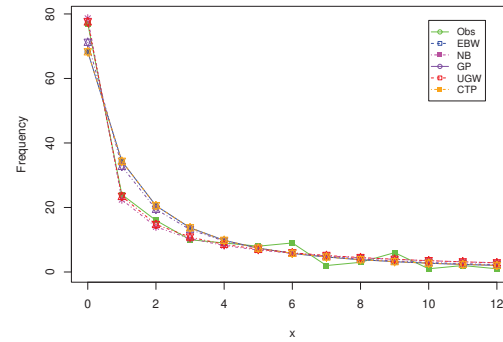
- [1] P. C. Consul and F. Famoye. Maximum likelihood estimation for the generalized Poisson distribution when sample mean is larger than sample variance. *Communications in Statistics: Theory and Methods*, 17:299–309, 1988.
- [2] J.O. Irwin. The generalized Waring distribution applied to accident theory. *Journal of the Royal Statistical Society. Series A*, 131:205–225, 1968.

<i>Bayern Munich</i>	<i>Borussia Dortmund</i>	<i>Manchester United</i>	<i>Liverpool</i>
<i>NB</i>		<i>GP</i>	
$\hat{\theta} = 0.368(0.033)$	$\hat{\theta} = 0.296(0.042)$	$\hat{\theta} = 0.295(0.033)$	$\hat{\lambda} = 1.410(0.117)$
$\hat{\mu} = 11.792(1.50)$	$\hat{\mu} = 7.586(1.019)$	$\hat{\mu} = 7.886(1.02)$	$\hat{\theta} = 0.761(0.030)$
<i>Barcelona</i>	<i>Real Madrid</i>	<i>Juventus</i>	<i>Milan</i>
<i>GP</i>			
$\hat{\lambda} = 1.089(0.091)$	$\hat{\lambda} = 1.088(0.092)$	$\hat{\lambda} = 1.411(0.117)$	$\hat{\lambda} = 0.956(0.074)$
$\hat{\theta} = 0.861(0.023)$	$\hat{\theta} = 0.862(0.024)$	$\hat{\theta} = 0.861(0.023)$	$\hat{\theta} = 0.853(0.024)$

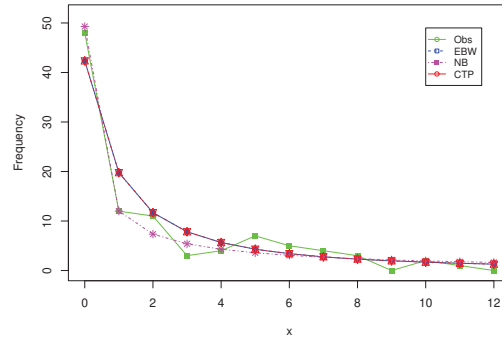
Table 5: MLEs and standard errors (in brackets) for the best fit of goals scored data



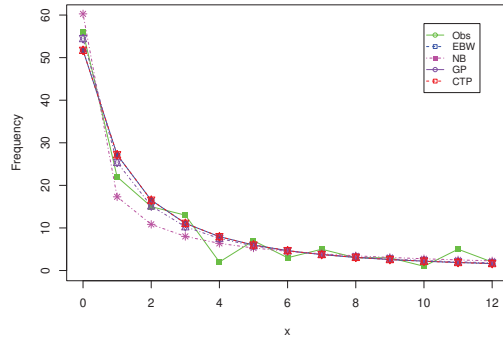
(a) Bayern Munich



(b) Borussia Dortmund



(c) Manchester United



(d) Liverpool

Figure 3: Observed and expected frequencies for data about the number of goals scored

- [3] N. L. Johnson, A. W. Kemp, and S. Kotz. *Univariate discrete distributions*. Wiley, New York, 3rd edition, 2005.
- [4] J. Rodríguez-Avi, A. Conde-Sánchez, and A. J. Sáez-Castillo. A new class of discrete distributions with complex parameters. *Statistical Papers*, 44:67–88, 2003.
- [5] J. Rodríguez-Avi, A. Conde-Sánchez, A. J. Sáez-Castillo, and M. J. Olmo-Jiménez. A triparametric

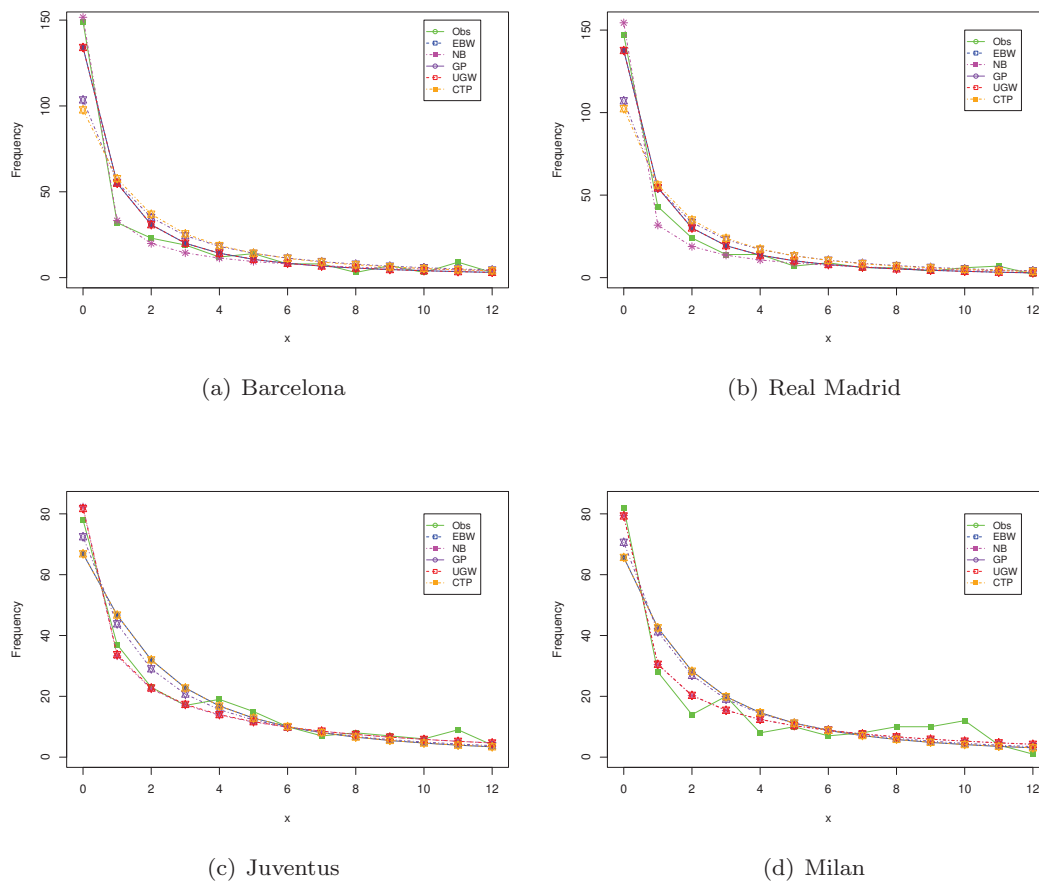


Figure 4: Observed and expected frequencies for data about the number of goals scored

- discrete distribution with complex parameters. *Statistical Papers*, 45:81–95, 2004.
- [6] J. Rodríguez-Avi, A. Conde-Sánchez, A.J. Sáez-Castillo, M.J. Olmo-Jiménez, and A.M. Martínez-Rodríguez. A generalized Waring regression model for count data. *Computational Statistics & Data Analysis*, 53(10):3717–3725, 2009.
- [7] J. Rodríguez-Avi and M.J. Olmo-Jiménez. A regression model for overdispersed data without too many zeros. *Statistical Papers*, 58:749–773, 2017.
- [8] J. Rodríguez-Avi, M.J. Olmo-Jiménez, and V. Cueva-López. A review of the CTP distribution: a comparison with other over- and underdispersed count data models. *Journal of Statistical Computation and Simulation*, 2018.
- [9] S. Vilchez-López, A.J. Sáez-Castillo, and M.J. Olmo-Jiménez. GWRM: An R package for identifying sources of variation in overdispersed count data. *PLoS ONE* 11(12): e0167570, 11, 2016.
- [10] E. Xelakaki. The univariate generalized Waring distribution in relation to accident theory: prone-ness, spells or contagion? *Biometrics*, 39:887–895, 1983.

Randomness of Play Calling in College Football

Brian Curley*, Gretchen Hopkirk*, Ryan Lokhorst* and A. Pilkington**

*University of Notre Dame, Notre Dame, IN 46556: bcurley@nd.edu, ghopkirk@nd.edu, rlokhors@nd.edu

** Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556: Pilkington.4@nd.edu

Abstract

In American Football it is widely accepted that the ability to predict the next play of the offensive team gives an advantage to their opponent. The purpose of this study was to examine the effect of a random versus a non-random (and hence more predictable) sequence of offensive plays on outcomes in American College Football. Surprisingly, differences in outcomes were found only for away teams and in this case, the outcomes were seen to depend on the nature of the non-random pattern of play. We saw that non-random play with very few switches between runs of rushing and passing plays did not significantly reduce the chances of winning for either team. In fact, away teams having this style of play showed a significant advantage in some key game statistics. Away teams with non-random play that had many switches between rushing and passing plays had a significantly lower proportion of wins and a significant disadvantage in some game statistics. Our conclusions are that predictability itself does not lead to a serious disadvantage on the playing field in American College Football, rather away teams who exhibit a particular type of non-random play involving many switches between running and passing plays are at a disadvantage. The asymmetry between the results for home and away teams suggest that this type of play plays some role in home advantage. Since play calling is, to a large degree, a factor under the control of the offensive team, knowledge of this association may be useful in forming a strategy for the away team.

1 Introduction

Recent studies on models for predicting plays in the NFL [2] point to the commonly held belief that the ability to predict plays in American Football is a significant factor in determining the outcome of the game. The assumptions of game theory tell us that predictability should be a disadvantage for any team, but we find no papers in the literature examining the effect of this factor on wins and losses or other important statistics in football. This study examines the effect of randomness of play on wins and losses and various performance statistics in college football with the available public data. One would expect that sequences of offensive play that differ significantly from a random pattern increase the ability of the opposing team to predict plays and thus lead to a disadvantage for the offensive team. On the other hand patterns of offensive play that are not significantly different from randomly generated plays would tend to render attempts at prediction ineffective. Unlike game statistics such as yards gained or time of possession, the sequence of offensive plays executed is to a large degree under the control of the offensive team and can be changed strategically to improve outcomes.

As with many statistical studies of college football, we found that the relationship between randomness of play and wins and losses is more complex than what we expected. Using the Wald Wolfowitz test for

randomness, we categorized the sequence of offensive plays for each game for both home and away teams as either significantly different from a randomly generated sequence or not. Teams with sequences of offensive play that had a significantly large or small number of runs of play types (or equivalently switches between play types) in comparison to what is expected are deemed significantly different than a randomly generated sequence by the test. For the sequences of play deemed significantly different from a random pattern, we also kept track of whether the number of runs in the sequence of offensive plays was significantly larger or significantly smaller than the expected number. The Wald Wolfowitz test does not make any assumption about the percentage of rushing or passing plays in the data. Thus it recognizes the subtle difference between a team who plays a large percentage of rushing plays, switching between passing and rushing plays in a manner resembling random play calling, and a team who plays a large percentage of rushing plays, but has too many or too few switches between play types for the play calling to be considered random.

The results were not as expected and somewhat counterintuitive. For the home team there was no significant disadvantage to executing a sequence of plays that was significantly different from a randomly generated sequence. There was a surprising asymmetry between the results for the home team and the away team. Away teams for whom the sequence of offensive plays had a significantly higher number of runs of rushing and pass plays than that expected in a randomly generated sequence had a significantly lower probability of winning than the other groups. On the other hand, away teams demonstrating non-random offensive play sequences with fewer runs (and hence fewer switches between play types) did not have a significantly lower probability of a win when compared with the other groups, in fact they had an advantage in rushing yards and the number of first downs from rushing plays when compared with the other groups of away teams.

Since this is an observational study, it is unclear whether the pattern of play resulting in a large number of switches between offensive play types is a symptom of the confusion and disorientation that an away team experiences, or if it is a poor strategy choice that leads to poor outcomes for the away team. Our results seem to indicate that an away team can significantly reduce the home field advantage by switching between offensive play types less often.

2 Data, Methodology and Notation

In the study, we looked at 7,220 Division I college football games, played between 2005 and 2013. The data sets were compiled by <http://www.cfbstats.com/> and are publicly available. For a given game and team, we looked at the sequence of offensive play choices of the form "RUSH" and "PASS" for the entire game. If the number of runs (or equivalently the number of switches between play types) of "RUSH" and "PASS" plays in such a sequence was either too large or too small, one would expect that the play calling was not similar to random play calling.

2.1 Wald Wolfowitz (WW) Test for Randomness

The Wald Wolfowitz test is a test for randomness in binary data with two values success (S) and failure (F). The test statistic is the number of runs of Ss and Fs in the data. The Wald Wolfowitz test does not make any assumptions about the probability of success or failure on any trial.

Given a sequence with two values, success (S) and failure (F), with N_s success' and N_f failures, let X denote the number of runs (of both S's and F's). Wald and Wolfowitz determined that for a random sequence

of length N with N_s success' and N_f failures (note that $N = N_s + N_f$), the number of runs has mean and standard deviation given by

$$E(X) = \mu = \frac{2N_s N_f}{N} + 1, \quad \sigma(X) = \sqrt{\frac{(\mu - 1)(\mu - 2)}{N - 1}}.$$

The distribution of X is approximately normal if N_s and N_f are both bigger than 10. Therefore the Z - value:

$$Z = \frac{x - \mu}{\sigma} = \frac{X - \left(\frac{2N_s N_f}{N} + 1\right)}{\sqrt{\frac{(\mu - 1)(\mu - 2)}{N - 1}}}$$

has a standard normal distribution.

For example to test the hypothesis that the sequence of "RUSH"'s and "PASS"'s shown below is generated randomly, we let N_P denote the number of "PASS"'s and let N_R denote the number of "RUSH"'s in the sequence.

RUSH RUSH PASS PASS PASS RUSH PASS RUSH RUSH RUSH PASS RUSH PASS PASS RUSH PASS PASS PASS
PASS RUSH PASS RUSH PASS RUSH RUSH RUSH PASS PASS RUSH RUSH RUSH PASS PASS PASS RUSH PASS
RUSH RUSH PASS RUSH RUSH RUSH PASS PASS RUSH RUSH RUSH PASS RUSH PASS PASS RUSH RUSH PASS PASS
RUSH PASS PASS RUSH RUSH RUSH PASS RUSH PASS PASS RUSH PASS RUSH PASS PASS RUSH PASS RUSH PASS
RUSH PASS PASS RUSH PASS PASS RUSH PASS RUSH PASS PASS PASS

The sequence of plays has length $N = 87$. The number of "PASS" 's in the sequence is $N_P = 45$ and the number of "RUSH" 's is $N_R = 42$. We have underlined the runs of "RUSH" 's in the sequence and we see that the total number of runs in the sequence is $X = 52$ Thus we have

$$\mu = \frac{2(45)(42)}{87} + 1 = 44.44828, \quad \sigma \approx \sqrt{\frac{(43.44828)(42.44828)}{86}} \approx 4.630918$$

and the value of our test statistic is

$$Z = (X - \mu) / \sigma = 1.630718, \quad \text{p-value} = 0.1029497.$$

Thus we do not have sufficient evidence to reject the null hypothesis at a 95% level of confidence in this case; in other words, this sequence of plays is not significantly different from what we would expect to see in a randomly generated sequence of plays. Note that a sequence can lead to a rejection of the null hypothesis in either of two ways; the test statistic, Z , might be high in absolute value and negative, or it might be high in absolute value and positive.

2.2 Notation

For each game in our set of data we classified the teams as either the Home (H) or Away (A) team. We calculated the Wald-Wolfowitz (WW) test statistic for the sequence of passing and rushing plays made by the team's offense throughout the course of the game and classified the test statistic, Z , resulting from that sequence as; Low (L), indicating that it was not large enough in absolute value to reject the Null Hypothesis, Significant and Positive (SP), indicating that the test statistic was positive and led to a rejection of the null

hypothesis at a 5% level of confidence, Significant and Negative (SN), indicating that the test statistic was positive and led to a rejection of the null hypothesis at a 5% level of confidence. We amalgamated both of the above classifications to get 9 categories for the teams involved in any given game: HSN (Home Team with significant negative WW Z-score), HL (Home Team with non-significant WW Z-score), HSP (Home Team with significant positive WW Z-score), ASN (Away Team with significant negative WW Z-score), AL (Away Team with non-significant WW Z-score) and ASP (Away Team with significant positive WW Z-score).

With this classification, we have a cross classification of each of the 7,220 games studied in terms of the two categorical variables, the status of the WW test statistic for the away team, with categories ASN, AL and ASP and the status of the WW test statistic for the home team, with categories HSN, HL and HSP. Table 1 below shows the number of games in each of the resulting nine game categories. As you can see the vast majority of games had at least one team with an offensive play calling sequence that was not statistically distinguishable from random play calling.

Home Team(H)	Away Team(A)			Total
	ASN	AL	ASP	
HSN	N = 19	N = 384	N = 15	N = 418
HL	N = 327	N = 6073	N = 224	N = 6624
HSP	N = 11	N = 163	N = 4	N = 178
Total	N = 357	N = 6620	N = 243	N = 7220

Table 1: N = Number of games in each category.

A = Away Team, H = Home Team .

WW Test Statistic (SP = Significant and Positive, SN = Significant and Negative, L = Not significant).

3 Wins and Losses

In the contingency table below, Table 2, we show the number of wins for the home team (H), the number of wins for the away team (A) and the proportion of wins for the away team ($p(A)$) for the games in each category, where the games are cross classified as in Table 1, according to the status of the WW Z-score of the home and away teams. Chi-square tests of independence was performed to examine the relation between the status of the WW Z-statistic and wins.

The chi-squared test of independence on the marginal distribution of the away teams showed a significant difference in the success rate of the categories (χ -squared = 13.755, $df = 2$, p -value = 0.001031). Looking at the standardized residuals, we see that the success rate of the away teams in the category ASP is significantly lower than that of the away teams in the categories ASN and AL, and there is not a significant difference between the proportion of wins for away teams in the categories ASN and AL. This indicates that predictability itself does not lead to a significant disadvantage for the away team, rather the nature of unpredictability may lead to a disadvantage. In particular, away teams for which the number of switches between runs of passing plays and rushing plays is significantly higher than what one would expect in a randomly generated sequence, are at a significant disadvantage when compared to their less predictable and predictable counterparts in the

AL and ASN categories respectively.

A chi squared test of independence on the marginal distribution for the home teams does not reveal any significant difference between the probability of success for each of the three categories HSN, HL and HSP (χ -squared = 2.3137, df = 2, p-value = 0.3145), thus indicating that the overall probability of a win for the home team is independent of whether the sequence of offensive plays appears random or not.

The conditional distributions obtained by fixing a category of play for the home team give the following results:

- When the home team falls in the category HSN, a χ -squared test of independence reveals no significant relationship between the status of the WW Z-statistic for the away team and wins (χ -squared = 1.4006, df = 2, p-value = 0.4964).
- When the home team falls in the category HL, a χ -squared test of independence show that there is a relationship between the status of the WW Z-statistic for the away team and wins for the away team (χ -squared = 11.242, df = 2, p-value = 0.003621). An examination of the standardized residuals reveals that the success rate of the away teams in the category ASP is significantly lower from that of the away teams in the categories ASN and AL and there is not a significant difference between the proportion of wins for away teams in the categories ASN and AL.
- When the home team falls in the category HSP, Fisher's exact test (used when the counts in some categories are very small) reveals no significant relationship between the status of the WW Z-statistic for the away team and wins (p-value = 0.1553).

The conditional distributions of wins and losses for the home team reveal also that given any of the three categories of play for the away team (ASN, AL, ASP), the probability of a win for the home team is independent of whether the sequence of offensive plays appears random or not.

Thus we conclude that the away team can significantly reduce their chances of winning by switching between offensive play types very often. For example, if a team expects to play about 60 offensive plays in a game and wants to play roughly 1/2 ($N_R = 30$) rushing plays and 1/2 ($N_P = 30$) passing plays, then they should keep the number of switches of play types below $\mu + 1.96\sigma \approx 38.53$. where

$$\mu = \frac{2(30)(30)}{60} + 1 = 31, \quad \sigma = \sqrt{\frac{(\mu - 1)(\mu - 2)}{N - 1}} \approx 3.84.$$

3.1 Point Differential

An analysis of variance also shows significant differences between the average point differential across levels of the WW Z-score factor for the away team but not for the home team. The average point differentials (PD) for the three different groups are given by :

WW Away Team	Mean PD(Home Pts. - Away Pts.)
ASN	4.37
AL	7.16
ASP	14.70

The output of an analysis of variance applied to the means for the three groups is shown in Table 3 and shows that there is a significant difference between the means.

Home Team	Away Team			
	ASN	AL	ASP	Total
HSN	H 9 A 10 p(A) 0.52	H 226 A 158 p(A) 0.41	H 10 A 5 p(A) 0.33	H 245 A 173 p(A) 0.41
HL	H 193 A 134 p(A) 0.40	H 3767 A 2306 p(A) 0.38	H 162 A 62 p(A) 0.28	H 4122 A 2502 p(A) 0.38
HSP	H 5 A 6 p(A) 0.54	H 99 A 64 p(A) 0.39	H 4 A 0 p(A) 0.00	H 108 A 70 p(A) 0.39
Total	H 207 A 150 p(A) 0.42	H 4092 A 2528 p(A) 0.38	H 176 A 67 p(A) 0.27	H 4475 A 2745 p(A) 0.38

Table 2: Percentage of Away wins for each combination of factors.

H = # Wins for Home Team, A = # Wins for Away Team,

p(A) = Proportion Wins for Away Team.

Pairwise comparisons reveal significant differences (at a 95% level of confidence) between the average point differential in all three pairwise comparisons with the Holm adjustment, and significant differences between the mean for ASP and the other two levels of the WW Z score status for the Away team with the Bonferroni adjustment. The p-values for both follow up tests are shown in Table 4 below.

Note that the p-value for the difference in the average point differential between the ASN and AL levels is borderline at 0.06 with the Bonferroni adjustment, this is still significant at a 90% level of confidence. The barplot in Figure 1 below shows the the means for all three categories along with error bars.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
WW Away Team	2	16496	8248	17.01	4.26e-08 ***
Residuals	7217	3499141	485		

Table 3: Output of ANOVA applied to the mean of the PD for three levels of WW Away Team.

	Holm Adj.			Bonferroni Adj.	
	ASN	AL		ASN	AL
AL	0.02	-	AL	0.06	-
ASP	5.3e-08	3.2e-07	ASP	5.3e-08	4.9e-07

Table 4: p-values for pairwise comparisons of the mean of the PD for three levels of WW Away Team.



Figure 1: Mean of PD for each category for Away Team.

4 Comparison of Means for Game Statistics

In [3], Wagner lists a number of game statistics which are significant variables in predicting the point differential for the game. In order to get more insight into how the number of switches between types of offensive plays for the away team affects the point differential, we compared the means of several of these statistics across the three categories of the status of the WW Z-score for the away team.

The variables we looked at were:

- DTOP: Difference in time of possession (Home TOP - Away TOP).
- DRY: Difference in Rushing Yards (Home RY - Away RY).
- DPY: Difference in Passing Yards (Home PY - Away PY).
- DFDR: Difference in First Downs from Rushing (Home FDR - Away FDR).
- DFDP: Difference in First Downs from Passing (Home FDP - Away FDP).
- DFDPen : Difference in First Downs from Penalties (Home FDPen - Away FDPen).

We ran an analysis of variance for each statistic above and found significant differences in the means across the categories ASP, AL, ASN only for DRY and DFDR. Tables 5 and 6 show the results of our analysis of variance and the follow up comparison of means with Bonferroni adjustment, for the variables DRY and DFDR respectively. Figure 2 shows the barplots for the means of the other variables on our list across the categories ASP, AL, ASN. Despite the fact that the differences between means are not statistically significant, they do show an upward trend as the number of switches in play types increases across the three categories, most likely contributing to the significant increase in the point differential in favor of the home team when the away team is in the category ASP.

5 Conclusion and Further Study

The most surprising result in this study was that teams who switched between offensive play types less often were at no disadvantage, despite having what would appear to be a more predictable sequence of plays. In fact, for the away team, this style of play seems to be advantageous.

The asymmetry between the effect of randomness of play calls for the home and away teams is also something of a surprise. For home teams, there was no significant difference in the probability of win between those teams whose play sequences fitted the random profile and those who had play sequences that did not. On the other hand for the away teams, teams who had offensive play sequences that did not fit the random profile and had relatively large numbers of switches between play types had a significantly smaller proportion of wins than the away teams in other categories. In fact we saw that away teams with non-random offensive play sequences but fewer switches between play types had a significantly more favorable rushing yard differential than the away teams in the other categories.

We see that non-random offensive play sequences with high numbers of switches between play types puts the away team at a disadvantage and is not a good choice of strategy. On the other hand, teams with a significantly lower number of switches in offensive play types than that expected in random sequences are not at a disadvantage because of their increased predictability. In fact, they have a borderline advantage in the point differential and a significant advantage in rushing yard and first downs from rushing differentials (DRY and DFDR resp.) when compared with the away teams in other groups.

The fact that play sequences with high numbers of switches between play types lead to poor outcomes for the away team indicates that it is a factor that plays some role in home advantage. The nature of that role may be cause and effect, or it may be part of a more complex dynamical system that leads to the disadvantage for the away team. However, since the sequence of offensive plays is a factor which can be controlled to a large extent by the offensive team, it is an association worthy of attention. Overall, the study seems to indicate that it is in the interest of the away team to make sure that the number of times they switch between offensive play types is not significantly higher than that expected in a random sequence. The results on the point differential and key game statistics suggest that it may well be advantageous to keep switched between play types at a

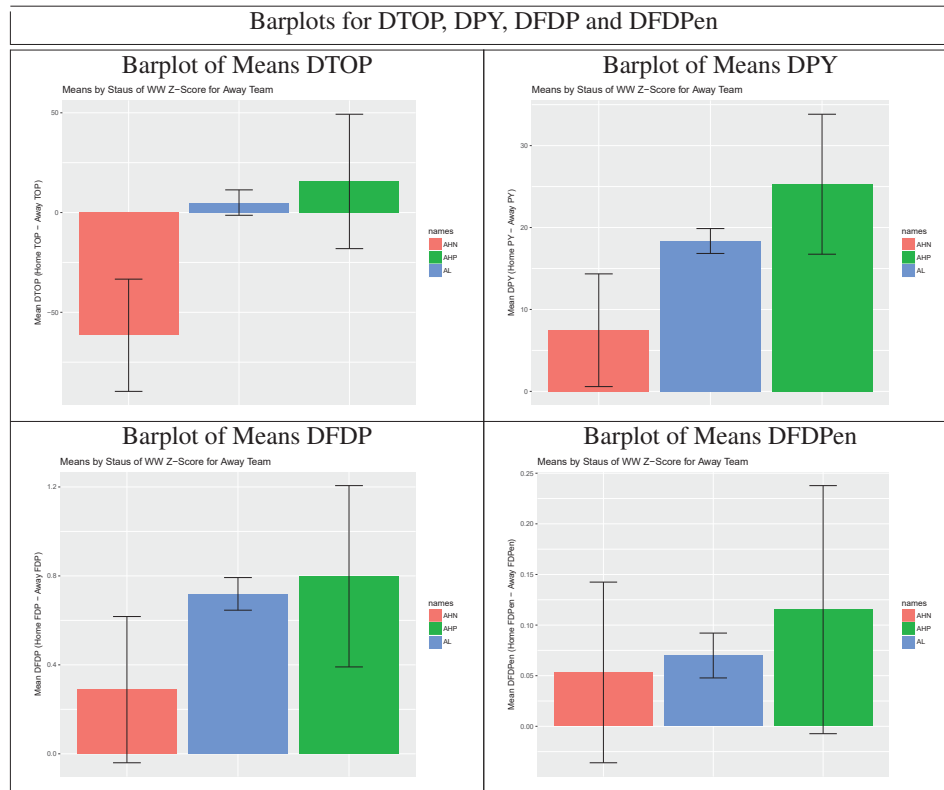


Figure 2: Means of variables DTOP, DPY, DFDP and DFDPen for all three levels of WW Away Team.

minimum. Both the home and away teams should note that teams with increased predictability resulting from a reduction in the number of switches between offensive play types to levels below those expected in randomly generated sequences are no less likely to win than their less predictable counterparts.

Further study plans for this project include examining the effects studied above when the sample is restricted to the top 50 teams (selected using an average of a number of reputable rankings) from each season. Clearly considering the interaction between offensive and defensive strategies would improve the study greatly, however no defensive play-by-play data is publicly available.

References

- [1] C. Barry Pfitzner, Steven D. Lang and Tracy D. Rishel (2014) *Factors Affecting Scoring in NFL Games and Beating the Over/Under Line* SJ Thesportjournal.org
- [2] W. Burton and M. Dickey. (2015) *NFL play predictions*. JSM Proceedings, Statistical Computing Section.

- [3] Wagner V, G. Oliver *College and Professional Football Scores: A Multiple Regression Analysis*. The American Economist, 1987, Vol.31(1), pp.33-37
- [4] Wald, A. and Wolfowitz, J. *On a test whether two samples are from the same population* (1940). Ann. Math. Statist. 11, 147-162.
- [5] Schilling, Mark F. *The surprising predictability of long runs* Mathematics Magazine, 2012, **2**, 141-149.
- [6] Alan Agresti *Categorical Data Analysis- 3rd ed.* Wiley series in probability and statistics, Wiley & Sons, 2013.
- [7] T. Gilovich, R. Vallone, and A. Tversky *The hot hand in basketball: on the misinterpretation of random sequences* Cognitive Psychology, 1985, Vol. 17, # 3, 295-314

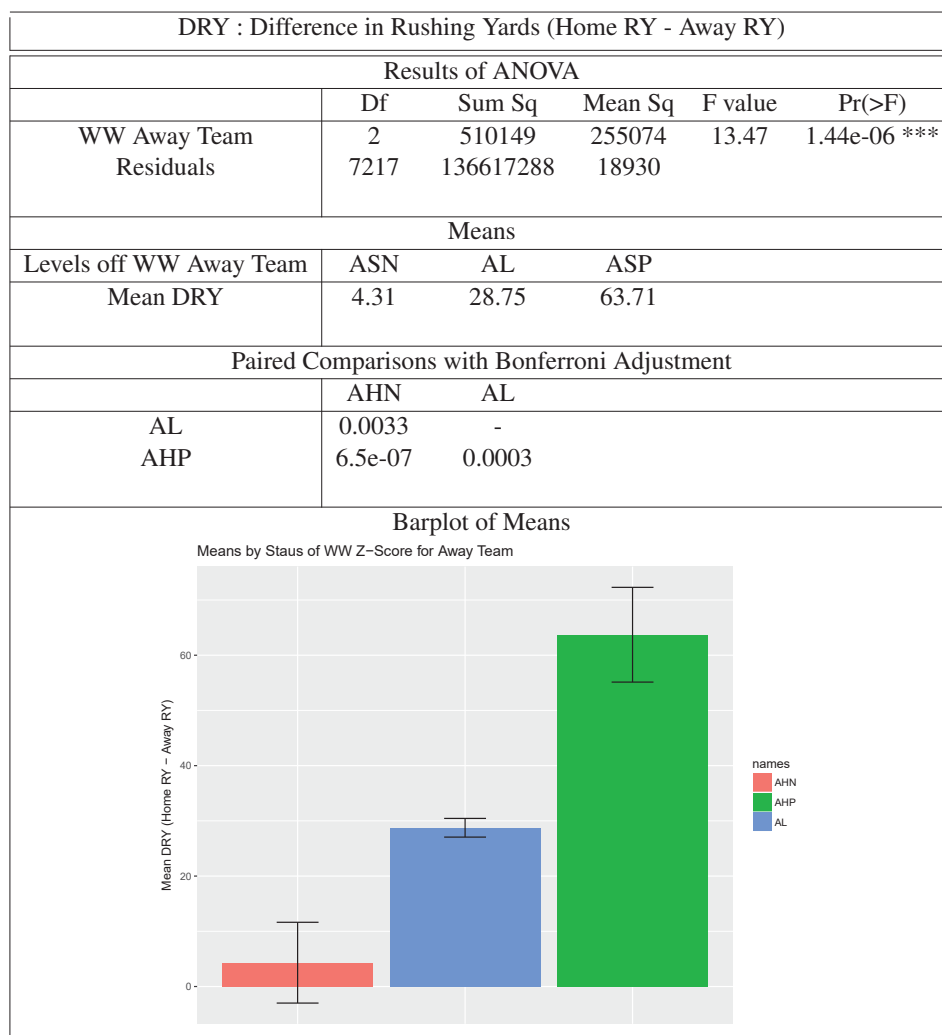


Table 5: Statistical Analysis for DRY for three levels of WW Away Team.

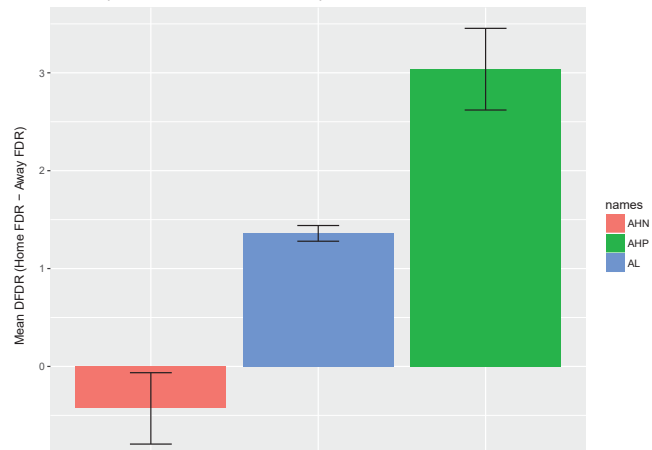
DFDR : Difference in First Downs from Rushing (Home FDR - Away FDR)					
Results of ANOVA					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
WW Away Team	2	1818	908.9	21.22	6.47e-10 ***
Residuals	7217	309112	42.8		
Means					
Levels off WW Away Team	ASN	AL	ASP		
Mean DFDR	-0.43	1.36	3.04		
Paired Comparisons with Bonferroni Adjustment					
	AHN	AL			
AL	1.5e-06	-			
AHP	6.1e-10	0.00026			
Barplot of Means					
<p>Means by Staus of WW Z-Score for Away Team</p>  <p>Mean DFDR (Home FDR - Away FDR)</p> <p>names</p> <ul style="list-style-type: none"> AHN AHP AL 					

Table 6: Statistical Analysis for DFDR for three levels of WW Away Team.

Complex 1 in Male Volleyball as a Markov Chain.

Sotirios Drikos, PhD.

National & Kapodistrian University of Athens, School of Physical Education and Sport
Science

sodrikos@phed.uoa.gr

Abstract

In Volleyball, complex 1 consists of the serve's pass (reception) - setting - attack skills in this specified order. This sequence is a stable pattern to win a point. Furthermore, it is important for the teams' success. Taking into account that this pattern is a first-order Markov chain, the creation of a probability transition matrix is feasible. Assuming multinomial likelihood with a Dirichlet prior on the transition probabilities a Markovian transition matrix can be constructed and the calculation of conditional success probabilities is, thus, achievable. Data from the performance analysis of the winning team from recent world championships in three age categories (U19, U21, Men) of male Volleyball is used. The findings lead to redefining target pass area and to shrinking the evaluation scale at least for the teams under study. Moreover, pass accuracy is necessary because it must give at least two options for attack, but not sufficient condition for the success of attack in all age categories for male Volleyball. In the U19 age category, there is a lack of stabilization in the complex 1 sequence after pass against jump spin serve.

1 Introduction

Volleyball consists of 3 stable patterns to win a point: pass-setting-attack after pass- outcome serve-outcome and block-dig- setting- attack after dig or counterattack-outcome (Florence, Fellingham, Vehrs, & Mortensen, 2008). For each pattern three are the possible outcomes: win a point, continuation of the action and a point for the opponent. In rally point system the pattern pass-setting-attack after the pass is the necessary condition to claim the victory because in terms of probability winning a point when receiving is easier than winning a point when serving in male volleyball (Calhoun, Dargahi-Noubary, & Shi, 2002; Ferrante & Fonseca, 2014).

Winning teams were significantly better in attack after pass than losing teams (Hayrinen, Hoivala, & Blomqvist, 2004) and attack after pass emerged as a decisive factor for team's success (Patsiaouras, Charitonidis, Moustakidis, & Kokaridas, 2009). It is crucial for a team to organise a tactically well structured and highly synchronised offensive game after receiving opponents serve. It is the hierarchical order of skills in Volleyball that makes the performance in one skill depends on the performance in the previous one. The precise pass is a powerful aggressive tool for high-level teams and is a good predictor for winning (Zetou, Moustakidis, Tsigilis, & Komninakidou, 2007). For many coaches receiving well is a guarantee for a winning attack. The connection between the quality of pass and achievement in attack is undoubted for men age category in many types of research. A partial rejection of this belief is suggested by Lobietti, Michele, & Merni (2006) who proposed that passing accuracy does not appear so fundamental but it is important avoiding passing errors.

The assumption that pass-setting-attack after pass pattern is a first-order Markov chain allows the recording of these sequences in a transition probabilities matrix where data of the matrix represent the probability to move from one state to another and, finally, to an outcome. With the use of the Bayesian analysis, the past team's performance or the coaches' opinions about passing effects in the attack can be taken into consideration as a prior distribution in order to create with actual data the posterior distribution and, consequently, the conditional success probability.

Thus, the aim of this study is to determine the influence of each level of a pass to the success of attack in 3 different age categories (U19, U21, Men) for male high-level Volleyball.

2 Method

All recorded data refer to the performance in pass and attack after the pass of the winning team of the world championship for national teams in three age categories for male volleyball. All data record the performance on selected matches of the World national team champions (Poland in Men, 2014; Russia in U21 and in U19 for 2013). Thus the initial sample (N=) was 815 for Men, 525 for U21 and 407 for U19 passes respectively. For the evaluation of pass, a 6-level ordinal scale was used with the 1st level being a passing error and the 6th level to be a pass performed in an optimal way. In Table 1 the performance ratings and a brief description of each passing level are presented. Attack was evaluated with three possible outcomes: point for the team under observation, rally continuation and point for the opponent.

Table 1. Performance ratings for a pass (vs Jump Spin & Jump Float Serve)

Level (Symbol)	code	Level brief description
6(#)		The ball was passed accurately with suitable height, speed and parabolic trajectory in the target area (3m-4m from the right sideline and about 30-50 cm from the net or over 30-50 cm over the net if setter has the ability to jump setting). The setter could have all the options (location & type) for a set from the sidelines and the central lane without any adjustments in his approach to the ball.
5(+)		The ball was passed either away (1m. behind or 2m. in front of the target area), or travelled higher, or lower (setter's shoulder level). The setter could have all the options for attack (location & type) from the sidelines and the central lane with adjustments in his approach to the ball.
4(!)		The ball was passed with either 3m away from the net or near the sidelines or to the top of the net. The setter could have two options for attack only from the sidelines.
3(-)		The ball was passed with very poor parabolic trajectory or near the sidelines, end line or outside of the court. The setter could have just one mandatory option for attack or the setter could not approach the ball and another player sets the ball mandatory.
2(/)		The ball was passed directly to the serving team court. No option for attack for the receiving team.
1(=)		The ball hit the floor directly or after touched by a receiver. The rally was ended after 1st or 2nd contact.

The observer was a volleyball coach, expert in evaluation and recording of volleyball performance data and excellent user of the software. The interobserver reliability of the data collection and recording was checked by a test-retest procedure, with a one-week interval, from a random sample of 100 actions of stable pattern pass-set attack after pass-outcome for each one of the teams under observation. As the acceptable value of Adjusted Cohen's Kappa was set 0.80 (Altman, 1991). The interobserver reliability in evaluation and recording of data was well established because of acceptable Adjusted Cohen's Kappa values calculated after the test-retest procedure. The values were 0.91 and 0.90 for a pass against jump spin and jump float serve respectively.

Every time the opponent serves the ball on the side of the observed team a sequence of events takes place that follows a specific scheme: pass-set-attack after pass-outcome. An assumption that this scheme is a first-order Markov chain is stated. This sequence was recorded in a transition probability matrix where data of the matrix represent the probability to move from one state to another and finally to reach an outcome. In this way three (one for each team) transition probabilities matrices were created.

A simple Bayesian model $P(Y_{t+1} = S_k | Y_t = S_i)$ to estimate the transition probabilities, and through them, the success probabilities were made. A multinomial likelihood for each row (i.e. level of the pass)

$$f(y_{i1}, \dots, y_{i,n}, y_{i,n+1}, y_{i,n+2} | \pi_{i1}, \dots, \pi_{i,n}, \pi_{i,n+1}, \pi_{i,n+2}) \propto \prod_{k \in M_i} \pi_{ik}^{y_{ik}}$$

with $\sum_{k=1}^{n+3} \pi_{ik} = \sum_{k \in M_i} \pi_{ik} = 1$ for each i ; where M_i is the set of indexes corresponding to possible following

skill S_i , was assumed. Given that the interest was in what the data suggest on the relationship between the different states of the sequence, a minimally informative prior distribution is assumed. A conjugate Dirichlet prior distribution was used where each row of the prior parameters were all assumed to be equal to one (except those that were constrained to be zero). All conditional probabilities scores were calculated using a simple Monte Carlo scheme of 10,000 iterations to acquire a random sample from the posterior distribution. For a detailed description of the model see Drikos, Ntzoufras, & Apostolidis (2019).

3 Results

The posterior means of conditional probabilities for each no terminal level of the evaluation scale for jump spin and jump float serve are presented in Table 2. Level 1 of pass scale is a terminal level and its probability to move to another state or to reach a positive outcome is zero. For level 2 of the pass, there is a noticeable finding. After overpass against jump spin serve the receiving team keeps a sufficiently higher probability (0.45) to win a point than to keep the ball in its court and have a mandatory attack (level 3). As expected, the pass in level 4, 5, and 6 of the scales have higher conditional probabilities. An important increase of probability to win a point is obvious when the pass is evaluated as level 4 (two options from sidelines) contrary to evaluation as level 3 (one mandatory option for the setter). This increase is 0.21, 0.16, 0.28 against jump spin serve and 0.19, 0.19, 0.16 against jump float serve for Men, U21, and U19 respectively. For U19 against jump serve the probability to win a point with pass level 4 is higher than with more precise passes (levels 5&6). Comparing success probabilities between levels 5 & 6 it is clear that more precise pass (level 6) does not mean higher success probabilities. Taking into consideration the standard deviation of each posterior mean, it is clear that success probabilities of a pass in levels 5 & 6 are almost equal for each age category. Also in Table 2, the tail posterior probability level of differences across age categories for each level of pass evaluation scale is presented. It is remarkable that the U19 team has a significantly higher probability of taking a point after a pass level 4 against both types of serve (offensive options only from sidelines) than U21 and Men team. Also, the U19 team has a higher probability to gain a point after an overpass against jump spin serve than both U21 and Men. Meanwhile, the U19 team has a higher probability of winning a point compared to U21 when the pass from a float serve is accurate on the net (level 6).

Table 2. Posterior means (\pm sd) of conditional probabilities and summary of posterior differences across age categories for each no terminal level of pass evaluation scale

Skills (S_i)	Skills (sub)	Men	U21	U19	Posterior differences *
Pass in Jump	2(/)	0.274 (± 0.058)	0.266 (± 0.053)	0.454 (± 0.124)	Men, U21 < U19
	3(-)	0.308 (± 0.038)	0.337 (± 0.055)	0.307 (± 0.090)	
	4(!)	0.548 (± 0.022)	0.515 (± 0.033)	0.631 (± 0.045)	Men < U19, U21 < U19
	5(+)	0.593 (± 0.022)	0.548 (± 0.029)	0.605 (± 0.045)	
	6(#)	0.589 (± 0.0212)	0.545 (± 0.032)	0.565 (± 0.048)	
Pass in Jump	2(/)	0.256 (± 0.046)	0.188 (± 0.069)	0.281 (± 0.049)	
	3(-)	0.325 (± 0.039)	0.304 (± 0.052)	0.412 (± 0.079)	
	4(!)	0.539 (± 0.024)	0.513 (± 0.031)	0.603 (± 0.035)	Men < U19, U21 < U19
	5(+)	0.581 (± 0.022)	0.563 (± 0.027)	0.616 (± 0.031)	
	6(#)	0.569 (± 0.022)	0.558 (± 0.026)	0.629 (± 0.030)	Men < U19, U21 < U19

* Inequalities indicate important differences between age categories: Age category A has lower success rates than age category B with posterior probability less than 0.01 ("A < < B"), between 0.01 and 0.05 ("A < B"), between 0.05 and 0.10 ("A < B").

A detailed preview of success conditional probabilities are provided in Figures 1&2.

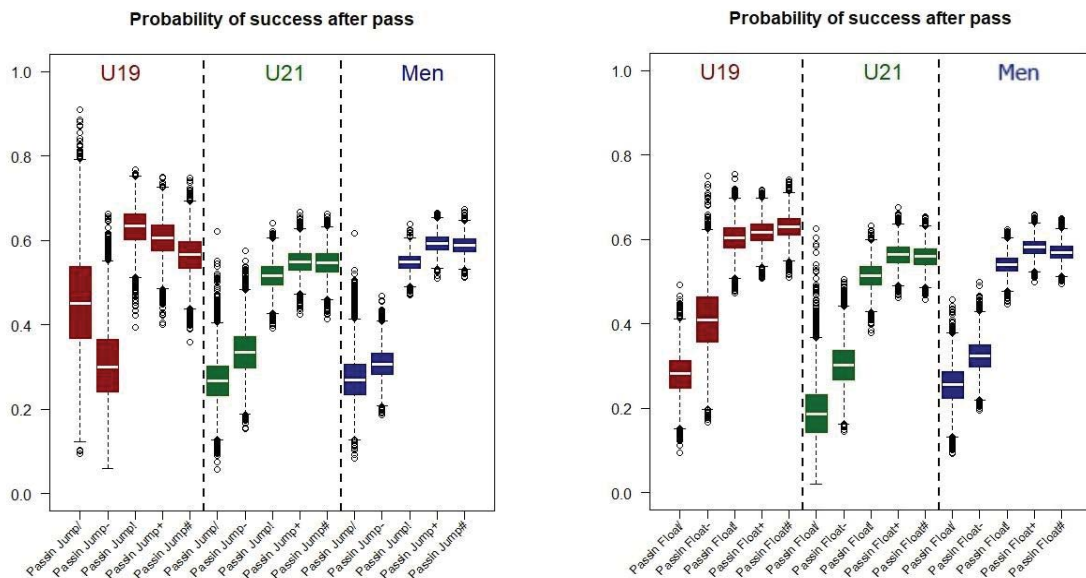


Figure 1 and 2.Box-plots (with outliers) of success conditional probability of each no terminal level of evaluation scale for a pass in Jump spin serve and for a pass in Jump float serve.

4 Discussion

The target for the receiver is an area close to the net or sometimes over it (3m-4m from the right sideline and about 30-50 cm from the net or over 30-50 cm over the net if the setter has the ability for jump setting). The pass that is directed to the court of the serving team (2nd level, that is to say, overpass) and the pass with the one-option setting (3rd level of the evaluation scale) have the same characteristics at all ages, with an exception of U19 only for a pass against jump spin serve. The penalty for the overpass is higher compared to this for a 3rd level pass. Also, the pass level 6 on the net or too close to the net does not present a higher probability compared to the 5th level. Silva, Lacerda, & Joao (2014) have mentioned the possible difficulty of the setter to handle a ball on the net. These findings follow the conclusions of Miskin et al. (2010) that, at least for the teams under consideration, the target area of a pass on the net must be more conservative.

In all age categories, the probability of winning a point in the stable pattern pass-set-attack after the pass is above 0.5 when the pass is evaluated on levels 4, 5, 6 of the evaluation scale. Thus, the first priority for a team should be to keep the ball in its court giving the setter the opportunity to choose at least between two attackers from the sidelines (outside hitter and opposite). The coaches' belief that a good pass is a guarantee for an effective attack can be more specified by pointing out that a pass which secures at least two attacking options increase the probability of a successful attack for all age categories in male Volleyball. This is in partial agreement with many studies about the relationship between pass and attack. The lack of discrimination between the 5th and 6th level of evaluation scale according to success probabilities ensures the finding of Lobiatti et al. (2006) that passing with high accuracy is not a necessary condition for a successful attack. Also, this means that, at least for the teams under examination, the passing rating system has to be changed. A possible junction of 5th and 6th level should be examined.

The large discrepancy of success probability from 1st, 2nd and 3rd in relation to 4th, 5th, 6th level of the pass evaluation scale is a clear message that the probability of success is not increasable as the evaluation grade gets higher. This phenomenon is observed in all age categories. There is no a fixed interval between

levels of the scale, thus the assumption of treating ordinal data as numerical data and the use of descriptive statistics, such as mean and standard deviation, for the evaluation of teams' or players' performance may be groundless. The same has been concluded by Florence et al. (2008) after examination of a college women's volleyball team.

It is difficult to explain the finding that the U19 team has higher probabilities after an overpass against jump spin serve instead of keeping the ball in its court with only one option for attack. It is clear that this analysis is applicable only to these teams, their level and their opponents and generalisations may be not applicable to other teams. In the model, only the next two touches of the team under observation were recorded, so it is highly likely that a point after an overpass is due to opponents' error. But even with this assumption, it is important to mention that the jump spin serve has a higher speed than jump float serve and the reaction time for receivers is reduced in <0.5s (Pena, Busca, Galceran, & Bauza, 2013).

Consequently, the reaction time is also limited to the serving team too, especially if they are not well prepared to play an opponents' overpass as a free ball.

Team U19 after pass level 4 against jump spin serve is more effective than Men & U21 teams. Also, it is noteworthy that there is not increased the probability to win a point when passing performance rises above level 4, contrary to Men and U21 teams. Performance of U19 team in pass-set-attack after pass pattern confirms the findings of Costa G. C. et al. (2011) that subsequent actions do not have high functional dependence in relation to the precedent ones in the age category of U19 due to the fact that because of lack of players' maturity the game is not well integrated.

To sum up, the present study is validating the six-level scale for evaluation of pass, it is developing a Bayesian model including prior distribution and is applying this model to performance data of world champion teams in three age categories. The conclusion reached is that for all ages the quality of pass is important to ensure at least two offensive options for the setter. Furthermore, the discrepancy of success probabilities among the levels of the scale makes it clear that for this ordinal scale it is unrealistic to use descriptive statistics, like a mean and standard deviation. Finally, the target area of the pass must be more conservative and the evaluation scale must be shrunk, at least for teams under observation.

References

- Altman, D. G. (1991). *Practical Statistics for Medical Research*. London: Chapman & Hall.
- Calhoun, W., Dargahi-Noubary, G. R., & Shi, Y. (2002). Volleyball Scoring Systems. *Mathematics and Computer Education*, 36 (1), pp. 70-79.
- Costa, G. d., Alfonso, J., Barbosa, R. V., Coutinho, P., & Mesquita, I. (2014). Predictors of attack efficacy and attack type in high-level Brazilian women's volleyball. *Kinesiology*, 46 (2), pp. 242-248.
- Drikos, S., Ntzoufras, I., & Apostolidis, N. (2019, June). Bayesian analysis of skill importance in world champions men's volleyball across ages. *International Journal of Computer Science in Sport*.
- Ferrante, M., & Fonseca, G. (2014, June). On the winning probabilities and mean duration of Volleyball. *Journal of Quantitative Analysis in Sports*, 10 (2), pp. 91-98.
- Florence, L., Fellingham, G., Vehrs, P., & Mortensen, N. (2008). Skill Evaluation in Women's Volleyball. *Journal of Quantitative Analysis in Sports*, 4 (2).
- Hayrinen, M., Hoivala, T., & Blomqvist, M. (2004). Differences between winning and losing teams in men's European top-level volleyball. In P. O'Donoghue, & M. Hughes (Ed.), *Performance Analysis of Sport* (pp. 194-199). Cardiff: Center for Performance Analysis, School of Sport, Physical Education and Recreation, University of Wales.
- Lobietti, R., Michele, R., & Merni, F. (2006). Relationships between performance parameters and final ranking in professional Volleyball. In H. Dancs, M. Hughes, & J. Ekler (Ed.), *World Congress of Performance Analysis in Sports 7*. Szombathely: Berzenyi College.
- Miskin, M., Fellingham, G., & Florence, L. (2010). Skill Importance in Women's Volleyball. *Journal of Quantitative Analysis in Sports*, 6 (2).
- Patsiaouras, A., Charitonidis, K., Moustakidis, A., & Kokaridas, D. (2009). Comparison of technical skills effectiveness of Men's National Volleyball teams. *International Journal of Performance Analysis in Sport* (9), pp. 1-7.

- Pena, J., Busca, B., Galceran, D.-M., & Bauza, J. (2013). The effect of aerodynamic drag in the service speed of high-level men's volleyball. *Proceedings of the 18th Congress European College of Sport Sciences*. Barcelona, Spain.
- Silva, M., Lacerda, D., & Joao, P. V. (2014, August). Match analysis of discrimination skills according to the setter defence zone position in high-level Volleyball. *International Journal of Performance Analysis in Sport*, 14 (2), pp. 463-472.
- Zetou, E., Moustakidis, A., Tsigilis, N., & Kominakidou, A. (2007). Does effectiveness of skill Complex 1 predict win in Men's Olympic Volleyball Games? *Journal of Quantitative Analysis in Sports*, 3 (4).

Modelling volleyball data using a Bayesian approach

Leonardo Egidi¹ and Ioannis Ntzoufras²

¹ University of Trieste, Trieste, Italy
`legidi@units.it`

² Athens University of Economics and Business, Athens, Greece
`ntzoufras@aueb.gr`

Abstract

Unlike what happens for other major sports such as football, basketball and baseball, modeling volleyball match outcomes has not been thoroughly addressed by statisticians and mathematicians. The main reason could be the game complexity: the total number of sets is a random variable which ranges from a minimum of three to a maximum of five; the number of points achieved by the two competing teams in each set varies depending on whether they are playing the fifth set or not; the number of final set points for two competing teams may exceed 25 when both the teams reach 24 points (24-deuce). We propose a Bayesian negative binomial model for the points achieved by the team losing the single set, modelling the probability to realize a point via some team-specific point abilities; the probability of winning a set depends on team specific set abilities. Both point and set abilities are assigned some weakly informative prior distributions. We used goodness of fit tools to compare our proposal with other competing models on Italian Superlega 2017-2018, and MCMC replications from the predictive distribution as a simulation device to reconstruct the league. Preliminary results show that our model outperforms Poisson and binomial models in terms of DIC and is adequate in replicating the final ranking of the league.

1 Introduction

Statistical modelling for sports outcomes is a trend topic and the community of scholars involved to this task is still growing. Unlike what happens for other major sports such as football [4], basketball and baseball [5], modeling volleyball match outcomes has not been thoroughly addressed by statisticians and mathematicians: early attempts date back to [1] and [2]. The goals and the points realized over football and basketball matches are cumulative from the beginning to the end of the game: in such situations a model for the total scores is required. The complexity of volleyball in terms of final scores may be essentially summarized by three arguments: the total number of sets of a volleyball match is a random variable which ranges from a minimum of three to a maximum of five; moreover, the number of points achieved by the two competing teams in each set varies depending on whether they are playing the fifth set or not; finally, the number of final set points for two competing teams may exceed 25 when both the teams reach 24 points (24-deuce). Volleyball outcomes consist of a natural hierarchy of points within sets, and both the measurements are random variables.

In our perspective, the task of modelling volleyball match results should follow a top-down strategy, from the sets to the single points. Thus, defining the probability of winning a set is the first step; building up a generative discrete model for the points realized in each set is the second step. Although following this order is not mandatory, we maintain with the idea to replicate the hierarchy of the game into our models. In this paper we propose a set-by-set negative binomial model for the points achieved by the team losing the single set: the distribution of the points is then conditional to the set result. Another aspect to consider is the strengths' difference

among the teams: weaker teams are of course not favoured when competing against stronger teams, and a parametric assumptions about teams' skills is needed. In the Bayesian approach teams' abilities are easily incorporated into the model by use of some weakly-informative prior distributions [3]: similarly to what happens for football models [4], the abilities may regard both attack and defense skills, and, moreover, be considered as dynamic over the season [6].

The rest of the paper is organized as follows. In Section 2 we introduce some discrete models for volleyball outcomes, such as Poisson and binomial. Model extensions are thoroughly presented in Section 3, whereas model estimation and goodness of fit tools are detailed in Section 4. MCMC replications for the negative binomial model, the final selected model, are used in Section 5 to assess its plausibility in comparison with the observed results and to reconstruct the final rank of the league. Section 6 concludes.

2 Models

Let Y_{1g} and Y_{2g} are the points for each set $g = 1, 2, \dots, G$ collected by the two competing teams, and W_g is the binary indicator for the win of the home team. Then we can calculate Y_g , the number of points for the team loosing the g -th set using the equation $Y_g = W_g Y_{2g} + (1 - W_g) Y_{1g}$. We describe three different models to address the problem.

2.1 Truncated negative binomial model

We denote by $\text{NegBin}(r, p_g)I(Z < c)$ the right truncated Negative Binomial distribution, where: p_g ($1 - p_g$) denotes the probability of realizing a point for the team winning (loosing) the set g ; $r = 25$ because each set is played until the winning team achieves 25 points; the right truncation has been fixed at $c = 23$ points since this is the highest number of points that can be achieved by the loosing team (under the assumption of no ties).

The model for the total points realized by the team loosing the g -th set, for $g = 1, \dots, G$, is specified as a mixture as follows:

$$\begin{aligned} Y_g &= W_g Y_{2g} + (1 - W_g) Y_{1g} \\ Y_g &\sim \text{NegBin}(25, p_g)I(Y_g \leq 23) \\ W_g &\sim \text{Bernoulli}(\omega_g), \end{aligned} \tag{1}$$

where the number of points for the team loosing the set are realized by team B (A) if team A (B) wins the set, i.e. $W_g = 1$ ($W_g = 0$). Both set and point probabilities ω_g and $1 - p_g$ depend on some team specific abilities. Team A wins the set with probability ω_g :

$$\text{logit}(\omega_g) = H_s + \alpha_{A(g)} - \alpha_{B(g)}, \tag{2}$$

where H_s is the set home advantage for the hosting team, and $\alpha_{A(g)}, \alpha_{B(g)}$ the set teams abilities for teams $A(g)$ and $B(g)$, respectively. The logit probability of realizing a point when loosing the set is defined as:

$$\log \frac{1 - p_g}{p_g} = \mu + (1 - W_g)H_p + (\beta_{A(g)} - \beta_{B(g)})(1 - 2W_g) \tag{3}$$

where μ is a common baseline parameter, H_p is the point home advantage for the hosting team, $\beta_{A(g)}, \beta_{B(g)}$ are the point teams abilities for teams $A(g)$ and $B(g)$, respectively. The sampling distribution in (1) is an upper truncated negative binomial, with upper truncation at $c = 23$.

Let us focus for a moment on the untruncated negative binomial, for which the average number of points for team A (evaluated if $W_g = 0$) and team B (evaluated if $W_g = 1$) in the g -th set are, respectively:

$$\begin{aligned} E[Y_{1g}] &\equiv \mu_{A,g} = 25 \frac{1-p_g}{p_g} = 25 \exp \{ \mu + H_p + \beta_{A(g)} - \beta_{B(g)} \} \\ E[Y_{2g}] &\equiv \mu_{B,g} = 25 \frac{1-p_g}{p_g} = 25 \exp \{ \mu - \beta_{A(g)} + \beta_{B(g)} \}. \end{aligned} \quad (4)$$

Consider the first equation: the larger is the difference $\beta_{A(g)} - \beta_{B(g)}$, and the higher is the expected number of points team A will win when loosing a set. Equivalently, the lower will be the number of points team B is winning when loosing a set.

However, in our model the loosing-set team can reach at most 23 points (in case of no extra points), then we need to reconsider the expected number of points of the loosing team (i.e. equations (4)) in light of the upper truncation at $c = 23$. From [7] we know that if $X \sim \text{NegBin}(r, p)I(X \leq c)$, then its mean value is $E(X|X \leq c) = \mu - \frac{1}{p} \frac{(c+1)f(c+1)}{P(X \leq c)}$, where the untruncated mean is $\mu = r(1-p)/p$, whereas $f(c+1)$ and $P(X \leq c)$ represent the probability mass function and the cdf of the untruncated negative binomial distribution, $\text{NegBin}(p, r)$, evaluated in c , respectively. Then, the truncated means are:

$$\begin{aligned} E[Y_{1g}|Y_{1g} \leq 23] &= \mu_{A,g} - \frac{24f(24)}{p_g P(Y_{A,g} \leq 23)} = \\ &= 25 \exp \{ \mu + H_p + \beta_{A(g)} - \beta_{B(g)} \} - c^* \\ E[Y_{2g}|Y_{2g} \leq 23] &= \mu_{B,g} - \frac{24f(24)}{p_g P(Y_{A,g} \leq 23)} = \\ &= 25 \exp \{ \mu - \beta_{A(g)} + \beta_{B(g)} \} - c^*, \end{aligned} \quad (5)$$

where $c^* = \frac{24f(24)}{p_g P(Y_{A,g} \leq 23)}$, $c^* > 0$. The interpretation is identical wrt the untruncated case: the higher is the set bility of a team, the higher will be the number of points when loosing a set. However, the untruncated mean is subtracted by the positive factor c^* , which forces the loosing team points to be lower or equal than 23 (see Figure 1 for a graphical comparison between the untruncated and the truncated negative binomial). In Section 3 we will extend the model to allow for eventual extra points after 25.

2.2 Binomial model

We denote by $\text{Bin}(n, 1-p_g)$ the binomial distribution, where $1-p_g$ denotes the loosing team probability of realizing a point in the g -th set and n is the total number of points $Y_{1g} + Y_{2g}$ realized in the same set. The model for the points Y_g realized by the team loosing the g -th set is the same as in (1), but the likelihood is binomial:

$$Y_g \sim \text{Bin}(n, 1-p_g) \quad (6)$$

As for the negative binomial case, equations (2) and (3) model the team A probability ω_g of winning a set and the loosing-set team probability of realizing a point, respectively.

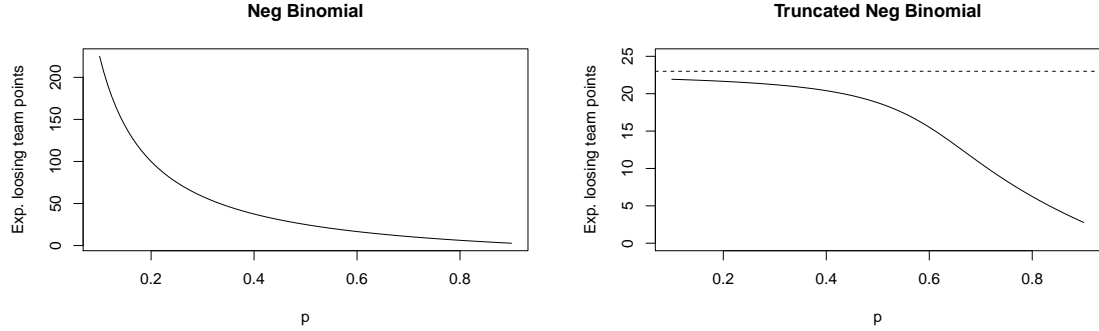


Figure 1: Expected number of points for a team losing a set according to the negative binomial (left plot) and the truncated negative binomial with upper truncation at $c = 23$ (right plot). As the point probability for the team winning the set increases, the expected number of points decreases.

2.3 Poisson model

We denote by $\text{Pois}(\lambda_g)$ the Poisson distribution, where the rate parameter λ_g represents the average number of points realized in the g -th set by the losing team. The model for Y_g is then:

$$Y_g \sim \text{Pois}(\gamma_g) \quad (7)$$

Team A probability ω_g of winning a set is defined by (2), whereas the logarithm of the average number of points realized by the losing-set team in the g -th set is modelled as:

$$\log \gamma_g = \mu + (1 - W_g)H_p + (\beta_{A(g)} - \beta_{B(g)})(1 - 2W_g). \quad (8)$$

3 Model extensions and further assumptions

3.1 Prior distributions and constraints

The Bayesian model is completed by assigning some weakly informative priors [3] to the set and point abilities, for each team $t = 1, \dots, \text{nteams}$:

$$\begin{aligned} \alpha_t, \beta_t &\sim \text{Normal}(0, 1) \\ \mu, H_p, H_s &\sim \text{Normal}(0, 10^3) \end{aligned} \quad (9)$$

In order to achieve identifiability, set and point abilities need to be constrained; in such a framework we impose a sum-to-zero constraint for both α and β .

3.2 Attacking and defensive abilities

In many sports, such as football, basketball, hockey, there is the need to separately model the abilities arising from attacking and those coming from the defence. A proper assumption for the point abilities could then be:

$$\begin{aligned}\beta_{A(g)} &= \text{att}_{A(g)} + \text{def}_{A(g)} \\ \beta_{B(g)} &= \text{att}_{B(g)} + \text{def}_{B(g)},\end{aligned}\tag{10}$$

where $\text{att}_{A(g)}, \text{att}_{B(g)}, \text{def}_{A(g)}, \text{def}_{B(g)}$ are the attack and the defence abilities for the two teams, respectively. Many extensions could be proposed: in the basic models of the previous sections, we assumed that $\text{att}_{A(g)} = \text{def}_{A(g)}, \text{att}_{B(g)} = \text{def}_{B(g)}$, with no need of distinguishing between attack and defence skills.

3.3 Zero inflated Poisson (ZIP) for the extra points

To allow for eventual extra points due to the 24-deuce, the three models in Section 2 may be extended specifying a zero-inflated Poisson (ZIP) model for the extra points collected by the loosing-set team. The number of extra points is zero if the loosing-set team does not reach 24 points, and greater than zero otherwise. The ZIP model is then defined for $g = 1, \dots, G$ as:

$$\begin{aligned}Y_g &= W_g Y_{2g} + (1 - W_g) Y_{1g} \\ Y_g &\sim O_g + W_g L_{2g} + (1 - W_g) L_{1g} \\ O_g &\sim \text{ZIPoisson}(p_{0g}, \lambda_g).\end{aligned}\tag{11}$$

The zero inflated Poisson (ZIP) distribution for the number of extra points O_g collected by the loosing-set team in the g -th set is then defined as:

$$f_{ZID}(o_g) = p_{0g} I(o_g = 0) + (1 - p_{0g}) f(o_g; \lambda_g),\tag{12}$$

where p_{0g} describes the proportion of extra zeros and $f(o_g; \lambda_g)$ is the probability mass function of a Poisson distribution with rate parameter λ_g . The probability to observe a zero should be strictly related to the abilities between the two competing teams, since the greater is their difference and the less likely should be the probability of a tie:

$$\begin{aligned}\text{logit}(p_{0g}) &= m + \delta(\alpha_{A(g)} - \alpha_{B(g)}) + \gamma(\beta_{A(g)} - \beta_{B(g)}) \\ \log(\lambda_g) &= \eta \\ \delta, \gamma &\sim \text{Normal}(0, 1); \quad \eta \sim \text{Normal}^+(0, 10^2)\end{aligned}\tag{13}$$

The log-linear model for λ_g is unstructured and η is assigned a weakly informative prior distribution, an half-normal distribution with location 0 and scale 10.

3.4 Dynamic abilities

Teams performance are likely to change over an entire season, and temporal trends may be helpful for modelling the ability of a given team in a given period. A dynamic structure assumption for the abilities is a step forward, a natural choice is an auto-regressive model for both the point and the set abilities. For each team t and match n , $n = 2, \dots, N$ we specify:

$$\begin{aligned}\alpha_{t,n} &\sim \text{Normal}(\alpha_{t,n-1}, \sigma_\alpha^2) \\ \beta_{t,n} &\sim \text{Normal}(\beta_{t,n-1}, \sigma_\beta^2),\end{aligned}\tag{14}$$

whereas for the first match we assume:

$$\begin{aligned}\alpha_{t,1} &\sim \text{Normal}(0, \sigma_\alpha^2) \\ \beta_{t,1} &\sim \text{Normal}(0, \sigma_\beta^2),\end{aligned}\tag{15}$$

As mentioned in Section 3.1, sum-to-zero constraints are required for each match-day to achieve identifiability. σ_α^2 and σ_β^2 are assigned two inverse Gamma priors with shape and rate parameters equal to 0.001.

3.5 Connecting the abilities

In equations (2) and (3) set and point abilities are separately modelled: conditionally on winning/loosing a set, point abilities are then estimated from the probability to realize a point. However, we could use them as jointly by defining a sort of global ability measure:

$$\begin{aligned}\log \frac{1-p_g}{p_g} = & \mu + (1 - W_g)[H_p + (\beta_{A(g)} - \beta_{B(g)}) + \gamma_1(\alpha_{A(g)} - \alpha_{B(g)})] + \\ & + W_g[(\beta_{B(g)} - \beta_{A(g)}) + \gamma_2(\alpha_{B(g)} - \alpha_{A(g)})],\end{aligned}\tag{16}$$

where parameters γ_1 and γ_2 summarize the effect of the set abilities. For illustration purposes only, just reason in terms of team A . If two teams are almost equally strong, then the set abilities difference $\alpha_{A(g)} - \alpha_{B(g)}$ is very small, and the point probability is entirely driven by the point abilities. Conversely, when two teams are expected to be quite far in terms of global performance, then the set abilities difference is expected to be high, and, consequently, the probability to realize a point is much affected by the parameter γ_1 .

4 Estimation

We used the `rjags` R package (MCMC sampling from the posterior distribution using the Gibbs sampling) to fit the models. Data come from the regular season of the Italian SuperLega 2017-2018 and consist of a seasonal sample of 680 set observations, for a total number of 182 matches and 14 teams.

In Table 1 we report the DIC for each model with the corresponding number of parameters: in this table we counted only the *primitive* parameters - for instance the set and point abilities α and β - and not the *transformed* parameters - such as the set and point probabilities w and p , specified in terms of some logit transformations involving the primitive parameters H_s, H_p, α, β . The ZIP truncated negative binomial model presented in Section 2.1 is the best fitted model. A dynamic structure seems not to improve over the fit, likewise considering the attack and the defense abilities as separately.

Posterior estimates for the set home advantage H_s , the point home advantage H_p , the grand intercept μ and the ZIP parameters η, m, δ, γ are reported in Table 2. There is a clear signal of home advantage both at the set and at the single point level; a small positive association is observed concerning the differences in terms of point (parameter δ) and set (parameter γ) abilities.

Set and point abilities for each team are displayed in Figure 2, in terms of posterior means \pm standard errors. The estimates are displayed following the final actual rank of the Italian

<i>Model distribution</i>	<i>Additional model details</i>	<i># param.</i>	<i>DIC</i>
1. Neg. binomial	$r = 25$	31	4779.3
2. Poisson	log-linear model for γ_g	31	4574.5
3. Binomial	$n_g \sim \text{Pois}(46)$	31	8031.6
4. ZIP Tr. Neg. bin.		35	4544.1
5. ZIP Tr. Poisson		35	4565.3
6. ZIP Tr. Binomial	$n_g \sim \text{Pois}(46)$	35	8408.7
7. ZIP Tr. Neg. bin.	att \neq def	49	—
8. ZIP Tr. Neg. bin.	α_t, β_t dynamic	737	4721.7

Table 1: Details of the fitted models: Italian SuperLega 2017-2018 season (MCMC sampling, 500 iterations).

	Mean	Median	sd	2.5%	97.5%
H_s	0.16	0.16	0.08	-0.00	0.31
H_p	0.20	0.19	0.07	0.07	0.34
μ	0.36	0.36	0.05	0.26	0.46
η	1.38	1.38	0.07	1.26	1.50
m	2.13	2.10	0.12	1.90	2.39
δ	-0.20	-0.20	0.76	-1.69	1.24
γ	0.09	0.09	0.18	-0.26	0.41

Table 2: ZIP truncated negative binomial model: Posterior estimates for the following parameters: the set home H_s , the point home H_p , the grand intercept μ ; η , m , γ , δ (ZIP part).

SuperLega 2017-2018: the global pattern mirrors almost perfectly the final rank. BCC Castelfranco, the worst team in the league, is associated with the lowest abilities, whereas Sir Safety Perugia, the league winner, registers the highest set and point abilities.

5 League reconstruction and predictive measures of fit

To assess the in-sample predictive accuracy of our model we try to reconstruct the league in terms of final points and rank positions. Table 3 reports the expected final points obtained from MCMC sampling along with the observed points and the final teams rank. The prediction is quite good for the majority of teams, only the positions of a few of them are switched, however the pattern is quite close to the observed one.

For each set g , we denote by d_g the set points difference $Y_{1g} - Y_{2g}$, and with $\tilde{d}_g^{(s)}$, $s = 1, \dots, S$ the corresponding points difference arising from the s -th MCMC replication, $\tilde{y}_{1g}^{(s)} - \tilde{y}_{2g}^{(s)}$. Once we replicate new existing values from our model, it is of interest to assess how far they are if compared with the actual data we observed. Figure 3 displays the predictive distribution of each $\tilde{d}_g^{(s)}$ (light blue) against the true observed distribution for d_g : the replicated distribution seem to perfectly mirror the observed distribution.

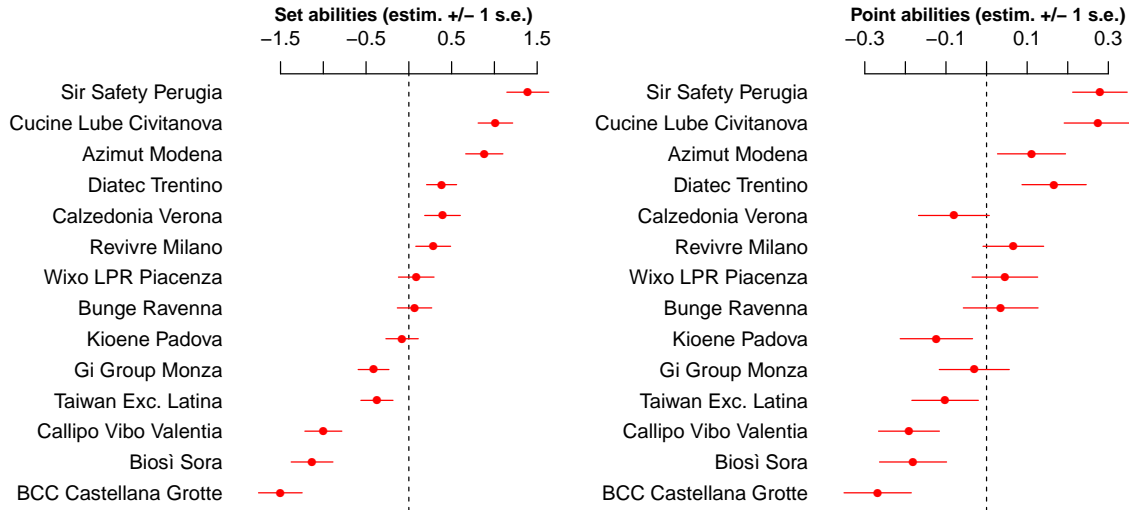


Figure 2: ZIP truncated negative binomial model: posterior means \pm s.e. for the set abilities α and the point abilities β , ordered by the actual final rank of the Italian SuperLega 2017-2018.

<i>Teams</i>	<i>Exp. Points</i>	<i>Actual points</i>	<i>Actual rank</i>
Sir Safety Perugia	70	70	1
Cucine Lube Civitanova	63	64	2
Azimut Modena	60	60	3
Diatec Trentino	52	51	4
Calzedonia Verona	50	50	5
Revivre Milano	44	44	6
Bunge Ravenna	41	41	8
Wixio LPR Piacenza	41	42	7
Kioene Padova	35	35	9
Gi Group Monza	28	28	10
Taiwan Exc. Latina	26	25	11
Callipo Vibo Valentia	13	13	12
Biosì Sora	13	13	13
BCC Castellana Grotte	10	10	14

Table 3: ZIP truncated negative binomial model: final league reconstruction from MCMC sampling along with the actual points and the actual final rank for each team.

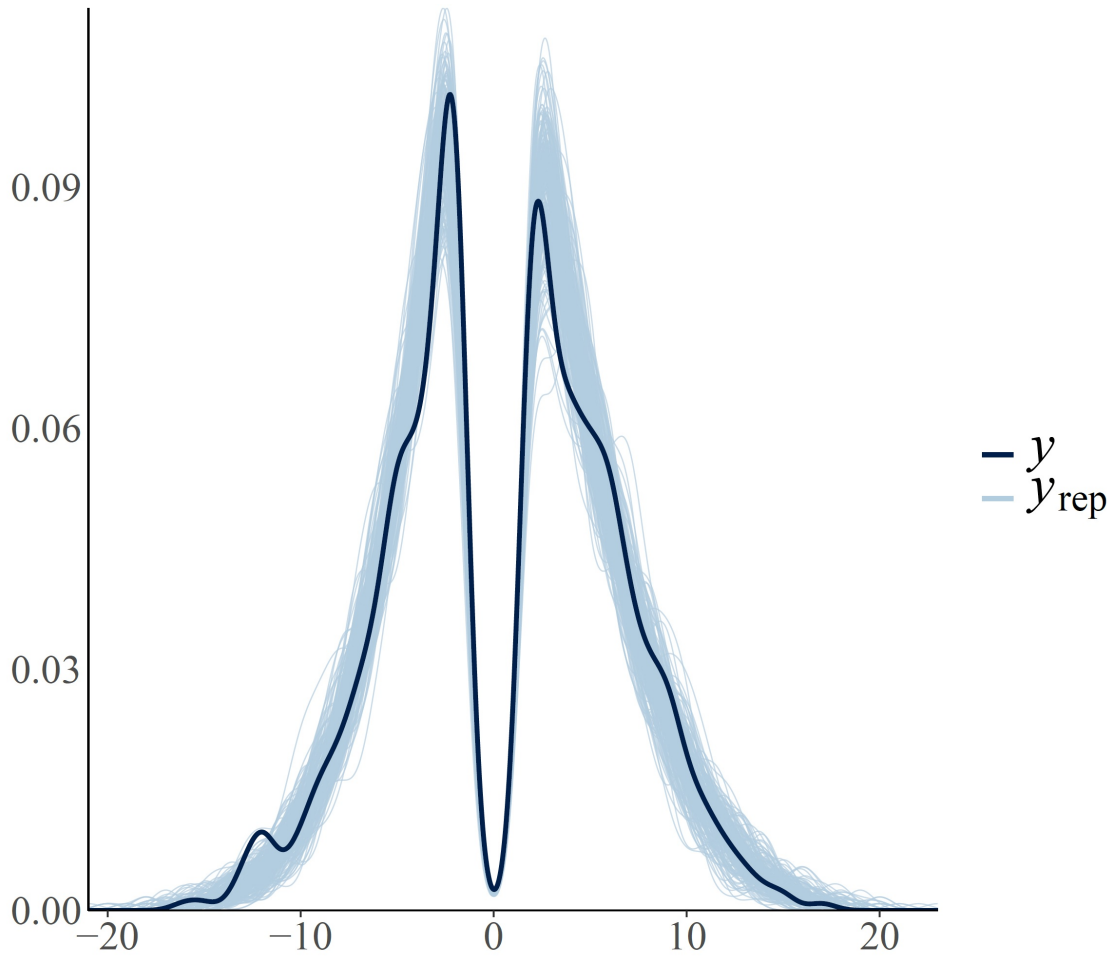


Figure 3: ZIP truncated negative binomial model: distribution of the observed set points differences $d_g = Y_{1g} - Y_{2g}$ (dark blue) against the MCMC simulated distribution $\tilde{d}_g^{(s)} = \tilde{y}_{1g}^{(s)} - \tilde{y}_{2g}^{(s)}$ (light blue).

6 Discussion

We end up to select a ZIP truncate negative binomial model for the volleyball match outcomes; with such a choice, we allow for eventual extra points after the 24-deuce. Preliminary results show a good plausibility of the model estimates in comparison of the observed results, and an overall good ability to replicate the final rank of the league.

Further work should be done to formulate an overall measure of goodness of fit, both at point and at set levels. Moreover, the inclusion of some game-covariates is of future interest.

References

- [1] Tristan Barnett, Alan Brown, Karl Jackson, et al. Modelling outcomes in volleyball. In *9th Australasian Conference on Mathematics and Computers in Sport (9M&CS)(Tweed Heads, Australia, 2008)*, pages 130–137, 2008.
- [2] Marco Ferrante and Giovanni Fonseca. On the winning probabilities and mean durations of volleyball. *Journal of Quantitative Analysis in Sports*, 10(2):91–98, 2014.
- [3] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, Yu-Sung Su, et al. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008.
- [4] Dimitris Karlis and Ioannis Ntzoufras. Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):381–393, 2003.
- [5] Gary Koop. Modelling the evolution of distributions: an application to major league baseball. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 167(4):639–655, 2004.
- [6] Alun Owen. Dynamic bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. *IMA Journal of Management Mathematics*, 22(2):99–113, 2011.
- [7] JS Shonkwiler. Variance of the truncated negative binomial distribution. *Journal of Econometrics*, 195(2):209–210, 2016.

Extreme value prediction: an application to sport records

Giovanni Fonseca¹ and Federica Giummolè²

¹ Dipartimento di Scienze Economiche e Statistiche,
Università di Udine
via Tomadini 30/A, I-33100 Udine, Italy.
`giovanni.fonseca@uniud.it`

² Dipartimento di Scienze Ambientali, Informatica e Statistica,
Università Ca' Foscari Venezia,
via Torino 155, I-30172 Venice, Italy.
`giummole@unive.it`

Abstract

Extreme value theory studies the extreme deviations from the central portion of a probability distribution. Results in this field have considerable importance in assessing the risk that characterises rare events, such as collapse of the stock market, or earthquakes of exceptional intensity, or floods. In the last years, application of extreme value theory for prediction of sport records have received increased interest by the scientific community. In this work we face the problem of constructing prediction limits for series of extreme values coming from sport data. We propose the use of a calibration procedure applied to the generalised extreme value distribution, in order to obtain a proper predictive distribution for future records. The calibrated procedure is applied to series of real data related to sport records. In particular, we consider sequences of annual maxima for different athletic events. Using the proposed calibrated predictive distribution, we show how to correctly predict the probability of future records and we discuss the existence and interpretation of ultimate records.

Keywords: athletic records, bootstrap, generalised extreme value distribution, prediction.

1 Introduction

From the very beginning, a big effort has been put into understanding the limits of human being capabilities: how fast can we run or swim? How far can we jump? In the last decades interest has mainly regarded the application of mathematical or statistical results in order to describe the progression of records in several sport events and in particular for track and field competitions.

Different approaches are used for assessing the probability of a new record or eventually the determination of an ultimate record, that is a measure that will not be overcome ever. [8] and [7] apply the theory of records to best annual performances. [10] propose a model for series of records, based on a random walk structure. In [11] a nonlinear regression model is introduced for fitting the progression of best annual results. Extreme value theory is applied in [9] to model the tail of the distribution for annual best records. [4] also takes advantage of the theory of extremes, enlarging the sample dimension by considering the personal best performance of as many athletes as possible over a period of several years.

In this work we apply the generalised extreme value (GEV) model to best annual results in the period from 2001 to 2018 for different athletic competitions. Depending on the data, the estimated model may comprise an end point that depends on the estimated parameters, or

not. We propose a bootstrap procedure that allows the computation of a calibrated predictive distribution for best annual performances. The proposed predictive distribution works well in regular cases, i.e. when the estimated model is unbounded, but can also be useful when the end point of the support depends on the estimated parameters. Being calibrated, it allows to compute the correct probability of improving a world record in regular cases and to assess the quality of the endpoint of the estimated model in non regular cases.

The paper is organised as follows. In Section 2 we briefly describe the bootstrap calibrating procedure and in Section 3 we define the family of GEV distributions. In Section 4 we apply the proposed predictive procedure to athletic records.

2 Calibrated distributions for prediction

In this section we briefly review the calibrating approach proposed by [5], that provides a predictive distribution function whose quantiles give prediction limits with well-calibrated coverage probability.

Suppose that $\{Y_i\}_{i \geq 1}$ is a sequence of continuous random variables with probability distribution specified by the unknown d -dimensional parameter $\theta \in \Theta \subseteq \mathbf{R}^d$, $d \geq 1$; $Y = (Y_1, \dots, Y_n)$, $n > 1$, is observable, while $Z = Y_{n+1}$ is a future or not yet available observation. For simplicity, we consider the case of Y and Z being independent random variables and we indicate with $G(z; \theta)$ the distribution function of Z .

Given the observed sample $y = (y_1, \dots, y_n)$, an α -prediction limit for Z is a function $c_\alpha(y)$ such that, exactly or approximately,

$$P_{Y,Z}\{Z \leq c_\alpha(Y); \theta\} = \alpha, \quad (1)$$

for every $\theta \in \Theta$ and for any fixed $\alpha \in (0, 1)$. The above probability is called coverage probability and it is calculated with respect to the joint distribution of (Z, Y) .

Consider the maximum likelihood estimator $\hat{\theta} = \hat{\theta}(Y)$ for θ , or an asymptotically equivalent alternative, and the estimative prediction limit $z_\alpha(\hat{\theta})$, which is obtained as the α -quantile of the estimative distribution function $G(\cdot; \hat{\theta})$. The associated coverage probability is

$$P_{Y,Z}\{Z \leq z_\alpha(\hat{\theta}(Y)); \theta\} = E_Y[G\{z_\alpha(\hat{\theta}(Y)); \theta\}; \theta] = C(\alpha, \theta) \quad (2)$$

and, although its explicit expression is rarely available, it is well-known that it does not match the target value α even if, asymptotically, $C(\alpha, \theta) = \alpha + O(n^{-1})$, as $n \rightarrow +\infty$, see e.g. [1]. As proved in [5], function

$$G_c(z; \hat{\theta}, \theta) = C\{G(z; \hat{\theta}), \theta\}, \quad (3)$$

which is obtained by substituting α with $G(z; \hat{\theta})$ in $C(\alpha, \theta)$, is a proper predictive distribution function, provided that $C(\cdot, \theta)$ is a sufficiently smooth function. Furthermore, it gives, as quantiles, prediction limits $z_\alpha^c(\hat{\theta}, \theta)$ with coverage probability equal to the target nominal value α , for all $\alpha \in (0, 1)$.

The calibrated predictive distribution (3) is not useful in practice, since it depends on the unknown parameter θ . However, a suitable parametric bootstrap estimator for $G_c(z; \hat{\theta}, \theta)$ may be readily defined. Let y^b , $b = 1, \dots, B$, be parametric bootstrap samples generated from the estimative distribution of the data and let $\hat{\theta}^b$, $b = 1, \dots, B$, be the corresponding maximum likelihood estimates. Since $C(\alpha, \theta) = E_Y[G\{z_\alpha(\hat{\theta}(Y)); \theta\}; \theta]$, we define the bootstrap-calibrated

predictive distribution as

$$G_c^b(z; \hat{\theta}) = \frac{1}{B} \sum_{b=1}^B G\{z_\alpha(\hat{\theta}^b); \hat{\theta}\}_{|\alpha=G(z; \hat{\theta})}. \quad (4)$$

The corresponding α -quantile defines, for each $\alpha \in (0, 1)$, a prediction limit having coverage probability equal to the target α , with an error term which depends on the efficiency of the bootstrap simulation procedure.

3 Generalised extreme value distribution

The previous result can be applied, with some care, to the context of extreme value prediction. Indeed, assume that $\{X_t\}_{t \geq 1}$ is a discrete-time stochastic process with probability distribution specified by an unknown parameter. Furthermore, let $Y_i = \max_{k \in T_i} X_k$ be the maximum of the process over time interval T_i , $i \geq 1$. It is a well known result in extreme value theory that, under suitable conditions and if the number of observations in each period is big enough, the Y_i 's are approximately independent and with the same generalised extreme value (GEV) distribution; see for instance [2].

Now, assume that $Y = (Y_1, \dots, Y_n)$, $n > 1$, is observable, while $Z = Y_{n+1}$ is a future or not yet available observation of the maximum of the process over the next time interval. Then Y_1, \dots, Y_n and $Z = Y_{n+1}$ can be considered as independent random variables with the same GEV distribution function

$$G(z; \mu, \sigma, \xi) = \exp \left\{ - \left[1 + \xi \left(\frac{z - \mu}{\sigma} \right) \right]^{-1/\xi} \right\}, \quad (5)$$

with z such that $1 + \xi(z - \mu)/\sigma > 0$ and $\sigma > 0$.

The GEV distribution has three parameters: a location parameter μ , a scale parameter σ and a shape parameter ξ . In particular, the values of ξ determine the type of GEV distribution:

- $\xi \rightarrow 0$ corresponds to the Gumbel distribution (type *I*);
- $\xi > 0$ corresponds to the Fréchet distribution (type *II*);
- $\xi < 0$ corresponds to the (negative) Weibull distribution (type *III*).

It is important noticing that when $\xi > 0$ or $\xi \rightarrow 0$ the support of the distribution is not limited from above. Only in the case when $\xi < 0$ the support has an upper bound equal to $\mu - \sigma/\xi$.

Inverting (5) we can achieve an explicit expression for the quantiles of the distribution:

$$z_\alpha = z_\alpha(\mu, \sigma, \xi) = \begin{cases} \mu - \frac{\sigma}{\xi} [1 - \{-\log(\alpha)\}^{-\xi}] & \text{if } \xi \neq 0 \\ \mu - \sigma \log\{-\log(\alpha)\} & \text{if } \xi = 0 \end{cases} \quad (6)$$

with $G(z_\alpha; \mu, \sigma, \xi) = \alpha$. The value z_α is also called return level and it indicates the value that is expected to be exceeded on average once every $1/(1 - \alpha)$ time intervals.

4 An application to athletic records

In this section we apply the calibrated procedure to athletic records for two different purposes. First, we estimate the probability of observing a new record in the next year and we predict the expected time for a future record. This also allows to evaluate the goodness of a current world record. Secondly, we discuss the existence of the ultimate record and we give the correct interpretation to its estimate.

We have collected data from the web site of the International Association of Athletics Federations (IAAF) [6]. Starting from 2001, we have registered the annual records for males and females, for the following events: 100 m, 200 m, 400 m, 10,000 m, long jump and javelin. We have transformed times into mean speeds so that, for each event, the higher the best.

4.1 Parameter estimation

The first step consists of estimating the unknown parameters of each distribution. In particular, the estimates obtained for the shape parameters are very important because they determine the particular distribution to be used inside the GEV family.

Here we consider three different methods of estimation: maximum likelihood, L-moments and generalised maximum likelihood. In spite of its optimal asymptotic properties, the method of maximum likelihood does not perform very well for small sample sizes. Instead L-moments and generalised maximum likelihood estimates ensure a better fit, especially for the shape parameter. In particular, estimation based on the generalised maximum likelihood retains large sample properties of the maximum likelihood but improves on its small sample performance (see, for instance, [3]).

Table 1 and Table 2 show estimates for the shape parameters ξ obtained using the three different estimating methods, for men and women, respectively. The first row of each table reports maximum likelihood estimates (mle), the second row contains estimates obtained by the method of L-moments (Lmom) and the third row is for generalized maximum likelihood (gmle). All the estimated values for the shape parameters are negative, with an exception for women long jump, for which the three estimates are positive. Thus, the corresponding estimative distributions are reverse Weibull distributions for all events with negative shape parameter and a Fréchet distribution for women long jump. In the sequel we will use generalised maximum likelihood estimates.

estimate	100 m	200 m	400 m	10,000 m	long jump	javelin
mle	-0.1618	-0.0826	-0.1781	-0.2157	-0.3984	-0.3231
Lmom	-0.1281	-0.0553	-0.2246	-0.0819	-0.3104	-0.3755
gmle	-0.1281	-0.0553	-0.2246	-0.0819	-0.3104	-0.3755

Table 1: Men: estimates of the shape parameters for different events

estimate	100 m	200 m	400 m	10,000 m	long jump	javelin
mle	-0.3343	-0.4416	-0.1658	-0.1269	0.1116	-0.3033
Lmom	-0.3069	-0.3006	-0.1330	-0.1803	0.1126	-0.1864
gmle	-0.3069	-0.3006	-0.1330	-0.1803	0.3311	-0.1864

Table 2: Women: estimates of the shape parameters for different events

4.2 Prediction

In this section we compare the estimative distribution function obtained from the generalised maximum likelihood estimator with the bootstrap calibrated one, for each of the considered events.

We will see that, using the bootstrap calibrated predictive distribution, we can properly calculate probabilities related to the variable Z which represents the best performance in the year to come. In particular we can predict the probability of having a new world record in the next year as $\alpha_{WR} = P(Z > WR)$, where WR represents the present world record. This probability can also be used to evaluate the goodness of the world record. Moreover, from α_{WR} we can calculate the expected number of years for the next record, $T_{WR} = 1/\alpha_{WR}$.

In all the cases when the estimate of the shape parameter ξ is negative, the estimative distribution is a (negative) Weibull distribution. It has a bounded upper tail at $UL = \hat{\mu} - \hat{\sigma}/\hat{\xi}$. In the analysis of sport data UL is the estimate of what is called the ultimate record, which is a value that cannot be exceeded by any performance. Instead, using the calibrated predictive distribution we can show that the probability of exceeding UL , $\alpha_{UL} = P(Z > UL)$, is different from 0. Unfortunately, in non regular cases when the support of the distribution depends on unknown parameters, formula (4) is only useful for calculating the bootstrap calibrated predictive distribution inside the estimated domain. As a consequence, when the present world record exceeds the estimated upper bound, we cannot calculate α_{WR} . This drawback is not present for women long jump, since in this case the estimated shape parameter is positive, giving rise to a Fréchet estimative distribution whose upper tail is unbounded.

Table 3 and Table 4 summarise the main results obtained for each considered event. In particular they report for men and women, respectively: the estimate of the ultimate record UL , the probability α_{UL} of exceeding that estimate, the present world record WR , the probability α_{WR} of exceeding it and the expected time T_{WR} for improving it.

	100 m	200 m	400 m	10,000 m	long jump	javelin
UL	10.797	12.052	9.431	6.804	8.871	95.364
α_{UL}	0.009	0.008	0.011	0.008	0.011	0.013
WR	10.438	10.422	9.296	6.339	8.95*	98.48*
α_{WR}	0.031	0.057	0.029	0.054	-	-
T_{WR}	31.79	17.51	33.91	18.62	-	-

Table 3: Men's summary results. * means that the corresponding world record is not included in the data.

	100 m	200 m	400 m	10,000 m	long jump	javelin
UL	9.477	9.368	8.449	5.897	-	78.318
α_{UL}	0.012	0.011	0.009	0.010	-	0.010
WR	9.533*	9.372*	8.403*	5.690	7.52*	72.28
α_{WR}	-	-	0.009	0.028	0.056	0.084
T_{WR}	-	-	105.55	36.05	17.93	11.83

Table 4: Women's summary results. * means that the corresponding world record is not included in the data.

Three different possible situations are illustrated and commented using data from men's 400 m, women's 100 m and women's long jump.

4.2.1 Men's 400 m

Figure 1 shows the estimative (red dashed) and bootstrap calibrated (black solid) distribution functions for men's 400 m data. The bootstrap procedure is based on 5000 replications. The present world record (blue dash-dotted) and the estimated ultimate record (red dotted) are also represented. Here the original time data (sec) have been transformed into mean speeds (m/sec) since the GEV model fits to maxima data.

For the transformed data, the estimate of the shape parameter is negative. This implies that the estimative distribution function is a reverse Weibull distribution with upper bound $UL = \hat{\mu} - \hat{\sigma}/\hat{\xi} = 9.431$ m/sec. This corresponds to a time of 42.41 sec. The value UL is usually interpreted as an estimate of the ultimate record which is the best possible performance in the event. It is important noticing that this is just an estimate and, of course, it is subject to variability. To account for this variability, we can use the calibrated predictive distribution and correctly predict the probability of exceeding UL . As one can see in the plot, in fact, this probability is the difference between the value 1 of the estimative distribution at UL and the value of the bootstrap calibrated distribution at UL . Thus, $\alpha_{UL} = P(Z > UL) = 0.011$. Similarly, we can calculate the probability of improving the present world record of 43.03 sec, $WR = 9.296$ m/sec, as $\alpha_{WR} = P(Z > WR) = 0.029$, meaning that we expect to improve the present world record about 3 times every 100 years. This can also be taken as a measure of goodness of a world record. Both probabilities α_{UL} and α_{WR} are wrongly underestimated by the estimative distribution function because in the estimation procedure the true parameters are substituted by their estimates without taking into account for the additional uncertainty introduced. In particular, the estimative distribution underestimates to 0.016 the probability of improving the current world record.

4.2.2 Women's 100 m

Figure 2 shows the estimative (red dashed) and bootstrap calibrated (black solid) distribution functions for women's 100 m data. The bootstrap procedure is based on 5000 replications. The present world record (blue dash-dotted) and the estimated ultimate record (red dotted) are also represented. As in the previous example, the original time data (sec) have been transformed into mean speeds (m/sec) since the GEV model fits to maxima data.

For the transformed data, the estimate of the shape parameter is negative, thus the estimative distribution function is a reverse Weibull distribution with upper bound $UL = \hat{\mu} - \hat{\sigma}/\hat{\xi} = 9.477$ m/sec. This corresponds to a time of 10.55 sec. We can use the calibrated predictive distribution to correctly predict the probability of exceeding UL . As one can see in the plot, in fact, this probability is the difference between the value 1 of the estimative distribution at UL and the value of the bootstrap calibrated distribution at UL . Thus, $\alpha_{UL} = P(Z > UL) = 0.012$. In this example, the present world record $WR = 9.533$ m/sec (10.49 sec) exceeds the upper limit UL , as can be seen in figure 2. This may occur when the data used for estimation do not include the world record. Indeed, the present world record dates back to 1988, while we have considered data from 2001 to 2018. A methodological problem arises in this situation, since we are not able to calculate the values of the bootstrap calibrated predictive distribution (4) in points that exceed the upper bound of the estimative distribution. The upper tail of the calibrated predictive distribution can be estimated using non linear regression, but this issue requires further research. At the moment, we can only conclude by saying that the probability of improving the present world record is $\alpha_{WR} = P(Z > WR) < P(Z > UL) = 0.012$. Actually, the present world record seems to be an exceptional result that can be hardly improved at the moment.

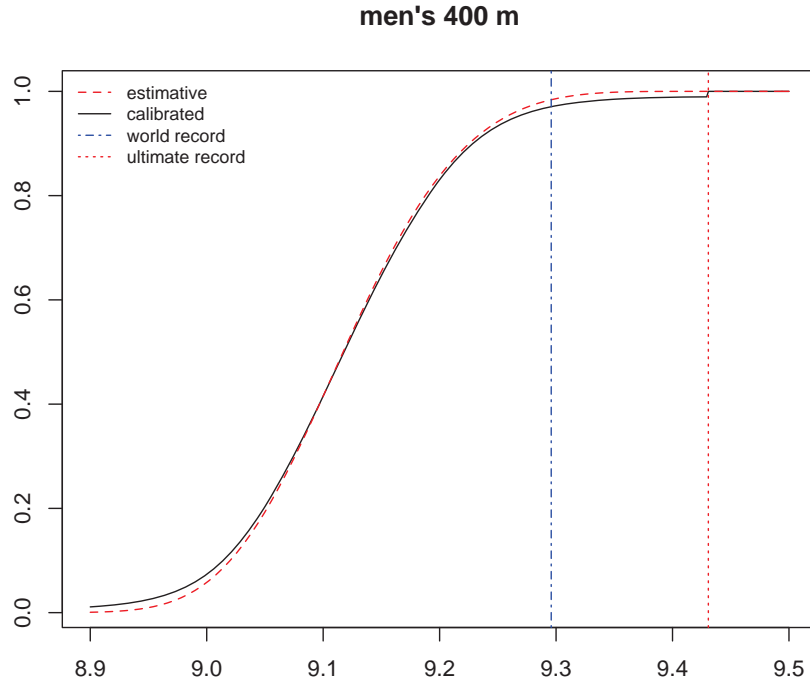


Figure 1: Men's 400 m. Plot of estimative (red dashed) and bootstrap calibrated (black solid) distribution functions for men's 400 m data. Bootstrap procedure is based on 5000 replications. World record (blue dash-dotted) and estimated ultimate record (red dotted) are also represented.

4.2.3 Women's long jump

Figure 3 shows the estimative (red dashed) and bootstrap calibrated (black solid) distribution functions for women's long jump data. The bootstrap procedure is based on 5000 replications. The present world record (blue dash-dotted) is also represented.

This is the only event for which the estimate of the shape parameter of the GEV distribution is positive, thus the estimative distribution function is a Fréchet distribution with no upper bound. The present world record, $WR = 7.52$ m, dates back to 1988 and is not included in the data. Anyway, this is not a problem, being the upper bound $UL = +\infty$. Using the bootstrap calibrated distribution, we can predict the probability of improving the present world record: $\alpha_{WR} = P(Z > WR) = 0.056$. Notice that the estimative distribution wrongly underestimates this probability to 0.040. The expected time for improving the current world record is about 18 years.

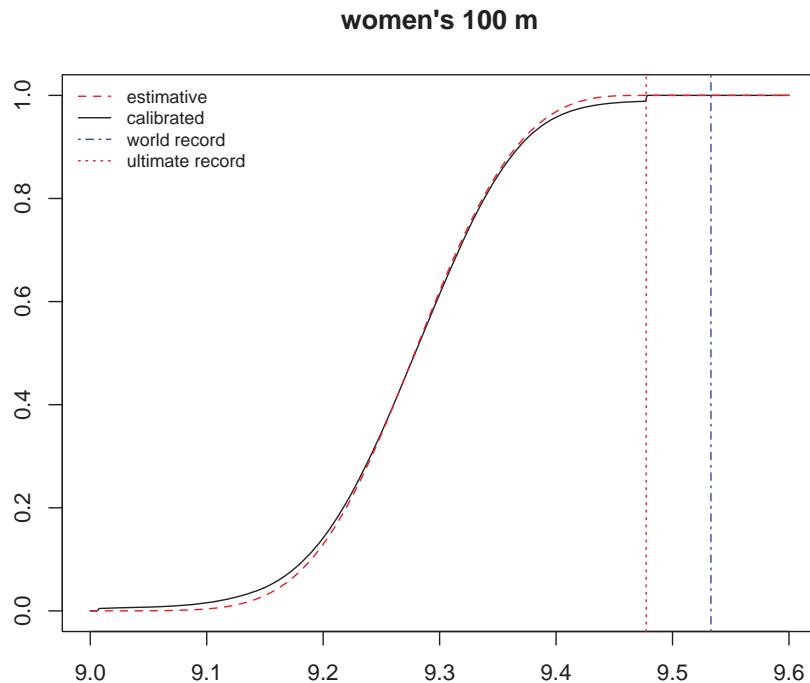


Figure 2: Women's 100 m. Plot of estimative (red dashed) and bootstrap calibrated (black solid) distribution functions for women's 100 m data. Bootstrap procedure is based on 5000 replications. World record (blue dash-dotted) and estimated ultimate record (red dotted) are also represented.

Acknowledgments

This research is partially supported by the Italian Ministry for University and Research under the PRIN2015 grant No. 2015EASZFS 003.

References

- [1] O.E. Barndorff-Nielsen and D.R. Cox. Prediction and asymptotics. *Bernoulli*, 2:319–340, 1996.
- [2] S. Coles. *An introduction to statistical modeling of extreme values*. Springer-Verlag, London, 2001.
- [3] S.J. Coles and M.J. Dixon. Likelihood-based inference for extreme value models. *Extremes*, 2(1):5–23, 1999.
- [4] J.H.J. Einmahl and J.R. Magnus. Records in athletics through extreme-value theory. *Journal of the American Statistical Association*, 103:1382–1391, 2008.
- [5] G. Fonseca, F. Giummolè, and P. Vidoni. Calibrating predictive distributions. *Journal of Statistical Computation and Simulation*, 84:373–383, 2014.
- [6] IAAF. International Association of Athletics Federations. <https://www.iaaf.org/home>.

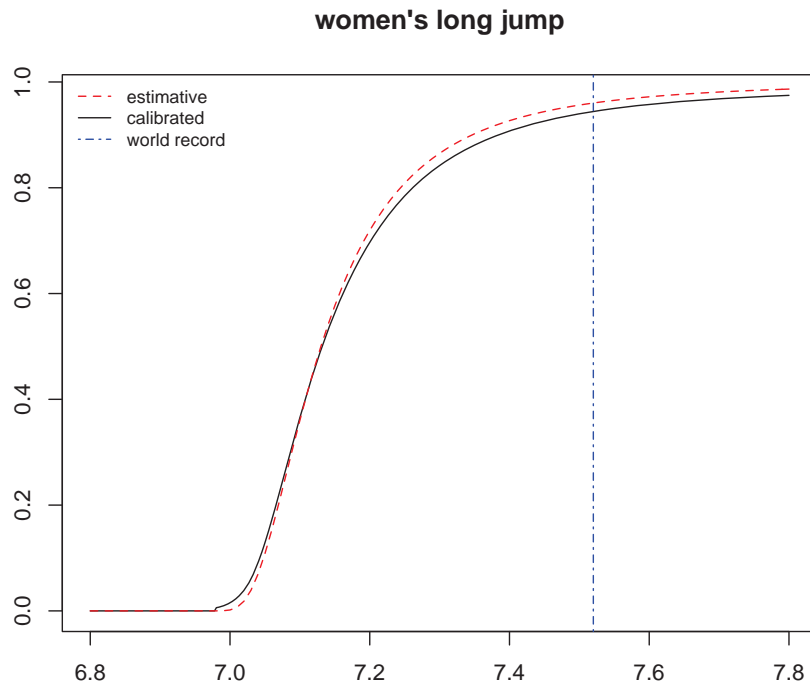


Figure 3: Women's long jump. Plot of estimative (red dashed) and bootstrap calibrated (black solid) distribution functions for women's long jump data. Bootstrap procedure is based on 5000 replications. World record is also represented (blue dash-dotted). Since the estimative density is not bounded from above, the ultimate record is $+\infty$.

- [7] F. Noubary and R. Noubary. Survival analysis of the men's 100 meter dash record. *Applications and Applied Mathematics*, 11(1):115–126, 2016.
- [8] R.D. Noubary. A procedure for prediction of sports records. *Journal of Quantitative Analysis in Sports*, 1(1):Article 4, 2005.
- [9] R.D. Noubary. Tail modeling, track and field records, and Bolt's effect. *Journal of Quantitative Analysis in Sports*, 6(3), 2010.
- [10] J.T. Terpstra and N.D. Schauer. A simple random walk model for predicting track and field world record. *Journal of Quantitative Analysis in Sports*, 3(3):Article 4, 2007.
- [11] P. Volf. A stochastic model of progression of athletic records. *Journal of Management Mathematics*, 22:157–169, 2011.

Predicting match outcome in professional Dutch soccer using tactical performance metrics computed from position tracking data

Floris Goes^{1*}, Matthias Kempe¹ and Koen Lemmink¹

¹ University of Groningen, University Medical Center Groningen (UMCG), Department of Human Movement Sciences, Groningen, The Netherlands

f.r.goes@umcg.nl

Abstract

Quality as well as quantity of tracking data have rapidly increased over the recent years, and multiple leagues have programs for league-wide collection of tracking data. Tracking data enables in-depth performance analysis, especially with regard to tactics. This already resulted in the development of several Key Performance Indicators (KPI's) related to scoring opportunities, outplaying defenders, numerical balance and territorial advantage. Although some of these KPI's have gained popularity in the analytics community, little research has been conducted to support the link with performance. Therefore, we aim to study the relationship between match outcome and tactical KPI's derived from tracking data. Our dataset contains tracking data of all players and the ball, and match outcome, for 118 Dutch premier league matches. Using tracking data, we identified 72.989 passes. For every pass-reception window we computed KPI's related to numerical superiority, outplayed defenders, territorial gains and scoring opportunities using position data. This individual data was then aggregated over a full match. We then split the dataset in a train and test set, and predicted match outcome using different combinations of features in a logistic regression model. KPI's related to a combination of off-the-ball features seemed to be the best predictor of match outcome (accuracy of 64.0% and a log loss of 0.67), followed by KPI's related to the creation of scoring opportunities (accuracy of 58% and a log loss of 0.69). This indicates that although most (commercially) available KPI's are based on ball-events, the most important information seems to be in off-the-ball activity. We have demonstrated that tactical KPI's computed from tracking data are relatively good predictors of match outcome. As off-the-ball activity seems to be the main predictor of match outcome, tracking data seems to provide much more insight than notational analysis.

* Presenting & Corresponding Author

1 Introduction

Soccer is one of the most popular global sports, and match performance analysis has been the subject of intensive research over several decades¹. Soccer, nowadays, is a multi-billion industry that embraces mathematical ideas as teams are constantly searching for ways to improve their odds at winning, while spectators are trying to predict the outcome of a game to win money on the gambling market². As a result, analyzing tactics and match performance in soccer is of particular interest to a broad and varied audience.

Traditionally, tactical analysis has been conducted based on observational assessment by experts or by means of notational assessment on-ball events like passes, dribbles, and tackles³. Despite the limitations of notational data, the focus on ball-events like passes in itself is understandable. A pass is the most frequent ball-event in a match and passing is a – or even the – key aspect of tactics in soccer. However, notational data only provides discrete low-level data, and thus only tells us what happens with the ball. Therefore it has limited practical value⁴. Teams might even dominate typical summary statistics like possessions, shots and number of passes, but still fail to score⁵. Nevertheless, notational analysis is still frequently used for tactical analysis by broadcasters, teams and scientists^{3,6}. One could argue however that it would be much more interesting to look at what is going on with the 21 players not carrying the ball during a ball-event like passing. Yet achieving this requires not only notational data, but also position tracking data.

As opposed to notational event data, automatically generated position tracking data provides the opportunity to derive high-level continuous data off all players and the ball at the same time⁷. As a result of technological innovation and the league-wide implementation of position tracking systems in for example the German Bundesliga and the Dutch Eredivisie, the quantity and quality of available tracking data rapidly increased over the recent years^{3,6}. Despite this increasing availability, the potential of position tracking data to analyze tactical performance has not been harnessed as tracking data is mostly used by analysts to monitor physical performance⁸. However, this data allows us to automatically study the complex interactions of all players on the field during every pass, and can therefore be regarded as a potential game changer for tactical analysis in soccer³.

The limited practical use of position tracking data for tactical analysis might be explained by two reasons. First of all, most scientific work on tactical analysis using position tracking data – although of great scientific importance – has relatively little practical implications. Only a minority of the work investigated a link between the features they used for tactical analysis and actual match performance, and most of them did not find a clear relationship. In order to derive practical meaning from these types of analyses, we therefore propose it is critical to study the link between tactical features and match performance. Secondly, one could argue that as position tracking data is characterized by a much higher complexity and volume in comparison to notational event data, it challenges the typical data management and data analytics methods⁹ commonly employed in sports science, and can therefore be considered big data. As a result, we propose that unlocking the potential of this data for tactical analysis requires the implementation of skills and techniques from other domains than sports science.

In conclusion, one could argue position tracking data harnesses the potential to provide in-depth insights in the complex tactics of soccer, and these insights can theoretically be used in the analysis and maybe even the prediction of performance. However, in order to achieve this and derive practical meaning from tactical analysis, the link between tactical features derived from position tracking data and actual match performance first has to be established. With the current paper, we therefore aim to study the relationship between tactical features derived from position tracking data and match outcome. To achieve this, we will use a match outcome prediction model based on tactical key performance indicators (KPI's). The results of our study could allow analysts to derive more practical meaning from tactical analysis using position tracking data, and scientists could use these KPI's to study the relationship between their tactical features and match success.

2 Quantifying Tactical Behavior

Tactics, often referred to in research as tactical behavior, can be defined as the management of space and time by a group of cooperating individuals, in interaction with the opponent while constantly adapting to the conditions of play, in order to achieve a common goal^{3,10}. This common goal is related to ball-possession status, as teams have different tactical objectives when attacking and defending¹¹. When in possession of the ball, teams aim to move the ball in the direction of the opponents goal, increase the effective play area through depth and width mobility, create numerical superiority in key offensive areas of the field, destabilize the defense, and ultimately create scoring opportunities¹². On the other hand, when defending, teams aim to keep the opponent away from the goal, keep the effective play area small, move in unity to prevent destabilization, and keep numerical superiority close to their own goal¹². These common goals are widely considered the general principles of play in soccer^{11,12}. Achieving these goals can be seen as successful tactical behavior, and a relationship between tactical behavior and match outcome is widely assumed.

In order to study the relationship between tactical behavior and match outcome, one first has to quantify successful tactical behavior. As we are mainly concerned with offensive tactical behavior, we focused on tactical features related to the offensive principles of play. For this purpose, we first need to discuss how existing tactical features (either commercially available or derived from scientific research) can be related to the offensive principles of play.

First of all, moving the ball towards the opponent's goal and subsequently creating scoring opportunities (*zone principle*) can be assumed to have the most direct relationship with scoring goals in comparison to the other principles of play. Existing features like *expected goals (xG)*¹³ (Optasports, London, United Kingdom), and Link's *dangerosity*¹⁴ feature directly quantify this tactical principle. Both features are computed using distance and angle between the goal and the ball carrier, and award higher values for locations closer to the goal. As *xG* is typically computed using only notational event data, it is relatively inaccurate and does not take the pressure of defenders or any other of-the-ball activity into account. Therefore, it provides low-level information. *Dangerosity* is computed in a somewhat similar fashion, yet it is computed based on position tracking data and takes defensive pressure as well as of the ball activity into account as moderating factors. Therefore, *dangerosity* could be regarded as a high-level expected goals model.

Secondly, gaining numerical superiority (*balance principle*) is often believed to be of key importance for creating high probability scoring opportunities, as it will contribute to space creation and destabilization of the defense¹⁵. Numerical superiority and outplaying defenders can be analyzed from an on-the-ball as well as an off-the-ball perspective: teams can try to outplay defenders through passing and dribbling, and they can position their off-the-ball players in key areas of the field. Existing features like *Packing-Rate*¹⁶ and *Impect*¹⁶ (Impect GmbH, Cologne, Germany) have gained popularity in especially the German Bundesliga. They quantify the number of outplayed (*packed*) opponents or defenders through passing, and can easily be derived from position tracking data. Off-the-ball superiority has gained considerably less attention in the literature, but can also be directly derived from the position tracking data.

Finally, it is often believed in soccer that keeping the effective play area large when in possession of the ball (*space mobility principle*) is another prerequisite for space creation and destabilization of the defense. The effective play area of the attacking team can be defined as the attacking team's surface, and can be derived from position tracking data using the *Convex Hull* method¹⁷. One can compute the *Convex Hull* for every timeframe in ball possession and take the average as an indicator of space mobility.

3 Feature Engineering

3.1 Features Related to Zone

To quantify tactical performance with regard to the *zone*, *balance* and *space mobility* principles, we constructed separate features for every principle of attacking play. For the current study we adapted features currently available in science and practice. As in most cases limited technical details underlying a certain feature are publically available, and in order to solve some feature-specific limitations, we choose to construct our own adaptation of these features rather than exactly replicate existing features. All feature construction was conducted in Python 3.6 using the NumPy, Pandas and SciPy libraries.

To quantify tactical performance on the *zone* principle, we constructed a low-level and high-level *zone* feature, partly adapted from the work by Link¹⁴. First, we determined the low-level *zone* value based on the position of the ball-carrier relative to the goal in every pass and reception (Figure 1). *Zone* values could range from 0 (furthest from the goal) to 1 (closest to the goal). The high-level feature was then computed by adding on-ball-pressure to the model. Pressure on the ball was computed using the model proposed in Andrienko et al¹⁸. This model computes a pressure value *PR* (0-100%) based on the distance off all defensive players to the ball carrier, and the angle of all defensive players towards the threat direction (in this case the direction from the ball carrier to the goal). In this model, 0 represents no pressure at all, while values of 100% represent high pressure from the defenders close to the ball-carrier. As we assume high pressure increases the difficulty of creating a scoring opportunity, the zone value *Z* is penalized by *PR* as shown in Eq. 1.

$$Z = Z * (1 - PR) \quad (1)$$

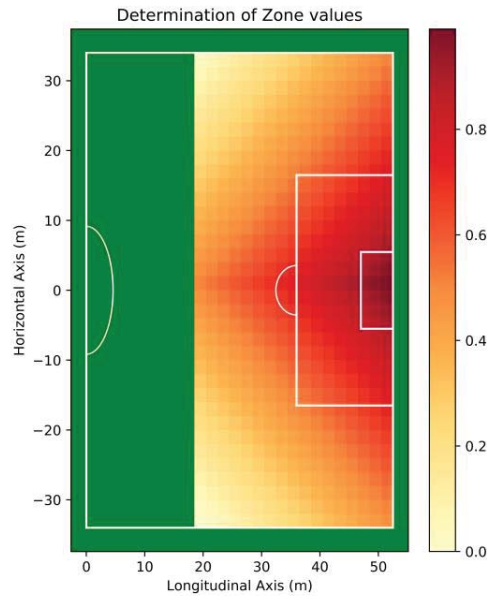


Figure 1 - Visual representation of zone values computed for every pass and reception. Color bar represents the zone values (range 0.0 - 1.0)

Both the low-level and high-level *zone* were computed for every successful pass and reception and then aggregated over the full match. This resulted in mean and total low- and high-level zone values for passers and receivers on a team.

3.2 Features Related to Balance

To quantify tactical performance on the *balance* principle, we constructed two passing features and three off-the-ball balance features. Our passing features follow the description of the *Packing-Rate*¹⁶ and *Impect*¹⁶. We computed the number of outplayed opponents based on the longitudinal coordinates of the pass, reception and all the opposing players, and we computed the number of outplayed defenders based on the longitudinal coordinates of the pass, reception and the last 6 players on the field plus the goalkeeper (Figure 2). Note that the number of outplayed opponents can also be negative in the case of a backwards pass. Furthermore, note that we only looked at the X-coordinates to determine what

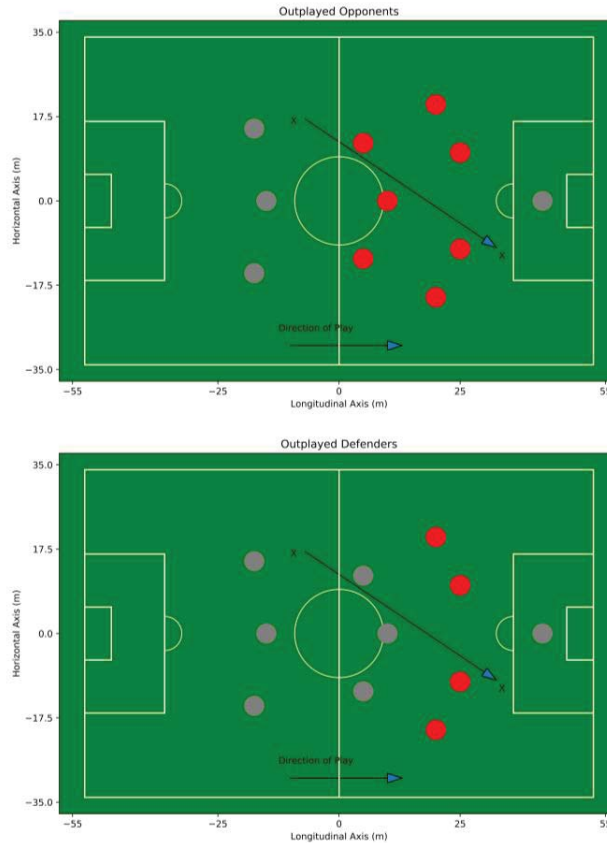


Figure 2 - Visual representation of outplayed opponents (top) and outplayed defenders (bottom) as a result of pass. Outplayed opposing players are shown in red, other opposing players are shown in grey. Note that in our approach the number of outplayed defenders is based on the last 6 outfield players and the goalkeeper.

As off-the-ball balance features we computed numerical superiority scores for the attacking team on the opposing half, in the final 3rd and in the score-box. To do so, we assessed numerical balance (by counting players of both teams) in a certain area (i.e. final 3rd or score-box) during every pass-reception window, and awarded points for every window in which the attacking team had numerical superiority in that area (+1 player = 1 point, +2 players = equal 2 points, etc.).

3.3 Features Related to Space Mobility

Finally, as a quantification of the *space mobility* principle, we computed the average attacking team's surface area for every attack during a game, over the duration of the complete possession. The attacking team surface area (S_A) on every timestamp t in the game was computed as the *Convex Hull* of an array P_t containing the positions of all n outfield players (the goalkeeper was excluded), using the QHull implementation in the SciPy library (eq. 2 & 3).

$$P_t = [[X_i^t + Y_i^t], [X_{i+1}^t + Y_{i+1}^t], [...], [X_n^t, Y_n^t]] \quad (2)$$

$$S_A = \text{ConvexHull} \parallel P_t \parallel \quad (3)$$

4 Modelling Match Performance in Soccer

To evaluate tactical performance of a team in relation to the different principles, and analyze the relationship between tactical performance and match outcome, we collected and processed position tracking data on both teams for matches played during 4 consecutive Dutch Eredivisie seasons. Players were tracked with a semi-automatic optical tracking system (SportVU; STATS LLC, Chigago, IL) that captures the X and Y coordinates of all players and the ball at 10 Hz. Our dataset contained 118 matches in which 26 unique teams played each other. As we were only concerned with the differences between winning and losing teams, we excluded matches that ended in a draw. This resulted in a final dataset that consists of 25 teams that played in 89 matches that resulted in a win or a loss and contained 98.718 pass attempts of which 60.524 passes were successful.

The data of every single match were first pre-processed with ImoClient software (Inmotio Object Tracking B.V., The Netherlands). Pre-processing consisted of filtering the data with a weighted Gaussian algorithm (85% sensitivity) and automatic detection of ball possessions and ball events based on the tracking data. Both the tracking data and the ball event data were then imported as individual data frames in Python 3.6 and automatically processed on a match-by-match basis. We then computed the low-level and high-level zone feature for every pass and reception, the number of outplayed opponents and outplayed defenders for every pass, the numerical superiority in 3 areas for every pass-reception window, and the team surface area of all outfield players for every timeframe the team was in possession of the ball. All features were computed according to the methods as described in section 3.

Table 1 - Descriptive statistics (mean \pm std.) of winning and losing teams on the various principles of play. * ($p < .05$) and ** ($p < .01$) denote significant differences between winning and losing teams.

	Wins (N = 89)	Losses (N = 89)	Mean Diff.	Effect Size (Cohen's d)
<i>Zone Principle</i>				
Low-level zone passer (Mean)	0.031 \pm 0.013	0.028 \pm 0.012	+10.7%	0.24
Low-level zone receiver (Mean)	0.040 \pm 0.014	0.037 \pm 0.014	+8.1%	0.24
High-level zone passer (Mean)	0.022 \pm 0.010	0.020 \pm 0.010	+10%	0.21
High-level zone receiver (Mean)	0.032 \pm 0.012	0.028 \pm 0.011	+14.3%	0.28*
Low-level zone passer (Total)	10.62 \pm 5.40	9.55 \pm 4.54	+11.2%	0.21
Low-level zone receiver (Total)	13.54 \pm 6.21	12.36 \pm 5.26	+9.5%	0.20
High-level zone passer (Total)	7.11 \pm 3.70	6.51 \pm 3.57	+9.2%	0.16
High-level zone receiver (Total)	10.10 \pm 4.52	9.14 \pm 4.01	+10.5%	0.22
<i>Balance Principle</i>				
Outplayed defenders (Mean)	0.23 \pm 0.10	0.21 \pm 0.09	+9.5%	0.19
Outplayed opponents (Mean)	0.39 \pm 0.17	0.39 \pm 0.16	+2.6%	0.13
Outplayed defenders (Total)	71.01 \pm 29.69	67.88 \pm 30.57	+4.7%	0.11
Outplayed opponents (Total)	119.69 \pm 49.46	121.91 \pm 50.88	-1.8%	-0.04
Half Superiority (Total)	2.82 \pm 7.67	1.87 \pm 5.78	+50.8%	0.14
Final 3 rd Superiority (Total)	3.11 \pm 3.52	2.22 \pm 3.04	+40.0%	0.27**
Score Box Superiority (Total)	0.84 \pm 1.51	0.76 \pm 3.39	+10.5%	0.03*
<i>Space Mobility Principle</i>				
Team Surface Area (mean)	979.76 \pm 99.12	966.41 \pm 96.70	+1.4%	0.14

To compare performance between winning and losing teams, we aggregated all feature scores into mean (values per pass), and total (sum over a full match) scores. We then took the means and standard deviations of all winning and losing teams for a between-group comparison (Table 1). As most features scores were not-normally distributed, and variances were heterogenic, we conducted Kruskal-Wallis tests to statistically compare both groups. We found that winning teams had a significantly increased mean high-level zone score for pass receivers ($H(176) = 4.16$, $p < 0.05$), and a significantly increased superiority score in the final 3rd ($H(176) = 6.90$, $p < 0.01$) and score box ($H(176) = 5.09$, $p < 0.05$) compared to losing teams.

As a next step, we predicted match outcome based on several combinations of performance features. To do so we first split the data set in a training set that contained 80% of the data, and a test set that contained 20% of the data, stratified on match outcome. Furthermore, we scaled our features to the same scale using a robust scaling algorithm. We then fitted a 5-fold cross-validated Logistic Regression model to our training dataset and predicted winning and losing probability for both teams in every match.

First, we fitted the model using only the features that had shown (significant) power to discriminate between winning and losing teams (Table 1), as we expected this model to perform the best. Based on the mean high-level zone receiver score (β_1), the total final 3rd superiority score (β_2), and the total score box superiority score (β_3), we were able to predict binary match outcome with an accuracy of 64% and a log loss of 0.67, based on the following regression equation (4):

$$\text{Outcome} = -0.0167 + 0.136 \beta_1 + 0.130 \beta_2 - 0.0162 \beta_3 \quad (4)$$

Then, we fitted models for all three discussed principles of play, to see what principle has the strongest relation with success. In cases where we had both mean and total values for a variable, we opted for the mean as this consistently proved to be a better discriminator. For performance on the *zone* principle, we fitted a model using the mean low-level zone for passers (β_4) and receivers (β_5), and the mean high-level zone for passers (β_6) and receivers (β_7). Based only on zone features, we were able to predict binary match outcome with an accuracy of 58% and a log loss of 0.69, using the following regression equation (5):

$$\text{Outcome} = -0.7e^{-6} + 0.00028 \beta_4 + 0.00035 \beta_5 + 0.00014 \beta_6 + 0.00054 \beta_7 \quad (5)$$

For performance on the *balance* principle, we fitted a model using the mean outplayed defenders (β_8) and opponents (β_9), and the total half superiority (β_{10}), final 3rd superiority (β_{11}), and score-box superiority scores (β_{12}). Based only on balance features, we were able to predict binary match outcome with an accuracy of 58% and a log loss of 0.70, using the following regression equation (6):

$$\text{Outcome} = 0.018 + 0.97 \beta_8 - 0.65 \beta_9 - 0.06 \beta_{10} + 0.38 \beta_{11} - 0.04 \beta_{12} \quad (6)$$

Finally, for performance on the *space mobility* principle, we fitted a model using the mean team surface area per attack (β_{13}). Based only on a space mobility feature, we were able to predict binary match outcome with an accuracy of 64% and a log loss of 0.69, using the following regression equation (7):

$$\text{Outcome} = 0.003 + 0.06 \beta_{13} \quad (7)$$

5 Discussion

The aim of this study was to analyze the relationship between tactical features derived from position tracking data and match outcome. To achieve this we constructed features that quantify performance on three main principles of attacking play in soccer^{11,12}, and studied the relationship between performance on these principles and binary match outcome (win or lose). Our results indicate differences between winning and losing teams are relatively small, but especially features that are either directly related to off-the-ball activity (numerical superiority) or at least incorporate off-the-ball activity (high-level zone for receivers) are able to discriminate between winning and losing teams and predict match outcome with fair accuracy. Based on these results we were able to confirm the relationship between tactical performance on the *zone* and *balance* principles, but not on the *space mobility* principle. Furthermore, our results indicate some of the features that have gained considerable popularity within the analytics community over the recent years seem to have limited practical value.

To study tactical performance on the *zone* principle, we constructed low-level and high-level zone features for both the passer and receiver in every pass. Our low-level feature has some resemblance with the popular expected goals (*xG*) feature¹³, and – while we derived it directly from the tracking data – could be approximated with notational analysis. Our high-level feature accounts for defensive pressure and therefore requires position tracking data of all players on the field. Both the high-level and low-level features showed some discriminative power between winning and losing teams, with low to medium effect sizes, yet only the mean high-level zone for receivers was significantly increased in winning teams in comparison to losing teams. Based on these results we conclude winning teams more often seem to bring the ball into a position from which scoring opportunities can be created. Both high- and low-level features seem capable of capturing this principle, yet high-level features seem to have more discriminative power. As Optasport's *xG* is typically only computed for actual shots, and we

computed *zone* values for every pass and reception, one has to be cautious in generalizing our results to interpret actual *xG* values.

To assess performance on the *balance* principle, we used both on-the-ball and off-the-ball features. Our on-the-ball-features are focused on outplayed opponents and defenders, and resemble the popular *Packing-Rate*¹⁶ and *Impect*¹⁶. Although these features have gained considerable popularity in especially the German Bundesliga over the recent years¹⁹, and multiple claims have been made about a possible link with match outcome, our research does not support such a relationship. Whereas winning teams did show a slightly higher mean number of outplayed defenders per pass, there was no difference in the mean and total number of outplayed opponents between winning and losing teams, and adding these features to the prediction model decreased prediction accuracy. Off-the-ball features on the other hand seemed to be a strong discriminator between winning and losing teams, as winning teams had significantly increased superiority scores in the final 3rd and the score box. Interestingly, the effect for score-box superiority was only small, but still significant, and leaving this feature of the prediction model harmed the accuracy of the prediction. The lack of a relationship between outplayed opponents/defenders and match outcome might be explained by methodologic limitations. One could for example argue that one should not only look at how many players were passed in the longitudinal direction but also in the lateral direction, and that in some areas of the field passing backwards can be more effective. However, to closely resemble existing approaches we choose not alter the approach for the current study.

Finally, performance on the *space mobility* principle did not seem to have a clear relationship with match outcome, despite the fact that space mobility is assumed to be a key aspect of offensive performance¹¹. The absence of a clear effect might be explained by the fact that we used the team's surface area to assess space mobility. One could argue that although the team's surface is a valid feature to describe the effective area of play, space mobility actually refers to attackers dynamically creating depth by moving away from the ball at the right moment. It is questionable whether this dynamic effect is captured by a collective variable that is aggregated over all timeframes in possession of the ball.

Although capturing performance in easily interpretable KPI's is popular within the analytics community as well as the media, the reality of soccer seems much more complex. One likely explanation for the absence of a strong relationship between most popular KPI's and match outcome might be the fact that these KPI's are typically related to frequent events like passing, that are then aggregated over the full match. As there is a large match-to-match variability and actual tactics depend heavily on the interaction with the opponent; features like the *Packing-Rate* might be more dependent on the playing style of both teams than the actual match outcome. Soccer is a low-scoring game, and one could argue that in order to accurately predict match outcome, one should capture the rare events that lead to offensive success. One such an example is our proposed superiority score. Although highly discriminative between winning and losing, achieving final 3rd superiority also proved to be a rare event. The average superiority score of 3.11 in winning teams indicates these teams only achieve a +1 numerical superiority in the final 3rd on 3 occasions during a match, and these occasions seem to have a big importance for match outcome.

6 Conclusion

With this study, we have shown that although soccer is a complex game that is often considered highly unpredictable, the outcome of a match can be modelled with a fair accuracy. However, despite popular belief, soccer is not really a numbers game that can be analyzed based on simple KPI's of frequent events aggregated over the full course of a match. Discriminating between winning and losing teams and understanding tactical performance requires advanced features that can only be derived from position tracking data and heavily focus on off-the-ball rather than on-the-ball performance.

Disclosure Statement

The authors of this paper reported no conflicts of interest

Acknowledgements

This work was supported by a grant of the Netherlands Organization for Scientific Research (project title: “The Secret of Playing Soccer: Brazil vs. The Netherlands”).

References

1. Dubitzky, W., Lopes, P., Davis, J. & Berrar, D. The Open International Soccer Database for machine learning. *Mach. Learn.* **108**, 9–28 (2019).
2. Constantinou, A. C., Fenton, N. E. & Neil, M. Profiting from an inefficient association football gambling market: Prediction, risk and uncertainty using Bayesian networks. *Knowledge-Based Syst.* **50**, 60–86 (2013).
3. Rein, R. & Memmert, D. Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *Springerplus* **5**, 1410 (2016).
4. Vilar, L., Araujo, D., Davids, K. & Bar-Yam, Y. Science of winning soccer: Emergent pattern-forming dynamics in association football. *J. Syst. Sci. Complex.* **26**, 73–84 (2013).
5. Brooks, J., Kerr, M. & Gutttag, J. Using machine learning to draw inferences from pass location data in soccer. *Stat. Anal. Data Min.* **9**, 338–349 (2016).
6. Brink, M. S. & Lemmink, K. A. P. M. Performance analysis in elite football: all in the game? *Sci. Med. Footb.* **2**, 253–254 (2018).
7. Memmert, D., Lemmink, K. A. P. M. & Sampaio, J. Current Approaches to Tactical Performance Analyses in Soccer Using Position Data. *Sport. Med.* **47**, 1–10 (2017).
8. Sarmiento, H. *et al.* Match analysis in football: a systematic review. *J. Sports Sci.* **32**, 1831–1843 (2014).
9. Gandomi, A. & Haider, M. Beyond the hype: Big data concepts, methods, and analytics. *Int. J. Inf. Manage.* **35**, 137–144 (2015).
10. Gréhaigne, J.-F., Godbout, P. & Bouthier, D. The Foundations of Tactics and Strategy in Team sports. *J. Teach. Phys. Educ.* **18**, 159–174 (1999).
11. Clemente, F. M., Martins, F. M. L., Mendes, R. S. & Figueiredo, A. J. A systemic overview of football game: The principles behind the game. *J. Hum. Sport Exerc.* **9**, 656–667 (2014).
12. Costa, I. T., Garganta, J., Greco, P. J. & Mesquita, I. Tactical Principles of Soccer: concepts and application Tactical Principles of Soccer. *Rev. Mot.* **15**, 657–668 (2009).
13. Expected Goals (xG). Available at: <https://www.optasports.com/services/analytics/advanced-metrics/>. (Accessed: 3rd April 2019)
14. Link, D., Lang, S. & Seidenschwarz, P. Real time quantification of dangerousity in football using spatiotemporal tracking data. *PLoS One* **11**, e0168768 (2016).
15. Rein, R., Raabe, D. & Memmert, D. “Which pass is better?” Novel approaches to assess passing effectiveness in elite soccer. *Hum. Mov. Sci.* **55**, 172–181 (2017).
16. Impect. Available at: <https://www.impact.com/en/#idea>. (Accessed: 3rd April 2019)
17. Moura, F. A., Barreto Martins, L. E., Anido, R. D. O., Leite De Barros, R. M. & Cunha, S. A. Quantitative analysis of Brazilian football players’ organisation on the pitch. *Sport. Biomech.* **11**, 85–96 (2012).
18. Andrienko, G., Andrienko, N., Budziak, G., Dykes, J., Fuchs, G., von Landesberger, T. &

- Weber, H. Visual analysis of pressure in football. *Data Min. Knowl. Discov.* **31**, 1793–1839 (2017).
19. Packing Rate – the Art of Outplaying - RWTH AACHEN UNIVERSITY - English. Available at: <http://www.rwth-aachen.de/cms/root/Die-RWTH/Aktuell/Pressemitteilungen/Juni/~ljyt/Packing-Rate-die-Kunst-des-Ueberspiele/?lidx=1>. (Accessed: 3rd April 2019)

Champions League or domestic league: a coach's choice

Dries Goossens¹, Martina Vandebroek², and Chang Wang³

¹ Faculty of Economics and Business Administration, Ghent University, Belgium

`dries.goossens@ugent.be`

² Faculty of Economics and Business, KU Leuven, Belgium

`martina.vandebroek@kuleuven.be`

³ Faculty of Economics and Business, KU Leuven, Belgium

`chang.wang@kuleuven.be`

Abstract

With the increase of the number of association football matches, at both club and national team level, professional football players are more exposed to fatigue and injuries than before. In order to keep players at a satisfactory fitness level, club coaches have a clear incentive to make player rotations in their starting line-up during the season. They need to decide in which matches to line-up their key players, and consequently on which competitions to focus. In this paper, we develop a new measure to quantify the relative commitment teams show for domestic and the UEFA Champions League. We compare the value of the starting line-ups to detect whether there is a difference in commitment between the top 10 UEFA associations, and we study the matches that deviate from the common strategy. Furthermore, we investigate whether commitment to the Champions League has implications on the results in the domestic league.

1 Introduction

It is safe to say that Europe is the center of professional association football (i.e. soccer). The success of European professional football is accompanied by a demand by broadcasters, sponsors and even clubs for increasingly more matches. The UEFA Champions League (CL) has gradually increased its number of participating teams (in the group stage) from 8 since its inception in 1992 to the current 32 teams. Whereas in 1992–1993, 25 matches sufficed to determine the winner, the 2016–2017 Champions League season required 125 matches to award the Champions League crown to Real Madrid (not including the 92 qualification matches). The UEFA Europa League (EL) has also been reformed several times in the past decades to include more teams playing more matches. At the same time, several European associations have expanded the number of teams of their first division leagues (e.g. France in 2002 and Italy in 2004), or increased the number of matches by redesigning the competition format (e.g. The Netherlands in 2005, Belgium in 2009). Furthermore, the top teams in Europe eagerly engage in lucrative summer tours, involving a series of friendly matches, usually in Asia, to prepare for the new season. On the other hand, the European Club Association (ECA) has been pushing to reduce the number of international matches and friendlies, since they have to release their players each time when called upon by their countries [12]. Despite these attempts, professional football players are potentially exposed to a considerable number of matches over the course of the competitive season [6, 11].

Confronted with congested match schedules, clubs face a tough task maintaining a satisfactory performance. Empirical evidence shows that it's almost unfeasible for a football player to play every minute in every match for his club throughout a season (goalkeepers can be exceptions). For instance, Ispirlidis et al. [15] study the effects of a single football match on indices of

performance, muscle damage, and inflammation during a 6-day recovery period. They find that anaerobic performance deteriorates for as long as 72 hours after a match. Their results clearly emphasize the need of sufficient recovery time for elite players after a match. Moreover, congested fixtures lead to fatigue and/or psychological exhaustion, which in turn may increase the chance of injuries. Dupont et al. [8] monitor 32 players' physical performance (total distance, high-intensity distance, sprint distance), injuries, and participation data during seasons 2007–08 and 2008–09. They find that the recovery time of 72 to 96 hours between 2 matches appears sufficient to maintain the physical performance level. However, they report that injury rates are over 6 times higher when players participated in 2 matches per week compared to only 1 match per week. A study by Bengtsson et al. [1] also demonstrates a strong relation between the recovery time available between successive matches and muscle injury rates. On the other hand, Carling et al. [4] find that injury incidence is not associated to the number of days separating games, and that an interval of 3 days or less between matches did not result in an increased injury rate or number of days lost to injury compared to a longer interval. The impact of fixture congestion on team performance has not so frequently been studied. Bengtsson et al. [1] find no differences in the distribution of matches won, lost or drawn between matches played with a preceding short recovery (≤ 3 days) and matches with long recovery (> 3 days) for domestic league and Champions League matches, except for Europa League matches where a recovery of three or less days did make a significant difference compared with four or more. Lago-Peñas [16] indicates that Spanish Champions League teams did not perform below their normal standard in the domestic league match following their midweek Champions League match. Champions League debutants even performed above their standard in the weekend matches following a Champions League midweek match.

Ekstrand et al. [10] demonstrate that the injury rate in professional football is considerable, amounting to on average 14% of the squad being unavailable due to injury at any point during the season. Hägglund et al. [14] show that the performance of a team is highly influenced by the injury situation; teams that can avoid injuries and keep the players on the pitch are more successful and win more matches. Player rotation has been suggested as a strategy to reducing injury rates as well as maintaining match performance during periods with a lot of matches ([8, 9]). Player rotation inevitably means that teams will not start each game with their best line-up. One common strategy is that the coach will not select some of his key players for the less important match to let them rest for the next, more important match, or let some bench players start in the less important match just after having played the more important match as a compensation. However, which matches are considered important can vary from team to team, or even from country to country. Each country has its own football history, culture and format, leading to different preferences with respect to certain competitions. As an example, it's often conjectured that the English Premier League clubs generally do not treat European club competitions as seriously as clubs from other associations. An explanation for this would be that the income from playing the Champions League is not very attractive for English clubs as they can already earn around £100 million per season on average by just staying in the English Premier League.

In this paper, we develop a method to measure the relative commitment of teams to the UEFA Champions League compared to their domestic league. By studying player rotation and the quality of their starting line-up in the domestic league match before and after the European match, we investigate on which competition clubs from the ten main associations in Europe (Spain, Germany, England, Italy, France, Portugal, Russia, Ukraine, The Netherlands and Bel-

gium) focus. As far as we are aware, no such measure has been described in the literature before. We discuss the main differences with respect to commitment to the European competitions between these countries, and we study the matches that deviate from the common strategy. Finally, we also study the impact on performance in the domestic league of the choices with respect to player rotation. Although some elite professional players also play international matches for their countries, in this study we ignore the impact of international matches on domestic league matches. This choice is to some extent supported by work by Carling et al. [5], who find that in domestic league matches following international matches, the risk injury is similar for players with or without national team obligations.

The paper is organized as follows. Section 2 gives a brief overview of European club football competitions and current scheduling practices. Section 3 provides details about the measure devised to evaluate a club's relative commitment to the Champions League and the method to investigate the impact of these competitions on domestic leagues. In Section 4 we discuss the results, including one-sample t-tests, boxplots, a comparison of associations, and interpretations of the outliers. We discuss the relative importance and attractiveness of the current club competitions and make some concluding remarks in Section 5.

2 European club football competitions

Club football competitions in Europe can be divided into two categories: domestic (national) competitions and European competitions. Domestic competitions include domestic leagues, domestic cups, and the domestic super cups. European club competitions organized by UEFA are the UEFA Champions League, the UEFA Europa League, and the UEFA Super Cup. In this paper, we focus on the domestic leagues and the UEFA Champions League.

2.1 Domestic leagues

Each UEFA national association has its own professional league, which usually consists of several divisions. For the top 10 countries according to the UEFA Country Ranking at the end of the 2014–15 season, the highest division leagues are played by 16 (Belgium, Russia and Ukraine¹), 18 (Germany, Portugal and The Netherlands), or 20 clubs (Spain, England, France and Italy). Each of these divisions are played according to a so-called double round robin tournament, i.e. each team faces each other team twice, possibly followed by play-offs. All of the leagues are played cross-year, except for the seasons in Russia before 2012–13, which were played within a year. Usually, one domestic league match is scheduled per weekend, which includes Friday and/or Monday matches in several cases. Most league schedules feature a few midweek matchdays (i.e. on Tuesdays, Wednesdays and Thursdays) on top of that, though only when no Champions League or Europa League matches are scheduled. We refer to Goossens and Spijksma [13] for a detailed overview of scheduling practices in Europe's most prominent domestic leagues.

2.2 The UEFA Champions League

The UEFA Champions League is the most prestigious club competition in European football, contested yearly by top European clubs. The number of teams each national association can

¹In seasons 2014–15 and 2015–16, Ukraine only had 14 clubs playing the first division league due to political issues

delegate to the UEFA Champions League is determined by the UEFA Country Ranking. The higher an association's ranking, the more of its clubs can compete in the Champions League, and the fewer qualification rounds they face. Currently, the UEFA Champions League consists of a group stage and a knock-out stage. At the group stage, 32 qualified clubs are divided into 8 groups, playing in a double round robin tournament. The first and second ranked teams from each group qualify for the knock-out stage; the third placed team in a group enters the UEFA Europa League (round of 32). In the knock-out stage, 3 two-legged rounds remain before the final, which is contested in a single match. Champions League matches are scheduled to be played on Tuesday and Wednesday evening. Since 2010, the final match has been held on weekend, usually in the final two weeks of May.

3 Material and methods

3.1 Estimating the strength of a starting line-up

A team's starting line-up consists of 11 players (during the match each team can substitute at most 3 players). We assume that the strength of a line-up is determined by the cumulative strength of the players. This is clearly a simplification, as there are indications that player interactions play an important role in team performance (see e.g. [18]). However, the assumption that players do not affect each other's performance is not uncommon in modelling (e.g. [2], [3], [17]). There are many ways to assess the strength (or quality) of a player. One could observe the playing minutes each player obtained, and assume that the better players will be the ones that collect the more playing minutes. However, this quality measure cause a problem when a player has been sustaining a long-time injury, a suspension, or has been transferred in or out during a season. Another measure of player quality would be his salary. Indeed, not surprisingly, there is evidence of a positive pay-performance relationship of soccer players (see e.g. [19]). However, clubs almost never publish data on their individual player's salaries for confidentiality reasons, making the collection of the reliable salary data for all clubs infeasible. Furthermore, a player's salary typically only changes with the closing of a renewed contract with his current team or a transfer to a new team, and consequently, may lag behind on his performance. We opted to use the market value of a player, as published by *Transfermarkt*, as his current strength or contribution to his club. *Transfermarkt* is a German company estimating each active soccer player's market value, which they publish on their website². This estimation is based on an undisclosed algorithm, which includes expert opinions as well as data analysis based on the player's age, position, nationality, past and current performance, injury history, etc. Note that, even though this market value is updated several times per year, it may still not be able to capture a player's form, confidence, fatigue, or other short-term influences.

We assume that a team's strength in a match is reflected by the average market value of its starting 11 players. Given this consideration, the strength of team i in match j can be modeled as:

$$S_{ij} = \frac{\sum_{k=1}^{11} M_{ijk}}{11}, \quad (1)$$

where M_{ijk} is the market value as estimated by *Transfermarkt* of the k -th player in the starting line-up of team i in its match against team j .

²See <https://www.transfermarkt.de/>

3.2 Measuring commitment and performance

As part of the UEFA Home Grown Player Rule (see e.g. [7]), teams can have no more than 25 players in their squad. Which 11 of these players make up the starting line-up of a match is up to the manager/coach. We assume that a coach will select his strongest team for those matches he wants to win most, and give his key players some rest for those matches that are less important, or that may be won with a weaker starting line-up as well. In this way, the strength of the starting line-up reflects the determination a team has to win the game, and the commitment it has towards performing at its best in a competition.

To measure a team's relative commitment towards two competitions, we compare the strength of this team's starting line-ups in two consecutive matches from both competitions. More precisely, we compute the ratio of a team's starting line-up's strength in the Champions League match to the starting line-up's strength in its domestic league match immediately preceding or succeeding this match. We call a club's domestic league match immediately preceding (succeeding) its Champions League match if it is played at most 6 days before (after) the Champions League match. It is important to restrict the analysis to comparing line-ups for matches that are closely followed by each other, because this reflects player rotation, i.e. a deliberate choice by the coach to give rest for one or more key players and to give a playing opportunity to a less valuable player. A ratio larger than 1 indicates that the team has been lining up its stronger players for the Champions League match, rather than for the domestic league match; a ratio smaller than 1 reflects a stronger relative commitment to the domestic league. A one sample t-test is performed to find out whether these ratio's differ significantly from one for each country.

Our method of measuring relative commitment includes a few limitations and assumptions. First of all, we ignore the impact of substitutions with respect to commitment, i.e. we make no distinction between a key player coming on to the pitch for the final half an hour compared to being left out of the selection. We believe commitment is reflected mainly in the starting line-up, and that the impact of substitutes is limited: at most 3 can be done, and usually a substitute plays far less than one third of the match. Second, we largely ignore the impact of injuries, which may have an impact on the measured commitment even though the coach did not make a deliberate choice not to play with his strongest line-up. To some extent, this can be explained by the fact that we don't have data on injuries. In fact, one might question the reliability of any injury, as it regularly happens that a mildly injured player is deemed unfit to play a match of low interest, and seriously injured players are patched together to play (as much as possible) of a highly important match. However, by comparing only matches that are at most 6 days apart, we reduce the impact of injuries on our analysis. Indeed, a team can be forced to use a weaker line-up if a key player is injured, but with this approach, it is likely that this player will still be injured for the next game as well and have no impact on the relative commitment. Thirdly, the strength of a team's starting line-up is clearly impacted by player transfers. A club can considerably enhance its team strength by signing a highly gifted player (which is usually expensive), or oppositely, see their team strength reduced when selling out their best players. The fact that transfers happen only during a pre-season window ending in August, and a mid-season window in January, combined with the fact that we are only comparing matches that are played within 7 days, make that the impact of transfers on our approach is minimal. Finally, we would like to point out that our measure is strictly speaking not a measure for player rotation. Indeed, a team consisting of 25 players with the same market value could rotate maximally after each match, and have a ratio that never deviates

from 1. From the perspective of commitment, this seems fair, as this teams starts each match with a team of equal strength. In practice however, teams have players with very different market values, and we expect our commitment measure to be highly correlated with player rotation.

To assess the impact of commitment on match results, we compare for each team its average points obtained in the league matches that were immediately followed (preceded) by a Champions League match with the average points in league matches that were not immediately followed (preceded) by a Champions League match. We perform a t-test to verify whether these differences differ significantly from zero. Note that we opted not to use any advanced model to predict expected match outcomes; instead we assume that home advantage and opponent strength do not differ between matches played before or after a Champions League confrontation and other matches.

3.3 Data collection

We are interested in the commitment that teams show for the European competitions relative to their domestic league, and whether there are differences between countries. We study the 10 major European associations with respect to club football according to the 2016 UEFA Country Ranking: Spain, Germany, England, Italy, France, Portugal, Russia, Ukraine, Netherlands and Belgium. We have collected the average market value of each side’s starting line-up, the date, and the final score for all matches in these 10 domestic leagues and the UEFA Champions League in seasons 2010–11 to 2014-15 from the *Transfermarkt* website.

4 Results

4.1 Relative commitment to UEFA Champions League

In this section, we compare the quality of the starting line-up of a team in the Champions league and in its domestic league. More precisely, we compute the ratio of a team’s strength (i.e. average market value of the team’s starting players) in its Champions League match to its team strength in the domestic league match immediately preceding this Champions League match (Table 1, left part). We compute a similar ratio for the Champions League match with the domestic league match immediately succeeding it (Table 1, right part). The data has been grouped per country, and a one sample t-test for these ratio’s has been performed to find out whether they differ significantly from one.

Looking at the domestic game preceding the European match, Table 1 shows that teams from France, Germany, Italy, The Netherlands, Russia, Spain, and Ukraine in general give their key players some rest before the Champions League match. On the other hand, England is the only country with a ratio smaller than one (though not significantly). Also when comparing the starting line-up between the Champions League match and the subsequent domestic league match, England is the only association with a ratio significantly smaller than 1. This is a clear indication that the commitment of English teams for the Champions League is lower than their commitment to the Premier League. Teams from from Germany, Italy, The Netherlands, Russia, Spain, and Ukraine tend to give their better players some rest after a Champions League match. Although teams from France do not use their best team line-up in the domestic league match before the Champions League, they do line-up their key players in the domestic match following the Champions League. Recall that Champions League matches are played on Tuesdays and Wednesdays, which are typically closer to the domestic league matches preceding

Table 1: One sample t-test of the ratios of team strength in the UEFA Champions League match to team strength in the domestic league match played before (left) or after (right) it for 10 countries. We test the null hypothesis that the mean ratio is 1 against a two-tailed alternative.

Country	Before UEFA CL					After UEFA CL				
	mean	st.dev.	#obs	t-value	p-value	mean	st.dev.	#obs	t-value	p-value
Belgium	1.030	0.124	24	1.17	0.255	1.009	0.122	24	0.36	0.721
England	0.981	0.185	156	-1.28	0.204	0.964	0.162	155	-2.77	0.006
France	1.094	0.220	101	4.29	<0.001	1.016	0.149	103	1.09	0.278
Germany	1.098	0.363	158	3.40	0.001	1.100	0.385	158	3.26	0.001
Italy	1.060	0.201	108	3.10	0.003	1.064	0.189	105	3.44	0.001
Netherlands	1.050	0.127	36	2.37	0.023	1.077	0.140	36	3.30	0.002
Portugal	1.026	0.201	52	0.92	0.362	1.014	0.206	78	0.62	0.538
Russia	1.061	0.195	49	2.12	0.033	1.046	0.128	43	2.36	0.023
Spain	1.073	0.273	185	3.65	<0.001	1.035	0.242	190	1.98	0.049
Ukraine	1.113	0.153	38	4.55	<0.001	1.120	0.189	33	3.66	0.001

than succeeding them. Hence, the impact of player rotation and rest may be larger for the match preceding the Champions League match; the impact on fatigue of the Champions League match on the following domestic league match would be less pronounced. Finally, for Belgium and Portugal, we find no significant differences in commitment: they appear to start all matches with a team of similar strength.

For a better illustration, we present boxplots of the commitment ratios for each country in Figure 1 (domestic match preceding CL match) and in Figure 2 (domestic match following CL match). Each observation corresponds to a particular Champions League match. Germany and Spain show a lot of dispersion in both boxplots, while several other countries also have some upper/lower outliers. A full list of these outliers is provided in Table 2 and Table 3. Both tables show that the upper outliers are for the most part semifinals and quarterfinals (finals are missing, as they don't have a preceding or succeeding domestic league match). Giving players some rest in the domestic league games before and after these matches makes perfect sense, as these matches are crucial for winning the Champions League, at a stage where hopes of winning it are substantial. The state of affairs in the domestic league also matters. For instance, the two largest upper outliers in Table 2 and Table 3 are both corresponding to semi-finals in season 2012-13. In this season, Bayern managed to secure their *Bundesliga* championship already after 28 matchdays. Dortmund also had already booked a Champions League group stage position in the next season with a huge advantage in points over the 4th ranking team in the *Bundesliga* before the first leg semifinal match against Real Madrid. Therefore, both Bayern and Dortmund could fully focus on the CL semifinals without any worry about their situations in the *Bundesliga* at all. Spanish clubs also appear a lot in the table, and most of their appearances can be explained in a similar way as Germany. However, in the 2011-12 *La Liga* season, an encounter between Real Madrid and Barcelona was scheduled exactly between the two legs of the semi-finals against Bayern and Chelsea respectively. Given the high importance of all the three successive matches, neither Real Madrid nor Barcelona managed to rotate their starting line-up (which is reflected in commitment ratio's close to 1). Finally, both sides were eliminated in the semi-finals. Most of the knock-out entries in Table 2 are second leg matches, which is the more decisive match. UEFA schedules the two legs of the quarterfinals and semifinals in two successive weeks. Hence, the domestic league match right in between both legs is a highly interesting opportunity to give key players rest. This then leads to entries in Table 3 for the first leg match as well as entries for the second leg match in Table 2. The group stage matches

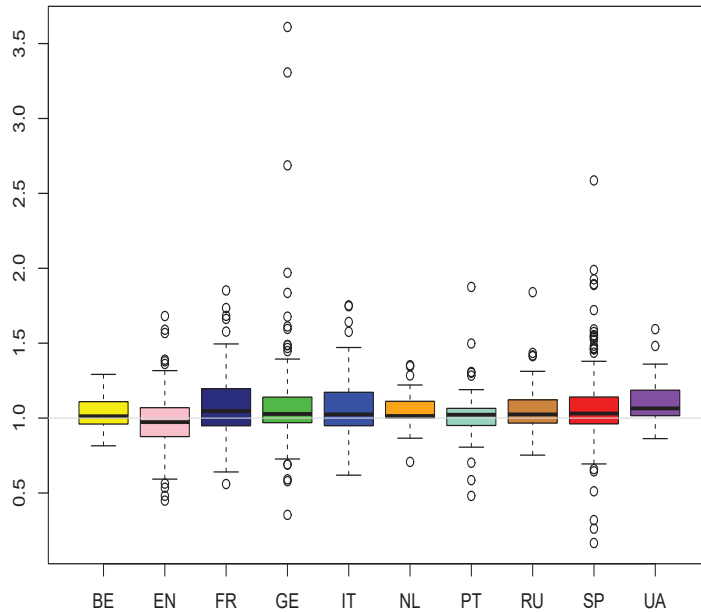


Figure 1: Boxplots of the ratios of the strength of the starting line-up in the UEFA Champions League compared to the domestic league match played before it. Results for the 10 countries.

appearing in Table 2 and Table 3 nearly all concern matches that were crucial for qualification. Table 2 even contains 3 matches from the same group in season 2013-14, which was particularly tough featuring Barcelona, Milan and Ajax.

The second part of Tables 2 and 3 represents the matches where very little relative commitment was displayed. Not surprisingly, nearly all these matches are group stage matches played on the final two matchdays, which concern a team that had already qualified for the next stage or a team that had already been eliminated for the next round. The only match not from the group stage is the match between Juventus and Celtic Glasgow in the round of 16. However, having defeated Celtic 0-3 in the first leg, Juventus preferred to rotate their starting line-up, as their spot in the quarterfinals seems beyond doubt.

4.2 Impact of UEFA Champions League on performance in domestic league

Since we found clear indications of differences in commitment to the Champions League compared to the domestic league, the question arises whether player rotation has an impact on the results in the domestic league. There is some anecdotal evidence of teams underestimating their opponents in the group stage, with serious consequences. For instance, Arsenal faced Shaktar Donetsk in season 2010-11 with a starting line-up which was considerably weaker than in its last domestic league match. They lost the game and with that the group winning position,

Table 2: Outlier ratios of the market value in the Champions League match to the market value in the domestic league match played before it.

Season	Round	Home team	Away team	Ratio
2012-13	Semifinal - second leg	Real Madrid	Borussia Dortmund	3.61
2012-13	Semifinal - second leg	FC Barcelona	FC Bayern München	3.31
2012-13	Quarterfinal - second leg	Borussia Dortmund	FC Málaga	2.69
2012-13	Quarterfinal - second leg	Borussia Dortmund	FC Málaga	2.59
2012-13	Semifinal - second leg	Real Madrid	Borussia Dortmund	1.99
2011-12	Semifinal - second leg	Real Madrid	FC Bayern München	1.97
2010-11	Semifinal - second leg	FC Barcelona	Real Madrid	1.93
2012-13	Round of 16 - second leg	FC Barcelona	AC Milan	1.90
2010-11	Semifinal - first leg	Real Madrid	FC Barcelona	1.89
2014-15	Quarterfinal - second leg	FC Bayern München	FC Porto	1.88
2014-15	Round of 16 - second leg	FC Chelsea	FC Paris Saint-Germain	1.85
2012-13	Group stage	Zenit Sint-Petersburg	FC Málaga	1.84
2013-14	Quarterfinal - second leg	FC Bayern München	Manchester United	1.84
2014-15	Semifinal - second leg	Real Madrid	Juventus FC	1.75
2013-14	Group stage	Ajax	AC Milan	1.75
2014-15	Group stage	Benfica Lissabon	AS Monaco	1.73
2010-11	Semifinal - first leg	Real Madrid	FC Barcelona	1.72
2014-15	Quarterfinal - second leg	AS Monaco	Juventus FC	1.68
2011-12	Group stage	Otelul Galati	Manchester United	1.68
2013-14	Quarterfinal - first leg	Manchester United	FC Bayern München	1.68
2011-12	Group stage	Real Madrid	Olympique Lyon	1.66
2010-11	Group stage	AS Roma	FC Bayern München	1.64
2014-15	Semifinal - first leg	FC Barcelona	FC Bayern München	1.61
2013-14	Group stage	Shakhtar Donetsk	Bayer 04 Leverkusen	1.59
2014-15	Group stage	BATE Borisov	Shakhtar Donetsk	1.59
2012-13	Round of 16 - second leg	FC Málaga	FC Porto	1.59
2012-13	Group stage	Manchester United	Galatasaray Istanbul	1.59
2012-13	Quarterfinal - second leg	FC Barcelona	FC Paris Saint-Germain	1.58
2014-15	Round of 16 - second leg	Borussia Dortmund	Juventus FC	1.58
2012-13	Semifinal - first leg	FC Bayern München	FC Barcelona	1.57
2013-14	Semifinal - second leg	FC Chelsea	Atlético Madrid	1.57
2012-13	Round of 16 - second leg	Manchester United	Real Madrid	1.55
2013-14	Round of 16 - first leg	FC Schalke 04	Real Madrid	1.55
2010-11	Semifinal - second leg	FC Barcelona	Real Madrid	1.54
2012-13	Group stage	FC Barcelona	Spartak Moskou	1.54
2010-11	Quarterfinal - first leg	Real Madrid	Tottenham Hotspur	1.52
2011-12	Group stage	FC Porto	Shakhtar Donetsk	1.50
2011-12	Semifinal - first leg	FC Bayern München	Real Madrid	1.49
2012-13	Group stage	FC Valencia	FC Bayern München	1.49
2011-12	Group stage	FC Barcelona	AC Milan	1.48
2010-11	Quarterfinal - second leg	Shakhtar Donetsk	FC Barcelona	1.48
2014-15	Quarterfinal - first leg	Atlético Madrid	Real Madrid	1.47
2013-14	Round of 16 - first leg	FC Schalke 04	Real Madrid	1.47
2010-11	Round of 16 - second leg	Real Madrid	Olympique Lyon	1.46
2013-14	Group stage	AC Milan	FC Barcelona	1.46
2012-13	Semifinal - first leg	FC Bayern München	FC Barcelona	1.45
2010-11	Quarterfinal - first leg	FC Barcelona	Shakhtar Donetsk	1.44
2013-14	Group stage	Atlético Madrid	Zenit Sint-Petersburg	1.44
2012-13	Group stage	FC Málaga	Zenit Sint-Petersburg	1.42
2013-14	Group stage	CSKA Moskou	Manchester City	1.41
2014-15	Round of 16 - first leg	FC Paris Saint-Germain	FC Chelsea	1.39
2014-15	Group stage	CSKA Moskou	Manchester City	1.38
2010-11	Round of 16 - second leg	FC Barcelona	FC Arsenal	1.36
2013-14	Group stage	FC Barcelona	Ajax	1.35
2010-11	Group stage	AC Milan	Ajax	1.35
2014-15	Quarterfinal - first leg	FC Porto	FC Bayern München	1.31
2014-15	Round of 16 - first leg	FC Basel 1893	FC Porto	1.30
2011-12	Group stage	Ajax	GNK Dinamo Zagreb	1.28
2010-11	Group stage	FC Arsenal	SC Braga	1.28
2010-11	Group stage	Real Madrid	Ajax	0.71
2010-11	Group stage	Shakhtar Donetsk	SC Braga	0.70
2013-14	Group stage	Steaua Boekarest	FC Schalke 04	0.69
2014-15	Group stage	Benfica Lissabon	Bayer 04 Leverkusen	0.69
2011-12	Group stage	Real Madrid	GNK Dinamo Zagreb	0.66
2012-13	Group stage	Real Madrid	Ajax	0.64
2012-13	Group stage	Borussia Dortmund	Manchester City	0.59
2014-15	Group stage	FC Porto	Shakhtar Donetsk	0.59
2012-13	Group stage	HSC Montpellier	FC Schalke 04	0.58
2011-12	Group stage	Olympiakos Piraeus	FC Arsenal	0.56
2012-13	Group stage	BATE Borisov	LOSC Lille	0.56
2012-13	Group stage	Olympiakos Piraeus	FC Arsenal	0.53
2011-12	Group stage	Ajax	Real Madrid	0.51
2010-11	Group stage	Shakhtar Donetsk	FC Arsenal	0.48
2014-15	Group stage	Benfica Lissabon	Bayer 04 Leverkusen	0.48
2010-11	Group stage	FC Chelsea	MSK Zilina	0.45
2011-12	Group stage	Manchester City	FC Bayern München	0.35
2010-11	Group stage	FC Barcelona	Roebin Kazan	0.32
2012-13	Group stage	FC Barcelona	Benfica Lissabon	0.26
2011-12	Group stage	FC Barcelona	BATE Borisov	0.17

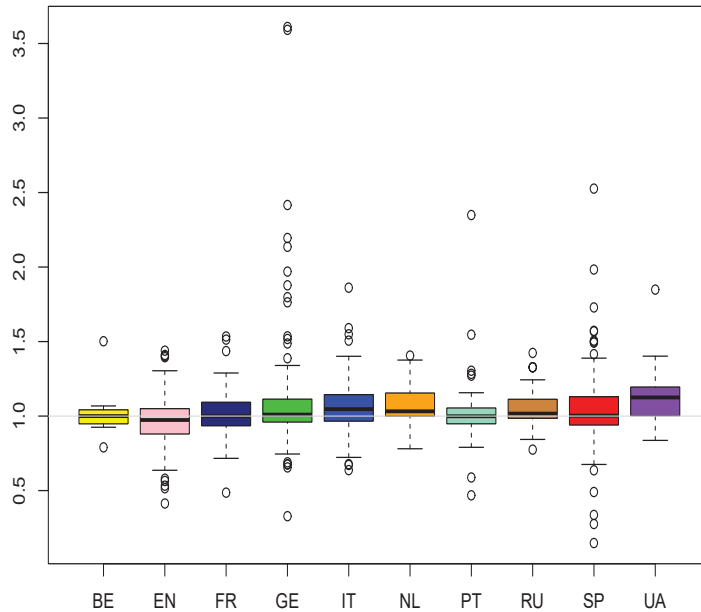


Figure 2: Boxplots of the ratios of the strength of the starting line-up in the UEFA Champions League compared to the domestic league match played after it. Results for the 10 countries.

resulting in a confrontation with Barcelona in the round of 16, and a subsequent elimination.

Table 4 (left part) shows the difference between the average points obtained in domestic league matches played before a Champions League match and the average points obtained in domestic league matches that were not followed by a Champions League confrontation, for teams from each of the 10 major European football leagues. Table 4 (right part) does the same for domestic league matches following a Champions League match. Table 4 shows that most countries where teams start the domestic match after their Champions League confrontation with a weaker line-up get away with this in term of domestic league performance. Teams from Germany and Spain, however, generally perform significantly worse in domestic league matches following a Champions League³. Several countries (e.g. The Netherlands, Russia) also show a negative, however not significant, impact of the Champions League on domestic league performance. The impact on the domestic league match preceding the Champions League is much less pronounced, and even significantly positive in the case of Spain. No significant impact on the domestic results is found in countries where the commitment towards the Champions League is not different from (i.e. Belgium, Portugal), or even falls behind (i.e. England) the commitment the domestic league receives.

³Notice that for a one-sided alternative, the p-value has to be divided by 2

Table 3: Outlier ratios of the market value in the Champions League match to the market value in the domestic league match played after it.

Season	Round	Home team	Away team	Ratio
2012-13	Semifinal - first leg	Borussia Dortmund	Real Madrid	3.61
2012-13	Semifinal - first leg	FC Bayern München	FC Barcelona	3.59
2012-13	Quarterfinal - first leg	FC Málaga	Borussia Dortmund	2.53
2012-13	Quarterfinal - first leg	FC Málaga	Borussia Dortmund	2.42
2014-15	Quarterfinal - first leg	FC Porto	FC Bayern München	2.35
2012-13	Quarterfinal - second leg	Juventus FC	FC Bayern München	2.20
2011-12	Semifinal - second leg	Real Madrid	FC Bayern München	2.14
2012-13	Semifinal - first leg	Borussia Dortmund	Real Madrid	1.98
2011-12	Semifinal - first leg	FC Bayern München	Real Madrid	1.97
2012-13	Semifinal - second leg	FC Barcelona	FC Bayern München	1.88
2014-15	Semifinal - second leg	Real Madrid	Juventus FC	1.86
2010-11	Quarterfinal - first leg	FC Barcelona	Shakhtar Donetsk	1.85
2012-13	Semifinal - second leg	Real Madrid	Borussia Dortmund	1.80
2013-14	Quarterfinal - first leg	Manchester United	FC Bayern München	1.76
2010-11	Semifinal - first leg	Real Madrid	FC Barcelona	1.73
2010-11	Group stage	AC Milan	Real Madrid	1.59
2012-13	Quarterfinal - second leg	FC Barcelona	FC Paris Saint-Germain	1.57
2010-11	Semifinal - first leg	Real Madrid	FC Barcelona	1.57
2010-11	Round of 16 - first leg	AS Roma	Shakhtar Donetsk	1.55
2014-15	Round of 16 - first leg	FC Basel 1893	FC Porto	1.55
2013-14	Group stage	SSC Napoli	Borussia Dortmund	1.53
2014-15	Group stage	AS Monaco	Benfica Lissabon	1.53
2013-14	Group stage	FC Schalke 04	FC Basel 1893	1.52
2012-13	Quarterfinal - first leg	FC Paris Saint-Germain	FC Barcelona	1.51
2011-12	Round of 16 - first leg	AC Milan	FC Arsenal	1.51
2011-12	Semifinal - second leg	FC Barcelona	FC Chelsea	1.51
2013-14	Group stage	RSC Anderlecht	FC Paris Saint-Germain	1.50
2012-13	Semifinal - first leg	FC Bayern München	FC Barcelona	1.50
2012-13	Quarterfinal - first leg	FC Paris Saint-Germain	FC Barcelona	1.49
2014-15	Round of 16 - second leg	Atlético Madrid	Bayer 04 Leverkusen	1.49
2013-14	Semifinal - second leg	FC Chelsea	Atlético Madrid	1.44
2014-15	Quarterfinal - first leg	Juventus FC	AS Monaco	1.44
2012-13	Group stage	Zenit Sint-Petersburg	RSC Anderlecht	1.42
2010-11	Round of 16 - second leg	FC Schalke 04	FC Valencia	1.42
2010-11	Group stage	FC Valencia	Manchester United	1.41
2010-11	Group stage	AC Milan	Ajax	1.41
2013-14	Semifinal - first leg	Atlético Madrid	FC Chelsea	1.40
2013-14	Quarterfinal - first leg	FC Paris Saint-Germain	FC Chelsea	1.39
2014-15	Group stage	FC Bayern München	Manchester City	1.39
2011-12	Group stage	Zenit Sint-Petersburg	Shakhtar Donetsk	1.33
2011-12	Group stage	Trabzonspor	CSKA Moskou	1.33
2011-12	Group stage	FC Porto	Shakhtar Donetsk	1.30
2010-11	Group stage	FC Arsenal	SC Braga	1.28
2014-15	Group stage	FC Porto	BATE Borisov	1.27
2014-15	Group stage	Borussia Dortmund	RSC Anderlecht	0.79
2013-14	Group stage	FC Porto	Zenit Sint-Petersburg	0.78
2014-15	Group stage	Benfica Lissabon	Bayer 04 Leverkusen	0.69
2012-13	Group stage	HSC Montpellier	FC Schalke 04	0.68
2011-12	Group stage	FC Viktoria Pilsen	AC Milan	0.68
2012-13	Group stage	Borussia Dortmund	Manchester City	0.67
2011-12	Group stage	FC Internazionale	CSKA Moskou	0.67
2013-14	Group stage	Steaua Boekarest	FC Schalke 04	0.65
2012-13	Round of 16 - second leg	Juventus FC	Celtic Glasgow	0.64
2011-12	Group stage	Real Madrid	GNK Dinamo Zagreb	0.64
2014-15	Group stage	FC Porto	Shakhtar Donetsk	0.59
2012-13	Group stage	Galatasaray Istanbul	Manchester United	0.58
2011-12	Group stage	Olympiakos Piraeus	FC Arsenal	0.56
2010-11	Group stage	Shakhtar Donetsk	FC Arsenal	0.53
2012-13	Group stage	Olympiakos Piraeus	FC Arsenal	0.52
2011-12	Group stage	Ajax	Real Madrid	0.49
2012-13	Group stage	BATE Borisov	LOSC Lille	0.49
2014-15	Group stage	Benfica Lissabon	Bayer 04 Leverkusen	0.47
2010-11	Group stage	FC Chelsea	MSK Zilina	0.41
2010-11	Group stage	FC Barcelona	Roebin Kazan	0.34
2011-12	Group stage	Manchester City	FC Bayern München	0.33
2012-13	Group stage	FC Barcelona	Benfica Lissabon	0.28
2011-12	Group stage	FC Barcelona	BATE Borisov	0.15

5 Conclusion

In this paper, we devised a new measure to compare the commitment of a football club to various competitions. This measure is based on a comparison of the value of the starting line-up in two successive matches. We analyzed the commitment to the UEFA Champions League matches relative to the matches from the domestic league for teams from the 10 major European associations in seasons 2010-11 to 2014-15. At the same time, we studied the impact of the Champions League competition on the results in the domestic leagues.

We found that the UEFA Champions League is clearly the most attractive club football

Table 4: Paired-sample t-test for the difference between the average points obtained in domestic league matches played before (left)/after (right) a Champions League match and the average points obtained in domestic league matches not followed/preceded by a Champions League match for 10 countries. We test the null hypothesis that the mean difference is zero against a two-tailed alternative.

Country	Before UEFA CL					After UEFA CL				
	mean	st.dev.	#obs	t-value	p-value	mean	st.dev.	#obs	t-value	p-value
Belgium	0.105	0.789	4	0.267	0.807	-0.091	0.190	4	-0.953	0.411
England	-0.061	0.697	20	-0.393	0.699	0.063	0.528	20	0.531	0.602
France	0.018	0.534	13	0.124	0.903	0.050	0.304	13	0.592	0.565
Germany	0.059	0.442	17	0.554	0.587	-0.206	0.287	17	-2.948	0.009
Italy	-0.075	0.375	13	-0.725	0.482	0.006	0.373	13	0.062	0.951
Netherlands	0.058	0.350	6	0.403	0.704	-0.246	0.522	6	-1.153	0.301
Portugal	-0.160	0.575	12	-0.964	0.356	0.030	0.285	12	0.370	0.719
Russia	-0.324	0.339	6	-2.339	0.066	-0.286	0.538	6	-1.305	0.249
Spain	0.216	0.354	19	2.657	0.016	-0.207	0.468	19	-1.923	0.070
Ukraine	-0.176	0.338	6	-1.277	0.258	-0.002	0.531	6	-0.008	0.994

competition for all the main UEFA associations, except England. This can be explained by the fact that the *Premier League* is the richest domestic league in the world, and at the same time a highly contested competition. For English clubs, the earnings from the Champions League are not as attractive as for clubs in other countries. Teams from the other countries (apart from Belgium and Portugal) in general side-line their key players in the domestic league in order to have them fully rested at the start of the Champions League, although there is a lot of variation between clubs and seasons. This decision is rewarding, since these teams generally do not suffer from worse results in their domestic league, Germany excepted.

Acknowledgement

The authors wish to thank Frederik Christiaens for his contribution with collecting the data on market values and match results.

References

- [1] Bengtsson, H., Ekstrand, J. and Hagglund, M. (2013). Muscle injury rates in professional football increase with fixture congestion: an 11-year follow-up of the UEFA Champions League injury study. *British Journal of Sports Medicine*, **47**:743–747.
- [2] Boon, B.H. and Sierksma, G. (2013). Team formation: Matching quality supply and quality demand. *European Journal of Operational Research* **148**:277–292.
- [3] Budak, G., Kara, İ., İc, Y.T. and Kasimbeyli, R. (2017). New mathematical models for team formation of sports clubs before the match. *Central European Journal of Operations Research*. In press, <https://doi.org/10.1007/s10100-017-0491-x>.
- [4] Carling, C., Orhant, E. and LeGall, F. (2010). Match Injuries in Professional Soccer: Inter-Seasonal Variation and Effects of Competition Type, Match Congestion and Positional Role. *International Journal of Sports Medicine* **31**(4):271–276.
- [5] Carling, C., McCall, A., Le Gall, F. and Dupont, G. (2015). The impact of in-season national team soccer play on injury and player availability in a professional club. *Journal of Sports Sciences*, **33**(17):1751–1757.

- [6] Carling, C., McCall, A., Le Gall, F. and Dupont, G. (2015). What is the extent of exposure to periods of match congestion in professional soccer players? *Journal of Sports Sciences*, **33**(20):2116–2124.
- [7] Dalziel, M., Downward, P., Parrish, R., Pearson, G., Semens, A. (2013). Study on the Assessment of UEFA's "Home Grown Player Rule". University of Liverpool and Edge Hill University.
- [8] Dupont, G., Nedelec, M., McCall, A., McCormack, D., Berthoin, S. and Wisloff, U. (2010). Effect of 2 soccer matches in a week on physical performance and injury rate., *The American journal of sports medicine*, **38**:1752–1758.
- [9] Ekstrand, J., Waldén, M. and Häggglund, M. (2004), A congested football calendar and the well-being of players: correlation between match exposure of European footballers before the World Cup 2002 and their injuries and performances during that World Cup., *British Journal of Sports Medicine*, **38**:493–497.
- [10] Ekstrand, J., Häggglund, M., Kristenson, K., Magnusson, H. and Waldén, M. (2013). Fewer ligament injuries but no preventive effect on muscle injuries and severe injuries: an 11-year follow-up of the UEFA Champions League injury study. *British Journal of Sports Medicine* **47**:732–737.
- [11] Folgado, H., Duarte, R., Marques, P. and Sampaio J. (2015) The effects of congested fixtures period on tactical and physical performance in elite football. *Journal of Sports Sciences*, **33** (12):1238–1247.
- [12] Gibson, O. (2012). European clubs clinch deal with Uefa to cut international matches. *The Guardian*, Feb. 28, 2012.
- [13] Goossens, D. and Spieksma, F. (2012), Soccer schedules in Europe: An overview., *Journal of Scheduling*, **15**(5):641–651.
- [14] Häggglund, M., Waldén, M., Magnusson, H., Kristenson, K., Bengtsson H. and Ekstrand, J. (2013). Injuries affect team performance negatively in professional football: an 11-year follow-up of the UEFA Champions League injury study. *British Journal of Sports Medicine* **47**:738–742
- [15] Ispirlidis, I., Fatouros, I.G., Jamurtas, A.Z., Nikolaidis, M.G., Michailidis, I., Douroudos, I., Margonis, K., Chatzinikolaou, A., Kalistratos, E., Katrabasas, I., Alexiou, V. and Taxildaris, K. (2008), Time-course of changes in inflammatory and performance responses following a soccer game. *Clinical Journal of Sport Medicine*, **18**(5):423–431.
- [16] Lago-Penã, C. (2009), Consequences of a busy soccer match schedule on team performance: empirical evidence from Spain., *International SportMed Journal*, **10**(2):86–92.
- [17] Özceylan, E. (2016) A mathematical model using AHP priorities for soccer player selection: a case study. *South African Journal of Industrial Engineering* **27**(2):190–205.
- [18] Ramos, J., Lopes, R.J., Marques, P. and Araújo, D. (2017) Hypernetworks Reveal Compound Variables That Capture Cooperative and Competitive Interactions in a Soccer Match. *Frontiers in Psychology* **8**:1379.
- [19] Torgler, B. and Schmidt, S.L. (2007). What shapes player performance in soccer? Empirical findings from a panel analysis. *Applied Economics* **39**:2355–2369.

Play-by-play data analysis for team managing in basketball

Luca Grassetti¹, Ruggero Bellio¹, Giovanni Fonseca¹, and Paolo Vidoni¹

Department of Economics and Statistics, University of Udine, Italy
Via Tomadini, 30/A – 33100 Udine (UD)
`paolo.vidoni@uniud.it`

Abstract

The sports analytics literature regarding basketball is vast but the analyses based on disaggregated data, such as the play-by-play match data, are not very common. The analysis of the whole sequence of play-by-play match events has an undeveloped potential, yet most of the available methods focus on the final match results. The present work illustrates a model-based strategy for the analysis of the match progress, built upon the literature of Adjusted Plus Minus for the evaluation of player efficiency. This approach is extended in two main directions. The first extension consists in the adoption of a response variable which considers the most relevant events in the game, and not only the number of scored points. This offers some useful advantages, including the possibility of obtaining separate estimates about different complementary aspects. Further, next to player efficiency effects, the efficiency of five-man lineups is estimated. The model fitting procedure follows an empirical Bayes approach, which provides a suitable regularization. For the empirical analysis, we consider a dataset regarding the Italian Basketball League (Serie A1), focusing on the matches of the first round of the current championship 2018/2019. The dataset collects the play-by-play information along with the matches box scores, which are made available by the league website (www.legabasket.it). The results of the analysis could support the decision-making process of team management, and some illustrations on this point are provided.

Keywords: Basketball Analytics, Statistical Model, Play-by-play data, Web-crawling, Data-driven decision process.

1 Introduction

Basketball has a long history of describing the performance of players using statistics. Box scores data, such as points, assists and rebounds, are regularly provided by newspapers and websites in order to highlight the impact of a player in a match or during a full season. Nowadays, data are collected in real time during the game and play-by-play outcomes are readably available on the web. Team managers and coaches use such information to build effective lineups, and from this point of view, the need for more specialized measures arises. In the first decade of our century, regression-based player performance indices were proposed, e.g. Adjusted Plus-Minus (APM) method [9] and Regularized versions of it (RAPM) [see 10, 4]. These kind of metrics are also our starting point. These measures are computed using play-by-play data aggregated in shifts, where a shift is defined as a period of playing time without any substitution for either team. The APM is obtained by fitting a linear regression model where the response variable is the point differential for each shift, computed as the difference between the average points scored by the home team and the away team. In either case, the average is with respect to the number of possessions for each team. The regressors are instead given by signed dummy variables for every player involved in the shift. This usually gives a large number of terms, a typical setting where regularization is called for. To this end, RAPM performs regularized estimation of the model parameters using ridge regression, providing estimates with better properties. A useful feature of the method is that the estimated coefficients can be interpreted

as net player efficiency measures, i.e. they are adjusted for the other players on the field for both teams.

The overall aim of the work is to deepen the analysis of play-by-play data, providing some useful information for team management. In particular, we build upon the RAPM setting, and we extend it to encompass the analysis of the performance of entire lineups, defined as five-man units on the field for a given team. The idea underlying this extension is that player performances may depend on the interaction with teammates and on the counteraction of the opposite lineup on the field. From a statistical viewpoint, regularization is even more crucial for the estimation of lineup efficiency than for individual players, since the dimensions involved are higher. Beyond ridge regression, other approaches may be used for the task, such as empirical Bayes, boosting or full Bayes; see, for instance, [3]. Here we adopt an empirical Bayes approach, which turns out to be quite convenient.

Another distinctive feature of our proposal is the adoption of a more general and complete index rating, called *score* hereafter. This is the response variable that will be used in the regression models for player and lineup efficiency, and it is obtained as a suitable modification of the efficiency index commonly used in basketball; details are given in Section 2. A further aspect investigated in this paper is the multi-dimensionality of such score measure, which may be usefully disentangled in three distinct contributions. In particular, the contributions pertain to outside scoring capacity (three-pointers and mid-range shots), inside scoring capacity (lay-ups, dunks and free shots) and complementary abilities (such as assists, rebounds, blocks, steals), respectively.

The paper is organized as follows. Section 2 illustrates the data used for the analysis along with a few details about the data wrangling process. Some data exploration is also given. Section 3 introduces the model adopted for the estimation of lineup and player efficiency, together with possible usages of the results. Applications to the case study of interest are also provided. Finally, Section 4 contains a brief discussion and some concluding remarks.

2 Data wrangling and data exploration

In this section, the data wrangling process is briefly described and the results of a preliminary data analysis are presented. The data analyzed concern the matches of the first round of the championship 2018/2019 of the Italian Basketball League (Serie A1). The league website provides the play-by-play information along with the box scores.

2.1 Data wrangling

In order to gather the required data from the Italian Basketball League website (www.legabasket.it), we use the R statistical software [7] and, in particular, some specific add-on packages such as `rvest` [11], `scrapeR` [1] and `Rcrawler` [6]. For every single match, both the box scores data and the play-by-play information are collected. A play is defined as an event during the possession involving a positive or negative value for the attacking team, and deemed as the most relevant for the outcome of the game. In particular, as introduced in [5], the values of the events used in the computation of the outcome measure are (numeric contribution to the score in brackets): missed free-throw, turnover or offensive foul (-1); missed shot (-0.5); assist (0.5); steal, offensive or defensive rebound, block, scored free-throw or received foul (1); scored shot (2); scored three-pointer (3). The score is assigned to the offensive team, and the opposite score is assigned to the defensive team. The more traditional outcome given by the points scored in each play is also gathered. Moreover, other features are collected, such as

information concerning the time of the event and game status. In particular, the last piece of information has been employed to remove all the *garbage time* plays from the data, referred to game instances with a point gap larger than 20. Finally, the box score information has been used to identify the five-man unit involved in each play. This information is essential to aggregate the plays in shifts, as required for subsequent analyses.

2.2 Data exploration

The cleansed dataset consists of 3849 shifts and 19943 possessions, from 120 matches played by the 16 teams of the Italian Serie A. The total number of lineups in the dataset is 1886 and the players involved in the games are 212. A few players changed team during the season. The effect of these changes is negligible, hence they were considered as different players.

The following preliminary analyses aim at describing the characteristics of the different teams in terms of lineup and players usage. Table 1 shows that some teams present a homogeneous distribution for the number of possessions at the lineup level, such as Milano and Reggio Emilia. For other teams, such as Varese and Pesaro, the number of possessions varies substantially among different lineups.

Table 1: Summary statistics for the number of possessions by lineups.

Team	No. of	1 st			3 rd			
	Lineups	Min.	quart.	Mean	Median	S.D.	quart.	Max.
Avellino	91	1	4.00	25.10	12.00	50.65	23.50	374
Bologna	121	1	5.00	20.56	12.00	30.67	25.00	252
Brescia	115	1	5.00	21.61	11.00	32.02	25.50	238
Brindisi	73	1	6.00	37.58	14.00	69.73	39.00	431
Cantù	100	1	5.00	24.78	9.50	55.23	18.25	459
Cremona	80	1	7.00	33.67	13.00	54.73	37.00	369
Milano	172	1	4.00	14.90	9.50	15.82	18.00	89
Pesaro	60	1	5.75	40.63	10.50	116.91	33.00	888
Pistoia	99	1	6.50	24.00	11.00	62.07	21.00	603
R. Emilia	155	1	5.00	15.04	9.00	18.05	18.50	124
Sassari	180	1	3.00	14.17	7.00	27.47	15.00	224
Torino	137	1	6.00	19.19	11.00	26.03	21.00	214
Trentino	128	1	5.00	19.52	13.00	22.15	29.25	171
Trieste	139	1	5.00	18.22	10.00	23.59	20.50	140
Varese	65	1	5.00	39.35	16.00	118.83	42.00	957
Venezia	171	1	3.50	13.09	7.00	24.85	12.50	244

Table 2 shows that teams generally use their players in different ways, so that the team-specific distributions of the number of player possessions vary greatly. For instance, the minimum number of possessions exceeds 100 possessions for some teams, otherwise, this number is much lower. The teams presenting the largest variability in the number of possessions are Brindisi, Pesaro and Varese, which exhibit the same behaviour also for what concerns lineup usage.

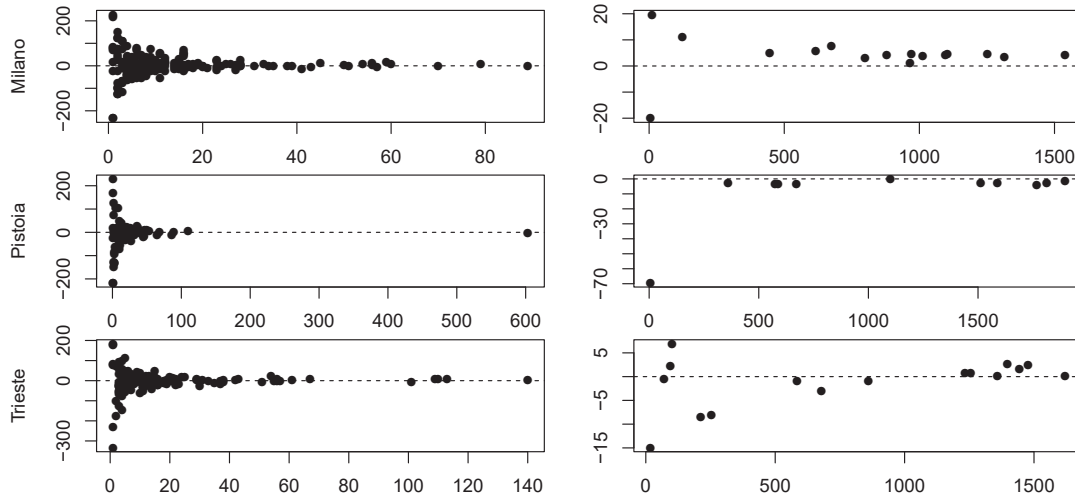
Figure 1 displays the average performance of players and lineups for three teams, measured as the average score differential between home and away teams, as already defined in Section 1. The averages are represented as a function of the number of possessions played, and they

Table 2: Summary statistics for the number of possessions by players.

Team	No. of Players	Min.	Mean	Median	S.D.	Max.
Avellino	14	3	815.71	657.50	666.79	1927
Bologna	13	21	956.92	880.00	608.99	1761
Brescia	13	2	955.77	1001.00	505.51	1730
Brindisi	12	3	1142.92	1128.50	805.88	2223
Cantù	12	117	1032.50	879.00	791.59	2010
Cremona	12	5	1122.50	1307.00	742.83	2055
Milano	16	5	800.62	922.50	461.15	1539
Pesaro	10	98	1219.00	1383.00	845.07	2188
Pistoia	11	6	1080.00	1100.00	668.12	1897
R. Emilia	18	1	647.50	652.50	430.69	1276
Sassari	13	105	981.15	1122.00	606.12	1830
Torino	14	104	938.93	883.50	498.43	1911
Trentino	12	135	1041.25	1219.00	505.45	1648
Trieste	16	18	791.25	769.00	602.47	1620
Varese	12	1	1065.83	821.50	810.66	2224
Venezia	14	17	799.29	745.00	501.91	1482

are multiplied by 100. This is customary in the APM literature, since an NBA game is roughly made of 100 possessions.

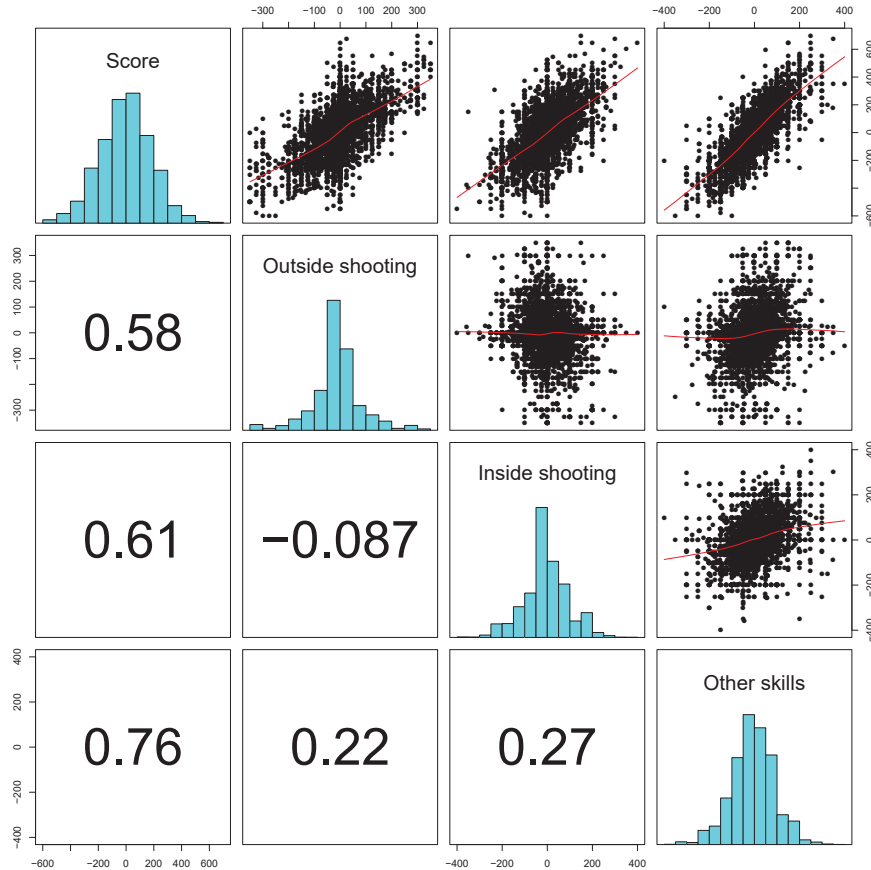
Figure 1: Average score of each lineup (left panels) and each player (right panel) for three teams as a function of the number of possessions.



A further interesting visualization, given in Figure 2, is about the relationship between the overall score and its three components, corresponding to outside shooting, inside shooting and complementary skills. While the linear relationship between these score contributions and the total measure is apparent, the pairwise relationship among the three components is weak. This

suggests that the different kinds of contributions may actually correspond to different aspects of player (lineup) performances.

Figure 2: Relationship between the overall score and its three components; pairwise correlations below the main diagonal.



3 Model-based analysis

The regression models for lineup and player effects are defined as follows. For the case of lineups, the starting point is a linear model for the score y_t of shift t , with $t = 1, \dots, T$,

$$y_t = \beta_0 + \mu_{h[t]} - \mu_{a[t]} + \varepsilon_t, \quad (1)$$

being ε_t a normal error term. Here we consider the data with all the matches of the first round, for which $T = 3849$, as stated in Section 2.2. The response in (1) is the aforementioned overall score, given by the difference between the mean outcome of the home team and the mean outcome of the away time for each shift. Where only one team produces a score in a given shift, the mean outcome of the other team is replaced by the grand mean over the entire sample,

as customary in the APM analysis. The vector of lineup effects $\boldsymbol{\mu}$ has length N , equal to the total number of lineups ($N = 1886$ in the data at hand). The notation $h[t]$ and $a[t]$ defines the lineup for the home and away team for shift t , respectively, i.e. $h[t]$ and $a[t]$ take a value in the set $\{1, \dots, N\}$.

The model for player effects is very similar, with the difference that each shift entails ten different players, five for each team, rather than just two teams. Equation (1) is then replaced by

$$y_t = \beta_0 + \sum_{j=1}^5 \gamma_{h_j[t]} - \sum_{j=1}^5 \gamma_{a_j[t]} + \eta_t, \quad (2)$$

with η_t denoting a normal error term, and where $\boldsymbol{\gamma}$ is the vector of player effects, with length M (with $M = 212$ in the data at hand). Here the two functions $h_j[t]$ and $a_j[t]$ identify the j -th player involved in shift t , for home and away team respectively, so that each of these functions takes value in the set $\{1, \dots, M\}$. This model specification is very similar to the one adopted by [2] in a different framework.

3.1 Estimation of lineup and player effects

The estimation of two vectors $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ requires a regularization technique, due to their large dimensions. Moreover, it should be noted that in either model each shift is computed over a certain number of possessions, with the implication that the actual sample size is much larger than the number of shifts, so that suitable observation weights must be employed in the estimation procedure.

The RAPM method of [10] is based on the estimation of $\boldsymbol{\gamma}$ based on ridge regression, but here we adopt instead an empirical Bayes approach, which achieves regularization by treating the lineup (or player) effects as normal random effects. The two methods are indeed related (e.g. [3]) and essentially differ only in the approach used to select the regularization parameter. Whereas for ridge regression the tuning parameter is usually estimated by cross-validation, in the empirical Bayes approach the variance of random effects is estimated by REML. For the data at hand, the two methods give very similar results, with REML providing slightly less shrinkage. At the same time, the empirical Bayes approach allows for straightforward inclusion of additional covariates, which seems a useful possibility worth considering. The results that follow have been obtained by means of the `hglm` R package [8].

Figure 3 is the estimated counterpart of Figure 1 since it reports the estimated effects for lineups and players concerning the same three teams. Some adjustments are apparent since the estimates take into account the different players and lineups which are simultaneously present on the field. The shrinkage provided by regularization is also noteworthy, resulting in effects for small number of possessions shrunk towards zero.

3.2 Choice of response variable

The entire analysis has been developed considering the overall score, which seems more comprehensive and informative than the number of points used in the original APM and RAPM methodology. Figure 4, left panel, visualizes the relationship between the two variables. A strong linear relationship is apparent (correlation around 0.9), but the marginal distribution of the score variable is less discrete, thus better suited for linear regression analyses. The right panel of Figure 4 displays instead the estimated fitted values for model (1) with respect to the same quantity computed considering the response as given only by the number of points, as in

the classic RAPM. Linearity is again very strong, confirming that the two choices are largely comparable in terms of fit.

Figure 3: Estimated effects of each lineup (left panels) and each player (right panel) for three teams as a function of the number of possessions.

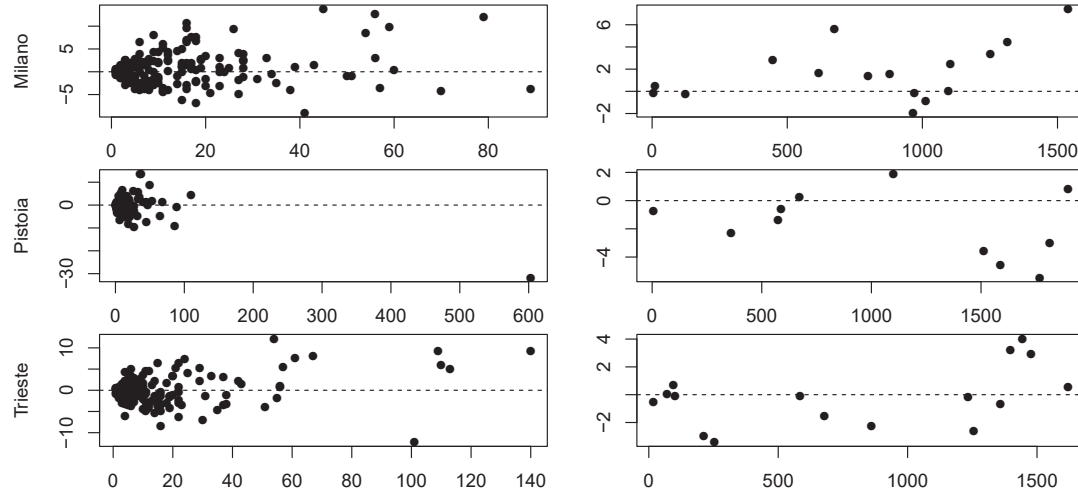
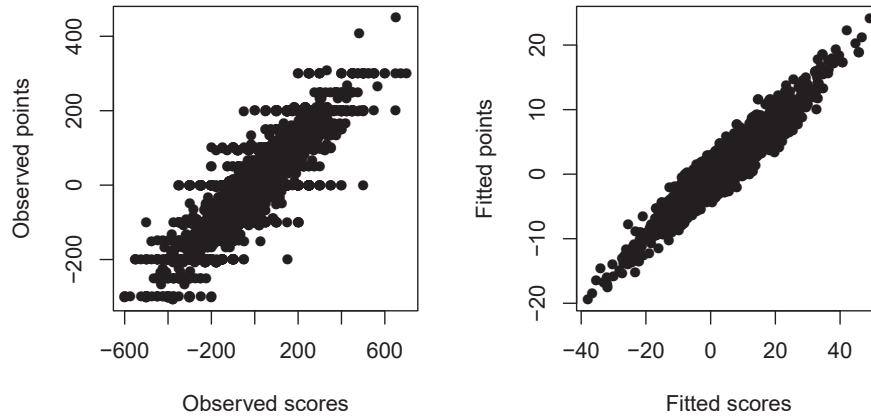


Figure 4: Observed points and scores for the shift data (left panel), and fitted points and values based on the estimated model for lineup effects (right panel).



3.3 Analysis of estimated lineup and player effects

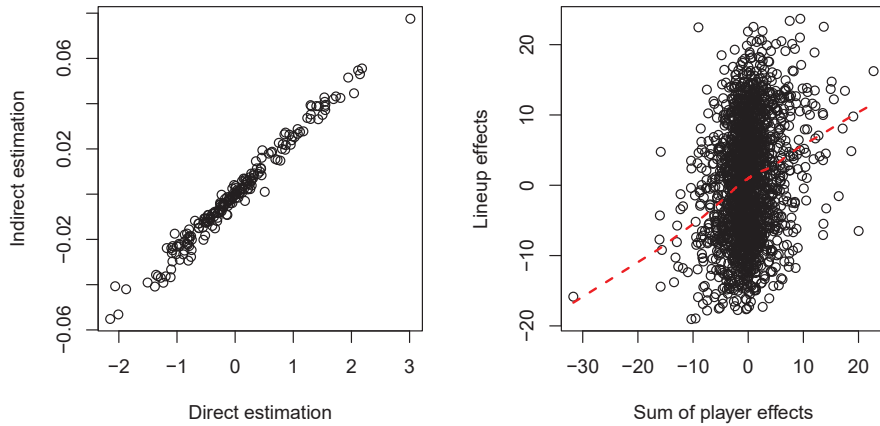
The estimates of the lineup effects result in a correlation between observed and fitted values which is higher than that one related to the estimated player effects. More precisely, the

correlation between the observed and the fitted values for the lineup model (1) is 0.34, very similar to the value 0.33 obtained by replacing the overall score by the number of points. The corresponding correlations related to model (2) are instead lower, since they are based on a much smaller number of regressors, and they are equal to 0.17 and 0.14, respectively. The latter corresponds to the original RAPM result.

Furthermore, it is of interest to investigate the connection between the two sets of estimates, $\hat{\mu}$ and $\hat{\gamma}$, considering the response given by the overall score, since it is slightly preferable. The information content of the latter is partially embodied in the former, so that a two-step analysis, where the estimated lineup effects $\hat{\mu}$ are employed as the response variable to estimate the player effect, leads to results very similar to those of model (2). At the same time, the lineup effects are able to extract more information, since they quantify also the interaction effects among different players, and not only their main effects.

These two facts are supported by the plots in Figure 5. The left panel compares the player effects estimated on shifts data using model (2) and those ones given by a two-step approach. For the latter case, the estimated lineup effects of model (1) are considered as response variables in order to get an indirect estimate of player effects. The plot suggests that the two approaches are largely equivalent, differing only by a change of scale. The right panel compares the estimated lineup effects $\hat{\mu}$ with the sum of the estimated effects for the players entering that lineup. The observed correlation is positive, as suggested by the smoother, yet only a limited portion of the lineup effect variability is explained by the sum of the player effects.

Figure 5: Indirect (two-step) and direct estimation of player effects (left panel) and estimated lineup effects and sum of the player effects of each lineup, with a robust smoother added (right panel).



3.4 Disentangling the score-based effects

With the objective of enhancing the preliminary analysis on the overall score reported in Section 2.2, further considerations could be made on the separate estimated effects associated with the three components of the score shown in Figure 2. This can be done for either the lineup effects model (1) and the player effects model (2). The relationships between the lineup effects

estimated when the response variable equals the overall score and the corresponding measures obtained when the response variable is replaced in turn by the three components lead to some plots very similar to those in Figure 2. In particular, the correlation coefficients between the estimates based on the overall score and those based on the three components are 0.60, 0.53, and 0.87, for outside shooting, inside shooting and other skills, respectively. The correlation between the estimated lineup effects based on outside shooting and inside shooting is negative (-0.25), whereas for the case of outside shooting and other skills the correlation is 0.34. Finally, for the case of inside shooting and other skills it corresponds to 0.48.

The estimation of player or lineup effects for the different classes of responses is summarized at the team level in Table 3. It reports the averages of the four different estimated lineup effects, together with the corresponding team ranking obtained from the averages. The results reported in Table 4 show how the general player effects can be described by considering different features of the game. For instance, Milano presents a positive global effect which is split into the three components, suggesting that outside shooting and other skills are the main components of it. Instead, Trieste presents a moderate negative global evaluation, but exploring the specific effects we can distinguish between a positive average effect for outside shooting and negative ones for the other two aspects.

Table 3: Averages of estimated lineup effects for the overall score and its components and the corresponding team rankings.

Teams	Score-based lineup effect	Rank	Outside shooting	Rank	Inside shooting	Rank	Other skills	Rank
Avellino	0.232	7	0.648	2	-0.290	13	0.029	7
Bologna	-0.217	11	0.204	7	-0.020	8	-0.154	15
Brescia	-0.115	10	-0.582	13	0.136	6	0.053	6
Brindisi	0.336	5	0.415	5	-0.449	16	0.227	1
Cantù	-0.592	13	-0.985	16	0.218	3	-0.075	11
Cremona	0.782	1	0.540	4	-0.058	10	0.226	2
Milano	0.646	2	0.564	3	0.163	4	0.086	5
Pesaro	-1.199	16	-0.600	14	-0.335	15	-0.279	16
Pistoia	-0.655	14	-0.381	12	-0.239	12	-0.128	14
R. Emilia	-0.107	9	-0.106	8	-0.016	7	-0.012	8
Sassari	0.477	3	-0.303	10	0.432	1	0.157	3
Torino	-0.527	12	-0.704	15	0.146	5	-0.101	12
Trento	-0.674	15	-0.259	9	-0.330	14	-0.125	13
Trieste	0.030	8	0.346	6	-0.213	11	-0.018	9
Varese	0.302	6	-0.314	11	0.410	2	0.105	4
Venezia	0.477	4	0.924	1	-0.032	9	-0.022	10

It seems worth noticing that the results of the two analyses are not the same, suggesting once again that the analyses of lineups and players supply a different kind of information.

4 Conclusions

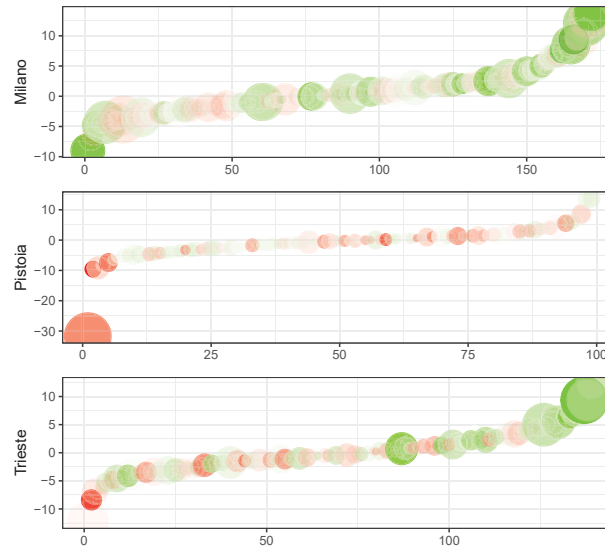
The literature on basketball analytics considers the RAPM measures as an important tool for the evaluation of single players. This article extends the RAPM methodology by considering in the model specification a response variable which is more comprehensive than the points

Table 4: Averages of estimated player effects for the overall score and its components and the corresponding team rankings.

Teams	Score-based player effect	Rank	Outside shooting	Rank	Inside shooting	Rank	Other skills	Rank
Avellino	0.300	7	0.569	4	-0.149	11	0.056	7
Bologna	-0.378	11	0.459	5	-0.042	7	-0.572	15
Brescia	-0.237	10	-0.763	14	0.129	6	0.192	6
Brindisi	0.382	6	0.351	6	-0.337	14	0.498	3
Cantù	-1.099	13	-1.316	16	0.264	4	-0.234	11
Cremona	1.241	4	0.644	3	-0.065	10	0.618	2
Milano	1.747	1	0.854	2	0.269	3	0.467	4
Pesaro	-1.803	16	-0.575	11	-0.332	13	-0.688	16
Pistoia	-1.694	15	-0.728	13	-0.423	16	-0.530	13
R. Emilia	-0.137	8	-0.000	8	-0.051	8	-0.024	8
Sassari	1.476	2	-0.687	12	0.695	1	0.895	1
Torino	-1.078	12	-1.086	15	0.176	5	-0.344	12
Trento	-1.468	14	-0.291	10	-0.412	15	-0.562	14
Trieste	-0.170	9	0.305	7	-0.205	12	-0.127	10
Varese	0.436	5	-0.262	9	0.398	2	0.233	5
Venezia	1.408	3	1.845	1	-0.063	9	-0.114	9

scored, and includes further important features. Moreover, the estimation of lineup effects is developed in addition to that of player effects.

Figure 6: Bubble plots for the sorted estimated lineup effects, with color scaling denoting the sum of estimated player effects (green for higher values, red for lower ones). The bubble size is proportional to the number of possessions played by lineups.



These two advances further exploit the information carried in the play-by-play data, defining new tools, potentially useful for team management.

As a final instance of the kind of output related to the play-by-play data analyses illustrated in this paper, we present a graphical summary of both kinds of estimated effects, namely for lineups and players. This is given in Figure 6, where the lineups of three teams already considered are sorted by the estimated lineup effects. Each lineup is represented by a bubble, with a color scale defined by the sum of estimated effects of all the players of the lineup. In particular, green bubbles correspond to higher sums, while red bubbles to lower values. The size of each bubble is scaled by the total number of possessions. The plot reveals the tendency that the most used lineups correspond to the players with higher performance, but some remarkable exceptions occur.

Acknowledgments

We are grateful to the Italian Basketball League for the permission of using the play-by-play data. This research is partially supported by the Italian Ministry for University and Research under the PRIN2015 grant No. 2015EASZFS.003.

References

- [1] R.M. Acton. *scrapeR: Tools for Scraping Data from HTML and XML Documents*, 2010. R package version 0.1.6.
- [2] S.K. Deshpande and S.T. Jensen. Estimating an NBA players impact on his teams chances of winning. *Journal of Quantitative Analysis in Sports*, 12(2):51–72, 2016.
- [3] B. Efron and T. Hastie. *Computer Age Statistical Inference*. Cambridge University Press, 2016.
- [4] J. Engelmann. Possession-based player performance analysis in basketball (adjusted +/- and related concepts). In *Handbook of Statistical Methods and Analyses in Sports*, pages 231–244. Chapman and Hall/CRC, 2017.
- [5] L. Grassetti, R. Bellio, G. Fonseca, and P. Vidoni. Estimation of lineup efficiency effects in basketball using play-by-play data. In Arbia G., Peluso S., Pini A., and Rivellini G., editors, *Book of Short Papers SIS2019*. Pearson, 2019.
- [6] S. Khalil. *Rcrawler: Web Crawler and Scraper*, 2018. R Package version 0.1.9-1.
- [7] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.
- [8] L. Rönnegård, X. Shen, and M. Alam. *hglm*: A package for fitting hierarchical generalized linear models. *The R Journal*, 2(2):20–28, 2010.
- [9] D.T. Rosenbaum. Measuring how NBA players help their teams win. *82Games.com* (<http://www.82games.com/comm30.htm>), 2004.
- [10] J. Sill. Improved nba adjusted +/- using regularization and out-of-sample testing. In *Proceedings of the 2010 MIT Sloan Sports Analytics Conference*, 2010.
- [11] H. Wickham. *rvest: Easily Harvest (Scrape) Web Pages*, 2016. R package version 0.3.2.

Will Groups of 3 Ruin the World Cup?*

Julien Guyon¹²

¹ Department of Mathematics, Columbia University
jg3601@columbia.edu

² Courant Institute of Mathematical Sciences, New York University
julien.guyon@nyu.edu

Abstract

In 2026, and maybe even as soon as 2022, the FIFA World Cup will for the first time gather 48 men's national teams. It will consist of a group stage made of 16 groups of three, with the best two teams in each group advancing to the knockout stage. Using groups of three raises several fairness issues, including risk of match fixing and schedule imbalance. In this article we examine the risk of collusion. The two teams who play the last game in the group know exactly what results will let them advance to the knockout stage. Suspicion of match fixing occurs when a result qualifies both of them at the expense of the third team of the group, and can seriously tarnish the tournament. We quantify how often this is expected to happen and explain how to build the match schedule so as to minimize the risk of collusion. We also quantify how the risk of collusion depends on competitive balance. Moreover, we show that forbidding draws during the group stage (a rule considered by FIFA) does not eliminate the risk of match fixing, and that surprisingly when draws are forbidden the 3-2-1-0 point system does not do a better job at decreasing the risk of collusion than the 3-0 point system.

1 Introduction

The soccer World Cup is the most popular sporting event in the world, even more widely viewed and followed than the Olympic Games. It is organized every four years by FIFA (Fédération

Internationale de Football Association), the sport’s world governing body. In 2026, and maybe even as soon as 2022, for the first time 48 senior men’s national teams will participate in the final tournament, based on their results in the two-year qualification process—except for the host nation(s), who may automatically qualify.

The final tournament will consist of a group stage followed by a knockout stage. For the group stage, the 48 finalists will be divided into 16 groups of three. Each group will play a single round-robin tournament, and the winner and runner-up will advance to the knockout stage.

Using groups of three may look harmless, but it actually raises several fairness issues. A first obvious issue is schedule imbalance. Let us denote by A the team that will play the first two group games, B the team that will play the first and last group matches, and C the remaining team, which will play the last two group games (see Table 1). Team B will enjoy more rest days between their two group matches than Teams A and C; Team A, if they advance to the knockout round, will enjoy more rest days than the other advancing team; Team C will have none of these benefits.

A more serious issue is the subject of this article: the suspicion of match fixing (or collusion). As soon as Match 2 is finished (see Table 1), Teams B and C will know what results of Match 3 will let them advance to the knockout stage. Suspicion of collusion occurs when a result lets both of them advance, at the expense of Team A. It can badly harm the tournament and more globally the game of soccer, *whether the match is actually fixed or not*, since outcome uncertainty is at the very root of sport’s popularity. The “disgrace of Gijón” is certainly the most famous example of match fixing in the history of soccer. It refers to the match between West Germany and Austria who refused to attack each other during 80 minutes, satisfied by the 1-0 Germany win that would let both teams advance to the second round of the 1982 FIFA World Cup, at the expense of Algeria, who had played its last group game the day before. To prevent this to happen again, FIFA decided that all teams in a given group would play their last group match at the same time, which of course is not possible with groups that have an odd number of teams, in particular with groups of three.

Even in traditional groups of four, playing the last two group games at the exact same time does not fully prevent collusion. Denmark-France (0-0 on June 26, 2018 during the 2018 FIFA World Cup) is a recent example of tacit collusion in this context: both teams knew that a draw would let them both advance to the knockout stage whatever the result of the other game in the group, Australia-Peru. They did very little to attack each other, which resulted in a very boring game and the only goalless match of the 2018 World Cup. The crowd made its displeasure known, as well as football fans around the world on social media [11]. Denmark’s manager Åge Hareide said after the game: “We just needed one point, we were up against one of the best teams in the world at counterattacks, so we would have been stupid to open up a lot of space. We stood back and got the result we needed, it was a 0-0 and we’re very pleased with that” [10]. Denmark-Sweden at UEFA Euro 2004 is another example of a tacit collusion situation: a 2-2 tie would qualify both teams at the expense of Italy, whatever the result of Italy against Bulgaria. The game indeed ended as a 2-2 draw, raising complaints from the Italian team and fans, even though Sweden and Denmark seemed to attack each other without restraint and try to win the game.

Kendall and Lenten [7] provide other examples of tacit collusion in sports, and more generally examples where the rules of sports have led to unforeseen and/or unwanted consequences. Csató [1] also investigates an example of tacit collusion in soccer.

In this article, we quantify the risk of suspicion of match fixing in groups of three, when two teams advance to the next phase. Section 2 describes the situations in which collusion will be

Match 1	Match 2	Match 3
A–B	A–C	B–C

Table 1: Match schedule of a group

suspected. In Section 3 we compute the probability of occurrence of those situations, first at the level of a group, then at the level of the tournament, which is made of 16 groups. Section 4 investigates the impact of the match schedule on the risk of collusion. In particular, we show that in order to minimize this risk, the team that plays the first two group games should be the *a priori* strongest team in the group. In Section 5 we measure the impact of competitive balance on the risk of collusion. Section 6 and 7 quantify by how much the risk of collusion would decrease if FIFA does not use the traditional 3-1-0 point system but adopts alternate point systems that forbid draws, the 3-0 and 3-2-1-0 point systems. Finally, we discuss our results in Section 8.

2 Occurrences of suspicion of match fixing

We use the notation of Table 1. In a group of three, suspicion of match fixing occurs when it is known after Match 2 if there exists a result of Match 3 (B vs C) which lets both Teams B and C advance at the expense of Team A. We assume that teams have an incentive to finish first of the group. For instance, this happens when group winners play the runners-up of another group in the first round of the knockout stage, as it has been the case since World Cup 1998. FIFA is likely to continue to implement this rule. We say that the suspicion of match fixing is *aggravated* when Team B or C can win the group even after losing its last game.

We assume that, like for the most recent World Cups, wins are worth 3 points, draws 1 point, losses 0 point, and that ties in the ranking table of the group are decided using the following ordered criteria: (1) overall goal difference, (2) overall goals scored; for the purpose of this study we only need to consider the further criterion (3): if exactly two teams are still even after criteria (1) and (2) are applied, the winner (if any) of the match between these two teams is ranked higher. The following proposition describes all the possible situations after Match 2 raising suspicion of match fixing. We denote by GD_A the goal difference of Team A after Match 2.

Proposition 1. *Suspicion of match fixing occurs exactly in the following cases:*

1. Team A has one draw and one loss.
2. Team A has two draws.
3. Team A has one win and one loss and $GD_A \leq 0$.

Aggravated suspicion of match fixing occurs if and only if Team A has one win and one loss and $GD_A < 0$.

Proof. See [6]. □

3 Probability of suspicion of match fixing

Here we consider a simple model to estimate the probability of the situations where collusion will be suspected. We assume that the result of Match 2 is independent of the result of Match

Win prob.	A	B	C
A		p_{AB}	p_{AC}
B	p_{BA}		p_{BC}
C	p_{CA}	p_{CB}	

 Table 2: Win probabilities: p_{XY} is the probability that Team X wins against Team Y.

Situation of Team A after Match 2	Probability	SMF	SMF*
Two wins	$p_{AB}p_{AC}$		
One win and one draw	$p_{AB}d_{AC} + d_{AB}p_{AC}$		
One win and one loss, $GD_A > 0$	$p_{>0}(p_{AB}p_{CA} + p_{BA}p_{AC})$		
One win and one loss, $GD_A = 0$	$p_0(p_{AB}p_{CA} + p_{BA}p_{AC})$	✓	
One win and one loss, $GD_A < 0$	$p_{<0}(p_{AB}p_{CA} + p_{BA}p_{AC})$	✓	✓
Two draws	$d_{AB}d_{AC}$	✓	
One draw and one loss	$p_{BA}d_{AC} + d_{AB}p_{CA}$	✓	
Two losses	$p_{BA}p_{CA}$		

Table 3: Summary of all the possible situations of Team A after Match 2, their probabilities, and whether they lead to suspicion of match fixing (SMF) and aggravated suspicion of match fixing (SMF*)

1. We denote by p_{XY} the probability that Team X wins against Team Y (see Table 2) and by $p_{<0}$ (resp. p_0 , $p_{>0}$) the probability that Team A has negative (resp. null, positive) goal difference, i.e., $GD_A < 0$ (resp. $= 0$, > 0) given that Team A has one win and one loss in the group stage.¹ For simplicity, we denote by

$$d_{XY} = 1 - p_{XY} - p_{YX}$$

the probability that Teams X and Y draw and by $p_{\leq 0} = p_{<0} + p_0$. Table 3 summarizes all the possible situations of Team A after Match 2, their probabilities, and whether they lead to suspicion of match fixing (SMF) and aggravated suspicion of match fixing (SMF*). The following proposition gives the probability of suspicion of match fixing for a given group of three in this model. It immediately follows from Proposition 1 and Table 3.

Proposition 2. *The probability of suspicion of match fixing in a given group of three is*

$$p_{\text{SMF}} := d_{AB}p_{CA} + p_{BA}d_{AC} + d_{AB}d_{AC} + p_{\leq 0}(p_{AB}p_{CA} + p_{BA}p_{AC}). \quad (1)$$

The probability of aggravated suspicion of match fixing in a given group of three is

$$p_{\text{SMF}}^* := p_{<0}(p_{AB}p_{CA} + p_{BA}p_{AC}). \quad (2)$$

In the case of perfect competitive balance, $p_{AB} = p_{BA} = p_{AC} = p_{CA} = p_{BC} = p_{CB} \leq \frac{1}{2}$ which we denote by p , and $p_{\leq 0} > \frac{1}{2}$, typically close to $\frac{1}{2}$. Then $d_{AB} = d_{AC} = d_{BC} = 1 - 2p$ and

$$p_{\text{SMF}} = 2p(1 - 2p) + (1 - 2p)^2 + 2p_{\leq 0}p^2 = 1 - 2p + 2p_{\leq 0}p^2$$

¹Note that, ignoring collusion issues, Krumer and Lechner [8] have examined the role of the schedule in round-robin tournaments with sequential games between three and four contestants. A more complicated model could have the probabilities p_{XY} depend on the match schedule.

When $p = \frac{1}{3}$, $p_{\text{SMF}} = \frac{1}{3} + \frac{2}{9}p_{\leq 0}$. Assuming $p_{\leq 0} = 0.6$, we get $p_{\text{SMF}} = \frac{7}{15}$. For a slightly more reasonable value $p = \frac{3}{8}$, then $p_{\text{SMF}} = \frac{1}{4} + \frac{9}{32}p_{\leq 0} = \frac{67}{160} \approx 42\%$. Both values are very close to 50%! In the situation of perfect competitive balance, the risk of suspicion of match fixing is very high.

The next two corollaries, which are easy consequences of Proposition 2, give sufficient conditions under which the risk of collusion is maximal or minimal.

Corollary 3. *The probability of suspicion of match fixing is maximum, equal to 1, in the case where $d_{AB} = d_{AC} = 1$.*

This corollary somewhat explains why it has been reported that FIFA has considered banning draws during the group stage [2, 12]. All group stage matches would have a winner and a loser, possibly decided by a penalty shootout in the case where two teams are tied after 90 minutes. When draws are forbidden for Matches 1 and 2, then the first three terms in p_{SMF} are zero and

$$p_{\text{SMF}} = p_{\leq 0} (p_{AB}p_{CA} + p_{BA}p_{AC}), \quad p_{\text{SMF}}^* = p_{< 0} (p_{AB}p_{CA} + p_{BA}p_{AC}).$$

However, the values of p_{AB} , p_{BA} , p_{AC} , p_{CA} are inflated, compared with the case where draws are allowed, since the probability of a draw between Teams X and Y is redistributed to both win probabilities p_{XY} and p_{YX} . For instance, if we assume perfect competitive balance, then $p_{\text{SMF}} = p_{\leq 0}/2$ is typically greater than $\frac{1}{4}$, while $p_{\text{SMF}}^* = p_{< 0}/2$ is typically close to $\frac{1}{4}$. Hence forbidding draws does not eliminate the risk of collusion. The situations where A has one win and one loss and a nonpositive goal difference will still be prone to match fixing.

Corollary 4. *The probability of suspicion of match fixing is minimum, equal to 0, if one of those three conditions holds:*

- (i) $p_{AB} = 1$ and ($p_{CA} = 0$ or $p_{\leq 0} = 0$): A surely wins against B, and it cannot lose against C, or if it loses against C its global goal difference GD_A can only be positive.
- (ii) $p_{AC} = 1$ and ($p_{BA} = 0$ or $p_{\leq 0} = 0$): A surely wins against C, and it cannot lose against B, or if it loses against B its global goal difference GD_A can only be positive.
- (iii) $p_{BA} = p_{CA} = 1$: A surely loses against B and C.

This corollary indicates that in order to minimize the probability of suspicion of match fixing, Team A should be the *a priori* strongest team in the group (so it is close to satisfy one of the first two conditions above) or the *a priori* weakest team in the group, if very weak (so it is close to satisfy the last condition above). Team A should not be the middle team. However, conditions (i), (ii), or (iii) are never satisfied in practice: even when a soccer powerhouse meets an underdog, there is always a positive probability that the underdog draws or wins, even if it is small. This means that in practice suspicion of match fixing cannot be avoided. In particular, we have:

Corollary 5. *Assume one of the following conditions:*

- (i) All the probabilities p_{AB} , p_{BA} , p_{AC} , p_{CA} , $p_{\leq 0}$ are strictly positive.
- (ii) The probabilities d_{AB} and d_{AC} are strictly positive.

Then the risk of collusion cannot be avoided: $p_{\text{SMF}} > 0$.

Finally, this last proposition is also an immediate consequence of Proposition 2. It quantifies the risk that collusion be suspected in at least one of the 16 groups.

Proposition 6. *Let us assume that the same values of p_{AB} , p_{BA} , p_{AC} , p_{CA} , $p_{<0}$, and $p_{\leq 0}$ apply to all 16 groups of the World Cup, and that the results in the 16 groups are all independent. Let p_{SMF} and p_{SMF}^* be given by (1) and (2). Let N_{SMF} (resp. N_{SMF}^*) be the number of groups in which suspicion of match fixing (resp. aggravated suspicion of match fixing) occurs. Then for all $k \in \{0, 1, \dots, 16\}$,*

$$\begin{aligned}\mathbb{P}(N_{\text{SMF}} = k) &= \frac{16!}{k!(16-k)!} p_{\text{SMF}}^k (1 - p_{\text{SMF}})^{16-k} \\ \mathbb{P}(N_{\text{SMF}}^* = k) &= \frac{16!}{k!(16-k)!} (p_{\text{SMF}}^*)^k (1 - p_{\text{SMF}}^*)^{16-k}.\end{aligned}$$

In particular, the probability that there is suspicion of match fixing for at least one group is

$$p_{\text{SMF}}(16) = 1 - (1 - p_{\text{SMF}})^{16}$$

and the probability that there is aggravated suspicion of match fixing for at least one group is

$$p_{\text{SMF}}^*(16) = 1 - (1 - p_{\text{SMF}}^*)^{16}.$$

There are on average

$$\mathbb{E}[N_{\text{SMF}}] = 16 p_{\text{SMF}} \quad (\text{resp. } \mathbb{E}[N_{\text{SMF}}^*] = 16 p_{\text{SMF}}^*)$$

groups in which suspicion of match fixing (resp. aggravated suspicion of match fixing) occurs.

4 Impact of the match schedule on the risk of collusion

Let us consider the realistic example of a 2026 World Cup group given in Table 4, with a strong team S, a middle team M, and a weak team W. There are three possible choices for Team A: S, M, and W, corresponding to three possible match schedules.² We naturally assume that the stronger Team A is, the smaller $p_{\leq 0}$ and $p_{<0}$ are (see Table 5). The corresponding values of p_{SMF} and p_{SMF}^* are given in Table 5. For this plausible example, it is apparent that in order to minimize the risk of collusion, Team A (the team that plays the first two group games) should be the *a priori* strongest team in the group: the risk of collusion is about 15% in any given group, if Team A is the *a priori* strongest in the group, but it climbs to around 50% otherwise. Indeed, if Team A is the *a priori* strongest in the group, it would likely be already qualified after Match 2 (first three lines of Table 3). However, arbitrarily deciding which team will play the first two games in a group is unfair, as it is the only team that may be the victim of collusion.

Note that if this schedule ($A = S$) is implemented:

- The *a priori* strongest team in the group, if it is not already qualified after Match 2, might be the victim of a collusion between the two other teams.
- The *a priori* strongest team in the group, if it advances to the knockout stage, will enjoy more rest days than the other qualified team before the round of 32.
- In all groups, the third match will oppose the two *a priori* weakest teams in the group.

²The order of the first two games is irrelevant as regards the risk of collusion.

Win prob.	S	M	W
S (Strong)		$p_{SM} = 50\%$	$p_{SW} = 80\%$
M (Middle)	$p_{MS} = 20\%$		$p_{MW} = 50\%$
W (Weak)	$p_{WS} = 5\%$	$p_{WM} = 20\%$	

Table 4: Win probabilities: p_{XY} is the probability that Team X wins against Team Y.

A	S	M	W
$p_{\leq 0}$	30%	60%	90%
$p_{< 0}$	10%	40%	80%
p_{SMF}	14.6%	47.4%	52.7%
p_{SMF}^*	1.9%	11.6%	14.8%
$p_{\text{SMF}}(16)$	91.9%	99.997%	99.999%
$p_{\text{SMF}}^*(16)$	25.8%	86.1%	92.3%
$\mathbb{E}[N_{\text{SMF}}]$	2.3	7.6	8.4
$\mathbb{E}[N_{\text{SMF}}^*]$	0.3	1.9	2.4

Table 5: Probabilities of suspicion of match fixing p_{SMF} , probabilities of suspicion of match fixing $p_{\text{SMF}}(16)$ in at least one of the 16 groups, and average number $\mathbb{E}[N_{\text{SMF}}]$ of groups with suspicion of match fixing for the example of Table 4, depending on the order of matches (A = S, M, or W). Aggravated suspicion of match fixing is denoted with a * superscript

The probabilities $p_{\text{SMF}}(16)$ (resp. $p_{\text{SMF}}^*(16)$) that there is suspicion of match fixing (resp. aggravated suspicion of match fixing) for at least one of the 16 groups, as well as the expected numbers of groups in which suspicion of match fixing will occur, are also given in Table 5. Note how large $p_{\text{SMF}}(16)$ is, even in the most favorable case where in all groups Team A is the strongest team (more than 90%!). It is almost certain that there will be a risk of collusion for at least one group. Even in this most favorable case, it is actually expected that suspicion of match fixing will occur in 2.3 groups. For the other schedules (A = M or W), match fixing will be suspected in eight groups on average! The “disgrace of Gijón” will not only be made possible again, the risk of its repetition will be very high, which is a terrible step back in the history of the World Cup.

5 Impact of competitive balance on the risk of collusion

Tables 6 and 7 compare three situations of competitive balance within a group: perfect balance (the three teams are equally skilled), imbalance (there is a strong team, a middle team, and a weak team), and strong imbalance (the strong team is much stronger than the weak team). Of course, only the last two cases are realistic for the World Cup.

As can be seen from these tables, when Team A is the strongest team in the group, the stronger the imbalance, the smaller the risk of collusion. This is because A is more likely to be already qualified after Match 2. However, when Team A is the weakest team in the group, the risk of collusion is not a monotonic function of imbalance.

Note that, even in the most favorable case where all groups are highly imbalanced and in all groups Team A is the strongest team, the risk of a collusion in at least one of the 16 groups is still very high, larger than 60%, and match fixing will be suspected in 0.9 group on average.

	Perfect balance			Imbalance			Strong imbalance		
Win prob.	S	M	W	S	M	W	S	M	W
S (Strong)		37.5%	37.5%		50%	80%		70%	90%
M (Middle)	37.5%		37.5%	20%		50%	10%		70%
W (Weak)	37.5%	37.5%		5%	20%		2%	10%	

Table 6: Win probabilities: p_{XY} is the probability that Team X wins against Team Y.

	Perfect balance	Imbalance			Strong imbalance		
A	S/M/W	S	M	W	S	M	W
$p_{\leq 0}$	60%	30%	60%	90%	30%	60%	90%
$p_{< 0}$	40%	10%	40%	80%	10%	40%	80%
p_{SMF}	41.9%	14.6%	47.4%	52.7%	5.9%	50.0%	34.6%
p_{SMF}^*	11.3%	1.9%	11.6%	14.8%	1.0%	20.0%	8.3%
$p_{\text{SMF}}(16)$	100.0%	91.9%	100.0%	100.0%	62.3%	100.0%	99.9%
$p_{\text{SMF}}^*(16)$	85.2%	25.8%	86.1%	92.3%	15.4%	97.2%	75.1%
$\mathbb{E}[N_{\text{SMF}}]$	6.7	2.3	7.6	8.4	0.9	8.0	5.3
$\mathbb{E}[N_{\text{SMF}}^*]$	1.8	0.3	1.9	2.4	0.2	3.2	1.3

Table 7: Probabilities of suspicion of match fixing p_{SMF} , probabilities of suspicion of match fixing $p_{\text{SMF}}(16)$ in at least one of the 16 groups, and average number $\mathbb{E}[N_{\text{SMF}}]$ of groups with suspicion of match fixing for the three examples of Table 6, depending on the order of matches ($A = S, M, \text{ or } W$). Aggravated suspicion of match fixing is denoted with a * superscript

6 Impact of forbidding draws on the risk of collusion

Like in the previous section, Tables 8 and 9 compare the three situations of competitive balance within a group, but now in the case where draws are forbidden: $p_{XY} + p_{YX} = 1$. Assuming that in the case of a draw both teams have equal chances to win the penalty shootout, we have equally reallocated the draw probabilities of Table 6 to both teams.

Let us compare Tables 7 and 9. Banning draws would indeed decrease the risk of collusion, but not much: for a reasonably unbalanced group, the risk of collusion would be around 10% (down from 15%) if the strongest team plays the first two group games, around 30% (down from 50%) otherwise. For a strongly unbalanced group, the risk of collusion would be around 7% (up from 6%) if the strongest team plays the first two group games, around 20% (down from 35%) if the weakest team plays the first two group games, and around 40% (down from 50%) if the middle team plays the first two group games. The probability that at least one group faces suspicion of collusion would still be very high, at about 69% (up from 62%) in the most favorable case (strong imbalance, $A = S$), and close to 100% in many cases. Even in the most favorable case, match fixing will be suspected in at least one group.

Moreover, note that while forbidding draws decreases the risk of collusion, it actually increases the risk of aggravated collusion, the most dangerous form of match fixing, since aggravated collusion can only occur when Team A has one win and one loss, even when draws are allowed. When draws are forbidden, all the win probabilities p_{XY} are larger, and as a consequence so is the probability that Team A has one draw and one loss. In the most favorable case (strong imbalance, $A = S$), the probability that at least one group faces aggravated suspicion

	Perfect balance			Imbalance			Strong imbalance		
Win prob.	S	M	W	S	M	W	S	M	W
S (Strong)		50%	50%		65%	87.5%		80%	94%
M (Middle)	50%		50%	35%		65%	20%		80%
W (Weak)	50%	50%		12.5%	35%		6%	20%	

Table 8: Win probabilities: p_{XY} is the probability that Team X wins against Team Y. Here, draws are forbidden, so $p_{XY} + p_{YX} = 1$

	Perfect balance	Imbalance			Strong imbalance		
A	S/M/W	S	M	W	S	M	W
$p_{\leq 0}$	60%	30%	60%	90%	30%	60%	90%
$p_{< 0}$	40%	10%	40%	80%	10%	40%	80%
p_{SMF}	30.0%	11.6%	32.7%	30.6%	7.1%	40.8%	21.2%
p_{SMF}^*	20.0%	3.9%	21.8%	27.2%	2.4%	27.2%	18.9%
$p_{\text{SMF}}(16)$	99.7%	86.2%	99.8%	99.7%	69.1%	100.0%	97.8%
$p_{\text{SMF}}^*(16)$	97.2%	46.9%	98.0%	99.4%	31.8%	99.4%	96.5%
$\mathbb{E}[N_{\text{SMF}}]$	4.8	1.9	5.2	4.9	1.1	6.5	3.4
$\mathbb{E}[N_{\text{SMF}}^*]$	3.2	0.6	3.5	4.4	0.4	4.4	3.0

Table 9: Probabilities of suspicion of match fixing p_{SMF} , probabilities of suspicion of match fixing $p_{\text{SMF}}(16)$ in at least one of the 16 groups, and average number $\mathbb{E}[N_{\text{SMF}}]$ of groups with suspicion of match fixing for the three examples of Table 8 (draws forbidden), depending on the order of matches (A = S, M, or W). Aggravated suspicion of match fixing is denoted with a * superscript

of collusion would be around 32% (up from 15%).

7 Impact of the point system on the risk of collusion

Let us assume that draws are forbidden, and that the winner of a tied game decided by a penalty shootout wins 2 points, instead of 3 points, while the loser wins 1 point, instead of 0 point. In this case, in all group matches, 3 points are distributed to the teams: either $3 + 0$, if there is a winner at the end of the 90 minutes of play, or $2 + 1$ if the game is tied and is decided by a penalty shootout. At first sight it seems that this new 3-2-1-0 point system, which is very natural and plausible, would significantly reduce the risk of collusion by increasing the number of possible point scenarios after Match 2. This was in particular argued by Ignacio Palacios-Huerta in [9] after we published two articles in The New York Times [3, 4] based on a first version of this work. Let us check to what extent this is true. To ease comparisons with the 3-0 point system, we still speak of draw for a match that is tied after 90 minutes and is decided by a penalty shootout, and of win and loss for games whose result is decided after 90 minutes. We recall that we say that the suspicion of match fixing is *aggravated* when Team B or C can win the group even after losing its last game. In the 3-2-1-0 point system, we introduce another notion of aggravated suspicion of match fixing: we say that the suspicion of match fixing is *aggravated* of type II when Team B or C can win the group and eliminate Team A even after drawing its last game and losing on penalties. In such a case, Teams B and C may agree

(explicitly or not) on a draw, say 0-0, and the team leading in the rankings can at no expense *decide* to eliminate Team A by losing the penalty shootout – a situation FIFA surely wants to avoid by all means.

Proposition 7. *In the 3-2-1-0 point system, suspicion of match fixing occurs exactly in the following cases:*

1. *Team A has one draw and one loss, and wins the penalty shootout.*
2. *Team A has two draws, and loses at least one of the two penalty shootouts.*
3. *Team A has one win and one loss and a goal differential $GD_A \leq 0$.*

Aggravated suspicion of match fixing occurs if and only if Team A has one win and one loss and $GD_A < 0$. Aggravated suspicion of match fixing of type II occurs if and only if Team A has one draw and one loss and wins the penalty shootout.

Proof. See [6]. □

Table 10 summarizes all the possible situations of Team A after Match 2 in the 3-2-1-0 point system, their probabilities, and whether they lead to suspicion of match fixing (SMF), aggravated suspicion of match fixing (SMF*), and aggravated suspicion of match fixing of type II (SMF*_{II}). Compared with the traditional 3-1-0 point system (Table 3), two situations that always led to suspicion of match fixing are now split into two:

- **Team A has two draws.** There is no more suspicion of match fixing if A wins both penalty shootouts – in this case A is already qualified. This decreases the probability of match fixing by $\frac{1}{4}d_{AB}d_{AC}$.
- **Team A has one draw and one loss.** There is no more suspicion of match fixing if A loses on penalties – in this case A is already eliminated. This decreases the probability of match fixing by $\frac{1}{2}(p_{BA}d_{AC} + d_{AB}p_{CA})$.

This makes suspicion of match fixing a little less likely compared with the traditional 3-1-0 point system. However, there is no change regarding aggravated suspicion of match fixing: it still occurs if and only if A has one win and one loss, and $GD_A < 0$. Moreover, the 3-2-1-0 point system introduces a new, problematic aggravated suspicion of match fixing (type II), in which a team can decide to eliminate Team A at no cost by losing the penalty shootout if the last group game ends in a draw after 90 minutes. This happens if and only if Team A has one loss and one draw won on penalties. As a consequence we have

Proposition 8. *In the 3-2-1-0 point system, the probability of suspicion of match fixing in a given group of three is*

$$p_{\text{SMF}} := \frac{1}{2}(d_{AB}p_{CA} + p_{BA}d_{AC}) + \frac{3}{4}d_{AB}d_{AC} + p_{\leq 0}(p_{AB}p_{CA} + p_{BA}p_{AC});$$

the probability of aggravated suspicion of match fixing in a given group of three is

$$p_{\text{SMF}}^* := p_{< 0}(p_{AB}p_{CA} + p_{BA}p_{AC});$$

and the probability of aggravated suspicion of match fixing of type II in a given group of three is

$$p_{\text{SMF},\text{II}}^* := \frac{1}{2}(d_{AB}p_{CA} + p_{BA}d_{AC}).$$

The various probabilities of suspicion of match fixing and average number of groups with suspicion of match fixing in the 3-2-1-0 system are reported in Table 11.

Situation of Team A after Match 2	Probability	SMF	SMF*	SMF _{II} *
Two wins	$p_{AB}p_{AC}$			
One win and one draw	$p_{AB}d_{AC} + d_{AB}p_{AC}$			
One win and one loss, $GD_A > 0$	$p_{>0}(p_{AB}p_{CA} + p_{BA}p_{AC})$			
One win and one loss, $GD_A = 0$	$p_0(p_{AB}p_{CA} + p_{BA}p_{AC})$	✓		
One win and one loss, $GD_A < 0$	$p_{<0}(p_{AB}p_{CA} + p_{BA}p_{AC})$	✓	✓	
Two draws, two wins on penalties	$\frac{1}{4}d_{AB}d_{AC}$			
Two draws, at least 1 loss on penalties	$\frac{3}{4}d_{AB}d_{AC}$	✓		
One draw and one loss, win on penalties	$\frac{1}{2}(p_{BA}d_{AC} + d_{AB}p_{CA})$	✓		✓
One draw and one loss, loss on penalties	$\frac{1}{2}(p_{BA}d_{AC} + d_{AB}p_{CA})$			
Two losses	$p_{BA}p_{CA}$			

Table 10: Summary of all the possible situations of Team A after Match 2 in the 3-2-1-0 point system, their probabilities, and whether they lead to suspicion of match fixing (SMF), aggravated suspicion of match fixing (SMF*), and aggravated suspicion of match fixing of type II (SMF_{II}*)

	Perfect balance	Imbalance			Strong imbalance		
A	S/M/W	S	M	W	S	M	W
$p_{\leq 0}$	60%	30%	60%	90%	30%	60%	90%
$p_{< 0}$	40%	10%	40%	80%	10%	40%	80%
p_{SMF}	30.9%	11.2%	34.7%	35.8%	4.9%	41.0%	22.4%
p_{SMF}^*	11.3%	1.9%	11.6%	14.8%	1.0%	20.0%	8.3%
$p_{SMF,II}^*$	9.4%	9.3%	10.5%	15.8%	5.2%	8.0%	11.8%
$p_{SMF}(16)$	99.7%	85.0%	99.9%	99.9%	55.4%	99.98%	98.7%
$p_{SMF}^*(16)$	85.2%	25.8%	86.1%	92.3%	15.4%	97.2%	75.1%
$p_{SMF,II}^*(16)$	79.4%	78.8%	83.0%	93.6%	57.5%	73.7%	86.6%
$\mathbb{E}[N_{SMF}]$	5.0	1.8	5.6	5.7	7.9	6.6	3.6
$\mathbb{E}[N_{SMF}^*]$	1.8	0.3	1.9	2.4	0.2	3.2	1.3
$\mathbb{E}[N_{SMF,II}^*]$	1.5	1.5	1.7	2.5	0.8	1.3	1.9

Table 11: Probabilities of suspicion of match fixing p_{SMF} , probabilities of suspicion of match fixing $p_{SMF}(16)$ in at least one of the 16 groups, and average number $\mathbb{E}[N_{SMF}]$ of groups with suspicion of match fixing for the three examples of Table 6 in the 3-2-1-0 system, depending on the order of matches (A = S, M, or W). Aggravated suspicion of match fixing is denoted with a * superscript; aggravated suspicion of match fixing of type II is denoted with a II subscript

8 Discussion

Figure 1 compares the probability of (a) suspicion of match fixing for a given group; (b) aggravated suspicion of match fixing for a given group; (c) suspicion of match fixing in at least one of the 16 groups; (d) aggravated suspicion of match fixing in at least one of the 16 groups, in the three different point systems 3-1-0, 3-0, 3-2-1-0, for the three competitive balance assumptions of Tables 6 and 8, and the three schedules A = S, M, or W. It shows that:

- Clearly the most important factor impacting suspicion of match fixing is the schedule:

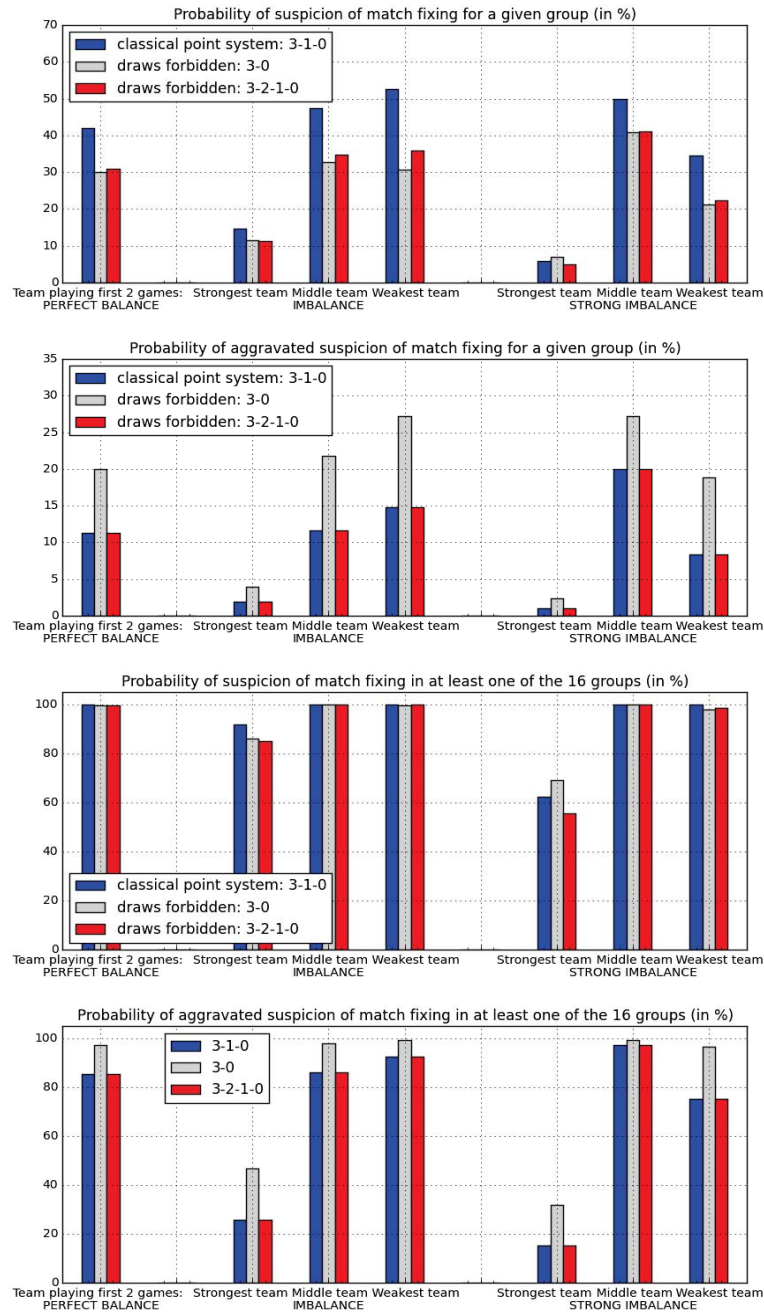


Figure 1: Comparison of probability (in %) of (a) suspicion of match fixing for a given group; (b) aggravated suspicion of match fixing for a given group; (c) suspicion of match fixing in at least one of the 16 groups; (d) aggravated suspicion of match fixing in at least one of the 16 groups, in three different point systems, for three competitive balance assumptions, and the three schedules $A = S, M$ or W .

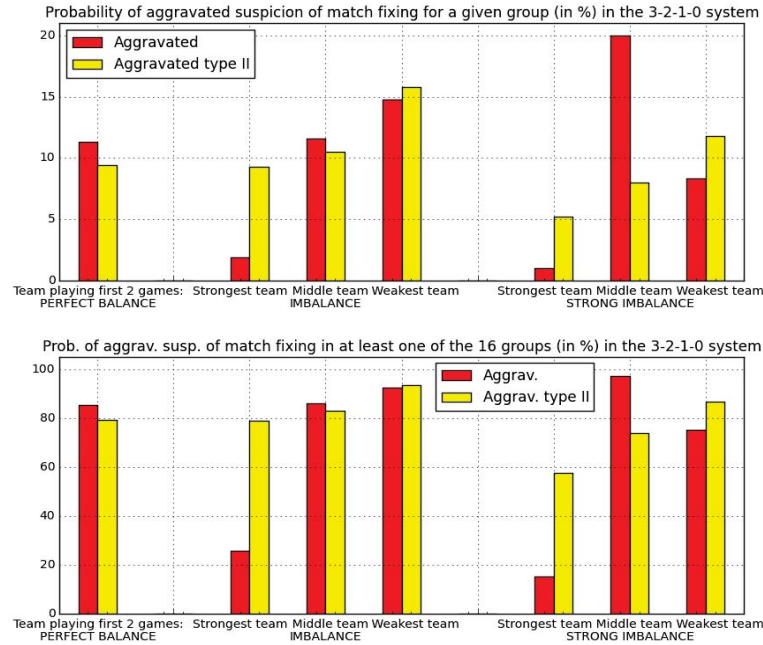


Figure 2: Comparison of probability (in %) of (a) aggravated suspicion of match fixing for a given group; and (b) aggravated suspicion of match fixing in at least one of the 16 groups, in the 3-2-1-0 point system, including type II, for three competitive balance assumptions, and the three schedules $A = S, M, \text{ or } W$.

the probability of suspicion of match fixing is minimized when it is the *a priori* strongest team that plays the first two games in the group.

- Forbidding draws (3-0 point system) decreases the probability of suspicion of match fixing, but increases the probability of aggravated suspicion of match fixing. Note however that in the case of strong imbalance and Team A being the *a priori* strongest team, forbidding draws actually increases the probability of suspicion of match fixing.
- Surprisingly, compared to the 3-0 point system, the probability of suspicion of match fixing is usually slightly larger in the 3-2-1-0 point system, except when Team A is the *a priori* strongest team.

The probability of aggravated suspicion of match fixing in the 3-2-1-0 point system is exactly the same as in the classical 3-1-0 point system. However, when we include aggravated suspicion of match fixing of type II, among the three point systems, it is the 3-2-1-0 point system that has the largest probability of aggravated suspicion of match fixing, in all cases (see Figure 3). In particular, the probability of aggravated suspicion of match fixing of type II is much larger than its “type I” equivalent when Team A is the *a priori* strongest team (see Figure 2).

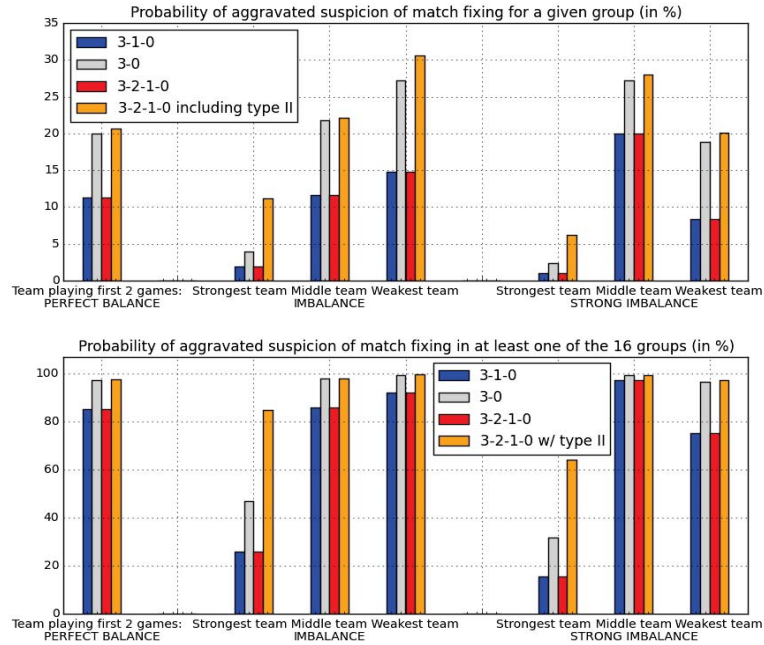


Figure 3: Comparison of probability (in %) of (a) aggravated suspicion of match fixing for a given group; (b) aggravated suspicion of match fixing in at least one of the 16 groups, in three different point systems, for three competitive balance assumptions, and the three schedules $A = S, M, \text{ or } W$. The fourth bar represents aggravated suspicion of match fixing of both types (including type II) in the 3-2-1-0 point system

9 Conclusion

We have quantified the risk of collusion in a group of three teams playing a single round-robin tournament, where two teams advance to the next phase. We have shown that the best way to minimize the risk of collusion is to enforce that the team that plays the first two group matches is the *a priori* strongest team in the group, especially if the group is strongly imbalanced. However this may be deemed unfair to that team as it would be the only one vulnerable to collusion. This would also mean that Match Day 3 of the World Cup would feature none of the seeded teams.

We have quantified how competitive imbalance within a group impacts the risk of collusion. We have also quantified by how much the risk of collusion would decrease if FIFA does not use the traditional 3-1-0 point system but adopts alternate point systems that forbid draws, the 3-0 and 3-2-1-0 point systems. Even though it looks appealing on paper, the 3-2-1-0 point system does not in general do a better job at decreasing the risk of collusion than the 3-0 point system.

Actually, whatever the rule that FIFA will use to rank the teams in a group of three, where only the group winner and the runner-up advance to the next phase, there will always be situations where Team A, the team that plays the first two group games, is neither qualified nor eliminated after Match 2, e.g., if A has one 1-0 win and one 2-0 loss. In these situations,

some results of Match 3 will qualify Team A and others will eliminate it, raising the risk of collusion between Teams B and C to eliminate Team A. As a consequence, if FIFA wants to keep groups of three with the best two teams advancing, Format 5 of [6, Section 9] seems the best solution to minimize the risk of collusion, where the knockout bracket is seeded based on performance across groups (see also [5]).

The fact that there will be 16 groups of three makes the risk of collusion in at least one group very high, even in the most favorable case where all groups are strongly imbalanced and in every group Team A is the *a priori* strongest team in the group. This proves, by the numbers, that the introduction of groups of three is a terrible step back in the history of the World Cup. Not only it makes the “disgrace of Gijón” possible again, but it makes the risk of its repetition very high. Of course, not all teams would collude if given the opportunity, but even suspicion of match fixing may seriously tarnish the World Cup, as unpredictability of the outcome is fundamental to its popularity, and to sport’s popularity in general.

It is FIFA’s responsibility to build a fair World Cup. It is not too late for FIFA to review the format of the 2022 and 2026 World Cups. Let us encourage the FIFA Council to realize the danger posed by groups of three, and, if it really wants a 48 team World Cup, opt for one of the formats described in [6, Section 9].

References

- [1] Csató, L.: *It may happen that no team wants to win: a flaw of recent UEFA qualification rules*. Preprint, 2018. Available at arxiv.org/pdf/1806.08578.pdf
- [2] Daily Mail article: “World Cup madness: 48 countries, 16 groups (yes, that’s 3 teams in each) and penalty shoot-outs in group matches... FIFA chiefs vote for bloated new format”. January 10, 2017. Accessed on May 23, 2018. Available at <http://www.dailymail.co.uk/sport/sportsnews/article-4104944/FIFA-announce-48-team-World-Cup-start-2026-idea-receives-unanimous-approval.html>
- [3] Guyon, J.: *Why Groups of 3 Will Ruin the World Cup (So Enjoy This One)*. The New York Times, June 11, 2018. Available at <https://www.nytimes.com/2018/06/11/upshot/why-groups-of-3-will-ruin-the-world-cup-so-enjoy-this-one.html>
- [4] Guyon, J. and Monkovic, T.: *FIFA, We Fixed Your World Cup Collusion Problem for You*. The New York Times, June 26, 2018. Available at <https://www.nytimes.com/2018/06/26/upshot/world-cup-fifa-collusion-readers.html>
- [5] Guyon, J.: *What a fairer 24 team UEFA Euro could look like*. Journal of Sports Analytics 4:297–317, 2018.
- [6] Guyon, J.: *Will Groups of 3 Ruin the World Cup?*. Preprint, 2018. Available at ssrn.com/abstract=3190779
- [7] Kendall, G, Lenten, L.: *When sports rules go awry*. European Journal of Operational Research 257:377-394, 2017.
- [8] Krumer, A., Lechner, M.: *First in first win: Evidence on schedule effects in round-robin tournaments in mega-events*. European Economic Review 100:412–427, 2017.
- [9] Palacios-Huerta, I.: *Penalties for Fair Play*. New Scientist, June 30, 2018.
- [10] The Guardian: *Denmark join France in last 16 after first goalless draw of World Cup*. June 26, 2018. Accessed on April 21, 2019. Available at <https://www.theguardian.com/football/2018/jun/26/denmark-france-world-cup-group-c-match-report>
- [11] The Sun: *“So Obviously Fixed” World Cup 2018: France and Denmark spark “fix” claims on social media after first 0-0 draw of the tournament*. June 27, 2018. Accessed on April 21, 2019. Available at <https://www.thesun.co.uk/world-cup-2018/6631385/france-denmark-world-cup-2018-fix-social-media>

- [12] World Soccer article: “Penalty shootouts may be used to settle drawn World Cup matches”. January 18, 2017. Accessed on May 23, 2018. Available at <http://www.worldsoccer.com/news/penalty-shootouts-may-be-used-to-settle-drawn-world-cup-matches-394315>

Analysing the Effect of a Change of Transition Probabilities Related to Possession on Scoring a Goal in a Football Match

Nobuyoshi Hirotsu^{1*} and Ayako Komine^{2†}

¹ Juntendo University, Inzai, Japan

² Japan Sport Council, Tokyo, Japan

nhirotsu@juntendo.ac.jp, ayako.komine@jpnnsport.go.jp

Abstract

In this paper, we use a Markov process model of a football match to analyse the effect of a change of transition probabilities related to possession on scoring a goal. In the model, we divide the pitch into 9 areas, and collect the data in terms of the change of location of the ball, together with the change of possession of the ball. Annual data from J League Division 1 in the 2015 season is used to estimate the transition rates between the states. Using these transition rates, we calculate the probability distribution of scoring goals and the probability of winning under the Markov process model. Using the model, we make a change of the transition rate of a team and calculate the effect of the change. We provide a numerical example of the effect of the change of transition data based on the averaged data of the league.

1 Introduction

Modelling an association football match is a topic of interest for evaluating teams' characteristics, predicting the outcome of a match, or analysing optimal tactical changes. For evaluating teams' characteristics, the factors of teams' offensive and defensive strengths are estimated by Maher (1982). He analysed three consecutive seasons from four English soccer league divisions starting with the 1971-72 season, and estimated the factors in a Poisson regression model using the maximum likelihood method. Lee (1997) similarly estimated teams' strengths in the 1995-96 season of the English Premier League. Hirotsu and Wright (2003a) estimated the factors relating to not only the transition rates of scoring and conceding goals but also the rates of gaining and losing possession. For predicting the outcome of a match, Dixon and Coles (1997) estimated teams' strengths to earn profits in the betting

* Masterminded EasyChair and created this document

† Assisted this research, especially in terms of modelling

market using English league and cup data from 1992 to 1995. They introduced a time-dependent effect to a Poisson regression model based on Maher's model. Dixon and Robinson (1998) proposed a more complicated statistical model that incorporated the goal scoring rate given the game's current time and score using English league and cup data from 1993 to 1996. For analysing the optimal tactical changes in a match, Hirotsu and Wright (2003b) applied a Markov process model together with the log-linear model, and discussed the optimal formulation changes of a team. More recently, Liu and Hohmann (2013) used a Markov chain model in which the state is determined by the player who is in possession of the ball and analysed the impact of changes in the transition probability on the attack in the front 35 meters, using data from the 2011 European Champions League final between FC Barcelona and Manchester United.

In a previous paper (Hirotsu et al., 2017), we extended the model of Hirotsu and Wright (2002, 2003a, 2003b), by considering the location of the ball on the pitch, in order to analyse teams' characteristics. Hirotsu et al. (2017) divided the pitch up to 9 areas, and collected the data in terms of the change of location of the ball, together with the change of possession of the ball.

In this paper, we use their Markov process model in which the pitch is divided to 9 areas, and analyse the effect of a change of transition probabilities related to possession on scoring a goal or winning a game. We use the annual data from the J League Division 1 in the 2015 season. Using these transition rates, we calculate the probability distribution of scoring goals and the probability of winning under the Markov process model. We make a change of the transition rates of a team in the model, and measure the effect of the change. We provide a numerical example of the effect of the change of transition data based on the averaged data of the league.

2 The Markov Process Model

A football match 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. Hirotsu and Wright (2002, 2003a, 2003b) assumed a Markov property in these transitions and proposed a Markov process model, which seems appropriate to a football match as an approximation. As a level of division of the pitch, we use a Markov process model in which the pitch divided into 9 areas 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" area ($I=1, \dots, 9$);
- State A_I : Away team is in possession of the ball and the ball is located in the "I" area ($I=1, \dots, 9$);
- State A_G : Away team scores a goal.

The "I" area ($I=1, \dots, 9$) on the pitch is also 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:

- T_i is the total time for which the game is in state i in a game ($i=H_1, H_2, \dots, A_1$);
- N_{iG} is the total number of goals scored by home team from state i in a game ($i=H_1, H_2, \dots, H_6$);
- N_{ij} is the total number of transitions from state i to state j in a game ($i, j=H_1, H_2, \dots, A_1$);
- N_{iG} is the total number of goals scored by for away team from state i in a game ($i=A_1, A_2, \dots, A_6$).

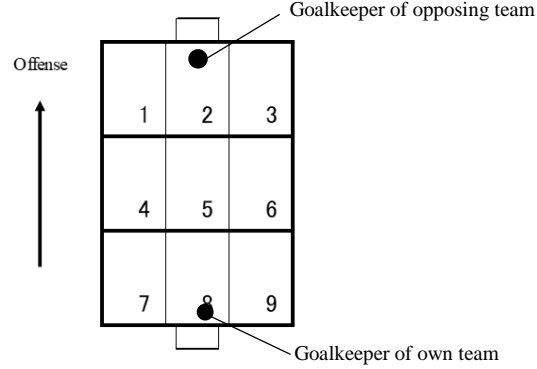


Figure 1: The areas on the pitch.

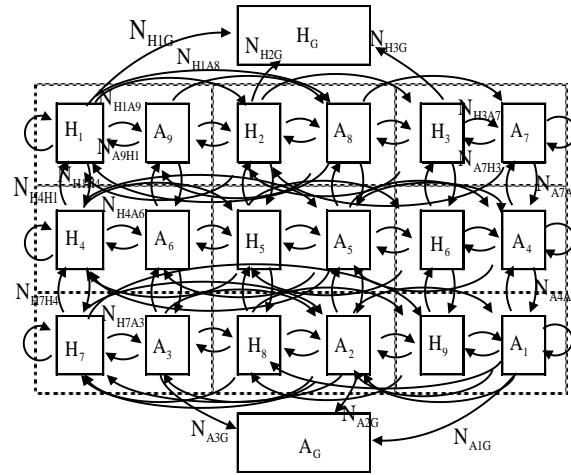


Figure 2: The Markov process model of a football match

In Figure 2, N_{H1G} represents the total number of goals scored by the home team in a game from state H_1 . N_{H1H2} represents the total number of the transition from the “1” area to the “2” area without the change of possession by the home team in the game. Here, under the assumption of the Markov property, N_{H1H2} follows Poisson distributions whose means are proportional to T_{H1} , the total possession time of the home team in the “1” area. Other total numbers of transitions such as N_{H1A9} are also defined in a similar manner.

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” area 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 $a_{H1G} \cdot dt$ and $a_{H1H2} \cdot dt$, respectively. Other transitions are also expressed in a same manner. The probability of remaining in state H_1 is thus $1 - (a_{H1G} + a_{H1H2} + \dots + a_{H1A1})dt$. Similarly, a_{ij} ($i, j = H_1, H_2, \dots, A_1$) is defined as the transition rate from state i to state j . Strictly, it may not be entirely accurate to describe the states H_G and A_G as “states”, because they instantaneously transit to state A_5 or state H_5 . However, we continue to refer to them as “states”, since this model gives us a clearer image of the basic idea underlying the model, and equations are formulated accordingly. While this is clearly

a simple model to represent a very complex process, the model does reflect the most fundamental aspects of a game (goals and possession) and therefore may nonetheless be useful.

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

Table 1: Definition of transition probabilities in a football match

The Markov process model could be expressed as a genuine Markov process model by taking into account n , the number of goals scored, or r , the number of goals home team leads by, as other states.

Using the transition probabilities shown in Table 1, we can obtain the probability distributions of the number of goals scored. Let $R_i(n|t)$ be the probabilities of home team scoring n goals in the remaining time t minutes, starting from state i ($i = H_G, H_1, H_2, \dots, A_G$). Then it can be seen that:

$$\begin{aligned}
 R_{HG}(n|t+dt) &= R_{A5}(n|t) \\
 R_{H1}(n|t+dt) &= R_{HG}(n-1|t) \cdot a_{H1G}dt + R_{H2}(n|t) \cdot a_{H1H2}dt + \dots + R_{A1}(n|t) \cdot a_{H1A1}dt \\
 &\quad + R_{H1}(n|t) \cdot \{1 - a_{H1G} - a_{H1H2} - \dots - a_{H1A1}\}dt \\
 R_{H2}(n|t+dt) &= R_{H2G}(n-1|t) \cdot a_{H2G}dt + R_{H2H1}(n|t) \cdot a_{H2H1}dt + R_{H3}(n|t) \cdot a_{H2H3}dt + \dots \\
 &\quad + R_{A1}(n|t) \cdot a_{H2A1}dt + R_{H2}(n|t) \cdot \{1 - a_{H2G} - a_{H2H1} - a_{H2H3} - \dots - a_{H2A1}\}dt \\
 &\quad \dots \\
 R_{AG}(n|t+dt) &= R_{H5}(n|t)
 \end{aligned} \tag{1}$$

By solving the equations expressed in (1), the probability distribution of goals scored by home team in the remaining time t can be obtained with the boundary conditions at the end of the game ($t = 0$), $R_{H1}(n|0) = R_{H2}(n|0) = \dots = R_{A1}(n|0) = 1$ if $n = 0$, zero otherwise.

The above method is also extended to obtain the probability of winning. Let $W_i(r|t)$ be the probability of home team winning from a position of leading by r goals with time t remaining, starting from state i ($i = H_G, H_1, H_2, \dots, A_G$). Then it can be seen that:

$$\begin{aligned}
 W_{HG}(r|t+dt) &= W_{A5}(r|t) \\
 W_{H1}(r|t+dt) &= W_{HG}(r+1|t) \cdot a_{H1G}dt + W_{H2}(r|t) \cdot a_{H1H2}dt + \dots + W_{A1}(r|t) \cdot a_{H1A1}dt \\
 &\quad + W_{H1}(r|t) \cdot \{1 - a_{H1G} - a_{H1H2} - \dots - a_{H1A1}\}dt \\
 W_{H2}(r|t+dt) &= W_{HG}(r+1|t) \cdot a_{H2G}dt + W_{H1}(r|t) \cdot a_{H2H1}dt + W_{H3}(r|t) \cdot a_{H2H3}dt + \dots \\
 &\quad + W_{A1}(r|t) \cdot a_{H2A1}dt + W_{H2}(r|t) \cdot \{1 - a_{H2G} - a_{H2H1} - a_{H2H3} - \dots - a_{H2A1}\}dt \\
 &\quad \dots \\
 W_{AG}(r|t+dt) &= W_{H5}(r|t)
 \end{aligned} \tag{2}$$

In order to obtain the probability of winning, we need to solve this equation with the boundary conditions at the end of the game such that $W_{H1}(r/0) = W_{H2}(r/0) = \dots = W_{A1}(r/0) = 1$ if $r > 0$ and 0 if $r < 0$. In this paper, we set $W_{H1}(r/0) = W_{H2}(r/0) = \dots = W_{A1}(r/0) = 0.5$ only if $r = 0$ in the case of drawing.

3 Estimation of Transition Rates

By solving Equations (1) or (2), the probability distributions for scoring goals and the probability of winning the match can be derived using the estimators for the parameters such as a_{H1G} and a_{H1H2} . If appropriate data are available, it is possible to deduce an estimate for these parameters for the game. If the total numbers of transitions and the time spent in each state are all known, the transition rates can be estimated thus:

$$\begin{aligned} a_{iG} &= N_{iG} / T_i \quad (i=H_1, H_2, \dots, H_6) \\ a_{ij} &= N_{ij} / T_i \quad (i, j=H_1, H_2, \dots, A_1) \\ a_{iG} &= N_{iG} / T_i \quad (i=A_1, A_2, \dots, A_6) . \end{aligned} \quad (3)$$

We can obtain the total numbers of transitions between states with the total time spent in each state for each game from the real data of the 2015 season of the J-League Division 1. In J-League Division 1, there are 18 teams and 306 matches played in a season. For this study, Data Stadium Inc., which helps to supply official data to the J-League, provided a large amount of data for the 306 games in the season. We extracted several data for our analysis, in which the events occurred during the game are recorded with the time. Time is measured from the beginning of the game and the location of the ball is identified as a x and y coordinate. For each game, around two thousand events are recorded, and we can use this basic information regarding to goals and possession of the ball with time and location for our analysis using Markov process models.

Table 2 shows the observed numbers of goals, transitions, and time for each game. We counted the numbers based on the annual data. For instance, we counted the numbers in the game No.1 in Table 2, such that V.Sendai scored 2 goals ($N_{H2G}=2$), or N_{H1H2} is 13, etc. As there are a lot of transitions between states, we show a part of them in Table 2.

Home	Away	Goal						Transition										Time (min.)												
No.		N _{H1G}	N _{H2G}	N _{H3G}	N _{AG}	N _{ANG}	N _{ATG}	N _{H1H2}	N _{H1A2}	N _{H2H2}	N _{H2H1}	N _{H2A2}	...	N _{A9A8}	N _{A8H8}	N _{A7A8}	N _{A5A8}	N _{A5H5}	T _{H1}	T _{H2}	T _{H3}	T _{H4}	T _{H5}	...	T _{A5}	T _{A4}	T _{A3}	T _{A2}	T _{A1}	
1	V.Sendai	M.Yamagata	0	2	0	0	0	0	13	2	10	3	2	12	11	22	6	6	1.4	1.3	2.1	3.2	2.9		3.1	3.7	2.5	3.0	2.0	
2	M.Yamagata	V.Sendai	0	1	0	0	1	0	13	2	6	5	3	12	14	10	5	6	3.0	2.1	1.8	4.3	2.1		2.1	3.1	1.3	3.7	2.0	
3	S.Hiroshima	V.Sendai	0	2	0	0	0	0	12	4	17	3	3	7	15	11	8	3	2.9	1.9	2.8	3.4	3.4		6.7	6.4	1.1	3.5	1.2	
4	V.Sendai	S.Hiroshima	0	3	0	0	3	1	22	2	19	8	2	7	8	9	5	4	4.2	3.5	3.5	6.6	6.9		3.3	2.8	3.3	4.0	2.1	
5	S.Hiroshima	M.Yamagata	0	5	0	0	1	0	13	4	14	9	6	17	18	16	9	4	2.3	2.2	1.4	4.3	3.3		4.8	3.7	1.4	2.8	0.8	
6	M.Yamagata	S.Hiroshima	0	1	0	0	2	1	19	4	11	10	2	9	13	14	3	5	2.5	2.0	2.6	4.1	5.9		4.1	4.8	3.3	6.7	2.0	
...
305	A.Niigata	V.Kofu	0	0	0	0	2	0	16	5	11	12	3	10	32	6	6	2	3.3	4.1	1.7	6.4	6.4		3.6	4.6	2.0	4.7	1.9	
306	S.Shimizu	V.Kofu	0	0	0	1	1	0	27	8	22	12	3	8	13	3	2	5	3.5	5.3	2.6	7.3	5.3		1.9	2.7	2.4	4.2	2.1	
Total			14	404	12	6	364	18											830.3 790.4 751.9 1426.1 1243.4 1220.5 1304.7 574.5 1110.7 564.5											

Table 2: Observed number of goals, transitions, and time for each game

From Table 2, we can obtain the total number of goals as 818, and total possession time as 16,833 minutes in the season. This total possession time corresponds to 55.2 (=16,833/306) minutes per game. This is not 90 minutes because in the measurement of possession time, we extracted consecutive possession time from the data, and did not count the following events toward the time of possession: Ball-out, Foul, Penalty, Offside, Substitution, and Goal.

Table 3 shows the estimates of the transition rates between states using (3). We present them as the mean and standard deviation in the season. For example, the transition rate from H_2 to H_1 appears as

3.3±1.8 times/minute in the third row and second column in Table 3. As we do not use the transition rates between same states in our calculation, we omit the numbers appearing on the diagonal in Table 3.

	H ₁	H ₂	H ₃	H ₄	H ₅	H ₆	H ₇	H ₈	H ₉	A ₉	A ₈	A ₇	A ₆	A ₅	A ₄	A ₃	A ₂	A ₁
H ₁		5.6 (2.3)	0.5 (0.4)	3.1 (1.5)	0.5 (0.5)	0.0 (0.1)	0.1 (0.2)	0.0 (0.1)	0.0 (0.0)	3.1 (1.3)	1.7 (0.9)	0.1 (0.1)	0.2 (0.3)	0.0 (0.1)	0.0 (0.1)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
H ₂	3.3 (1.8)		3.3 (1.7)	0.6 (0.6)	1.5 (1.0)	0.6 (0.5)	0.0 (0.1)	0.0 (0.1)	0.0 (0.1)	0.3 (0.4)	6.8 (2.0)	0.3 (0.3)	0.1 (0.2)	0.6 (0.5)	0.0 (0.1)	0.0 (0.1)	0.1 (0.2)	0.0 (0.1)
H ₃	0.5 (0.5)	5.8 (2.3)		0.1 (0.2)	0.5 (0.5)	3.1 (1.7)	0.0 (0.0)	0.0 (0.1)	0.0 (0.1)	0.0 (0.1)	2.1 (1.1)	3.2 (1.3)	0.0 (0.1)	0.0 (0.1)	0.2 (0.3)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
H ₄	4.1 (1.4)	0.7 (0.5)	0.2 (0.2)		3.9 (2.1)	0.4 (0.4)	1.2 (0.7)	0.6 (0.4)	0.0 (0.1)	0.6 (0.4)	0.3 (0.2)	0.0 (0.1)	1.8 (0.8)	0.2 (0.2)	0.0 (0.1)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)
H ₅	1.2 (0.7)	1.6 (1.0)	1.1 (0.6)	5.0 (2.5)		4.6 (2.3)	0.3 (0.3)	1.0 (0.7)	0.3 (0.3)	0.1 (0.2)	0.4 (0.3)	0.2 (0.2)	0.2 (0.2)	1.2 (0.6)	0.1 (0.2)	0.0 (0.0)	0.0 (0.1)	0.0 (0.0)
H ₆	0.3 (0.3)	0.8 (0.5)	4.0 (1.4)	0.4 (0.4)	3.8 (2.0)		0.0 (0.1)	0.6 (0.5)	1.3 (0.9)	0.0 (0.1)	0.3 (0.3)	0.7 (0.4)	0.0 (0.1)	0.2 (0.2)	1.9 (0.9)	0.0 (0.0)	0.0 (0.0)	0.1 (0.1)
H ₇	0.5 (0.6)	0.1 (0.2)	0.0 (0.2)	6.5 (2.4)	1.3 (0.9)	0.1 (0.3)		4.1 (2.4)	0.2 (0.4)	0.2 (0.3)	0.1 (0.2)	0.0 (0.1)	1.0 (0.7)	0.3 (0.4)	0.0 (0.1)	1.3 (1.0)	0.1 (0.2)	0.0 (0.0)
H ₈	0.3 (0.4)	0.3 (0.4)	0.3 (0.4)	2.2 (0.9)	3.4 (1.4)	2.3 (1.0)	2.5 (1.4)		2.5 (1.3)	0.1 (0.2)	0.2 (0.3)	0.1 (0.2)	0.3 (0.3)	0.5 (0.4)	0.3 (0.3)	0.1 (0.1)	0.2 (0.2)	0.1 (0.1)
H ₉	0.0 (0.2)	0.2 (0.4)	0.5 (0.7)	0.1 (0.3)	1.3 (1.0)	6.5 (2.7)	0.3 (0.5)	4.0 (2.5)		0.0 (0.1)	0.2 (0.3)	0.3 (0.4)	0.0 (0.1)	0.2 (0.4)	1.0 (0.8)	0.0 (0.1)	0.0 (0.2)	1.3 (1.0)
A ₉	1.5 (1.1)	0.1 (0.2)	0.0 (0.1)	1.2 (0.9)	0.3 (0.5)	0.0 (0.1)	0.3 (0.4)	0.1 (0.3)	0.0 (0.1)		3.7 (2.3)	0.3 (0.4)	6.3 (2.4)	1.2 (0.8)	0.1 (0.3)	0.6 (0.6)	0.2 (0.4)	0.0 (0.2)
A ₈	0.1 (0.1)	0.3 (0.3)	0.1 (0.1)	0.3 (0.4)	0.5 (0.3)	0.3 (0.3)	0.1 (0.2)	0.2 (0.2)	0.1 (0.2)	2.4 (1.1)		2.6 (1.3)	2.2 (1.0)	3.2 (1.3)	2.3 (1.0)	0.3 (0.4)	0.3 (0.4)	0.3 (0.4)
A ₇	0.0 (0.0)	0.1 (0.2)	1.5 (1.0)	0.0 (0.1)	0.3 (0.4)	1.1 (0.8)	0.0 (0.1)	0.1 (0.2)	0.2 (0.4)	0.3 (0.4)	3.8 (2.2)		0.2 (0.3)	1.3 (0.9)	6.6 (2.5)	0.0 (0.1)	0.1 (0.2)	0.4 (0.5)
A ₆	0.1 (0.1)	0.0 (0.0)	0.0 (0.1)	2.1 (0.9)	0.2 (0.2)	0.0 (0.1)	0.7 (0.4)	0.3 (0.3)	0.1 (0.1)	1.3 (0.8)	0.6 (0.5)	0.0 (0.1)		3.7 (2.1)	0.4 (0.3)	4.1 (1.4)	0.8 (0.5)	0.2 (0.3)
A ₅	0.0 (0.0)	0.0 (0.1)	0.0 (0.0)	0.2 (0.2)	1.3 (0.5)	0.2 (0.2)	0.1 (0.2)	0.4 (0.3)	0.2 (0.2)	0.2 (0.2)	1.0 (0.7)	0.3 (0.3)	4.6 (2.3)		4.6 (2.4)	1.0 (0.7)	1.6 (1.0)	1.1 (0.6)
A ₄	0.0 (0.0)	0.0 (0.0)	0.1 (0.1)	0.0 (0.1)	0.2 (0.2)	2.0 (0.8)	0.0 (0.1)	0.2 (0.2)	0.7 (0.4)	0.0 (0.1)	0.6 (0.4)	1.3 (0.7)	0.3 (0.3)	3.8 (2.1)		0.2 (0.2)	0.7 (0.5)	3.8 (1.4)
A ₃	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.2 (0.3)	0.0 (0.1)	0.0 (0.1)	3.3 (1.4)	2.0 (1.0)	0.1 (0.2)	0.0 (0.1)	0.0 (0.0)	0.0 (0.0)	3.2 (1.7)	0.6 (0.5)	0.1 (0.2)		5.5 (2.3)	0.5 (0.5)
A ₂	0.0 (0.1)	0.2 (0.3)	0.0 (0.1)	0.1 (0.2)	0.6 (0.5)	0.0 (0.1)	0.3 (0.3)	6.6 (2.0)	0.4 (0.4)	0.0 (0.1)	0.1 (0.2)	0.0 (0.1)	0.6 (0.5)	1.6 (1.0)	0.5 (0.5)	3.3 (1.6)		3.1 (1.6)
A ₁	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.1)	0.0 (0.1)	0.2 (0.3)	0.0 (0.1)	1.9 (1.0)	3.5 (1.5)	0.0 (0.0)	0.0 (0.1)	0.1 (0.1)	0.1 (0.2)	0.5 (0.5)	3.3 (1.7)	0.5 (0.4)	5.4 (2.2)	

Table 3: Transition rates between states (Mean & (SD))

Numerical Result

Table 4 represents the summary of the calculation results in terms of the probability distribution of scoring goals, the expected number of goals scored by home team, and the probability of home team winning.

Goals	H ₂	H ₅	H ₈	+ISD		-ISD	
				H ₄ H ₁	H ₇ H ₄	H ₄ A ₉	H ₇ A ₆
				H ₅	H ₅	H ₅	H ₅
0	0.244	0.250	0.251	0.240	0.247	0.247	0.248
1	0.346	0.348	0.349	0.344	0.347	0.347	0.348
2	0.244	0.241	0.240	0.245	0.242	0.242	0.242
3	0.114	0.110	0.109	0.116	0.112	0.112	0.111
4	0.039	0.038	0.037	0.041	0.038	0.038	0.038
Exp. Num.	1.401	1.376	1.370	1.417	1.388	1.388	1.382
Prob. Win.	0.540	0.535	0.534	0.543	0.538	0.539	0.538

Table 4: Calculation result in terms of the probability distribution of scoring goals, the expected number of goals, and the probability of winning

In Table 4, from state H₂ the probability of the home team scoring no goals is 0.244 in a match. This probability increases a little in the case from state H₅, and further increases a little from state H₈ to 0.251. That is, when the home team is in possession of the ball, the difference of the location between “2” and “8” area affects the difference of probability of scoring no goal by 0.007. In the same situation, the

expected number of goals scored decreases from 1.401 to 1.370, and the probability of winning the game also decreases from 0.540 to 0.534.

As an advantage of using the Markov process model, we can calculate the probabilities in terms of scoring or winning, and evaluate the effect of the change of transition rates on them. Concretely to see the sensitivity of the transition rate, we here look at the transition from the “4” to “1” area and from the “7” to “4” area. We change the transition rates by the amount of its 1SD. The calculation result of this effect has been presented in the right side of Table 4.

As shown in Table 4, if we increase the transition rate from H_4 to H_1 from 4.1 to 5.5 ($=4.1+1.4$, shown in Table 3) times/minutes (i.e. the increase of 1SD of the transition rate), the probability of scoring no goals decreases from 0.250 to 0.240 by 0.010, and the expected number of goals scored in a game increases from 1.376 to 1.388 by 0.012, when home team kick off in state H_5 . The probability of winning the game also increase from 0.535 to 0.543. Similarly, the case of the change of transition rate from H_4 to A_9 , is also calculated by changing the transition rate from 0.6 to 0.2 ($=0.6-0.4$). We note that A_9 is the “1” area from the aspect of the home team (corresponding to the “9” area from aspect of the away team). Decreasing this transition rate by 1SD results the increase of the expected number of goals from 1.376 to 1.388 by 0.012. We also present the effect of the change of transition rate from H_7 to H_4 and H_7 to A_6 , as shown in Table 4.

As shown in Table 2, the time spent in state H_4 is 1426.1 minutes which corresponds to 4.66 minutes/game, and the change of the transition rate is just 1.4 or 0.4 times/minutes as 1SD. Although the effect of the changes looks small, we demonstrated how to calculate the effects by this type of approach, which would be useful for analysis of the match.

5 Further study

In this paper, we have used the Markov process model of a football match to analyse the effect of a change of transition probabilities on scoring goals and the probability of winning. In the model, we have divided the pitch into 9 areas, and collected the data in terms of the change of location of the ball, together with the change of possession of the ball. Annual data from the J League Division 1 in 2015 was used to estimate the transition rates. Using these transition rates, we have calculated the probability distribution of scoring goals, the expected number of goals, and the probability of winning under the Markov process model. We also presented how 1SD change of the transition rate affects the probability distribution of scoring goals and so on.

In this paper, we have just shown the calculation result of the change of transition rate from state H_4 and H_7 by 1SD. As this study is still in progress, we plan to present more in the conference. Further, we plan to estimate the transition rates using log-linear models which explain such factors as home advantage, offensive and defensive strength, in terms of goals and possession, according to the location, and discuss the effect quantitatively.

Acknowledgements

This study was supported by Grants-in-Aid for Scientific Research (C) of Japan (No.26350434). The play-by-play data on J1 games used in this study was provided by Data Stadium Inc.

References

- Dixon, M.J. & Coles, S.G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Applied Statistics*, 46, 245-280.
- Dixon, M.J. & Robinson, M.E. (1998). A birth process model for association football matches. *The Statistician*, 47, 523-538.
- Hirotsu, H., Inoue, K., & Yoshimura, M. (2017). An analysis of characteristics of soccer teams using a Markov process model considering the location of the ball on the pitch, In *MathSport International 2017 Conference Proceedings* (pp.176-183).
- Hirotsu, N. & Wright, M. (2002). Using a Markov process model of an association football match to determine the optimal timing of substitution and tactical decisions. *Journal of the Operational Research Society*, 53, 88-96.
- Hirotsu, N. & Wright, M. (2003a). An evaluation of characteristics of teams in association football by using a Markov process model. *The Statistician*, 52, 591-602.
- Hirotsu, N. & Wright, M. (2003b). Determining the Best Strategy for Changing the Configuration of a Football Team. *Journal of the Operational Research Society*, 54, 878-887.
- Lee, A.J. (1997). Modeling scores in the Premier League: is Manchester United really the best? *Chance*, 10, 15-19.
- Liu, T. & Hohmann, A. (2013). Applying the Markov Chain theory to analyze the attacking actions between FC Barcelona and Manchester United in the European Champions League final. *International Journal of Sports Science and Engineering*, 7, 79-86.
- Maher, M.J. (1982). Modelling association football scores. *Statistica Neerlandica*, 36, 109-118.

Score-based soccer match outcome modeling

– an experimental review

Ondřej Hubáček*, Gustav Šourek, and Filip Železný

Czech Technical University in Prague

Abstract

In this experimental work, we propose to investigate the state-of-the-art in score-based soccer match outcome prediction modeling to identify the top-performing methods across the diverse classes of existing approaches to the problem. Namely, we bring together statistical methods based on Poisson distribution, a general ranking algorithm (Elo), domain-specific rating system (pi-ratings) and a graph-based approach to the problem (PageRank). We experimentally compare these diverse competitors altogether on a large database of soccer results to identify the true leaders in the domain.

1 Introduction

Soccer, being arguably the most popular sport in the world, continues to attract researches and practitioners competing for the design of the most accurate game result forecasting models. Indeed, there has been a plethora of such models published in the past 20 years. However, due to a lack of a standardized dataset, it has been difficult to draw conclusive statements about relative performance of the diverse approaches.

In order to gain more advantage, many of the works utilized detailed granularity of match and background information. For instance, in the top European leagues, a complete information about the game, including player-tracking data, can be obtained. However, such data are often proprietary and rather expensive, rendering them incompatible for use in academic benchmarks. Moreover, such an approach does not generalize onto the vast amount of the lower leagues, where merely the results with basic metadata is all that is being stored for each match.

To target the widest possible scope of the domain, we intersect the input information to the most common subset containing merely the match results. Such a score-based modelling paradigm allows us to predict virtually all possible matches and, consequently, unify the training and evaluation protocol across the diverse approaches.

Conveniently, a large dataset containing 218 916 match results from 52 leagues since the season 2000/01 was released recently by Dubitzky et al., 2019. The records in the dataset consist merely of the league names, dates, team names and the resulting scores. The availability of such a large dataset provides an ideal opportunity to finally shed some light onto the relative performance of the respective score-based state-of-the-art methods. For that purpose, we have started with reimplementation of the most promising models to analyze their performance under a unified protocol.

The rest of the paper is organized as follows. In Section 2 we summarize relevant research, Section 3 provides a brief description of implemented models, Section 4 explains fitting and evaluating the models, preliminary results are compiled in Section 5, conclusions and next steps can be found in Section 6.

*Corresponding author's email: hubacon2@fel.cvut.cz

2 Related Work

The research in the domain of score-based soccer modelling has traditionally been dominated by statistical approaches. In his pioneering work, Maher (1982) came up with a double Poisson model and bivariate Poisson model. The bivariate Poisson model provided a better fit for the data. Maher also introduced the notion of teams' attacking and defensive strengths and how to use them for forecasting of the match results. This notion is still used in the current research nowadays. Dixon and Coles (1997) extended Maher's ideas, as he introduced a dependency between home and away teams' goals scored for the double Poisson model, increasing the probabilities of low-scoring draws. While Maher considered the strength of the team to be time invariant, here the idea of likelihood weighting while fitting the model was introduced. Particularly, the authors used exponential time weighting to discount the effects of past results. A different approach to the time evolution was used in Rue and Salvesen (2000), where the authors used a brownian motion to tie together the teams' strength parameters in consecutive rounds. Karlis and Ntzoufras (2003) noticed, that the bivariate Poisson models tend to underestimate the probabilities of draws and introduced a diagonal-inflated bivariate Poisson model. Karlis and Ntzoufras (2008) eliminated the need to explicitly model the scores dependency via utilization of Skellam distribution. The evolution of the teams' strengths was implemented using Bayesian updates. A static hierarchical model based on double Poisson distribution was introduced in Baio and Blangiardo (2010), claiming performance not inferior to the bivariate Poisson model (Karlis and Ntzoufras, 2003). Koopman and Lit (2015) introduced time dynamics into the bivariate Poisson model using a state space model representation. Authors pointed out that the dependency between scores had a little effect on out-of-sample forecasting performance of the model. Angelini and De Angelis (2017) investigated another technique for implementing the time-dynamics with a PARX model (Agosto et al., 2016). The PARX model outperformed Dixon and Coles (1997) in forecasting number of scored goals.

The most recent novelty in statistical approaches is the use of bivariate Weibull count model (Boshnakov et al., 2017). Unlike in the Poisson distribution, where the mean is equal to the variance, the Weibull count distribution is determined by two parameters, allowing for better handling of both under and over dispersed data. The bivariate model is constructed using a copula function. The model provides a better fit for the data than the Poisson model at the expense of a higher computational time, as the computation of the probability density function of the Weibull count model is computationally demanding. A great review of the statistical approaches can be found in Ley et al. (2019).

Another technique to estimate the strength of an individual or a team are the so-called rating systems. The world's best known rating system is the Elo rating (Elo, 1978), originally used for assessing the strength of chess players. The player's performance is assumed to be drawn from a Gaussian distribution with fixed variance. The mean of such distribution is then the player's rating (skill). An application of Elo rating in the domain of soccer was shown in Hvattum and Arntzen (2010). While the authors have not provided a sufficient comparison against other models, a recent work by Robberechts and Davis (2018) demonstrated that the method is sound. Trueskill (Graepel et al., 2007) enhances the Elo rating as it operates not only with the variance of the player's performance but also with the variance of his skill (rating). This variance reflects the uncertainty about player's skills, when we have not observed enough data (performances). The author demonstrated faster convergence and better predictive performance in comparison with the Elo rating. One of the caveats of the Trueskill is that it does not propagate the newly obtained information backward to correct the ratings. In other words, it does filtering instead of smoothing. The work by Dangauthier et al. (2008) aimed to fix this issue. Also, the plain

version of Trueskill does not account for the score difference, as it only considers the win-draw-loss outcome of a match. Guo et al., 2012 proposed an extension to handle the score differences and claimed superior performance to the vanilla Trueskill, also on a soccer dataset. The current evolution of the Trueskill rating system is Trueskill2 (Minka et al., 2018), however most of the improvements are domain specific to matchmaking in online games, which is the primary focus of the system. A soccer domain-specific rating system called pi-ratings was introduced in Constantinou and Fenton (2013). The team’s strength is represented by its’ home and away ratings, that are updated after each match according to manually set learning rates. Another score-based rating system was developed by Berrar et al. (2019). The rating system parameters were tuned using particle swarm optimization and fed to a standard off-the-shelf learner.

Machine learning models are not very common in score-based modelling as they usually leverage on extra features besides the scores or ratings. Some recent exceptions were the models for the 2017 Soccer Prediction Challenge (Dubitzky et al., 2019), where the dataset contained merely the historical results with basic metadata on the matches. For the challenge, Constantinou (2019) extended his pi-ratings model with a Bayesian network to obtain the probability distribution over possible match outcomes from the rating difference. Tsokos et al. (2019) tested several variations of Bradley-Terry model and hierarchical Poisson model. In the end, the hierarchical Poisson model outperformed all the Bradley-Terry models. The inferiority of Bradley-Terry model to other methods was further confirmed by Ley et al. (2019).

The relational structure of the data was pointed out by Van Haaren and Van den Broeck (2015) where the authors achieved promising results. An advanced relational learner (Natarajan et al., 2012) was also tested in Hubáček et al. (2019), however with a little success. The same authors later proposed relational team embeddings (Hubáček et al., 2018), implemented in a framework for combining relational and neural learning (Sourek et al., 2018), with more promising results. The graph representation of the data was also utilized by Govan, Meyer, et al. (2008), who used the PageRank (Page et al., 1999) to estimate the teams’ strengths. The same author later proposed a so-called offense-defense model (Govan, Langville, et al., 2009), that can be seen as an analogy to the HITS algorithm (Kleinberg, 1999).

3 Models

In this section, we introduce the models we have reimplemented and tested so far. The selected models are considered to be very competitive in their respective categories. The Double Poisson model proved to be very competitive in the recent comparison of statistical models (Ley et al., 2019). Robberechts and Davis (2018) demonstrated effectiveness of the Elo ratings, while Constantinou and Fenton (2013) proposed their improvement – the pi-ratings. The PageRank model represents the category of models that utilize the graph structure of the data. This paper presents a work in progress, and we plan to broaden the portfolio of tested models further.

3.1 Double Poisson Model

Double Poisson model represents one of the earliest (Maher, 1982) and simplest models. However, as was shown in Ley et al. (2019), it is still very competitive nowadays. The model assumes the goals scored by the competing teams in a match to be independent. Therefore, the probability of a home team scoring x goals with the away team scoring y goals is given by

$$P(G_H = x, G_A = y | \lambda_H, \lambda_A) = \frac{\lambda_H^x e^{-\lambda_H}}{x!} \cdot \frac{\lambda_A^y e^{-\lambda_A}}{y!},$$

where λ_H and λ_A are the scoring rates of the teams (the means of the underlying Poisson distributions). The scoring rates for a match for the teams can be expressed in terms of Maher's specification as

$$\begin{aligned} \log(\lambda_H) &= Att_H - Def_A + H \\ \log(\lambda_A) &= Att_A - Def_H \end{aligned}$$

where H represents a home advantage, and Att and Def are respectively the defensive and offensive strengths of the teams (the actual model parameters).

Later, Ley et al. (2019) demonstrated that the number of the model's parameters can be effectively halved by considering only a single strength parameter for each team without any loss of predictive performance, i.e. reducing to

$$\begin{aligned} \log(\lambda_H) &= Str_H - Str_A + H \\ \log(\lambda_A) &= Str_A - Str_H \end{aligned}$$

3.2 Elo Ratings

The Elo (Elo, 1978) is a general rating system the modification of which is still used for evaluation of the strength of chess players. Hvattum and Arntzen (2010) proposed its modification for soccer and consequently Robberechts and Davis (2018) demonstrated effectiveness of the method. The modification involves the use of an ordered logit model (McCullagh, 1980) to obtain the probability distribution over the possible match outcomes. At the core, each the team's performance is assumed to be normally distributed around its true strength. The expected scores for both teams are then calculated as follows

$$\begin{aligned} E^H &= \frac{1}{1 + c^{(R^A - R^H)/d}} \\ E^A &= 1 - E^H \end{aligned}$$

where R^H and R^A are the ratings of the home and away teams, and c and d are metaparameters of the model. The actual outcome of the match is then numerically encoded as

$$S^H = \begin{cases} 1 & \text{if the home team won} \\ 0.5 & \text{if the match was drawn} \\ 0 & \text{if the home team lost} \end{cases}$$

Finally after the match, the ratings of both the teams are updated w.r.t.

$$\begin{aligned} R_{t+1}^H &= R_t^H + k(1 + \delta)^\gamma \cdot (S^H - E^H) \\ R_{t+1}^A &= R_t^A - k(1 + \delta)^\gamma \cdot (S^H - E^H) \end{aligned}$$

where δ is an absolute goal difference, k represent a learning rate and γ is a metaparameter scaling the influence of the goal difference on the rating change.

3.3 pi-ratings

The pi-ratings (Constantinou and Fenton, 2013) represent a domain-specific rating system. Each team is assigned two ratings, representing its strength when playing home and when playing away. For each match, expected goal difference is calculated, based on home team's *home rating* and away's team *away rating*. After the match is played, the *expected score* is compared to the actual outcome. If a team performs better than expected, its ratings are increased based on the discrepancy of the actual outcome and expected outcome and the learning rates (metaparameters of the model). Both team's home and away ratings get updated after a match, with both updates having a separate learning rate. We refer to the original paper for more details (Constantinou and Fenton, 2013). Finally, the probability distribution over the possible match outcomes is once again determined by an ordered logit model.

3.4 PageRank

The PageRank (Page et al., 1999) algorithm was originally designed for assessing importance of web pages. In the original algorithm, the directed edge (p_i, p_j) represents a link from page p_i to p_j . The importance of a webpage is proportional to the probability of a random walk over the webgraph visiting the page. As was shown by Govan, Meyer, et al. (2008) it can be similarly used for ranking of teams in a competition. The competition can be represented as a graph, where the nodes represent the teams and the edges represent the matches between them. For our model, the adjacency matrix M as defined as follows:

$$M_{ij} = \frac{\sum_m PTS_j(m) \cdot w_m}{\sum_m w_m}$$

where $PTS_j(m)$ is the number of points team j got from match m against team i and w_m is the weight of the match. This model represents a weighted version of the PageRank used by Hubáček et al. (2019).

4 Validation Framework

All the data used in this review came from the Open International Soccer Database v2 (Dubitzky et al., 2019). We limited the original database to seasons ranging from 2000/01 to 2005/06 to prevent data contamination in future experiments. Still, this subset provided us with nearly 60 000 of matches from 38 leagues and 27 countries. The first season of each league was only used as a warm-up season, omitted from model evaluation. Furthermore, the first 5 rounds of each season were also used as a burn-in period, omitted from the evaluation, too. We evaluated the models using ranked probability score (Epstein, 1969) and accuracy.

4.1 Model fitting

For fitting of models' free parameters we used common optimization routine based on the L-BFGS-B algorithm (Byrd et al., 1995). The fitting process and hyperparameter settings for each of the selected models is specified below.

Double Poisson Model Model's parameters are found by maximizing the weighted likelihood of the observed data

$$L = \prod (P(G_i^H = x, G_i^A = y | \lambda_i^H, \lambda_i^A) \cdot w_i)$$

Table 1: Comparison of the RPS and Accuracy of the tested models.

	RPS	Accuracy
Double Poisson	0.2082	0.4888
Elo	0.2088	0.4887
pi-ratings	0.2092	0.4897
PageRank	0.2128	0.4775

where w_i represents the weight of each observation. Since the first successful application (Dixon and Coles, 1997), exponential time weighting is being commonly used as

$$w_i = e^{-\alpha t_\Delta}$$

where t_Δ is the time passed since the match was played and α is a metaparameter. We use $\alpha = 0.0019$ as was done in Ley et al. (2019). The parameters are refitted after each league round to account for the newly obtained information.

Elo & pi-ratings Elo ratings and pi-ratings require 2 and 3 metaparameters respectively, and 3 parameters for the subsequent ordered logit model. These parameters are optimized jointly, minimizing the average RPS on previous seasons. The ratings are updated after each league round, while the (meta)parameters are refitted after each season.

PageRank The PageRank requires a setting of 1 metaparameter – the damping factor ($= 0.25$), which was tuned manually. The 3 parameters of the subsequent ordered logit model are optimized by minimizing the average RPS on previous seasons. The weight of each match is computed in the same way as in the double Poisson model. The ratings are recalculated after each league round. The parameters of the ordered logit model are refitted after each season.

5 Preliminary Results

The result are summarized in Table 1. The double Poisson model outperformed all the models in terms of RPS. The pi-ratings had a marginally higher accuracy. The PageRank trailed significantly behind other tested models both in RPS and accuracy.

The inferiority of the PageRank model could have its base in the fact that the other models leverage the scores of the teams, while the PageRank utilizes only the win/draw/loss outcome of the match. Here, we proposed a weighted version of the PageRank algorithm, which performed better than the original unweighted version (RPS of 0.2140). There are still countless ways how to advance construction of the adjacency matrix for the PageRank approach. For instance, integrating the scores into the adjacency matrix could lead to further improvements.

6 Conclusion

In this work we compared performance of diverse models for predicting soccer match outcomes based solely on historical results. Double Poisson model, one of the very oldest models in soccer forecasting, performed the best in terms of RPS. Pi-ratings, the newest model from our comparison, on the other hand outperformed the remaining models in terms of predictive

accuracy. While it was previously shown that the double Poisson model is, despite its simplicity, competitive among other statistical models (Ley et al., 2019), we can see it holds its ground against more diverse competitors as well.

Future work The work described in this paper is still in progress, and we plan to further extend on this review in various directions. Most importantly, we have merely compared 4 models so far, however we intend to update the portfolio of methods towards an extensive comparison of state-of-the-art in the domain. Regarding optimization of the models tested, we have optimized the metaparameters of the Elo and pi-ratings jointly with the parameters of subsequent ordered logit model, as was done by Robberechts and Davis (2018). Here we further plan to try out also a 2-step optimization protocol, where the optimizations of metaparameters and parameters are handled by two different optimizers. Moreover, we will investigate the influence of using a multinomial regression instead of the ordered logit model, as well as using more rating features as input covariates. Finally, with the complete set of models and optimization routines, we will extend our dataset to the full scope of available data.

References

- Agosto, Arianna et al. (2016). “Modeling corporate defaults: Poisson autoregressions with exogenous covariates (PARX)”. In: *Journal of Empirical Finance* 38, pp. 640–663.
- Angelini, Giovanni and Luca De Angelis (2017). “PARX model for football match predictions”. In: *Journal of Forecasting* 36.7, pp. 795–807.
- Baio, Gianluca and Marta Blangiardo (2010). “Bayesian hierarchical model for the prediction of football results”. In: *Journal of Applied Statistics* 37.2, pp. 253–264.
- Berrar, Daniel, Philippe Lopes, and Werner Dubitzky (2019). “Incorporating domain knowledge in machine learning for soccer outcome prediction”. In: *Machine Learning* 108.1, pp. 97–126.
- Boshnakov, Georgi, Tarak Kharrat, and Ian G McHale (2017). “A bivariate Weibull count model for forecasting association football scores”. In: *International Journal of Forecasting* 33.2, pp. 458–466.
- Byrd, Richard H et al. (1995). “A limited memory algorithm for bound constrained optimization”. In: *SIAM Journal on Scientific Computing* 16.5, pp. 1190–1208.
- Constantinou, Anthony C (2019). “Dolores: a model that predicts football match outcomes from all over the world”. In: *Machine Learning* 108.1, pp. 49–75.
- Constantinou, Anthony C and Norman Elliott Fenton (2013). “Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries”. In: *Journal of Quantitative Analysis in Sports* 9.1, pp. 37–50.
- Dangauthier, Pierre et al. (2008). “Trueskill through time: Revisiting the history of chess”. In: *Advances in neural information processing systems*, pp. 337–344.
- Dixon, Mark J and Stuart G Coles (1997). “Modelling association football scores and inefficiencies in the football betting market”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 46.2, pp. 265–280.
- Dubitzky, Werner et al. (2019). “The Open International Soccer Database for machine learning”. In: *Machine Learning* 108.1, pp. 9–28.
- Elo, Arpad E (1978). *The rating of chessplayers, past and present*. Arco Pub.
- Epstein, Edward S (1969). “A scoring system for probability forecasts of ranked categories”. In: *Journal of Applied Meteorology* 8.6, pp. 985–987.
- Govan, Anjela Y, Amy N Langville, and Carl D Meyer (2009). “Offense-defense approach to ranking team sports”. In: *Journal of Quantitative Analysis in Sports* 5.1.

- Govan, Anjela Y, Carl D Meyer, and Russell Albright (2008). “Generalizing Google’s PageRank to rank national football league teams”. In: *Proceedings of the SAS Global Forum*. Vol. 2008.
- Graepel, Thore, Tom Minka, and R TrueSkill Herbrich (2007). “A Bayesian skill rating system”. In: *Advances in Neural Information Processing Systems* 19, pp. 569–576.
- Guo, Shengbo et al. (2012). “Score-based bayesian skill learning”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 106–121.
- Hubáček, Ondřej, Gustav Šourek, and Filip Železný (2018). “Lifted Relational Team Embeddings for Predictive Sport Analytics”. In: *Proceedings of the 28th International Conference on Inductive Logic Programming*. CEUR-WS.org, pp. 84–91.
- Hubáček, Ondřej, Gustav Šourek, and Filip Železný (2019). “Learning to predict soccer results from relational data with gradient boosted trees”. In: *Machine Learning* 108.1, pp. 29–47.
- Hvattum, Lars Magnus and Halvard Arntzen (2010). “Using ELO ratings for match result prediction in association football”. In: *International Journal of forecasting* 26.3, pp. 460–470.
- Karlis, Dimitris and Ioannis Ntzoufras (2003). “Analysis of sports data by using bivariate Poisson models”. In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 52.3, pp. 381–393.
- Karlis, Dimitris and Ioannis Ntzoufras (2008). “Bayesian modelling of football outcomes: using the Skellam’s distribution for the goal difference”. In: *IMA Journal of Management Mathematics* 20.2, pp. 133–145.
- Kleinberg, Jon M (1999). “Authoritative sources in a hyperlinked environment”. In: *Journal of the ACM (JACM)* 46.5, pp. 604–632.
- Koopman, Siem Jan and Rutger Lit (2015). “A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League”. In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178.1, pp. 167–186.
- Ley, Christophe, Tom Van de Wiele, and Hans Van Eetvelde (2019). “Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches”. In: *Statistical Modelling* 19.1, pp. 55–77.
- Maher, Michael J (1982). “Modelling association football scores”. In: *Statistica Neerlandica* 36.3, pp. 109–118.
- McCullagh, Peter (1980). “Regression models for ordinal data”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 42.2, pp. 109–127.
- Minka, Tom, Ryan Cleven, and Yordan Zaykov (2018). “TrueSkill 2: An improved Bayesian skill rating system”. In: *Tech. Rep.*
- Natarajan, Sriraam et al. (2012). “Gradient-based boosting for statistical relational learning: The relational dependency network case”. In: *Machine Learning* 86.1, pp. 25–56.
- Page, Lawrence et al. (1999). *The PageRank citation ranking: Bringing order to the web*. Tech. rep. Stanford InfoLab.
- Robberechts, Pieter and Jesse Davis (2018). “Forecasting the FIFA World Cup—Combining result-and goal-based team ability parameters”. In: *Machine Learning and Data Mining for Sports Analytics ECML/PKDD 2018 workshop*. Vol. 2284. Springer, pp. 52–66.
- Rue, Havard and Oyvind Salvesen (2000). “Prediction and retrospective analysis of soccer matches in a league”. In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 49.3, pp. 399–418.
- Šourek, Gustav et al. (2018). “Lifted relational neural networks: Efficient learning of latent relational structures”. In: *Journal of Artificial Intelligence Research* 62, pp. 69–100.
- Tsokos, Alkeos et al. (2019). “Modeling outcomes of soccer matches”. In: *Machine Learning* 108.1, pp. 77–95.

Van Haaren, Jan and Guy Van den Broeck (2015). “Relational learning for football-related predictions”. In: *Latest Advances in Inductive Logic Programming*. World Scientific, pp. 237–244.

The explosive growth of eSports and the potential for research opportunities

Lyn Kee^{1,2}, Minh Hyunh^{1,2}, Denny Meyer¹, and Kelly Marshall¹

¹ Swinburne University of Technology

² Corresponding authors

ykee@swin.edu.au, mhyunh@swin.edu.au

Abstract

It is estimated that over 300 million people worldwide watch eSports events, both live and online. This viewership shows no sign of slowing down and is expected to grow by 12% each year. The growing popularity of eSports has thus transformed the industry into an excellent revenue-making opportunity. According to market research, the global revenue of eSports was over \$1.5 billion in 2017 alone. Consequentially, the opportunity for such a growing industry has attracted a wide range of investment opportunities. For example, applications of eSports within traditional sporting contexts, such as modelling and betting, are leading such applications and are one of the most prolific areas. According to Pinnacle Sports, one of the eSports betting leaders, betting on eSports events have surged from 100 thousand to five million only a five-year period. Given that research in eSports betting and modelling is still in its infancy, it warrants a promising research potential. This paper will outline the relatively untapped eSports betting industry and discuss opportunities for researchers to apply statistical methods and to collaborate within this growing field.

Keywords: eSports, modelling, betting

1 Introduction

[Commentator 1] “oh my god they are going to get control of this, they got to move quick right now”

[Commentator 2] “it is not over yet, they are going to walk right into a giant trap”

[Commentator 1] “oh man oh my god my headphones I don't even know what just happened...”

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

2 A Brief History of eSports

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

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

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

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

3 Motivations of eSports Consumptions

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

Wagner (2006) suggests that whether an activity can be considered as a sports changes as the value system in society changes. In the Industrial Age, physical fitness was considered as one of the most dominant values in society. Therefore, traditional sports mostly aimed at measuring the physical abilities of the athletes. The onset of the Information Age, however, indicate that changes are in place. The mastery of technology by different means is becoming one of the most fundamental values in

society. In youth culture, particularly, individuals who feel the need to demonstrate this mastery, may choose to showcase through succeeding in competitions such as computer gaming. The participation of eSports competition can, thus, be interpreted as one of the logical consequences of the transition from Industrial- to Information-based societies.

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

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

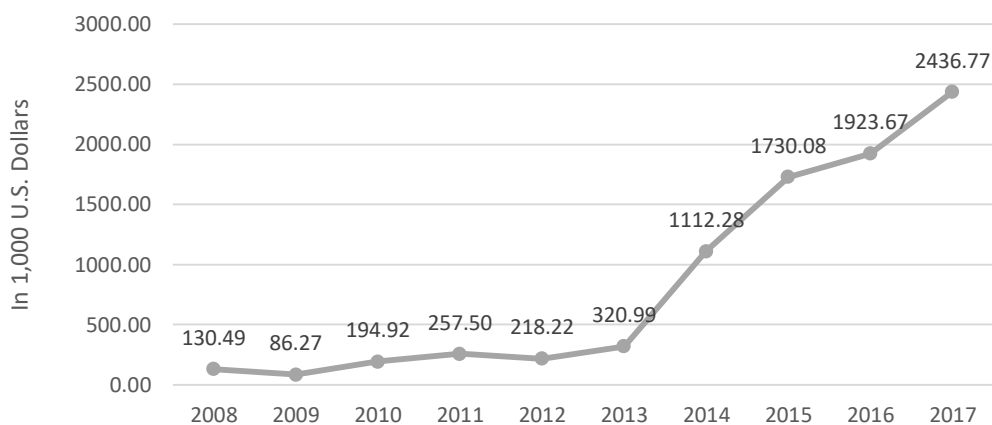


Figure 1: Highest earning of eSports players by years.

4 eSports Betting on the Rise

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

Today, cash gambling on eSports occurs through a mixture of traditional sportsbooks (e.g. Bet365, Pinnacle, and Ladbrokes) and eSports-only sportsbooks (e.g. Unikrn). Market research by Eilers and Krejci Gaming (Grove, 2016) has estimated that the amount fans wagered on eSports in 2016 was

close to \$5.58 billion (USD). This figure encapsulates both skin and cash gambling. And despite a highly controversial decision by game developer, Valve, to crackdown on unregulated skin gambling, it is expected that the eSports gambling market will reach \$12.92 billion (USD) by 2020. However, this growth (an increase of 234%) is not without challenges, one of which being the availability of data for bookmakers.

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

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

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

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

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

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

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

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

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

Where:

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

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

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

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

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

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

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

Table 1: Comparing Aligulac Glicko ratings to Pinnacle Sports Betting Odds (IEM championship 2017); Note: *u* = unknown

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

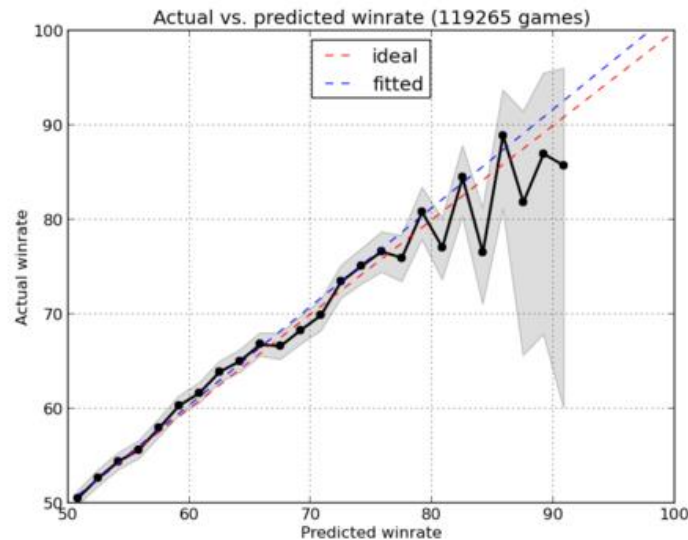


Figure 2: Comparing actual to predicted win rates for StarCraft 2 (image sourced from <http://aligulac.com>)

5 Model Comparison

For a comparison, the Elo rating system has been adopted to model one of the most popular games in the Multiplayer Online Battle Arena (MOBA) genre - the *League of Legends* (LoL). A battle in LoL consists of two teams competing against each other, with five players in each team, and the first team to destroy the opposing team wins (Gamepedia, 2018a). The authors of Gamepedia suggest that the number of kills, total teams' golds and towers taken are often good indicators of which team is ahead of the game. To verify this, data from a LoL tournament will be collected for analysis. The present research question is to identify which variables are the most important in predicting the chance of winning a battle so that the information can be incorporated into the Elo model for teams rating.

5.1 Methods

The relevant statistics were extracted using Microsoft Excel. This includes data from week one to week nine of the 2018 Summer Season Challengers Korea tournament (Gamepedia, 2018b). The data consist of the information of players in each team, number of kills, deaths, assists, golds and towers taken by each player, and the match outcomes. A decision tree will be produced as a preliminary approximation to identify the important variables. Decision tree was chosen because the outcome variable is binary (0 = lose & 1 = win) and it provides relevant parameters that could be incorporated into the Elo model.

5.2 Results and Discussion

The data was imported into R Studio and descriptive statistics was produced. On average, each team has participated in 28 matches. For analysis purpose, the number of kills, deaths, assists, golds and creep scores were aggregated as team totals (i.e., *Team_K*, *Team_D*, *Team_A*, *Team_G*, *Team_CS*). The first three variables are presented in Figure 3 below.

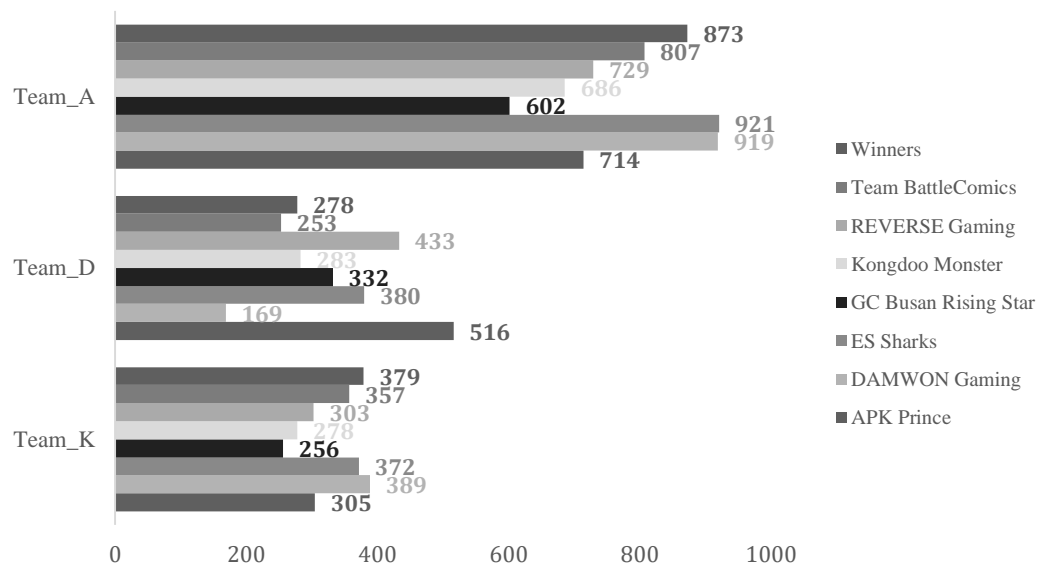


Figure 3: Total of kills, deaths, and assists by teams.

As shown in Figure 3, Team_K did not differ significantly across teams. However, Team_D has a quite distinct ranking, especially for the four teams with the highest number of deaths (i.e., APK Prince, Reverse Gaming, ES Sharks, & GC Busan Rising Star). While Damwon Gaming has the lowest number of deaths, the other three teams did not differ significantly. For Team_A, GC Busan Rising Star has the lowest assists, followed by Kongdoo Monster and APK Prince.

On the other hand, Team_G ranged from 1489 to 1705 ($M = 1601$). Teams Winners, Damwon Gaming, and Reverse Gaming have the highest Team_G (1705, 1641, & 1641, respectively). Similarly, teams Winners, Reverse Gaming, and APK Prince have the highest Team_CS (31535, 31336, & 31173, respectively).

To find out which variables are the most important in predicting the chance of winning a battle, a decision tree was produced using Rattle in R and is presented in Figure 4 below.

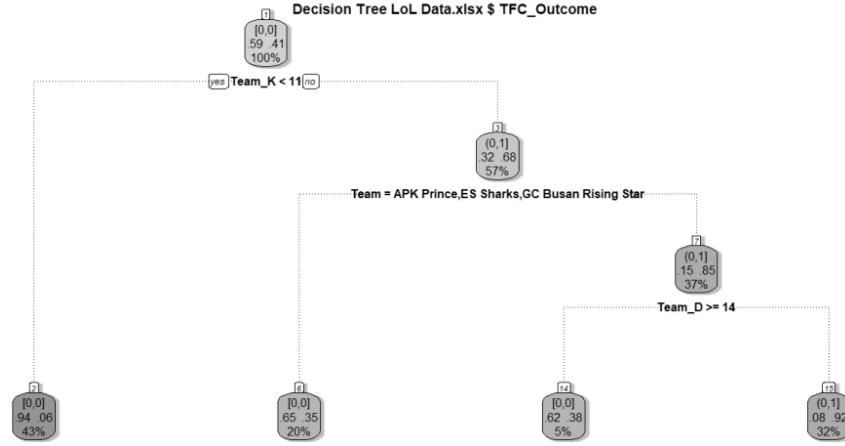


Figure 4: Decision tree for match outcome

As shown in Figure 4, the most important variables are *Team_K*, *Team*, and *Team_D*. Firstly, when *Team_K* is less than 11, there is a 94% chance of losing. Secondly, when the *Team* is APK Prince, ES Sharks, and GC Busan Rising Star, there is a 65% chance of losing. Lastly, when *Team* is not APK Prince, ES Sharks, and GC Busan Rising Star, but *Team_D* is greater than or equal to 14, there is a 62% chance of losing. However, when *Team_K* is greater than 10.5, *Team* is other than APK Prince, ES Sharks, and GC Busan Rising Star, and *Team_D* is less than 14, there is a 92% chance of winning. The Out-of-Bag (OOB) error was 10%, indicating when the resulting model is applied to new data, there will be around 10% errors. That is, the model has around 90% accuracy, which is reasonably high.

Using the information provided by the decision tree, a modified Elo model is created. The original Elo model can be operationalized as:

$$R_n = R_o + K(W - W_e) \quad (1)$$

Where R_n represents the new ratings; R_o represents the old ratings; K represents weight index; W represents match results (win = 1; lose = 0; draw = 0.5); W_e represents the expected results and can be operationalized as:

$$W_e = \frac{1}{10^{\frac{\text{diff}(\text{ratings})}{\text{average } N \text{ games} + 1}} + 1} \quad (2)$$

In the modified model, K is assigned with different values based on the information generated by the decision tree, where:

Condition 1: $Team_K \geq 10$

Condition 2: $Team = \text{Damwon Gaming, Kongdoo Monster, Reverse Gaming, Team Battlecomics}$

Condition 3: ≤ 13.5

When all three conditions are fulfilled, $K = 30$;

When two conditions are fulfilled, $K = 20$;

When one condition is fulfilled, $K = 10$;
 Otherwise, $K = 0$.

Therefore, to determine the new ratings for each team (R_n), the following formula is applied:

$$R_n = R_o + K(W - \frac{1}{10^{\frac{\text{diff}(\text{ratings})}{\text{average } N \text{ games}} + 1}}) \quad (3)$$

To illustrate the model through the LoL example, the formula was applied to the data sourced from Gamepedia (2018b). The results for the final week collected in the data (week 9) are presented in Table 2.

Day	Game	Team	Opp	Team r	Opp r	Prediction	Actual	Correct?
1	1	Winners	DAMWON	1108	1140	lose	lose	correct
1	2	DAMWON	Winners	1127	1108	win	win	correct
1	1	APK	GC	888	981	lose	lose	correct
1	2	GC	APK	981	888	win	win	correct
2	1	REVERSE	Kongdoo	979	1025	lose	lose	correct
2	2	Kongdoo	REVERSE	1025	979	win	win	correct
2	1	Team BC	ES	1039	981	win	win	correct
2	2	ES	Team BC	981	1039	lose	lose	correct
3	1	APK	Winners	888	1141	lose	lose	correct
3	2	Winners	APK	1116	893	win	draw	incorrect
3	3	APK	Winners	888	1141	lose	lose	correct
3	1	GC	DAMWON	961	1150	lose	lose	correct
3	2	DAMWON	GC	1150	961	win	win	correct

Table 2: Comparing Elo Rating with Actual Results (2018 Summer Season Challengers Korea Tournament); Note: DAMWON = Damwon Gaming; APK = APK Prince; GC = GC Busan Rising Star; REVERSE = Reverse Gaming; Kongdoo = Kongdoo Monster; Team BC = Team BattleComics; ES = ES Sharks.

Based on the results in week nine, the Elo rating correctly predicted the results 92% of the time. This is in line with the OOB error produced by the decision tree (around 10%). The only incorrect prediction was when the result was a draw, instead of win or lose. Percentages of correct prediction were lower and rather inconsistent in the previous weeks and only start to improve around week seven (see Figure 5 below). This indicates more data are needed to validate and improve the prediction accuracy.

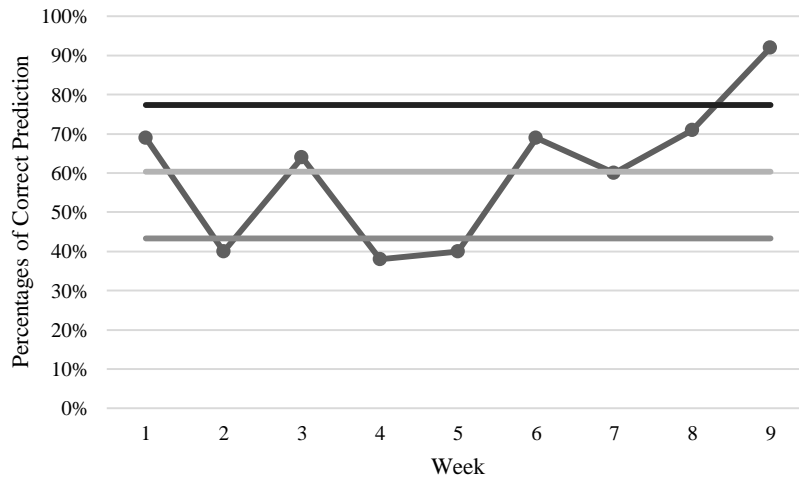


Figure 5: Control chart for Elo model prediction rate

As the outcome variable is binary (i.e., either win or lose), it is not possible to accurately identify all matches correctly, given the results can be “draw” sometime. To improve the model, it might be beneficial for future study to focus on this area.

6 Conclusions

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

References

- Aligulac (2018). StarCraft 2 Progaming Statistics and Predictions. Retrieved May 31, 2018 from <http://aligulac.com/>
- Ausretrogamer. (2015). *The Atari \$50,000 world championships fiasco*. Retrieved May 31, 2018, from <http://www.ausretrogamer.com/tag/1980s-gaming-tournaments/>
- Crystal, S., & Smith, J. (2017). *eSports betting: The past and future*, SCCG Management. Retrieved May 31, 2018, from <https://sccgmanagement.com/news/2017/10/20/esports-betting-past-future>
- ESports Earning. (2018). *Largest overall prize pools in eSports*, ESports Earning. Retrieved May 31, 2018, from <https://www.esportsearnings.com/tournaments>
- Gamepedia, (2018a). *New to League/Welcome*. Retrieved September 15, 2018, from https://lol.gamepedia.com/New_To_League

- Gamepedia, (2018b). *Challengers Korea/2018 Season/Summer Season/Scoreboards*. Retrieved September 15, 2018, from https://lol.gamepedia.com/Challengers_Korea/2018_Season/Summer_Season/Scoreboards
- Glickman, M. E. (2013). Example of the Glicko-2 system. Boston University. Retrieved May 31, 2018 from <http://glicko.net/glicko/glicko2.pdf>
- Grove, C. (2016). Esports & Gambling: Where's the action? *Eilers & Krejcik Gaming*, 2. Retrieved May 31, 2018 from <https://www.thelines.com/wp-content/uploads/2018/03/Esports-and-Gambling.pdf>
- Hamari, J., & Sjöblom, M. (2017). What is eSports and why do people watch it. *Internet Research*, 27, 211–232. doi: 10.1108/IntR-04-2016-0085
- Heitner, D. (2018). *A look inside riot games, from \$320,000 player salaries to using eSports as a catalyst for sales*, Forbes. Retrieved May 31, 2018, from <https://www.forbes.com/sites/darrenheitner/2018/05/02/a-look-inside-riot-games-from-320000-player-salaries-to-using-esports-as-a-catalyst-for-sales/#518f1d0f2c6a>
- Pannekeet, J. (2018). *Newzoo: Global eSports economy will reach \$905.6 million in 2018 as brand investment grows by 48%*. Newzoo. Retrieved May 31, 2018, from <https://newzoo.com/insights/articles/newzoo-global-esports-economy-will-reach-905-6-million-2018-brand-investment-grows-48/>
- Pinnacle. (2017). *The road to five million eSports bets*. Retrieved May 12, 2018, from <https://www.pinnacle.com/en/esports/betting-articles/educational/esports-betting-growth-at-pinnacle/ay22gtmplb93agpa>
- Popper, B. (2013). *Field of streams: How twitch made video games a spectator sport*. The Verge. Retrieved May 31, 2018, from <https://www.theverge.com/2013/9/30/4719766/twitch-raises-20-million-esports-market-booming>
- Sheer, I. (2014). *Player tally for 'League of Legends' surges*. The Wall Street Journal. Retrieved May 31, 2018, from <https://blogs.wsj.com/digits/2014/01/27/player-tally-for-league-of-legends-surges/>
- Statista. (2018). *Leading eSports players worldwide as of January 2018*, Statista. Retrieved May 31, 2018, from <https://www.statista.com/statistics/518010/leading-esports-players-worldwide-by-earnings/>
- Steam Charts (2018). An ongoing analysis of Steam's concurrent players. Retrieved May 31, 2018, from <http://steamcharts.com/app/570>
- Super Data Research. (2017). *Esports market report: courtside - Playmakers of 2017*. Retrieved May 12, 2018, from <https://www.superdataresearch.com/market-data/esports-market-report/>
- Wagner, M. (2006). On the scientific relevance of eSport. *Proceedings of the 2006 International Conference on Internet Computing and Conference on Computer Game Development*, CSREA Press, Las Vegas, NV, pp. 437-440.

Move it or lose it: Exploring the relation of defensive disruptiveness and team success.

Matthias Kempe^{1*} and Floris Goes¹

¹ University of Groningen, University Medical Center Groningen (UMCG), Department of Human Movement Sciences, Groningen, The Netherlands
m.kempe@umcg.nl

Abstract

Due to the increasing number of tracking data available for official matches in different leagues there are new ways to capture the performance of teams. To not rely on notational data, we previously introduced the D-Def (Goes et. al, 2018), an aggregated variable to quantify passing solely based on tracking data. This value captures the change of organisation by a pass (defensive disruptiveness). In this study, we updated the D-Def by including an automated classifier for subunits, instead of using starting formations, and investigated the relation of the D-Def on team success. Position tracking data of all players and the ball collected during 88 Dutch Premier League matches was used. Alignment of subunits was automatically identified, using a K-Means classifier, for every pass. D-Def was calculated for every pass (N= 63601) as an aggregate in the change in movement as a result of the pass-based team- and line centroids of subunits and surface and spread of the defending team. Team success was evaluated via wins and losses. We excluded 21 matches because they resulted in a draw. The predictive value of the D-Def for success was calculated using logistic regression analysis. The regression model achieved a R^2 of 0.69, which is high in comparison to other key performance indicators in the literature. This shows that the approach previously introduced as a proof of concept is related to match outcome. Therefore, D-Def can be a useful tool to evaluate team performance. This study highlights that performance is predictable through spatio-temporal aggregates based on player tracking data and we do not need to rely on notational data anymore.

* Presenting & Corresponding Author

1 Introduction

Performance analysis in soccer in general and tactical analysis, in particular, did take great strides in the last decade due to the availability of player position (tracking) data. The installment of optical tracking systems allows to capture game performance in different ways and opens up new opportunities for match analysis (Rein & Memmert, 2016). Previously, match analysis was only based on event data captured via notational analysis that evaluates performance via on-ball events of teams and players (Sarmiento et al., 2014). However, tactical performance in team sports should not just be seen as a chain of events but rather as the management of space, time and individual actions (Garganta, 2009; Rein & Memmert, 2016). By just using event data that does not capture the interaction of players, that is focused on the player with the ball and gives no insight in the behavior of off ball players, a quantification of this management is close to impossible. Combining this with the unclear reliability of event data, several authors advocate for the use of player tracking data to investigate tactical team performance (Gudmundsson & Horton, 2016; Rein & Memmert, 2016).

The use of tracking data enables approaches to investigate this management process in order to evaluate match performance. One approach which takes these spatial-temporal constraints into account is the team centroid method (Folgado, Lemmink, Frencken, & Sampaio, 2014; Frencken, Lemmink, Delleman, & Visscher, 2011). Here the behavior of the team centroid, the geometric center of the positions of all players from one team (C_x, y), is used to analyze the behavior of the entire team. Results from this line of research indicate a strong coupling between team centroids during gameplay (Frencken et al., 2011) and key game events like goals and shots on goal (Frencken, de Poel, Visscher, & Lemmink, 2012).

Besides the team centroid, aggregates like the line centroid, stretch index, team surface area, team spread, or regions of dominance are also used frequently to capture the complex spatiotemporal dynamics of soccer from tracking data (Rein and Memmert 2016; Memmert et al. 2017). In general, these aggregates have proven to be valid measures of behavior in small-sided games, yet in their current form, the ability to capture the complex tactical dynamics of full-sized matches can be questioned.

In a previous study, we were able to combine several of those spatio-temporal features in a new approach to measure pass performance of soccer players (Goes, Kempe, Meerhoff, & Lemmink, 2018). The evaluation of passes is one of the most common ways to assess tactical performance at individual and team levels in (scientific) performance analysis. Performing a “good” pass is a key skill for successful performance in team sports (Bush, Barnes, Archer, Hogg, & Bradley, 2015) and a main predictor for success in soccer (Liu et al. 2016). Multiple authors have already used tracking data in their analysis to model pass options (Spearman, Basye, Dick, Hotovy, & Pop, 2017), or objectively quantify pass effectiveness (Link, Lang, & Seidenschwarz, 2016; Rein, Raabe, & Memmert, 2017), that way increasing our insight into passing performance.

However, the aforementioned approaches are all biased in the same way as they overvalue passes that move the ball towards the goal or directly lead to goals or shots on goal. Our approach, in contrast, is based on the displacement of defending players (I-Mov) and the disruption of the organization of the defensive team (D-Def). Both performance indicators value passes higher if they induce a higher amount of total movement of defending players (I-Mov) or result in a larger change in defensive alignment and distance and space between team subunits (D-Def). In a validation study we could demonstrate that our measures are sensitive and valid in the differentiation between effective and less effective passes, as well as between the effective and less effective players (Goes et al., 2018). In addition, we could show in a second study that I-Mov relates to classic individual pass performance parameters like passing accuracy of key passes (passes that create goals or shots on goal) (Kempe, Goes, & Lemmink, 2018).

As we proved the relationship of our approach on an individual level, we are investigating its importance on a team level in this study. Therefore, we analyze if this approach is able to correctly predict wins and losses in official match play.

In addition, we addressed two major issues within our approach. In previous studies, we used a set time window of three seconds to evaluate passing performance. Although this time window yielded valid results, it is arbitrary and does not represent the variability of passes performed during a match. Therefore, we now calculate the effect of a pass on a normalized per second basis. The second issue we addressed, concerns the calculation of subunits and the allocation of players to those subunits. In both previous studies we used team starting formations to calculate subunits and in consequence intra-team distances and subunit centroids. However, in a fluid game like soccer, formations change often during a game. Furthermore, teams often implement different formations while attacking or defending. To tackle this problem, we used the idea to cluster players in formations based on tracking data that showed promising results in previous research (Bialkowski et al., 2016, 2015).

To sum up, this study tries to prove that game outcomes can be reliably predicted based on pass performance indicators derived from tracking data quantifying the disruptiveness of a pass.

2 Quantifying Defensive Disruptiveness

To quantify the effect of a pass, we implemented an updated versions of two previously proposed features that capture the disruption of the defensive organization as result of a pass (D-Def), and the movement of all opposing players in response to a pass (I-Mov). The theoretical rationale behind these features is based on the assumption that the attacking team tries to create space between the opposing lines through destabilization of the links between the opposing attacking, midfield, and defensive lines, as well as through forcing the opponent to move.

The disruption of the defensive organization as result of a pass was quantified using our previously published Defensive Disruptiveness (Def-D) feature (Goes et al., 2018). This feature is constructed based on the change in the average position of the attacking, midfield, and defensive line, the change in the average team position, and the change in team surface area and team spread. The D-Def measure is constructed out of three components that are derived from the scaled absolute change on all of the afore mentioned variables (eq. 1). The first component is related to disruption in the longitudinal direction of the field (PC1), the second component is related to disruption in the lateral direction of the field (PC2), and the third component is related to disruption of the team surface and spread area (PC3). The absolute scores on these three components then make up the total disruption (D-Def) score.

$$D-Def = |PC1| + |PC2| + |PC3| \quad (1)$$

In our previous publication, the different lines (attack, midfield, defensive line) were manually determined based on the starting formation of the team, and player roles were constant. However, for this analysis we improved our approach by using a K-Means clustering ($n_clusters = 3$) algorithm to automatically detect the defensive formation. Based on the defensive formation (i.e. [4, 3, 3]), we then automatically identified, for example, the defensive line based on the 4 last players (excluding the goalkeeper) in every timeframe, creating a much more robust and representative feature. For further details we refer to our previous publications (Goes et al., 2018; Kempe et al., 2018).

The movement of all opposing players in response to a pass was measured using our previously proposed individual movement (I-Mov) feature. This feature is constructed based on the sum of the absolute displacement along the longitudinal (I-Mov-X) and lateral (I-Mov-Y) axis off all opposing players in response to a pass (eq. 2). In our previous publication, we used the sum of the displacement of all players to make up the I-Mov feature for the team. However, for the current analysis we improved

this by using the mean I-Mov per player, as this method is much more reliable in case of possible missing or erroneous data that occur quite frequently in tracking datasets.

$$I-Mov = (|Disp. X_1| + |Disp. Y_1| + \dots + |Disp. X_n| + |Disp. Y_n|) / n \quad (2)$$

We computed both the D-Def and I-Mov feature for every pass received by a teammate during the entire match. This was conducted by computing the change/displacement on all feature components during the pass window (between the moment of the pass and reception), and then dividing this value by the duration of the pass window in seconds. This resulted in standardized displacement/disruption scores/second. In our previous paper, we used a window of 3 seconds after a pass, as we assumed this should be adequate to detect both the effect of the pass, and prevent the inclusion of effects of the next pass. However, we experimentally determined that the standardized pass-window as implemented in the current study was a better fit and therefore improved our feature.

3 Modelling Team Success based on Pass Disruptiveness

To evaluate tactical performance and analyze the relationship between tactical performance and match outcome, we collected and processed position tracking data on both teams for matches played during 4 consecutive Dutch Eredivise seasons. Players were tracked with a semi-automatic optical tracking system (SportVU; STATS LLC, Chicago, IL) that captures the X and Y coordinates of all players and the ball at 10 Hz. Our dataset contained 118 matches in which 26 unique teams played each other. As we were only concerned with the differences between winning and losing teams, we excluded matches that ended in a draw. This resulted in a final dataset that consists of 25 teams that played in 89 matches that resulted in a win or a loss and contained 98.718 pass attempts of which 60.524 passes were successful.

The data of every single match were first pre-processed with ImoClient software (Inmotiotec GmbH, Austria). Pre-processing consisted of filtering the data with a weighted Gaussian algorithm (85% sensitivity) and automatic detection of ball possessions and ball events based on the tracking data. Both the tracking data and the ball event data were then imported as individual data frames in Python 3.6 and automatically processed on a match-by-match basis. We then computed the separate components of both the D-Def as well as the I-Mov feature for every pass during a match. All features were computed according to the methods as described in section 2.

Table 1 - Descriptive statistics winning and losing teams (*: $p = .05$ **: $p < .05$, *: $p < .01$)**

	Wins (N = 89)	Losses (N = 89)	Mean Diff.	Effect Size (Cohen's d)
<i>Individual Movement (I-Mov)</i>				
I-Mov-X (Mean)	0.866m ± 0.673m	0.515m ± 0.675m	+68.1%	0.52**
I-Mov-Y (Mean)	0.772m ± 0.600m	0.451m ± 0.591m	+71.2%	0.54**
I-Mov (Mean)	1.638m ± 1.268m	0.966m ± 1.265m	+69.6%	0.53**
I-Mov-X (Total)	261.46m ± 222.14m	163.53m ± 219.69m	+59.9%	0.44*
I-Mov-Y (Total)	238.85m ± 213.81m	142.92m ± 191.86m	+67.1%	0.47**
I-Mov (Total)	500.31m ± 434.39m	306.45m ± 411.12m	+63.3%	0.46*
<i>Defensive Disruptiveness (D-Def)</i>				
PC1 (Mean)	0.018 ± 0.015	0.013 ± 0.022	+34.1%	0.24*
PC2 (Mean)	0.010 ± 0.013	0.014 ± 0.033	-23.6%	-0.13
PC3 (Mean)	-0.026 ± 0.022	-0.021 ± 0.022	-25.5%	-0.25*

D-Def (Mean)	0.474 ± 0.048	0.484 ± 0.072	-2.0%	-0.16
PC1 (Total)	4.88 ± 4.17	4.14 ± 3.57	+17.8%	0.19*
PC2 (Total)	2.99 ± 4.09	3.28 ± 4.55	-8.9%	-0.07
PC3 (Total)	-7.96 ± 6.85	-6.30 ± 6.65	-26.2%	-0.24*
D-Def (Total)	133.60 ± 54.33	130.50 ± 46.12	+2.4%	0.06

To compare performance between winning and losing teams, we aggregated all feature scores into mean (values per pass), and total (sum over a full match) scores. We then took the means and standard deviations of all winning and losing teams for a between-group comparison (Table 1). Effect sizes were determined based on the Cohen's d and between group differences were statistically tested using an independent t-test. For completeness, we not only displayed and tested the composite feature scores, but also the individual components, as this might provide additional information.

As a next step, we predicted match outcome based on the mean total movement feature ($I\text{-Mov}_{\text{Mean}}$, as it captures both the movement in longitudinal as well as lateral direction), mean longitudinal disruption feature ($PC1_{\text{Mean}}$), and mean surface disruption feature ($PC3_{\text{Mean}}$). We choose this combination of features based on their discriminative power and the fact that the combination of these features yielded the highest accuracy and lowest log loss scores. To do so we first split the data set in a training set that contained 80% of the data, and a test set that contained 20% of the data, stratified on match outcome. Furthermore, we scaled (Z-transformed) our features to the same scale using a Min-Max scaling algorithm. We then fitted a 5-fold cross-validated Logistic Regression model to our training dataset and predicted winning and losing probability for both teams in every match. Based only on the mean total movement per pass ($I\text{-Mov}_{\text{Mean}}$), the mean longitudinal disruption per pass ($PC1_{\text{Mean}}$), and the mean surface disruption per pass ($PC3_{\text{Mean}}$), we were able to predict binary match outcome with an accuracy of 69.4% and a log loss of 0.65, based on the following regression equation (3):

$$\text{Outcome} = -0.146 + 0.689 I\text{-Mov}_{\text{Mean}} + 0.172 PC1_{\text{Mean}} - 0.592 PC3_{\text{Mean}} \quad (3)$$

4 Discussion

The aim of this study was to further validate our approach of using changes in spatio-temporal features, derived of player tracking data, to evaluate (tactical) match performance. Our findings illustrate that this approach is capable to reliably distinguish between winning and losing teams. Therefore, we could prove that our approach is not just valid on an individual but also on a team level. In previous studies, we already showed that our performance indicators are able to evaluate players and passes (Goes et al., 2018), as well as relate to individual performance like passing accuracy and assists (Kempe et al., 2018).

Within this study, we now also showed that the $I\text{-Mov}$ clearly differentiates between winning and losing teams with a difference of mean induced movement of pass of 69,6% in favor of the winning teams. $D\text{-Def}$, as the more complex performance indicator that registers the changes in defensive organization, could not differentiate in the same way as the $I\text{-Mov}$. However, two of its three factors ($PC1$ & $PC3$) did yield statistical differences between winning and losing teams. One can assume from those results that changes in the longitudinal organization of the defending team, creating larger distances between the different lines of defense, and the surface of the team organization, shape and spread of the lines and the team in general, represent changes in overall organization while change in horizontal organization just adds noise to the equation.

In general, it is understandable that the $I\text{-Mov}$ is a more sensitive feature as teams are able to maintain their overall organization while moving. Therefore, changes in the $D\text{-Def}$ caused by a pass are way smaller than in the $I\text{-Mov}$. Following this line of assumption, the $I\text{-Mov}$ might be the better Key

performance Index to evaluate an overall or game performance whereas the D-Def might be more suitable to identifying the one or two key passes in a chain of events that led to a decrease in structural organization of the defending team. Therefore, the D-Def might rather be used to study passing or attacking sequences also referred to as “quality of possession” (Collet, 2013) and the I-Mov as a measure of overall team performance.

By combining the features of mean player movement (I-Mov), mean longitudinal disruption (PC1), and the mean surface disruption per pass (PC3) we are able to correctly predict the winning team in 69,4% of our test set. This results are especially promising as previous (pass) performance indicators just showed a weak relationship with success (Rein et al., 2017). By our knowledge, this is the first approach that is solely based on player tracking data that is able to predict game outcome better than pure chance with a prediction power better than previous models based on event data (Collet, 2013; Oberstone, 2009).

In order to achieve this prediction performance, we updated our previous model in two important ways. First, instead of a three second window, we now normalize the effect of a pass per second. In the previous model we undervalued longer passes as their effect might not be captured in total with the three second window. In a second step, we implemented a new way to register team formations which are the basis to calculate the changes in defensive organization. Therefore, we adapted the idea of Bialkowski et al. (2015 & 2016). They use a K-Nearest Neighbour like approach to cluster players in different playing positions and formations showing that this approach is able to predict playing formation with a maximal mean variation of 5.5 m. By applying this idea to our approach, although in a different form, instead of starting formations of a team, we now differentiate between offensive and defensive formation and are able to evaluate passes by taking the change of playing positions and formations into account. Both of those updates increase the validity of our approach by reflecting the high amount of variation in the game of soccer.

5 Conclusion

In this paper, we could further demonstrate that an approach solely based on spatiotemporal variables is able to capture tactical game performance on a team level and is able to reliably predict game outcomes. One of our performance indicators (I-Mov) could further highly differentiate between winning and losing teams. Therefore, the I-Mov might serve as a new tool to evaluate team performance instead of unreliable event data like pass accuracy, percentage of ball possession, or shots on goals.

Disclosure Statement

The authors of this paper reported no conflicts of interest

Acknowledgements

This work was supported by a grant of the Netherlands Organization for Scientific Research (project title: “The Secret of Playing Soccer: Brazil vs. The Netherlands”).

References

- Bialkowski, A., Lucey, P., Carr, P., Matthews, I., Sridharan, S., & Fookes, C. (2016). Discovering Team Structures in Soccer from Spatiotemporal Data. *IEEE Transactions on Knowledge and Data Engineering*, 28(10), 2596–2605. <https://doi.org/10.1109/TKDE.2016.2581158>
- Bialkowski, A., Lucey, P., Carr, P., Yue, Y., Sridharan, S., & Matthews, I. (2015). Identifying team style in soccer using formations learned from spatiotemporal tracking data. In *IEEE International Conference on Data Mining Workshops, ICDMW* (Vol. 2015-Janua, pp. 9–14). <https://doi.org/10.1109/ICDMW.2014.167>
- Bush, M., Barnes, C., Archer, D. T., Hogg, B., & Bradley, P. S. (2015). Evolution of match performance parameters for various playing positions in the English Premier League. *Human Movement Science*, 39, 1–11. <https://doi.org/10.1016/j.humov.2014.10.003>
- Collet, C. (2013). The possession game? A comparative analysis of ball retention and team success in European and international football, 2007-2010. *Journal of Sports Sciences*, 31(2), 123–136. <https://doi.org/10.1080/02640414.2012.727455>
- Folgado, H., Lemmink, K. A. P. M., Frencken, W., & Sampaio, J. (2014). Length, width and centroid distance as measures of teams tactical performance in youth football. *European Journal of Sport Science*. <https://doi.org/10.1080/17461391.2012.730060>
- Frencken, W., de Poel, H., Visscher, C., & Lemmink, K. (2012). Variability of inter-team distances associated with match events in elite-standard soccer. *Journal of Sports Sciences*. <https://doi.org/10.1080/02640414.2012.703783>
- Frencken, W., Lemmink, K., Delleman, N., & Visscher, C. (2011). Oscillations of centroid position and surface area of soccer teams in small-sided games. *European Journal of Sport Science*, 11(4), 215–223. <https://doi.org/10.1080/17461391.2010.499967>
- Garganta, J. (2009). Trends of tactical performance analysis in team sports: bridging the gap between research, training and competition. *Revista Do Porto de Ciencias Do Desporto*, 9(1), 81–89.
- Goes, F. R., Kempe, M., Meerhoff, L. A., & Lemmink, K. A. P. M. (2018). Not Every Pass Can Be an Assist: A Data-Driven Model to Measure Pass Effectiveness in Professional Soccer Matches. *Big Data*, 6(4). <https://doi.org/10.1089/big.2018.0067>
- Gudmundsson, J., & Horton, M. (2016). Spatio-Temporal Analysis of Team Sports -- A Survey. <https://doi.org/10.1145/3054132>
- Kempe, M., Goes, F. R., & Lemmink, K. A. P. M. (2018). Smart data scouting in professional soccer: Evaluating passing performance based on position tracking data. In *Proceedings - IEEE 14th International Conference on eScience, e-Science 2018* (Vol. 55, pp. 409–410). <https://doi.org/10.1109/eScience.2018.00126>
- Link, D., Lang, S., & Seidenschwarz, P. (2016). Real time quantification of dangerousity in football using spatiotemporal tracking data. *PLoS ONE*, 11(12), e0168768. <https://doi.org/10.1371/journal.pone.0168768>
- Oberstone, J. (2009). Differentiating the Top English Premier League Football Clubs from the Rest of the Pack: Identifying the Keys to Success. *Journal of Quantitative Analysis in Sports*, 5(3). <https://doi.org/10.2202/1559-0410.1183>
- Rein, R., & Memmert, D. (2016, December 24). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus*. <https://doi.org/10.1186/s40064-016-3108-2>
- Rein, R., Raabe, D., & Memmert, D. (2017). “Which pass is better?” Novel approaches to assess passing effectiveness in elite soccer. *Human Movement Science*, 55, 172–181. <https://doi.org/10.1016/j.humov.2017.07.010>
- Sarmiento, H., Marcelino, R., Anguera, M. T., Campaniço, J., Matos, N., & Leitão, J. C. (2014). Match analysis in football: a systematic review. *Journal of Sports Sciences*, 32(20), 1831–1843. <https://doi.org/10.1080/02640414.2014.898852>

Spearman, W., Basye, A., Dick, G., Hotovy, R., & Pop, P. (2017). Physics-Based Modeling of Pass Probabilities in Soccer, 1–14.

An evaluation of the three-point rule in association football

Wim Koevoets

ger4ec@gmail.com

Abstract

The three-point rule in association football replaced the two-point rule by awarding three points to teams that win a match instead of two. The Isthmian league introduced this rule in 1973 to make football more attractive to watch. Most national football associations implemented the rule in their competitions after the 1994 World Cup.

The widespread adoption of the three-point rule suggests that the associations are better off with this rule than they would have been with the two-point rule. We investigate this suggestion by describing general mechanisms through which we expect an effect of the introduction of the three-point rule on matches to operate.

We use information on goals to estimate the relation between the introduction of the three-point rule and match outcomes for matches in the Italian Serie A. The qualitative model and Maher (1982) form the basis of our statistical models. If you believe the qualitative model then you can interpret the estimated relation as a causal effect of the introduction of the three-point rule.

1 Introduction

The three-point rule in association football replaced the two-point rule by awarding three points to teams that win a match instead of two. The Isthmian league introduced this rule in 1973 to make football more attractive to watch. The English *Football Association* (FA) introduced it in its 1981/1982 season.¹ After the 1994 World Cup the FIFA recommended national football associations to introduce the three-point rule in their domestic championships. Many national football associations which did not yet use the three-point rule implemented the rule shortly after this World Cup.²

Because the associations introduced the three-point rule to increase the attractiveness of matches, the evaluation task is to estimate its effect on the entertainment value of matches. Entertainment value is a concept for which statistical information is not available. Instead, we use the following information that is arguably related to the entertainment value of matches: the number of goals in a match, whether a match ends in a draw or not and the time that matches are in an even-scores situation (e.g. 0-0,1-1,2-2 etc.), respectively.

This paper investigates whether it is possible to estimate an effect of the three-point rule. Its evaluation presents an empirical issue for the following reason. The change from two to three points for a win is an increase in the reward for teams that win. This has a likely effect on the incentives of teams in a football match.³ However, within a season, the incentives teams have may change and depend on their opponent from match to match. Teams also change between two seasons as football clubs hire managers and buy and sell players, respectively. This holds for any two consecutive seasons.

¹In 1980 the FA was worried about the decrease in the number of people watching football.

²By 1994, some national football associations already used the three-point rule. They include associations from New Zealand, Japan, Greece, Wales, Bulgaria, Norway, Sweden, Cyprus, Turkey, Iceland and Israel.

³For this reason, the introduction of the three-point rule has been adopted as a game theory application. Game-theoretic studies on the three-point rule predict, for example, how a change from two to three points for a win affects offensive play and defensive play, respectively.

The structure of this paper is as follows. In Section 2 we discuss the origins of the three-point rule in the early 1970s and propose a qualitative model that describes causal effects of the number of points on the value of entertainment in matches. We discuss measurement issues in Section 3. Using information on matches of the Italian Serie A we present some descriptive statistics in Section 4. We present estimation results in Section 5 and discuss our findings in Section 6.

2 A causal model to investigate the three-point rule

The three-point rule in football assigns three points to a team that wins a match. Both teams get one point if the match ends in a draw. A team that loses a match does not get points. In many association football championships the three-point rule in football replaced the two-point rule. The Isthmian league introduced the three-point rule in its 1973/1974 season. This league is an English association of teams in the greater London area.⁴

The reason for the introduction of the three-point rule was to influence the style of play during matches such that watching matches became more attractive. The purpose of rewarding three instead of two points for a win was to stimulate attack but more to prevent a match from ending with a spell of dull play after two teams scored the same amount of goals.

Jimmy Hill, the inventor of the three-point rule, explained the rule as follows: *‘It was designed to attack, but even more to prevent teams shutting up shop in negative fashion long before the game’s end, having captured 50 % of the afternoon’s cake’*.^[1] In 1980, a special report on the future of football quoted an interviewee as follows: *‘It encourages a team to go for goals instead of sitting back once a draw is on the cards’*.

Others noted a potential disadvantage of the rule. Another interviewee expressed *‘Once a side has got a goal, it could put down a blanket defence and make the game even more stupid than it is already’*.^[2] Terry Neill, Arsenal manager from 1976 to 1983, was quoted saying: *‘It could make a team a goal up want to sit on their lead bit more than at present’*.^[3]

These comments suggest that the main reasons for introducing the three-point rule were to stimulate attacks and to dissuade two teams from playing the ball around until the end of the match once they reached intermediate scores of 1-1, 2-2, 3-3, etc., respectively. They also suggest that it may lead to more defensive play.

As mentioned in the introduction, other incentives are also at play in each match and, because teams change between seasons, they may be different between seasons for the same match.

To deal with these different types of incentives, most unobservable to us, we benefit from structuring our evaluation. Suppose Figure 1 is a realistic description of how rewards and incentives have effects on entertainment in matches.

Figure 1 is a Directed-Acyclic-Graph (DAG), a tool which is useful to identify causal relations that can be estimated.⁵ A graph consists of edges and nodes. The nodes in Figure 1 have

⁴The Isthmian league introduced the three-point rule after seeking and receiving advice from Jimmy Hill on how to make the football game more attractive to spectators in the early 1970s. He was a manager of Coventry City and football commentator after a long professional career as a player for Brentford and Fulham. He invented the three-point rule. When the Isthmian league changed its point system from the two-point rule to the three-point rule it also split the division into two divisions. The 1973/1974 season was also the first season with sponsorship. These simultaneous changes were meant to increase the attractiveness of play in football matches.

⁵Judea Pearl and collaborators developed methods, *do*-calculus, which help to identify estimable causal relations. Morgan and Winship (2007) include a chapter on the use of graphs in empirical analysis. DAGitty is useful software for analyzing causal diagrams.^[4] It is available at dagitty.net See ^[4] and ^[5]

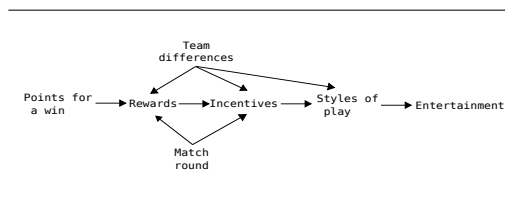


Figure 1: Effect of the number of points for a win on entertainment in a match

labels ‘Points for a win’, ‘Team differences’, ‘Rewards’, ‘Match rounds’, ‘Incentives’, ‘Style of play’ and ‘Entertainment’, respectively. Each edge links two separate nodes. The arrows on edges indicate the direction of causality, hence the name *directed* graph. Figure 1 is also an acyclic graph because it is not possible to leave one node and return to that node by following a path along connected directed edges.

The arrow from ‘Rewards’ to ‘Incentives’ means that a change in rewards changes incentives that are relevant for the competing teams in a match. The arrow from ‘Incentives’ to ‘Style of play’ means that a change in incentives for the competing teams changes their style of play. The arrow from ‘Style of play’ and ‘Entertainment’ means that the style of play affects the entertainment value in a match. The arrow from ‘Points for a win’ to ‘Rewards’ indicates a change in the rewards for teams if the points for a win change.

There are various reasons for why differences between two competing teams matter for the rewards and incentives in a match. Football teams are collections of players with different personalities. To prepare their team for the next match, football coaches and their coaching staff need to create a cohesive amalgam of players deemed best to face their opponent in the next match. Their opponent likely prepares the match similarly.

The rewards do not need to be the same for each match or monetary. They can range from bonuses for victories, claps on the shoulders of players, promises of playing at different positions or playing more matches. They can also be different for different players and vary between matches.

Team differences also have a direct effect on the style of play and not only through rewards and incentives. If a weak team plays against a stronger team, it may not have high hopes for a victory. It may opt for a defensive style of play to avoid ending the match without any point.⁶

Lastly, match rounds matter for rewards and incentives because it can make a difference whether teams are trying to fight off relegation or aiming at participation in a European competition towards the end of the season.

The model in Figure 1 is not complex, a result of the low granularity of the information we use for our analysis. It follows from Figure 1 that if we change the points for a win, we need to adjust for team differences and match rounds, in order to identify its effect on entertainment.

It is important to note that effects which are *not* in Figure 1 also reveal assumptions. For example, we assume in Figure 1 that there are no other measures that increase the value of entertainment in matches. This means Figure 1 is not adequate for the evaluation of the three-point rule in the Isthmian League, the English First Division or the World Cup 1994, for example. Other changes aimed at increasing entertainment took place simultaneously with the introduction of the three-point rule in these tournaments.⁷

⁶Guedes and Machado (2002) present a game-theoretic model in which being an underdog or a favorite in a match leads to different (optimal) strategies if the match is in an even-score situation.[6]

⁷We are not aware of other changes but the three-point rule at the time of its introduction in the Serie A.

3 Measurement issues

We do not have information measuring the match-specific values of entertainment to football associations. Therefore, we consider three match outcomes: the number of goals in a match, whether a match ends in a draw or not and the total time that a match is in a draw situation (0-0,1-1, 2-2 etc.). The first measure approximates entertainment, more goals make a match more entertaining to watch. The latter two measures are inversely related to entertainment.

If a match ends in a draw, then the supporters of *both* teams would have been happier with a win for their team. The amount of dissatisfied supporters is less with a winner of the match. A draw can then be interpreted as a measure of match dissatisfaction. We adopt this interpretation because without additional information we cannot find out the exciting matches that end in a draw or the boring matches that end with a winner.

For similar reasons, we can interpret an increase in even-score time during a match as being less attractive. It is the time supporters of *both* teams need to go through the situation in which their team is not winning the match. This situation ends with a goal or with the end of a match.

There is one arrow entering entertainment in Figure 1: ‘Style of play’ likely determines ‘Entertainment’. Returning to the discussions in the 1970s, 80s on the expected effects of the three-point rule on the style of play we can assess the possible effects on goals, draws and the even-score time.

Table 1 summarises the expected effects of the introduction of the three-point rule on these match outcomes. It reflects the expectations of commentators before the introduction of the three-point rule in the 1980/1981 season of the English First Division.

	Considered effects
Offensive play	+
Dull play	-
Defensive play by a leading team	+
	Possible effect
Goals	+/-
Probability of a draw	-
Time in even-score situations	-

Table 1: Considered effects by 1980 and possible effects on match outcomes

Table 1 shows an ambiguous possible effect for the number of goals in a match. While winning matches is more rewarding under the three-point rule, once ahead of their opponent it is possible that teams put all effort in defending their lead. The implication of this is that we would expect a lower probability of a draw. On even-score time, 0-0 situations may end sooner with an increase in offensive play while other even-score situations may end sooner by a decrease in dull play (by making scoring more attractive relative to the status-quo in these situations).

4 Data

We use information from the The Rec.Sport.Soccer Statistics Foundation on matches of the Serie A Championships in 1993/1994 and 1994/1995, respectively.⁸ Tables 6 and 7 in the

⁸Maurizio Mariani prepared and maintained these data for the Rec.Sport.Soccer Statistics Foundation. See <http://www.rsssf.com>

Appendix show the final ranking of teams for these seasons. Teams earned two points for a win in the 1993/1994 season. The three-point rule was introduced in the 1994/1995 season.

An advantage of considering two consecutive seasons over periods of seasons is that it is more likely that we are considering the same group of players when comparing the two-point rule and the three-point rule. Most players that played for a team in the 1994/1995 season also played for that team in the 1993/1994 season. With transfers of players and changes in managers between seasons this becomes less likely if we compare one season with a season many years earlier or later. Even so, one new player in a team in a new season may change the way a team plays.

Also, as a result of promotions and relegations, the competing teams are different between the seasons. Eight teams moved between the Serie A and the Serie B between the two seasons.⁹ The following 14 teams competed in both seasons: Cagliari, Cremonese, Foggia, Genoa, Inter, Juventus, Lazio, Milan, Napoli, Parma, Reggiana, Roma, Sampdoria and Torino. These teams played 182 matches against each other in each of the two seasons. In what follows, we use the information on these matches. Table 2 shows the differences in match outcomes between the two seasons.¹⁰

	Before (two-point rule)	After (three-point rule)		
	1993/1994	1994/1995	Difference	%
Goals	415	431	16	3.9
Draws	66	49	-17	-25.8
Even-score time (per 90 minutes)	103.5	96.3	-7.2	-6.9
Number of matches	182	182		

Table 2: Goals, draws and even-score times in the Serie A for teams competing in both seasons

Table 2 shows an increase of about four percent in the number of goals and draws, respectively. The even-score time shows a decrease of about seven percent between the two seasons. Table 3 shows the distribution of goals and the number of draws by score across matches for the data underlying .

Goals	1993/1994	1994/1995	Draws	1993/1994	1994/1995
0	26	18	0-0	26	18
1	30	36	1-1	37	26
2	57	56	2-2	2	4
3	31	33	3-3	1	1
4	19	22			
5	14	10	Draws	66	49
≥ 6	5	7	No draws	116	133
Matches	182	182		182	182

Table 3: Distribution of goals and draws by score and season

Table 3 shows that less 0-0 and 1-1 scores are behind the decrease in draws.

Figure 2 shows the failure probability, or the probability that a team scores a goal in an even-score situations within a given amount of time.¹¹ The gray and black line show this probability for 1993/1994 and 1995/1995, respectively.

Generally, Figure 2 shows a higher probability of a goal for the 1994/1995 season. The gap between the two seasons appears slightly larger from 75 minutes onwards.

⁹Piacenza, Udinese, Atalanta and Lecce left the Serie A after the 1993/1994 season. Bari, Padova, Brescia

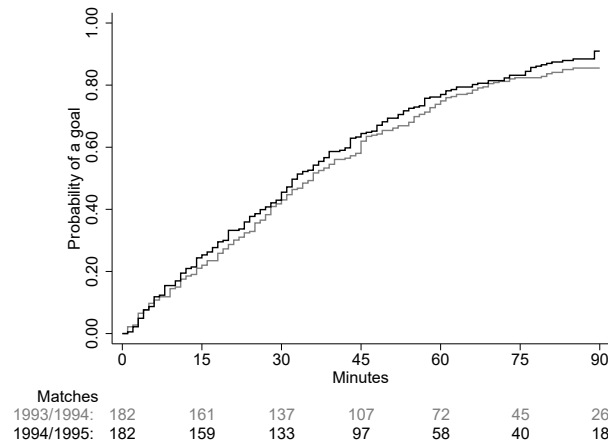


Figure 2: The probability of a goal in even-score situations within a given amount of time.

The differences between the seasons in this section appear consistent with the effects expected from the three-point rule (see Table 1). However, if we consider Figure 1 as the relevant causal model underlying our data then we cannot attribute these differences to the three-point rule. We need to adjust them for differences between teams and match rounds.

5 Estimation results

We follow Maher(1982) by allowing each match to have a different distribution of uncertainty across the possible match outcomes.[7]. We do this by including dummy variables for home teams and visiting teams in our models. We assume that the Poisson, Binomial and Weibull distributions describe the uncertainty across the possible number of goals in a match, the uncertainty about whether a match ends in a draw or not and the uncertainty across the possible number of minutes that matches are in even-score situations, respectively. Table 4 shows our statistical models.¹²

Each model adjusts the difference between the seasons for differences between teams and match rounds. The subscripts s , i , j and r indicate season s , home team i , visiting team j and match round r , respectively. Their corresponding parameters are δ_s , α_i , β_j and γ_r .

Table 9 in the Appendix shows the estimation results. Under the model in Figure 1 we can interpret the estimated coefficient of the 1994/1995 dummy variable as the estimated causal effects of the three-point rule. Table 5 summarizes the results of Table 9 in terms of relative

and Fiorentina entered the Serie A in the 1994/1995 season.

¹⁰Table 8 in the Appendix shows these comparisons for all matches.

¹¹The failure probability is calculated as one minus the Kaplan-Meier survival function. Figure 2 includes a risk table. The numbers at 15-minute intervals shows the number of matches in which teams did not score any goal. At 90 minutes this number corresponds to the number of 0-0 matches.

¹²The subscripts s, i, j and r indicate season s , home team i , visiting team j and match round r , respectively. δ_s , α_i , β_j and γ_r are their corresponding parameters.

Match outcome	Uncertainty distribution	Functional form
Goals	$g_{sijr} \sim \text{Poisson}(\lambda_{sijr})$	$\log(\lambda_{sijr}) = x_{sijr}$
Draws	$d_{sijr} \sim \text{Binomial}(p_{sijr})$	$\text{logit}(p_{sijr}) = x_{sijr}$
Even-score time	$t_{sijr} \sim \text{Weibull}(\lambda_{sijr}, p)$	$\text{hazard}(t_{sijr}) = p \lambda_{sijr} t^{p-1}$ $\lambda_{sijr} = \exp(x_{sijr})$
$x_{sijr} = \delta_s + \alpha_i + \beta_j + \gamma_r$		

Table 4: Statistical models

differences.¹³

Match outcome	Relative difference (%)		
	1994/1995-1993/1994	<i>P</i> -value	Distribution
Goals	4.0	0.537	Poisson
Probability of a draw	-20.6	0.068	Binomial
Time in ties	-7.6	0.414	Weibull

Table 5: Estimated differences (%) adjusted for team differences and match round

Table 5 shows an estimated increase of about four percent in the number of goals per match between 1994/1995 and 1993/1994. It further shows decreases of about 21% and 8% in the probability of a draw and the even-score-time, respectively. The signs of these estimates are according to the possible effects of the three-point-rule on these match outcomes (see Table 1).¹⁴ The *p*-values show the (approximate) probabilities of observing the estimated relative difference under the assumption that there is no difference. The strength of evidence for an effect of the three-point rule is larger for lower *p*-values. The *p*-values vary between the three match outcomes and the evidence for an effect of the three-point rule appears strongest (weakest) for the probability of a draw (goals).¹⁵

¹³The coefficients of the team variables in Table 9 measure differences relative to Cagliari. The coefficient of the 1994/1995 is the one of interest. It measures the difference with respect to the 1993/1994 season. The coefficients of the period dummy variables measure the difference with match rounds in the first half before New Year of the season. We classify the match round dummy variables into the following four within-season-period-dummy variables: first half before New Year, second half before New Year, first half after New Year and second half after New Year. The corresponding estimated relative differences for the models with 33 match round dummy variables (instead of three period dummy variables) are 3.9%, -35% and -8.6% with *p*-values of 0.589, 0.017 and 0.359, respectively.

¹⁴Deviance tests and mild overdispersion suggest that the Poisson and Logit models provide adequate descriptions of the number of goals and the probability of a draw, respectively. A Cox-Snell residual plot of the estimates for the Weibull model suggests the same for the time in even-score situations. The Weibull model estimates imply positive duration dependence: the longer an even-score situation lasts during a match the higher the probability that one of the two teams score a goal.

¹⁵The corresponding *p*-values of the untransformed coefficients are 0.566, 0.239 and 0.433, respectively (see on Table 9).

6 Discussion

What if the model in Figure 1 is not a correct description of how the points for a win might affect the entertainment in a match? For example, suppose that the three-point rule affected the transfer of players between 1993/1994 and 1994/1995? Figure 3 shows an adapted version of the model in Figure 1. It assumes that the number of points for a win may affect the differences between teams in a match. In this case, the number of points has both a direct and an indirect effect on rewards. The indirect effect operates through the differences between teams in a match.

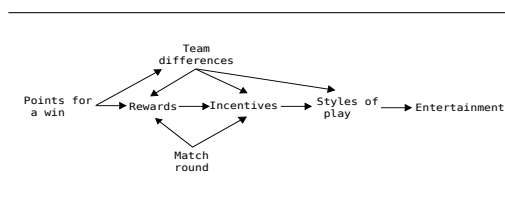


Figure 3: Figure 1 with points for a win affecting differences between teams in a match.

The model in Figure 3 may be relevant because, before introducing the three-point rule in the Serie A, the Italian football association ran two experiments with the three-point rule in its Italian Serie C1 and C2 championships, respectively. Another reason is the publicity of the 1994 World Cup in the United States and its attention to more attractive football.[8]

Without further information we cannot know whether teams changed between the two seasons anticipating one point more for a win in 1994/1995. As a result, should Figure 3 represent the causal story of the three-point rule, then we cannot interpret Table 5's estimates as causal effects of the rule.¹⁶

A final comment is that we only consider two causal models in this paper. They are relatively general because the information we have is not measuring how teams play during matches. With more detailed data we may be able to focus on an analysis of the three-point rule and the nature of play in a match once 1-1 is on the scoreboard.

References

- [1] B. Hill. *My Gentleman Jim*. The Book Guild Ltd., 2015.
- [2] Shoot. The future of football. <http://www.soccerattic.com/article/the-future-of-football>, 6 Dec 1980. Accessed: 2018-11-16.
- [3] J. Wilson. The question: Is three points for a win good for football? <https://www.theguardian.com/sport/blog/2009/feb/05/question-jonathan-wilson-three-points>, 5 Feb 2009. Accessed: 2018-11-07.
- [4] Textor, J., Hardt, J., and S. Knapp. Dagitty: A graphical tool for analyzing causal diagrams. *Epidemiology*, 22(5):745, 2011.

¹⁶Assuming the model in Figure 3 holds, the estimated coefficients of the team indicators and the 1994/1995 dummy variable both capture some effect of the three-point rule and we cannot interpret the estimated coefficient of the 1994/1995 dummy variable as the estimated effect of the rule anymore.

- [5] Morgan, S., L. and C. Winship. *Counterfactuals and Causal Inference*. Cambridge: Cambridge University Press, 2007.
- [6] Guedes, J.,C., and F. S. Machado. Changing rewards in contests: Has the three-point rule brought more offense to soccer? *Empirical Economics*, 27:607–630, 2002.
- [7] Maher, M., J. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 1982.
- [8] FIFA. Fifa World Cup USA '94. *Technical report*, 1994.

Appendix

Position, Team	Win	Draw	Loss	Goals for	Goals against	Points
1, Milan	19	12	3	36	15	50
2, Juventus	17	13	4	58	25	47
3, Sampdoria	18	8	8	64	39	44
4, Lazio	17	10	7	55	40	44
5, Parma	17	7	10	50	35	41
6, Napoli	12	12	10	41	35	36
7, Roma	10	15	9	35	30	35
8, Torino	11	12	11	39	37	34
9, Foggia	10	13	11	46	46	33
10, Cremonese	9	14	11	41	41	32
11, Genoa	8	16	10	32	40	32
12, Cagliari	10	12	12	39	48	32
13, Inter	11	9	14	46	45	31
14, Reggiana	10	11	13	29	37	31
15, Piacenza	8	14	12	32	43	30
16, Udinese	7	14	13	35	48	28
17, Atalanta	5	11	18	35	65	21
18, Lecce	3	5	26	28	72	11

Table 6: Serie A championship in 1993/1994 (two-point rule)

Position, Team	Win	Draw	Loss	Goals for	Goals against	Points
1, Juventus	23	4	7	59	32	73
2, Lazio	19	6	9	69	34	63
3, Parma	18	9	7	51	31	63
4, Milan	17	9	8	53	32	60
5, Roma	16	11	7	46	25	59
6, Inter	14	10	10	39	34	52
7, Napoli	13	12	9	40	45	51
8, Sampdoria	13	11	10	51	37	50
9, Cagliari	13	10	11	40	39	49
10, Fiorentina	12	11	11	61	57	47
11, Torino	12	9	13	44	48	45
12, Bari	12	8	14	40	43	44
13, Cremonese	11	8	15	35	38	41
14, Genoa	10	10	14	34	49	40
15, Padova	12	4	18	37	58	40
16, Foggia	8	10	16	32	50	34
17, Reggiana	4	6	24	24	56	18
18, Brescia	2	6	26	18	65	12

Table 7: Serie A championship in 1994/1995 (three-point rule)

	Before (two-point rule)	After (three-point rule)		
	1993/1994	1994/1995	Difference	%
Goals	741	773	32	4.3
Draws	104	77	-27	-26.0
Even-score time (per 90 minutes)	174.7	156.7	-27.2	-10.8
Number of matches	306	306		

Table 8: Goals, draws and even-score times - all matches in the Serie A

Distribution:	Poisson		Binomial		Weibull	
Match outcome:	Goals		Probability of a draw		Even-score time	
Functional form:	Expectation (in logs)		Logit		Hazard ratio (in logs)	
Variables	Coefficient	P-value	Coefficient	P-value	Coefficient	P-value
1994/1995	0.040	0.566	-0.467	0.051	1.083	0.433
Home team						
Cremonese	0.177	0.353	-0.116	0.847	0.940	0.827
Foggia	0.023	0.909	-0.066	0.913	0.916	0.757
Genoa	0.105	0.591	0.467	0.430	0.924	0.778
Inter	0.255	0.174	-0.135	0.825	1.202	0.497
Juventus	0.240	0.201	-1.122	0.106	1.087	0.761
Lazio	0.486	0.007	-0.885	0.183	1.435	0.193
Milan	0.022	0.910	-0.200	0.743	0.989	0.967
Napoli	0.002	0.991	-0.051	0.933	0.882	0.655
Parma	0.299	0.107	-1.921	0.025	1.327	0.298
Reggiana	-0.160	0.439	0.104	0.860	0.687	0.199
Roma	-0.012	0.952	0.356	0.547	1.176	0.561
Sampdoria	0.430	0.016	-0.354	0.570	1.585	0.078
Torino	0.170	0.372	0.527	0.373	0.964	0.895
Visiting team						
Cremonese	-0.295	0.107	-0.652	0.319	0.735	0.235
Foggia	0.055	0.741	-0.241	0.699	0.739	0.262
Genoa	-0.172	0.328	-0.028	0.964	0.713	0.205
Inter	-0.158	0.376	-0.133	0.830	0.746	0.262
Juventus	-0.087	0.613	0.080	0.896	0.738	0.242
Lazio	0.017	0.919	-0.379	0.545	0.894	0.661
Milan	-0.394	0.033	-0.022	0.972	0.592	0.067
Napoli	-0.146	0.404	-0.108	0.863	0.847	0.526
Parma	-0.046	0.787	0.042	0.945	0.970	0.907
Reggiana	-0.151	0.383	-1.544	0.044	0.837	0.498
Roma	-0.289	0.111	0.166	0.786	0.838	0.505
Sampdoria	-0.290	0.111	-0.215	0.729	0.704	0.182
Torino	-0.254	0.157	-0.381	0.548	0.910	0.717
Period within season						
2 nd < New Year	0.277	0.010	-0.728	0.049	1.150	0.358
1 st > New Year	0.106	0.318	-0.524	0.129	1.201	0.221
2 nd < New Year	0.158	0.128	-0.446	0.190	1.170	0.286
Intercept	0.671	0.001	0.253	0.709	0.010	0.000
Observations	364		364		518	
Weibull shape (in logs)					.187	0.000
Deviance test (<i>p</i> -value)	0.012		0.001			
Dispersion	1.019		1.062			

Table 9: Estimation results

A quantitative method for evaluating the skills of national volleyball teams:

Prediction accuracy comparisons of the official ranking system in the worldwide tournaments of 2010s.

Eiji Konaka¹

Meijo University, Nagoya, Aichi, Japan
konaka@meijo-u.ac.jp

Abstract

This paper proposes a quantitative skill-evaluation method for international volleyball teams. The main objective of this paper is to identify design flaws in the official FIVB (Federation Internationale de Volleyball Association) ranking by comparing its prediction performances with those of the proposed method in major worldwide tournaments (e.g., World Championships and Olympic Games, held in the 2010s). The detailed analysis showed that in the FIVB rankings, European teams were always under-evaluated, while the remaining teams were always over-evaluated. This is clear evidence of the design flaws in the FIVB ranking system.

1 Introduction

1.1 Main objective

This paper proposes a quantitative skill-evaluation method for international volleyball teams. The main objective of this paper is to identify design flaws in the official FIVB (Federation Internationale de Volleyball Association) ranking by comparing its prediction performances with those of the proposed method in major worldwide tournaments (e.g., World Championships and Olympic Games, held in the 2010s).

1.2 Background

Accurate ranking systems are important for players, event organizers, and sports enthusiasts. Players use rankings to evaluate and estimate their skill levels, while event organizers use rankings as a criterion in tournament design tasks such as group draws, player (team) seeding, and guest player (team) selection. Sports enthusiasts use rankings to evaluate the skill of a team and predict the results of matches. Inaccurate ranking systems confuse and disappoint event organizers, players, and enthusiasts by increasing the gap between predictions and match results. Therefore, accurate ranking systems are required to aid in the creation of attractive and consistent sporting events.

The number of wins and the percentage of victories are the most “fair” ranking criteria if all players are matched in a round-robin format. However, a fair round robin is not possible when the number of teams participating is larger than the number of schedulable matches. In particular, not all the national teams of major sports can compete in a fair round-robin format. As a result, teams have different opponents and play different numbers of matches, and volleyball is no exception. Major worldwide tournaments, such as Olympic Games and World Championships, conduct qualifying tournaments in each continental confederation.

To rank and order teams according to their abilities, the international association of each sport designs its own original ranking system. The most popular ranking system is based on an accumulative method[1]. This system calculates *ranking points* for each team as the sum of the points attributed to international tournaments and standings in the tournaments. The sum is calculated for a designated period, such as four years. World rankings of national volleyball teams are also defined using an accumulative method[2]. In an accumulative method, the overall design guideline for the attributed points for each standings, qualification and seeding processes, and tournament formula is necessary to construct accurate ranking systems.

However, FIVB rankings, the official worldwide volleyball rankings for national teams, has the following problems[3]:

- These ranking point attribution designs have no clear mathematical or statistical basis, and therefore the ranking points do not directly measure the scoring ability of the teams.
- Only a limited number of nations have right to enter a worldwide tournament as the host. The teams can get ranking points in the tournament, therefore, the teams are always overestimated in the FIVB ranking.
 - In addition, that tournament has a small spot for European nations. As a result, European nations in the second group cannot get ranking points in that tournament, and are thus always underestimated in the ranking.

In [3, 4], the authors proposed a quantitative skill-evaluation method based on scoring ratio in each match. The accuracy of the proposed method was verified in two Olympic Games[3] and five ball games in Rio Olympic Games[4]. Konaka[3, 4] showed that the prediction accuracy of his proposed method is significantly higher than that of the official rankings.

In the current study, the authors modified the method proposed in [3, 4], and then applied the proposed skill-evaluation and match-prediction model to 733 matches in World Championships and Olympic Games from 2010 to 2018. The prediction results were as follows:

- Match prediction accuracy of the proposed method is better than that of the official FIVB rankings, and
- Prediction of the qualifying teams from the first round by the proposed method is significantly better than the official FIVB rankings.

These results clearly show the design flaw in the official FIVB ranking system.

This paper is organized as follows: Section 2 describes the proposed skill-evaluation method. In Section 3, the rating calculation result and its prediction performance are shown. Section 4 provides a comparison between the proposed method and the official rankings. In addition, the reason for the differences in their prediction performances is discussed. Finally, Section 5 concludes this paper.

2 Rating: a quantitative skill evaluation method

Here, we define *ranking* and *rating* as follows:

- *ranking*: the order of teams.
- *rating*: a quantitative value associated with the ability of each team.

The objective of this study is to create a ranking based on ratings.

Assume that the following two elements affect the result of a match:

1. the stable and constant skill and ability of each team.
2. condition, form, luck, and other unstable and nonconstant elements.

The ranking points in the accumulative method include both sets of elements. In contrast, a point-exchange system estimates the first set of elements by denoising the effects of the second set. In this study, the rating is a quantitative value calculated using a statistical method based on the first set of elements.

2.1 Current ranking systems: FIVB rankings

The FIVB, the world governing body for volleyball, regularly reports the rankings of its member nations' teams. The FIVB Board of Administration has designed a system of point attribution for selected FIVB world and other official competitions[2]. Table 1 shows the points awarded in three international tournaments.

Table 1: FIVB Ranking Point System (2018)

Standing	Tournament name			
	Olympic Games	World Cup	World Championship	
			Men	Women
1	100	100	100	100
2	90	90	90	90
3	80	80	80	80
4	70	70	70	70
5	50	50	62	58
6	—	40	56	—
7	—	30	50	50
8	—	25	—	—
9	30	5	45	45
10	—	5	—	—
11	20	5	40	40
12	—	5	—	—
13 Tie			36	36
15 Tie			33	33
17 Tie			30	30
21 Tie			25	25

The design shows significant inconsistencies. For instance, there are no clear mathematical and statistical bases on the following attribution designs.

- The champions of several competitions are each awarded 100 points equally.
- Differences exist in points between standings.

Another inconsistency lies in the number of spots allocated to each confederations in World Cup. The World Cup is the second oldest event of the FIVB[5], and its champion is awarded 100 ranking points, which are equivalent to those awarded in equal the Olympic Games and World Championships. From 1977, this tournament was continuously hosted by Japan every four years. Two slots were allocated to the host and the latest World Champion. The remaining ten slots were equally allocated to five confederations. Therefore, only two European teams could appear this tournament. This seems unfair because Europe comprises many teams that can attain a better performance than the other confederations. Table 2 summarizes the final standings of the European teams in the World Championships from 2010. At least six teams were ranked 12th or above.

Table 2: European teams in World Championships

Year	Sex	Teams	Teams from Europe	Final standings
2018	M	24	10	1, 4, 5, 6, 7, 8, 10, 11, 12, 16
2018	W	24	8	1, 2, 4, 8, 10, 11, 12, 15
2014	M	24	9	1, 3, 4, 5, 9T(2), 13T(2), 17T
2014	W	24	10	4, 5T, 7T, 9T(2), 11T(2), 13T(2), 15T
2010	M	24	9	3, 4, 5, 7, 8, 10, 11, 12, 13T
2010	W	24	9	1, 5, 6, 7, 8, 9, 11, 15T, 17T

As a result, less ranking points went to Europe, and therefore the European teams could be underestimated in the current FIVB ranking system.

2.2 Proposed method: definition and calculation of rating values

As mentioned earlier, official ranking points do not directly measure the scoring ability of each team.

We propose a simple statistical estimation method of scoring ratios based on the score achieved in each match, which is always officially recorded and is common to all ball games.

Assume that the scoring ratio of team i in a match against team j (i and j are team indices), denoted as $p_{i,j}$, is estimated as

$$p_{i,j} = \frac{1}{1 + \exp(-(r_i + r_{adv} - r_j))}, \quad (1)$$

where r_i is defined as the *rating* of team i . Define $\vec{r} = (r_1, \dots, r_{N_T})^T$. r_{adv} denotes the effect of home advantage. Equation (1) is a model where team i hosts the corresponding match, and $r_{adv} = 0$ if the match is held in a neutral place. Here, assume that r_{adv} affects all teams equally.

Given (s_i, s_j) , the actual scores in a match between i and j ,

$$s_{i,j} = \frac{s_i}{s_i + s_j} = p_{i,j} + \epsilon_{i,j}, \quad (2)$$

where $s_{i,j}$ and $\epsilon_{i,j}$ are the actual scoring ratio and the estimation error, respectively.

This mathematical structure is a well-known *logistic regression model* and is widely used in areas such as in the winning probability assumption of Elo ratings in chess games[6] and the correct answer probability for questions in an item response theory[7, 8].

The update law is designed to minimize the sum of the squared error E^2 between the result and prediction, defined by the following equation:

$$E^2 \equiv \sum_{(i,j) \in \text{all matches}} (s_{i,j} - p_{i,j})^2. \quad (3)$$

$$r_i \leftarrow r_i - \alpha \cdot \frac{\partial E^2}{\partial r_i}, \quad (4)$$

$$r_{hadv} \leftarrow r_{hadv} - \alpha \cdot \frac{\partial E^2}{\partial r_{hadv}}, \quad (5)$$

where α is a constant.

By definition, the rating is an interval scale. Therefore, its origin, $r = 0$, can be selected arbitrarily and a constant value can be added to all r_i . For example,

$$\vec{r} \leftarrow \vec{r} - (\max \vec{r}) \cdot \vec{1} \quad (6)$$

implies that $r = 0$ always shows the highest rating, and $r < 0$ shows the distance from the top team.

2.2.1 Conversion to rating on winning probability

Rating r_i defined in (1) explains the scoring ratio via the logistic regression model.

Once we have scoring ratio $p_{i,j}$ given in (1), assume that the following independent Bernoulli process is executed N times, starting from $(s_i, s_j) = (0, 0)$.

$$\begin{cases} s_i \leftarrow s_i + 1 & \text{with probability } p_{i,j}, \\ s_j \leftarrow s_j + 1 & \text{with probability } 1 - p_{i,j} \end{cases} \quad (7)$$

At the end of the set, $s_i \geq s_j + 2, s_i \geq 25$ shows that team i wins a set against team j .

The probability of score difference or set count difference can be expressed by the cumulative distribution function for a normal distribution. In many applications, it is common to use a logistic regression model rather than a cumulative distribution[9].

Based on the discussions above, we convert the rating on the scoring ratio to that of a winning probability, as follows:

$$\hat{w}_{i,j} = \frac{1}{1 + \exp(-D_k(r_i + r_{hadv} - r_j))}, \quad (8)$$

$$D_k^* = \arg \min_{D_k} \sum (w_{i,j} - \hat{w}_{i,j})^2, \quad (9)$$

where

$$w_{i,j} = 1 \text{ (} i \text{ wins) or } 0 \text{ (} j \text{ wins)} \quad (10)$$

denote variables for match results (won/lost). Then, r_i is converted as follows:

$$\bar{r}_i = D_k^* r_i, \quad i = 1, 2, \dots, N_T. \quad (11)$$

Therefore, \bar{r}_i is a rating that explains the winning probability, and it can be utilized in match result predictions.

In Equations (3) and (9), the sum of squared errors are used as a loss function instead of the cross entropy. This is because these problems are regression problems and not classification ones.

2.3 Rating update during tournament

Elo rating[6] updates the rating values after every match according to the following update law:

$$\bar{r}_i \leftarrow \bar{r}_i + K(w_{i,j} - \hat{w}_{i,j}), \quad (12)$$

where

$$\hat{w}_{i,j} = 1 / \left(1 + 10^{-\frac{\bar{r}_i + \bar{r}_{adv} - \bar{r}_j}{400}} \right). \quad (13)$$

In this update, $K = 32$ is frequently used when the number of matches are small.

This short-term update is used in this study. The following update law is applied after every match.

$$r_i \leftarrow r_i + K(s_{i,j} - p_{i,j}), \quad K = \frac{32 \log_e 10}{400D_k^*}. \quad (14)$$

3 Rating calculation result and its prediction performance

In this section, the prediction performance is disclosed for the World Championships and Olympic Games after 2010 by using the proposed prediction method.

3.1 Predicted tournament and used datasets

In international volleyball, the World Championships and Olympic Games are the most authoritative tournaments.

Both tournaments are held every four years, and one is held two years after the other. The World Championships have the largest teams in both main (24) and qualifying tournaments. The Olympic Games are smaller (12 teams) but are considered as the most famous sporting event in the world.

In this study, the match results in the following five tournaments and ten events are predicted.

- World Championships: 2010, 2014, and 2018.
- Olympic Games: 2012, and 2016.

The match results in the following tournaments before two years from the prediction target event are used to calculate the ratings.

- World Cup: 2011 and 2015.
- Continental Championships. Every two years in odd-numbered years.
 - Asia (AVC), Africa (CAVB), Europa (CEV), South America (CSV), North America and Caribbean Nations (NORCECA), and Pan-American Games (Co-hosted by NORCECA and CSV).
- Qualifying tournaments for the World Championships and Olympic Games.
- World Grand Champions' Cup. An invitational event hosted by Japan (one year after the Olympic Games).

- World League (Men’s event. Until 2017), World Grand Prix (Women’s event. Until 2017), Nations League (Both men’s and women’s events inaugurated from 2018), and their associated qualifying events. (Every year.)

Table 3 summarizes the numbers of teams and matches in each tournament and the corresponding dataset.

Table 3: Dataset size for each tournament

Year	Name [*]	Sex	Teams	Matches	Teams in dataset	Matches in dataset
2018	WCh	M	24	94	133	999
2018	WCh	W	24	103	119	965
2016	OL	M	12	38	83	774
2016	OL	W	12	38	75	714
2014	WCh	M	24	100	149	974
2014	WCh	W	24	102	140	1057
2012	OL	M	12	38	85	713
2012	OL	W	12	38	79	574
2010	WCh	M	24	78	120	788
2010	WCh	W	24	104	107	686
Total				733		8244

[*]: WCh=World Championships, OL=Olympic games

3.2 Predicted items and prediction methods for comparison

The rating values are calculated using the datasets in Table 3. Every match results in a prediction target tournament is predicted using both the proposed rating and official FIVB rankings[2].

Prediction methods are as follows:

- Proposed rating method: the team with higher rating will win.
- Official FIVB ranking: the team with higher ranking (points) will win.

Then, both the prediction accuracies were compared.

Not only the matches, but the qualifying teams from the first round are also predicted. In both the World Championships and Olympic Games, the participating teams are divided into small pools, of each of which the first round is of the single-round robin format.

- Single round-robin by six teams. The top four teams will be qualified for the subsequent round (all tournaments except the 2010 Men’s event).
- Single round-robin by four teams. The top three teams will be qualified for the second round (2010 Men’s event).

Prediction methods are as follows:

- Proposed rating method: the team with the higher rating will qualify for the first round.

- Official FIVB ranking: the team with higher ranking (points) will qualify for the first round.

Then, the prediction accuracies obtained through both the proposed method and the official FIVB ranking were compared.

These prediction results can reveal the over/under-estimated teams in the official FIVB ranking. For the latest World Championships 2018, a detailed prediction using (8) was compared with the actual result.

3.3 Prediction result

Tables 4 and 5 show the prediction accuracy and classification table for the match results through ten tournaments, respectively.

Table 4: Prediction accuracy (all matches)

Year	Name[*1]	Sex	Matches	Corrects		Accuracy	
				[*2]	[*3]	[*2]	[*3]
2018	WCh	M	94	68	68	0.723	0.723
2018	WCh	W	103	86	86	0.835	0.835
2016	OL	M	38	28	29	0.737	0.763
2016	OL	W	38	32	29	0.842	0.763
2014	WCh	M	100	74	69	0.740	0.690
2014	WCh	W	102	80	78	0.784	0.765
2012	OL	M	38	30	31	0.789	0.816
2012	OL	W	38	31	29	0.816	0.763
2010	WCh	M	78	59	53	0.756	0.679
2010	WCh	W	104	77	72	0.740	0.692
Total		M	348	259	250	0.744	0.718
		W	385	306	294	0.795	0.764
		All	733	565	544	0.771	0.742

bold: better performance

[*1]: WCh=World Championships, OL=Olympic games

[*2]: Proposed method

[*3]: Official rankings

Table 5: Classification table (all matches)

Official					
		Corrects	Incorrects		
Proposed	Corrects	486	79	565	0.771
	Incorrects	58	110	168	
		544	189	733	
		0.742			

McNemar's p -value = 0.0875

Tables 6 and 7 show prediction accuracy and classification table for the qualifying teams from the first round for ten tournaments, respectively.

The p -value for a null hypothesis of “the prediction accuracies of both methods are equal” is provided in each classification table. A script obtained from [10] was used to calculate p -value.

Table 6: Prediction accuracy (qualification to second round)

Year	Name[*1]	Sex	Teams	Corrects		Accuracy	
				[*2]	[*3]	[*2]	[*3]
2018	WCh	M	24	22	18	0.917	0.750
2018	WCh	W	24	22	20	0.917	0.833
2016	OL	M	12	10	11	0.833	0.917
2016	OL	W	12	12	10	1.000	0.833
2014	WCh	M	24	24	20	1.000	0.833
2014	WCh	W	24	20	14	0.833	0.583
2012	OL	M	12	12	10	1.000	0.833
2012	OL	W	12	10	10	0.833	0.833
2010	WCh	M	24	16	18	0.667	0.750
2010	WCh	W	24	20	18	0.833	0.750
Total		M	96	84	77	0.875	0.802
		W	96	84	72	0.875	0.750
		All	192	168	149	0.875	0.776

bold: better performance

[*1]: WCh=World Championships, OL=Olympic games

[*2]: Proposed method

[*3]: Official rankings

Table 7: Classification table (qualification to 2nd round)

		Official			
		Corrects	incorrects		
Proposed	Corrects	143	25	168	0.875
	Incorrects	6	18	24	
		149	43	192	
		0.776			

McNemar’s p -value = 1.23×10^{-3}

For the latest World Championships 2018, a Monte-Carlo numerical simulation was performed along with the tournament formula[11, 12]. The final standings were predicted using the winning probability model (8).

The prediction results from 10^4 tournament simulations are shown in Figures 1 and 2, where the names of the nations are written in IOC three-letter code[13]. The numbers in parentheses show the latest FIVB ranking just before the start of the tournaments, and the bar length shows the probability of the final standings.

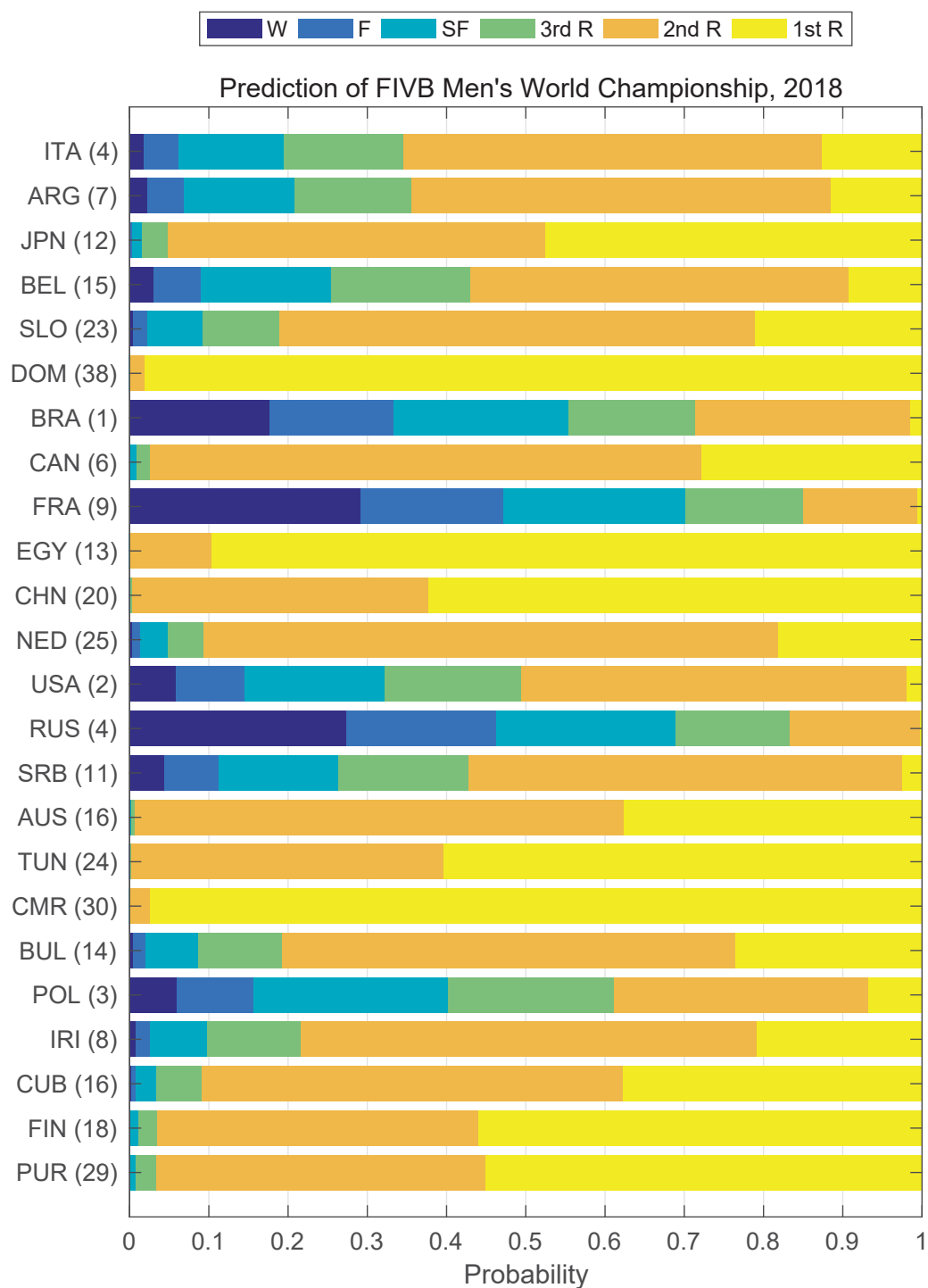


Figure 1: Tournament prediction, 2018 WCh, Men

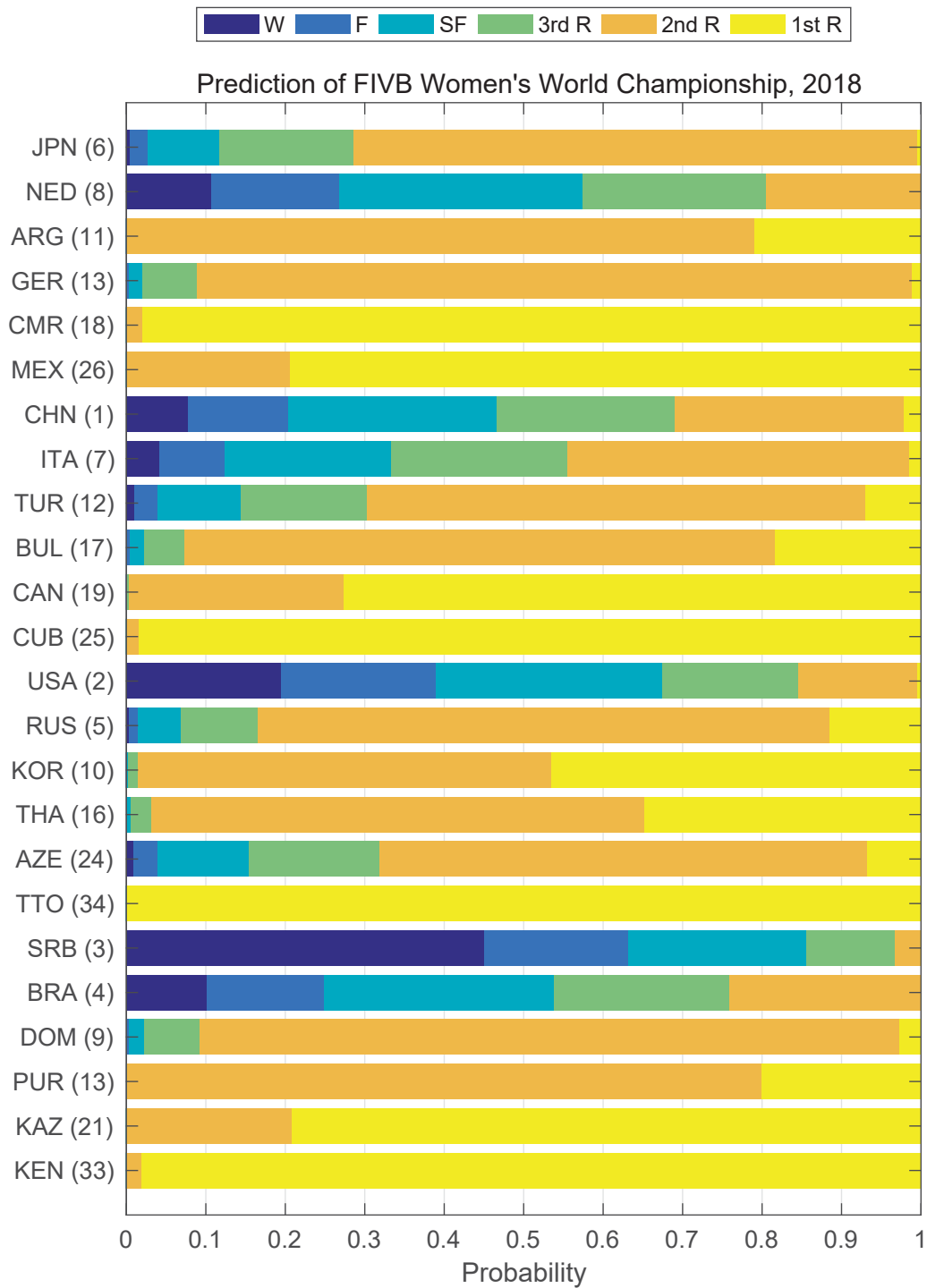


Figure 2: Tournament prediction, 2018 WCh, Women

4 Discussions

In this section, the validity of the proposed method is discussed based on the prediction results shown in the previous section.

4.1 Prediction accuracy

Tables 4 and 5 show that the proposed method can realize better prediction performance than the official FIVB ranking. However, there is no significant difference in the prediction accuracy for match results between the proposed method and official FIVB ranking, i.e., $p = 0.0875 > 0.05$.

In contrast, Tables 6 and 7 show a significant difference in the prediction accuracy for the qualification of the first round between the proposed method and official FIVB ranking, i.e., $p = 1.23 \times 10^{-3} < 0.05$.

These results show that there is small difference in the match-result-prediction accuracy between the proposed method and official FIVB ranking. The small difference is accumulated through five (or three) matches in the single-round robin format. Therefore, the proposed method is a better prediction method and can estimate each team's skill more accurately than the official FIVB rankings.

4.2 Over/under-estimation in official rankings

In the prediction of the qualification from the first round, the two prediction methods made different predictions for 31 teams, as detailed in the following:

Table 8: Different predictions between official ranking and proposed method

FIVB ranking	Proposed method	Result	Teams
Qualify	Not qualify	Qualify	2 teams. EGY(10M), KOR(10W)
Qualify	Not qualify	Not qualify	13 teams. JPN(18M), EGY(18M), KOR(18W), ITA(16W), TUN(14M), EGY(14M), ARG(14W), CAN(14W), CUB(14W), SRB(12M), SRB(12W), ALG(10W), PUR(10W).
Not qualify	Qualify	Qualify	12 teams. SLO(18M), NED(18M), AZE(18W), NED(16W), FIN(14M), CHN(14M), CRO(14W), BUL(14W), BEL(14W), GER(12M), PER(10W), TUR(10W).
Not qualify	Qualify	Not qualify	4 teams. FRA(16M), TUR(12W), IRI(10M), KEN(10W).

* **bold**: correct prediction.

The teams in the second group (FIVB ranking: qualify, proposed method: **not qualify**, result: not qualify) were overestimated teams in the FIVB ranking. Only two out of the thirteen teams are from Europe. The remaining, 11 out of the 13 teams are from other continents. Further, the teams in the third group were underestimated teams in the FIVB ranking; here, 10 out of the 12 teams are from Europa. These results show that European teams are underestimated and the other teams are overestimated in the official FIVB rankings.

Figures 3 and 4 illustrate the relationship between the ranking points and proposed rating values in August 2017 (latest before World Championships 2018). These figures show that some European teams are not awarded FIVB ranking points even when they can achieve similar rating values as the other teams.

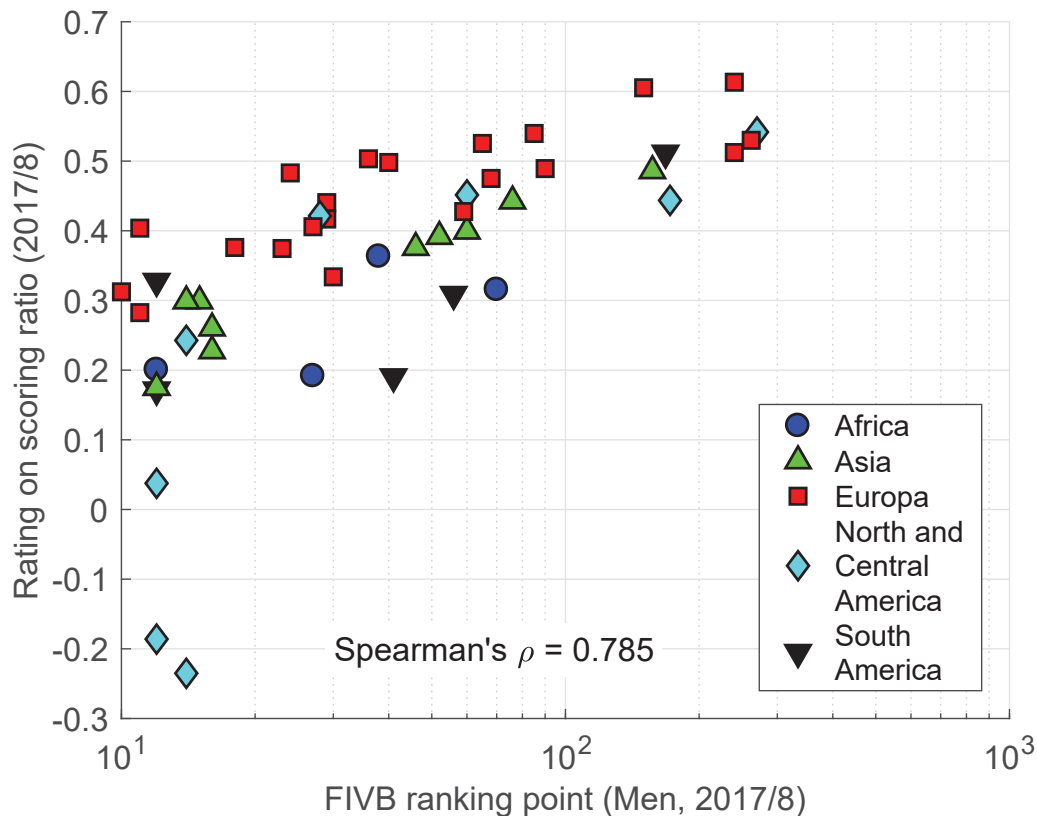


Figure 3: Ranking point and rating (Men, 2017/8)

4.3 Prediction and result: World Championships 2018

This section reviews the World Championships 2018 from the viewpoint of the Japan teams.

4.3.1 Men's tournament

The 24 qualified teams were divided into six pots based on their FIVB rankings. Teams in the top two pots were allocated to four pools through a serpentine system, while those in the remaining four pots were allocated according to ballots.

Japan was the first appearance since 2010 and its FIVB ranking was 12; it was the third in Pool A (see Figure 1). If the FIVB ranking was correct, Japan had enough chance to advance to the second round. In contrast, Japan was ranked fifth in Pool A by the proposed method. The predicted probability of qualification was approximately 0.5.

Japan ranked fifth and was out of the tournament after the first round. The set and score ratios were also fifth in the pool. This result supports that the proposed evaluation is better than the FIVB rankings. One of the reasons for this result was that Japan was in the same pool as Belgium and Slovenia, who are both underestimated European teams. However, the major reason is that Japan is always overestimated in the FIVB ranking system.

If rankings were determined by the proposed method, Japan would have been ranked 16th out of 24 teams. In this case, Japan could have been matched with two lower-ranked teams, and therefore could have higher qualifying possibility from the first round than the actual.

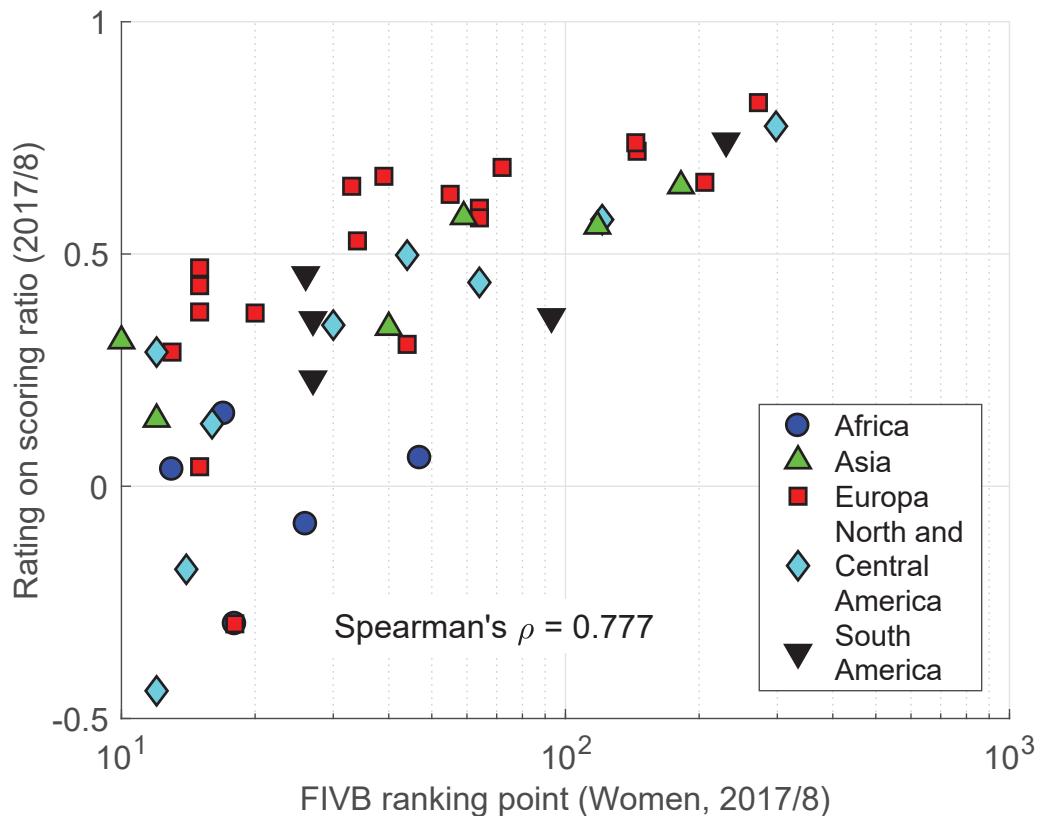


Figure 4: Ranking point and rating (Women, 2017/8)

4.3.2 Women's tournament

Japan was ranked 6th in the women's FIVB ranking. However, the proposed method ranked Japan at the 9th position out of 24 teams. The predicted probability was approximately 0.3 to advance to the third round. Therefore, their realistic goal was to achieve top 6 position (third round) and not semifinals.

In the women's tournament, the prediction accuracy is higher than in the men's tournament (see Tables 4 and 6). In other words, skill gap between teams are wide and lower-rated team could rarely make upset victory. Therefore, the prime concern of Japan in the first and second rounds was that how to win against the top three teams, i.e., Netherlands, Serbia, and Brazil in Pools A and D to reach the third round.

In the second round, Serbia could clinch the third round before the match against Japan. Serbia's top scorer, Tijana Boškovic, was out of the court after the second set in the match against Japan[14]. Taking this command could be understood as a strategic player selection looking on the subsequent rounds including semifinals and final.

Japan have won this match with sets 3 – 1 and advanced to the third round. However, Japan lost with sets 0 – 3 in the rematch against Serbia in the third round. In this rematch, Boškovic appeared in all three sets as a starter with no replacements and scored 24 points. She was the top-scorer in the match among both teams[15].

Based on the skill evaluation by using the proposed method, Japan could exploit their home advantage and they could reach their realistic goal (top 6).

5 Conclusion

This paper proposed a quantitative skill-evaluation method for international volleyball teams. The main objective of this study was to identify design flaws in the official FIVB ranking by comparing its prediction performances with the proposed method in major worldwide tournaments (e.g., World Championships and Olympic Games, held in the 2010s). By using a detailed analysis, in the FIVB rankings, European teams were always under-evaluated, while the remaining teams were always over-evaluated. This is clear evidence of the design flaws in the FIVB ranking system.

References

- [1] Stefani Ray. The methodology of officially recognized international sports rating systems. *Journal of Quantitative Analysis in Sports*, 7(4), 2011.
- [2] FIVB. FIVB volleyball world rankings. <http://www.fivb.org/en/volleyball/Rankings.asp>, 2016. accessed on 2016/6/14.
- [3] E. Konaka. Statistical rating method for volleyball national teams and its application to result prediction and competition format design. *Proceedings of the institute of statistical mathematics*, 65(2):251–269, 2017. (in Japanese).
- [4] Eiji Konaka. A statistical rating method for team ball games and its application to prediction in the Rio Olympic Games. In *proceedings of the MathSport International 2017 conference*, pages 204–216. Padova University Press, 2017.
- [5] Dr. Ary S. Graça F°. FIVB president’s message. <http://worldcup.2015.men.fivb.com/en/competition/presidentsmessages/fivbpresident>, 2015. accessed 2018/10/11.
- [6] Arpad E. Elo. *Ratings of Chess Players Past and Present*. Harper Collins Distribution Services, hardcover edition, 1979.
- [7] R. Hambleton, H. Swaminathan, and H.J. Rogers. *Fundamentals of Item Response Theory (Measurement Methods for the Social Science)*. Sage Publications, Incorporated, new. edition, 9 1991.
- [8] R. J. de Ayala. *The Theory and Practice of Item Response Theory (Methodology in the Social Sciences)*. Guilford Pr, 1 edition, 12 2008.
- [9] J. Lasek, Z. Szlávík, and S. Bhulai. The predictive power of ranking systems in association football. *International Journal of Applied Pattern Recognition*, 1(1):27–46, 2013.
- [10] G. Cardillo. McNemar test: perform the McNemar test on a 2x2 matrix. <http://www.mathworks.com/matlabcentral/fileexchange/15472>, 2007.
- [11] FIVB. Formula. Website of the 2018 FIVB Volleyball Men’s World Championship. <http://italy-bulgaria2018.fivb.com/en/competition/formula>, 2018. accessed 2018/9.
- [12] FIVB. Formula. Website of the 2018 FIVB Volleyball Women’s World Championship. <http://japan2018.fivb.com/en/competition/formula>. accessed 2018/9.
- [13] International Olympic Committee. List of all national olympic committees in IOC protocol order. <https://www.olympic.org/national-olympic-committees>, 2016. referred in 2016/6/15.
- [14] FIVB. Match - Japan-Serbia (in the second round). Website of the 2018 FIVB Volleyball Women’s World Championship. <http://japan2018.fivb.com/en/schedule/9233-japan-serbia/post>, 2018. accessed 2018/10/11.
- [15] FIVB. Match - Japan-Serbia (in the third round). Website of the 2018 FIVB Volleyball Women’s World Championship. <http://japan2018.fivb.com/en/schedule/9244-japan-serbia/post>, 2018. accessed 2018/10/11.

ELO or Coca Cola, which ranking is better?

Ruud H. Koning* Hidde Jan Goinga

January 24, 2019

In August 2018, the International Federation of Association Football (FIFA) published a new method for calculating the FIFA ranking. The ELO ranking model is used, which as self-proclaimed by the FIFA would be a better model than the old FIFA ranking model. The main aim of the new ELO ranking model was to identify an algorithm that was not only intuitive, but also easy to understand. Moreover, it should improve the overall accuracy of the ranking formula (FIFA, 2018b). The old FIFA ranking model is the ranking model used by the FIFA from straight after the FIFA World Cup 2006, i.e., August 2006, up until August 2018. The new ELO ranking model is used as the new ranking model for the Men's FIFA ranking. This ELO ranking method is used from August 2018 onwards. The ELO ranking model has been used before. For the Women's FIFA ranking it has already been in use since 2003. For chess, a sport which the creator of the ELO ranking played, uses the ELO ranking since 1960 (Schulz, 2018).

The FIFA ranking is an important method to allocate teams in qualification draws, to allocate teams in the available groups of a confederation tournament and to allocate teams in groups for the World Cup. The FIFA ranking decides which team will be put at which pot (FIFA, 2018a). All participating national teams will be ordered on current ranking. When considering the World Cup of 2018, the eight highest ranked teams that qualified will be allocated to pot 1, the eight teams below the highest 8 ranked teams are allocated to pot 2, and similarly for pot 3 and pot 4. Both for the qualification for a tournament as for the tournament itself, ranking plays an important role in allocating teams. A fair draw is important, as it could be the case that if a team has the wrong ranking, it could encounter tougher teams in either the qualification for a tournament, or the tournament itself. When a national team encounters a tougher opponent it is more likely that the national team fails to proceed the next round.

The further a national team proceeds a certain tournament, for example the

*Ruud H. Koning, Department of Economics, Econometrics and Finance, Faculty of Economics and Business, University of Groningen, The Netherlands, email: r.h.koning@rug.nl.

World Cup, the more a national federation will earn. One could argue that the money obtained from having a fair draw will be beneficial in the future, as national federations can invest the prize money into the next generation of football players. Hence, having a bad draw could result in a loss of earnings from a tournament in comparison to a fair draw, which in itself can have an impact for future generations of players. For the World Cup of 2018 a sum of 791 million dollars was available for all participants. A minimum of eight million dollars was available for each participant (FIFA, 2017). The existence of a ranking model that will fit reality in the best possible way is of utmost importance as it should be fair to every national team regarding the distribution of money. It should not be the case that, due to a bad model and thus a bad draw that national teams have a loss in a sufficiently large sum of earnings.

Regarding the modelling of football match outcomes, there are two possible models mostly used. Firstly, the Poisson type models which estimate the probability of every possible score line in a match (see, for example, Karlis and Ntzoufras (2003); Koning et al. (2003)). These types of models are becoming increasingly popular with the advent of betting markets on exact score-lines. Secondly, the ordinal outcome models are used, which predict the outcome of a match directly (see, for example, Goddard (2005)). The outcome of a match can either be a win for team 1, a draw or a win for team 2.

As the new ELO ranking is a new ranking method by the FIFA there has not been any comparison between the two FIFA ranking methods. As the ELO ranking model should be an upgrade to the old ranking model, it yet has to be tested if this is the case. In this research, both the old ranking method and the ELO method will be compared. The goal is to determine which model describes the actual outcomes most accurate. The World Cup of 2018 in Russia will be used as a point in time in order to compare both models.

In this paper the two ranking methods from FIFA will be evaluated and compared in order to determine which model is better. This will be checked by comparing the predicted outcomes with the actual outcomes for both the old ranking method and the ELO ranking method. In order to obtain these probabilities for the outcomes of all World Cup matches, an ordinal outcome model will be used to calculate the probabilities for the three possible outcomes of a match. The ordinal outcome model that will be used for this research is the ordered probit model. The method used consists of two parts. First of all, the parameters from the ordered probit model are calculated such that the outcome of a match can be predicted. Second, the predicted outcomes will be compared to the actual outcomes.

REFERENCES

- FIFA (2017). FIFA Council confirms contributions for FIFA World Cup participants. <https://www.fifa.com/about-fifa/news/y=2017/m=10/news=fifa-council-confirms-contributions-for-fifa-world-cup-participants.html>.
- FIFA (2018a). 2018 FIFA World Cup Russia™ - News - Close-up on Final Draw procedures. <https://www.fifa.com/worldcup/news/close-up-on-final-draw-procedures-2921440>.
- FIFA (2018b). Men's Ranking Procedure. <https://www.fifa.com/fifa-world-ranking/procedure/men>.
- Goddard, J. (2005). Regression models for forecasting goals and match results in association football. *International Journal of Forecasting* 21(2), 331–340.
- Karlis, D. and I. Ntzoufras (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)* 52(3), 381–394.
- Koning, R.H., M. Koolhaas, G. Renes, and G. Ridder (2003). A simulation model for football championships. *European Journal of Operational Research* 148(2), 268–276.
- Schulz, André (2018, Jun). Fifa introduces an elo system. <https://en.chessbase.com/post/fifa-fuehrt-elosystem-ein>.

Random Walks with Memory Applied to Grand Slam Tennis Matches Modeling

Tomáš Kouřim

Institute of Information Theory and Automation, Czech Academy of Sciences Prague, Czech Republic
kourim@outlook.com

Abstract

The contribution presents a model of a random walk with varying transition probabilities implicitly depending on the entire history of the walk, which is an improvement of a model with varying step sizes. The transition probabilities are altered according to the last step of the walker using a memory parameter to either reward or punish success by increasing or decreasing its probability in the next step. This walk is applied to model Grand Slam tennis matches and fitted on their entire history since 2009. The suitability of the model is thoroughly tested on a number of real datasets. The model seems to be robust and describe well the majority of matches, making it an useful tool to produce precise *in-play* odds.

1 Introduction

Tennis is one of the most popular sports both on professional and amateur level. Millions of people pursue tennis as their leisure time activity [3] and same numbers hold also for the people following and watching the professional tennis competitions. Tennis also plays a major role in the sports betting industry, which grows rapidly and becomes more and more important part of the global economy. In the Czech Republic only, the total sales in sports betting industry reached CZK 64.5 billion (2.9 billion USD) in 2017, representing 1.3% of Czech GDP [8]. The immense size of the betting market attracts also many fraudsters. The European Sports Security Association regularly reports on suspicious betting activities, the latest report (2018) contained 267 cases of such activity, 178 (67%) in tennis [1]. It is thus obvious that a precise model describing the game of tennis has many possible uses in real life.

Tennis is also a sport more than suitable to be modeled using random walks or random processes in general, as it naturally consists of many such processes. A series of tennis matches is a random walk, the sequence of sets within a match, games within a set, points within a game or even strokes within a point can be all considered a random process and modeled using a random walk. Additionally, these walks are well described by the tennis rules and there exist lots of data describing these random processes (i.e. various tennis result databases provided by the tennis federation as well as many private subjects). In this paper, the random walk consisting of a sequence of sets within a match is studied. Matches played as a *best-of-five*, i.e. the men Grand Slam tournaments, are considered in this paper. In these matches, up to 5 steps of the random walk can be observed, making them more suitable than the *best-of-three* games, where maximum 3 steps can occur.

The matches are modeled using a new type of a recently introduced random walk with varying probabilities [6], which is a modification of a random walk with varying step size introduced by Turban [9]. It seems more than suitable to model tennis matches as the data suggest that a success in tennis yields another success, or in other words, that winning one particular part of the match increases the chances of winning the next part as well. This behavior is well described by the new random walk model.

The paper is organized as follows. Next chapter introduces the new type of random walk used for tennis modeling. Section 3 provides general description of the data used, Section 4 shows how to obtain starting probabilities. In Chapter 5 the actual model is described and its performance is evaluated. Section 6 concludes this paper.

2 Random walk with varying probability

In 2010, Turban described [9] a new version of a random walk with memory, where the memory is introduced using variable step size. This idea was further extended by Kouřim [6, 7] and an alternative version of a random walk with memory was introduced, where the memory affects the walk through varying transition probabilities.

The walk evolves in a following way. Initial step is made following the result of a Bernoulli random variable with starting probability parameter p_0 , that is,

$$P(X_1 = \text{"right"}) = p_0.$$

From the second step on, the transition probability in the $t - th$ step is given by

$$X_{t-1} = \text{"right"} \implies P(X_t = \text{"right"}) = \lambda p_{t-1}$$

$$X_{t-1} = \text{"left"} \implies P(X_t = \text{"right"}) = 1 - \lambda(1 - p_{t-1})$$

for some $\lambda \in (0, 1)$. When the directions are formalized so that $\text{"right"} \approx 1$ and $\text{"left"} \approx -1$, the formula for the $t - th$ transition probability can be rewritten as

$$p_t = \lambda p_{t-1} + \frac{1}{2}(1 - \lambda)(1 - X_t). \quad (1)$$

This definition of a random walk means that the opposite direction is always preferred and that the walk tends to return back to the origin. Alternatively, inverse approach can be applied and the same decision can be supported. Formally, the expression for the $t - th$ transition probability is then

$$p_t = \lambda p_{t-1} + \frac{1}{2}(1 - \lambda)(1 + X_t). \quad (2)$$

For more details on the walk and its rigorous definition, see the original papers [6, 7].

3 Data description

For the purpose of this study, a database containing the results from all Grand Slam tournaments from 2009 to 2018 and corresponding Pinnacle Sports bookmaker's odds¹ was created based on the information publicly available from website www.oddsportal.com. There are 4 Grand Slam² tournaments each year, 40 tournaments together. Each Grand Slam has 128 participants playing in a single-elimination system (i.e. 127 games per tournament), making it a set of 5080 games together. However, the games where either one of the players retired were omitted from the dataset and so were the matches where no bookmaker's odds were available. Together there were 4255 matches with complete data available, presenting total 432 players.

¹This bookmaker is considered leading in the sports betting industry.

²Australian Open, French Open, The Wimbledon and US Open.

The most active player was Novak Djokovic, who participated in 188 matches. On average, each player played 19.7 matches, with the median value of 8 matches played. The most common result was 3:0, occurring 2138 times, on the other hand, 5 sets were played only 808 times.

The order in which the players are listed is rather random. The first listed players³ won 2201 in total, just slightly over the half. On the other hand, if the bookmaker's favorite (i.e. the player with better odds or the first listed player in case the odds are even) is considered, the situation changes significantly. The favorites won 3307 matches in total, mostly 3:0, and lost 311 times 0:3, 347 times 1:3 and 290 times 2:3. It suggests that bookmaker's odds can be used as a probability estimate, which is in accordance with previous results, for example [5].

4 Initial probability derivation

The model of a random walk with varying probabilities described in Section 2 takes two parameters, initial set winning probability p_0 and the memory coefficient λ . Finding the optimal value of λ is the main subject of this paper and is described in Section 5.

Estimating the initial set winning probability is a major task by itself and represents one of the elementary problems in tennis modeling. For the purpose of this article an estimation based on bookmaker's odds will be used. Specifically, the closing odds⁴ by Pinnacle Sports bookmaker for the first set result are used to estimate the probabilities of each player winning the first set, i.e. p_0 and $1 - p_0$. Such odds represent a good estimation of the underlying winning probability and are considered as a baseline in the sports betting industry. The odds, however, have to be transformed into probabilities. A method described in [4] is used to obtain probabilities, using a parameter $t \in [0, 1]$ set to the value $t = 0.5$. Obtained first set winning probabilities are then used as a given starting probability p_0 in the random walk.

5 Model description and evaluation

5.1 Model description

Original inspiration of the random walk described in Section 2 is based on intensive study of historical sport results and their development. The data suggest that the probability of success (i.e. scoring, winning a set or a point etc.) evolves according to the random walk with varying probabilities. Moreover, it follows from the data that sports can be very roughly divided into two categories. Sports played for a certain amount of time, such as soccer or ice-hockey, evolve according to the walk defined by expression 1. On the other hand, sports where there is necessary to achieve certain number of points, such as tennis or volleyball, appear to follow the pattern defined in equation 2. Therefore the later approach is used to model a tennis game.

The model is used to predict the winning probabilities of sets 2 through 5 and is constructed in a following manner. For each match, the first set winning probability of Player A⁵, p_0 , is given (see Section 4) and a coefficient λ is fixed for the entire dataset. In order to compute the second set winning probability⁶, the result of the first set is observed and second set winning

³Such player/team would be normally considered as "home", however, as there are (usually) no home players on the international tournaments, the order is based on the www.oddsportal.com data and/or the respective tournament committees.

⁴Closing odds means the last odds available before the match started.

⁵The player which is listed first in the database, see Section 3 for details.

⁶Winning probability of Player A is always considered as Player B winning probability is just the complement.

Year	Optimal lambda
2010	0.8074
2011	0.8497
2012	0.8142
2013	0.9162
2014	0.8523
2015	0.8429
2016	0.8920
2017	0.8674
2018	0.8333

Table 1: Optimal values of the coefficient λ for respective years.

probability is computed using equation 2. This procedure is repeated for all remaining sets played.⁷

5.2 Model evaluation

In order to verify the model's accuracy, several tests were performed. First, the dataset was divided into training and testing sets. The division can be done naturally by the order of games played. Given a specific time, past matches constitute to a training set, future matches to a testing set. For the purpose of this paper, the split was done on a yearly basis, the data from one previous tennis season were used as a training set to predict winning probabilities in the following season, considered the testing set (i.e. 2010 was the first season used as testing data, 2017 was the last season used as training data), making it 9 training/testing splits together. Another approach to dataset splitting is to consider data from all previous years as testing data and from one future year as training data, however, previous study shows that the difference between these two approaches is negligible [5].

Next step in model verification is the estimation of parameter λ . Training sets and maximal-likelihood estimates were used for this task. The likelihood function is defined as

$$L = \prod_{i=1}^{N_{train}} (x_i p_i + (1 - x_i)(1 - p_i)),$$

where N_{train} is the number of sets 2 thru 5 played in the training dataset, p_i is Player A's winning probability in the i -th set obtained using equation 2 for each match, and x_i is the result of the i -th set, $x_i = 1$ if Player A won the i -th set, $x_i = 0$ otherwise. For computational reasons the *log-likelihood* $L_l = \log(L)$ was used, i.e. the function

$$L_l = \sum_{i=1}^{N_{train}} \log(x_i p_i + (1 - x_i)(1 - p_i))$$

was maximized. Numerical methods implemented in Python library SciPy were used to obtain specific values of λ . The optimal values of the coefficient λ can be seen in Table 1.

Finally, the model was used to predict set winning probabilities of the unseen data from the training set using initial bookmaker derived odds, equation 2 and memory parameter λ

⁷There can be either 3, 4 or 5 sets played in total in a *best-of-five* tennis game.

obtained from the corresponding training set. In order to verify the quality of the model, the average theoretical set winning probability of Player A $\hat{p} = \frac{1}{n} \sum_{i=1}^{N_{test}} p_i$ and its variance $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{N_{test}} p_i(1 - p_i)$ were computed and so was the observed Player A winning ratio $\bar{x} = \frac{1}{n} \sum_{i=1}^{N_{test}} x_i$. Using the Lyapunov variant of Central Limit Theorem [2], the resulting random variable y follows the standard normal distribution

$$y = \frac{\sqrt{N_{test}}(\bar{x} - \hat{p})}{\hat{\sigma}} \sim \mathcal{N}(0, 1).$$

Then ow to verify the model accuracy, the the null hypothesis that the true average Player A set winning probability \bar{p} equals \hat{p} against the alternative hypothesis $\bar{p} \neq \hat{p}$ was tested. Formally,

$$H_0 : \bar{p} = \hat{p}$$

$$H_1 : \bar{p} \neq \hat{p}.$$

One of the assumptions of the CLT is that the observed random variables are independent. This is obviously not true in the case when N_{test} contains all sets from the testing data. Quite the opposite, the model assumes that the winning probability of a set directly depends on the winning probability of the previous set. This can be easily solved by splitting the testing dataset into 4 subsets containing only results from single set of each match, i.e. sets 2, 3, 4 and 5 (if they were played). The matches can be considered independent from each other and so can be the $i - th$ sets of respective matches.

Using this approach, there are 36 testing sets⁸ together. On a 95% confidence level, only on 2 out of the 36 available subsets provide strong enough evidence to reject the null hypothesis. On the other hand, the null hypothesis is relatively weak. It only says that the prediction is correct on average. In order to verify the quality of the predictions, more detailed tests have to be created. This can be done primarily by testing the null hypothesis on many subsets created according to some real life based criteria. The natural way how to create such subsets is dividing the matches to the 4 different tournaments. This refining yields 180 subsets⁹ altogether. Using 95% confidence level, only 6 of the 180 subsets have data strong enough to reject the null hypothesis. It is worth mentioning that the size of some of the datasets regarding fifth sets is only slightly above 20 observations, which can interfere with the assumptions justifying the use of Central Limit Theorem.

To further analyze the robustness of the model it is important to realize the structure of the data. So far, the player, whose winning probability was estimated, was chosen arbitrarily based on some external (more or less random) order. As such, the observed winning probability in every subset equals approximately to $\frac{1}{2}$, see further Section 3. In such a dataset it is not very difficult to estimate the average winning probability. The situation changes if the bookmaker's favorite is considered for predictions (more details on who is the favorite and how to choose him in Section 3). Performing the same tests as described in the previous paragraph the data allows to reject the null hypothesis (at 95% confidence level) on 5 subsets containing all tournaments and 8 single tournament subsets (out of 180 subsets total).

Finally, the testing data can be divided into groups using the initial probability p_0 . Such a division is based on an assumption that the matches with similar bookmaker odds should have similar development. The matches are divided into 5 groups, each containing 10 percentage points in first set winning probability. Except for the biggest favorites (with first set winning probability over 90%), this division seems reasonable. The data histogram can be seen on Figure

⁸Up to 4 sets considered in each match, 9 yearly testing datasets.

⁹4 sets in each match evaluated, 4+1 tournaments every year, 9 years for testing.

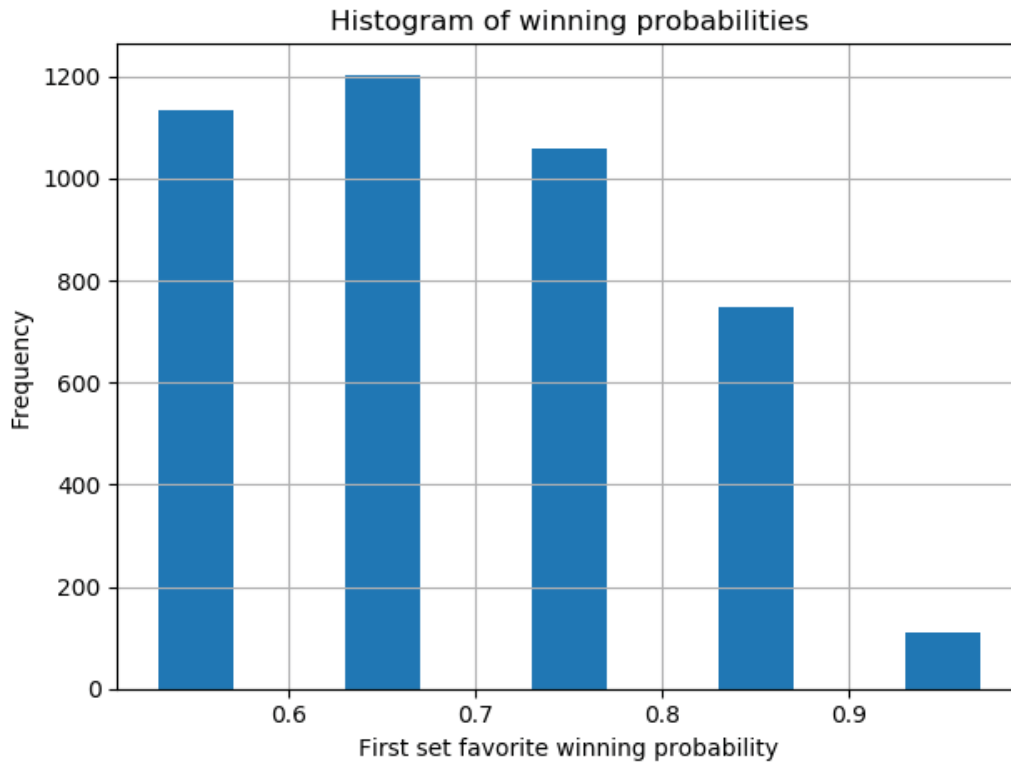


Figure 1: First set winning probability p_0 histogram.

1. Out of the 180¹⁰ newly created odds-based subgroups, only 9 have data strong enough to reject H_0 on a 95% confidence level. The entire results of the hypothesis testing (the p -values of respective tests) can be seen on Figure 2.

Overall, the model was tested on 360 different subsets and only 22 of them (6.1%) provided enough evidence to reject H_0 on 95% confidence level. These subsets are distributed randomly and there is no pattern among them, indicating there is no systematic bias in the model. The random walk with varying probabilities thus seems to be a robust model which can be used to precisely predict set winning probabilities in men tennis Grand Slam matches.

6 Conclusion

This paper describes the random walk with varying probabilities and its application on Grand Slam tennis data. A model describing the development of a single match is introduced and tested on a dataset containing all matches from seasons 2009-2018. The results show that the model is robust and performs well on the absolute majority of reasonable data subsets. This suggest that the model could be used as a tool to generate precise *in-play* odds during the

¹⁰Further division, i.e. by tournament and odds, was not performed as the resulting datasets would not contain enough data.

Year	Set number	Australian Open	French Open	Wimble don	US Open	0,6	0,7	0,8	0,9	1	All groups
2010	2	0,377	0,551	0,278	0,847	0,064	0,764	0,877	0,264	1,000	0,439
	3	0,981	0,629	0,302	0,703	0,738	0,091	0,927	0,644	1,000	0,561
	4	0,105	0,200	0,040	0,837	0,000	0,893	0,636	0,228	1,000	0,013
	5	0,808	0,270	0,156	0,509	0,567	0,060	0,076	0,930	1,000	0,084
2011	2	0,159	0,649	0,409	0,494	0,155	0,900	0,854	0,229	1,000	0,222
	3	0,893	0,696	0,622	0,553	0,195	0,108	0,875	0,697	1,000	0,689
	4	0,823	0,525	0,625	0,474	0,996	0,772	0,930	0,870	1,000	0,868
	5	0,496	0,329	0,014	0,427	0,144	0,130	0,622	0,206	1,000	0,127
2012	2	0,677	0,540	0,237	0,348	0,482	0,167	0,704	0,235	0,081	0,113
	3	0,752	0,304	0,264	0,621	0,245	0,852	0,440	0,313	0,928	0,138
	4	0,104	0,267	0,810	0,161	0,450	0,135	0,105	0,019	0,304	0,031
	5	0,223	0,184	0,412	0,359	0,279	0,452	0,358	0,019	0,137	0,192
2013	2	0,664	0,944	0,954	0,218	0,867	0,119	0,854	0,090	0,486	0,696
	3	0,306	0,629	0,320	0,476	0,359	0,476	0,001	0,197	0,175	0,373
	4	0,647	0,585	0,949	0,879	0,285	0,096	0,302	0,082	0,912	0,578
	5	0,488	0,501	0,510	0,385	0,357	0,907	0,377	0,161	1,000	0,579
2014	2	0,277	0,410	0,448	0,450	0,894	0,501	0,957	0,092	0,229	0,341
	3	0,244	0,612	0,511	0,987	0,908	0,181	0,404	0,894	0,885	0,253
	4	0,221	0,048	0,025	0,337	0,082	0,010	0,867	0,036	0,616	0,001
	5	0,191	0,142	0,495	0,792	0,117	0,852	0,240	0,170	1,000	0,636
2015	2	0,883	0,593	0,669	0,075	0,257	0,765	0,095	0,766	0,251	0,757
	3	0,084	0,223	0,565	0,272	0,227	0,798	0,447	0,237	0,294	0,081
	4	0,150	0,101	0,738	0,778	0,440	0,828	0,148	0,095	1,000	0,113
	5	0,025	0,316	0,428	0,454	0,063	0,694	0,520	0,907	1,000	0,089
2016	2	0,602	0,936	0,268	0,194	0,956	0,596	0,959	0,955	0,072	0,644
	3	0,563	0,021	0,697	0,341	0,411	0,929	0,867	0,169	0,986	0,442
	4	0,101	0,560	0,607	0,191	0,125	0,102	0,828	0,619	0,971	0,039
	5	0,240	0,311	0,352	0,035	0,197	0,067	0,074	0,562	0,593	0,008
2017	2	0,062	0,023	0,586	0,965	0,531	0,076	0,391	0,008	0,469	0,069
	3	0,677	0,901	0,154	0,146	0,852	0,053	0,636	0,654	0,504	0,110
	4	0,228	0,972	0,498	0,390	0,723	0,382	0,908	0,886	0,658	0,446
	5	0,526	0,381	0,542	0,465	0,783	0,613	0,729	0,466	1,000	0,915
2018	2	0,911	0,491	0,354	0,530	0,043	0,393	0,265	0,635	0,398	0,239
	3	0,765	0,233	0,404	0,165	0,481	0,730	0,473	0,965	0,444	0,320
	4	0,793	0,176	0,456	0,650	0,704	0,577	0,046	0,965	1,000	0,157
	5	0,258	0,216	0,171	0,060	0,841	0,052	0,399	0,806	1,000	0,190

Figure 2: p -values of hypothesis tests for different testing sets. Red are marked those allowing to reject H_0 on 99% confidence level, orange on 95% and yellow on 90% confidence level.

matches or to directly compete against the odds currently provided by the bookmakers.

7 Remarks

The source code containing all functionality mentioned in this article is freely available as open source at GitHub¹¹ together with a database containing all data that was used in this paper. Some results can be also obtained from the same repository.

References

- [1] European Sports Security Association. Essa 2018 annual integrity report. <http://www.eu-ssa.org/wp-content/uploads/ESSA-2018-Annual-Integrity-Report.pdf>, 2018. Accessed: 2019-05-12.
- [2] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 1995.
- [3] Physical Activity Council. 2019 physical activity council's overview report on u.s. participation. www.physicalactivitycouncil.com/pdfs/current.pdf, 2019. Accessed: 2019-05-12.
- [4] Tomáš Kouřim. Mathematical models of tennis matches applied on real life odds. *Doktorandské dny FJFI*, 2015. Available at <http://kmwww.fjfi.cvut.cz/ddny/historie/15-sbornik.pdf>.
- [5] Tomáš Kouřim. Predicting tennis match outcomes using logistic regression. *Doktorandské dny FJFI*, 2016. Available at <http://kmwww.fjfi.cvut.cz/ddny/historie/16-sbornik.pdf>.
- [6] Tomáš Kouřim. Random walks with varying transition probabilities. *Doktorandské dny FJFI*, 2017. Available at <http://kmwww.fjfi.cvut.cz/ddny/historie/17-sbornik.pdf>.
- [7] Tomáš Kouřim. Statistical analysis, modeling and applications of random processes with memory. *PhD Thesis Study, ČVUT FJFI*, 2019.
- [8] Ministry of Finance of the Czech Republic. Hazard games overview for 2017. <https://www.mfcr.cz/cs/soukromy-sektor/hazardni-hry/archiv-zakon-c-202-1990-sb/vysledky-z-provozovani/2017/hodnoceni-vysledku-provozovani-loterii-2016-32211>, 2018. Accessed: 2019-01-23.
- [9] Loïc Turban. On a random walk with memory and its relation with markovian processes. *Journal of Physics A: Mathematical and Theoretical*, 43(28):285006, 2010.

¹¹<https://github.com/tomaskourim/mathsport2019>

The Age Advantage in Youth Football

Stephen Lawrence, Laura Jonker and Jan Verbeek

steve.lawrence.ata@gmail.com laura@xoet.nl jan.verbeek@knvb.nl

Abstract

Average Team Age (ATA) and a Relative Age Index (RAEi) are variables against which performance outcomes in football can be measured and we consistently find that performance advantages are evident when measured against these variables. In ‘The Age Advantage in Association Football’, Lawrence, S., MSp2015, 6,389 matches played by males at U17 to adult team ages* were examined providing evidence of an age advantage. In this new paper analysis of additional match data from U12 to adult, providing a total dataset of 15,088 matches, is presented providing further insight into the development of relative age effects and signifying a causal connection between cut-off date eligibility rules and such effects. We conclude with a proposition that replacement of cut-off date rules with ‘average team age’ rules will assist with the elimination of such effects.

1 Introduction

For the purposes of this study the Average Team Age (ATA) is calculated as the mean of the chronological ages of the players composing the starting line-up in any given match.[†]

The Relative Age Index (RAEi) is calculated as the number of early-born players divided by the total number of players composing the starting line-up in any given match and is expressed as a decimal proportion between 0 and 1. Early-born players are those players born in the first six months of the competition year defined by the eligibility cut-off date. In our sample of youth players in the Netherlands this is January to June (following FIFA’s eligibility rule for youth football).[‡]

We use home team advantage as a comparator variable as a well-known and graphic comparison to age advantage and relative age advantage. An expected ratio in adult professional football of 46% - 24% - 30%, for Home win – Draw – Away win, translates into **1.62** points per game (PPG) for the home team and **1.14** PPG for the away team. We use this PPG parameter as an expression of competitive advantage.

* Data courtesy of Gracenote.

[†] ‘The Age Advantage in Association Football’, Lawrence, S., MSp2015

[‡] ‘The Age Advantage in Association Football’, Lawrence, S., MSp2015

2 Methodology

The accumulated data including the anonymised team-sheet, match date, date of birth for each player and match result, was compiled into sets for each competition. The percentages of wins, draws and losses accruing to each team according to the three variables (ATA, RAEi & Home/Away) were calculated and then converted into PPG for comparison. For each competition an average PPG score for home teams, away teams, older teams, younger teams, more biased teams and less biased teams was thus established along with the average age of the teams. The competitions were then aggregated into 10 age groupings for clarity. At youth age levels the aggregated group data was drawn from the same competition across 2 seasons in 2013/14 and 2014/15. Figure 1 shows how the data has been aggregated into 10 groupings.

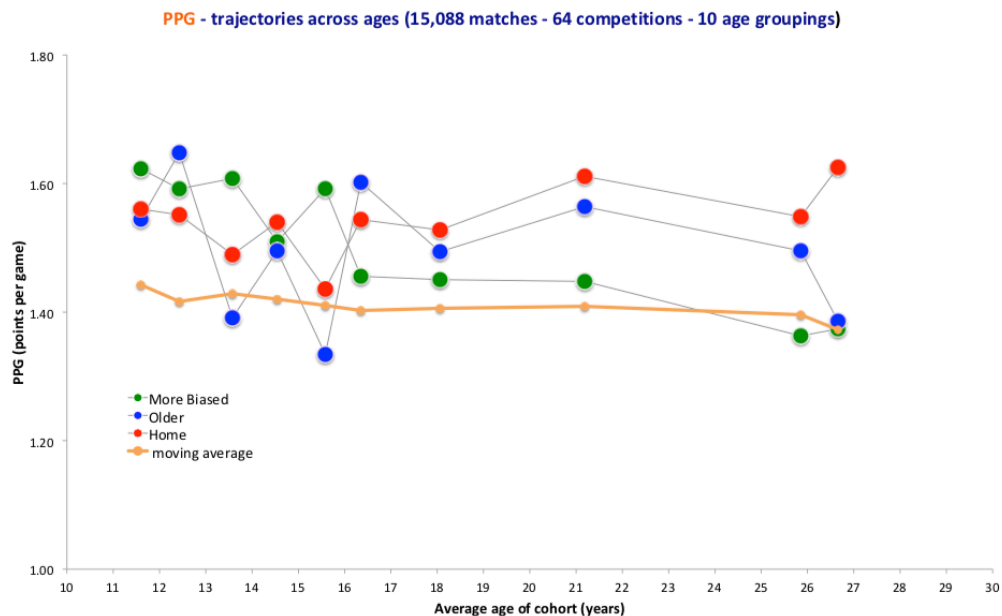


Figure 1. Graph of PPG accrued by the **Home** Team (red), **Older** Team (blue) and **More-biased** Team (green) with the moving average (calculated including away, younger and less-biased teams) shown (orange).

3 Comparing Home Team Advantage with Age Advantage and Relative Age Advantage

Three charts, each including moving average lines and peak height velocity (PHV) curves for reference purposes, show how our three chosen variables relate to the PPG parameter and the average cohort age.

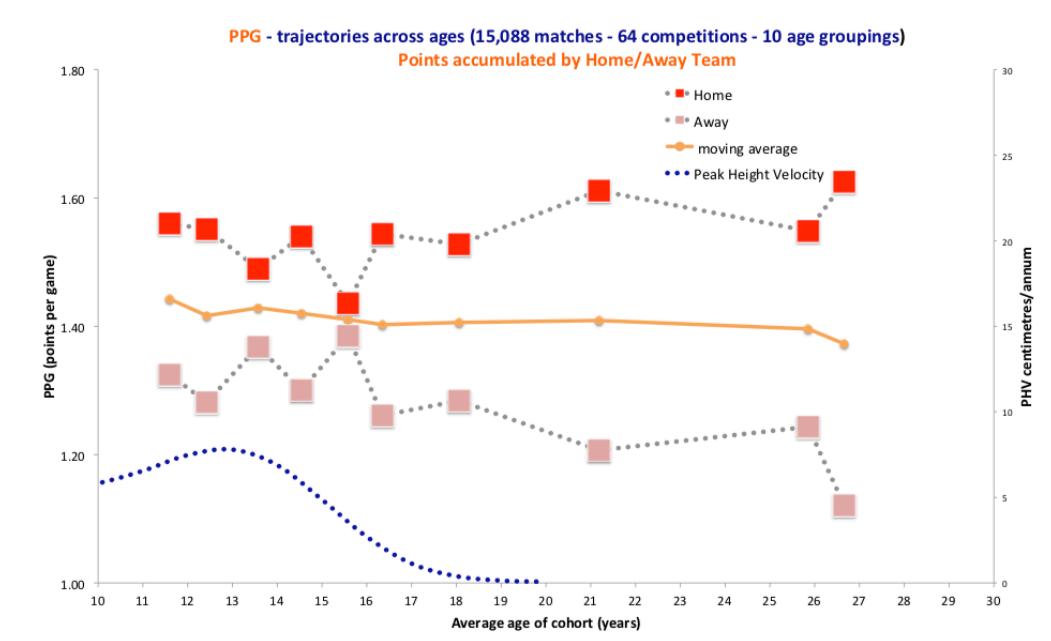


Figure 2. Graph of PPG accruing to the **Home** Team (red) or the **Away** Team (pink).

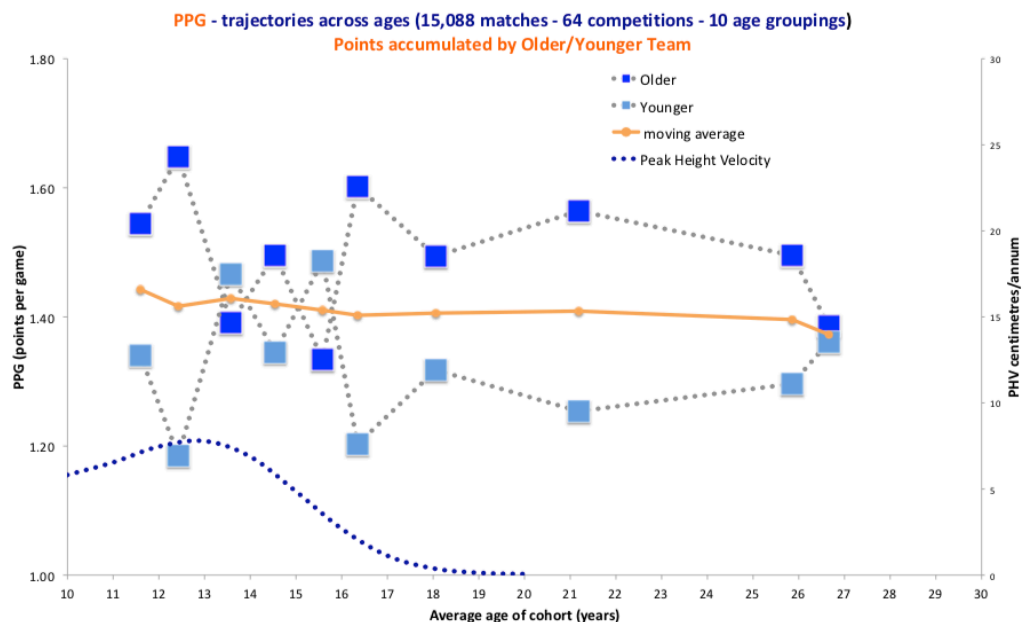


Figure 3. Graph of PPG accruing to the **Older** Team (blue) or the **Younger** Team (light blue).

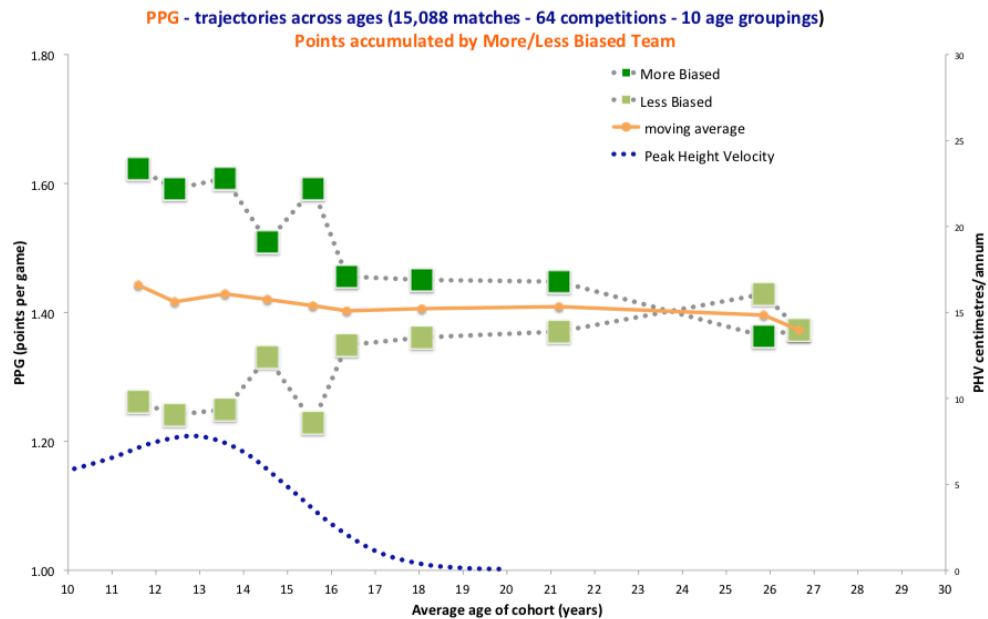


Figure 4. Graph of PPG accruing to the **More-biased** Team (green) or the **Less-biased** Team (light green).

Patterns of diminishing advantage for older or more biased teams, over time, are evident. No such diminution is evident in respect of home team advantage, which in our dataset is at its maximum at the cohort average age of 26.7. This corresponds with the age of minimum age advantage and minimum relative age advantage.

The p-values and chi-square values for the eight youth age groupings plus the semi-pro grouping are shown in Figure 5. The null hypotheses were that home or away teams, older or younger teams and more-biased or less-biased teams all had the same chance of winning matches.

Age Group	ATA	RAEi	n	Home/Away p-value	Home/Away chi-sq	ATA p-value	ATA chi-sq	RAEi p-value	RAEi chi-sq
Under 12	11.598	0.648	191	0.2485632482	1.331	0.3173105079	1.000	0.0646885438	3.413
Under 13	12.427	0.643	1,335	0.0003199830	12.950	0.0000000007	38.162	0.0000016930	22.915
Under 14	13.575	0.659	519	0.3194953451	0.991	0.5377235267	0.380	0.0013462033	10.278
Under 15	14.542	0.691	739	0.0179044401	5.605	0.1376077149	2.205	0.0559740195	3.653
Under 16	15.581	0.683	413	0.7038060054	0.145	0.2540516395	1.301	0.0024318454	9.191
Under 17	16.350	0.646	1,241	0.0002141461	13.703	0.0000001786	27.252	0.1248633764	2.355
Under 19	18.060	0.638	1,243	0.0014746617	10.110	0.0215535829	5.281	0.2003912073	1.640
U23 (2010-2013)	21.187	0.615	890	0.0000086877	19.780	0.0006502320	11.626	0.3424621581	0.901
Netherlands semi-pro	25.853	0.526	3,107	0.0000000002	40.592	0.0000327627	17.250	0.1331310074	2.256
Total			9,678						

Figure 5 Table of p-values and chi-square values for 8 youth age groupings and the semi-pro age grouping. All youth data derives from Netherlands[§] competition during seasons 2014/15 & 2015/16 except for U23 data which is from 2010/11, 2011/12 & 2012/13.

[§] Data courtesy of Koninklijke Nederlandse Voetbalbond.

The data also shows a steady decline in evident relative age bias, as measured by RAEi, as competition cohort ages increase, with parity being achieved in some competitions with cohort average ages of around 27 years. This is what we would expect if the importance of an age advantage and thus a relative age advantage diminished with increasing age. See Figure 6.

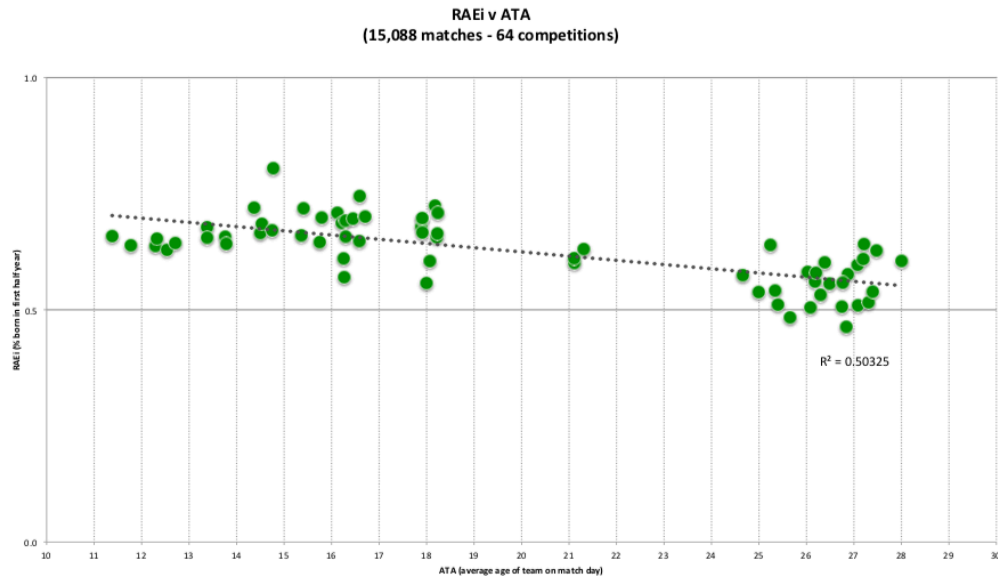


Figure 6. Graph of diminishing RAEi as the ATA of competition cohorts increases.

In respect of the ATA correlation (Figure 3) we observe volatility with higher p-values in the U12, U14 and U16 cohorts (Figure 5). We note that this occurs following a period of maximum peak height velocity and corresponds with a period of minimum evident mean age differences between competing teams (n.b. the mean team age difference is < 0.16 when RAEi is ~ 0.65). See Figure 7.

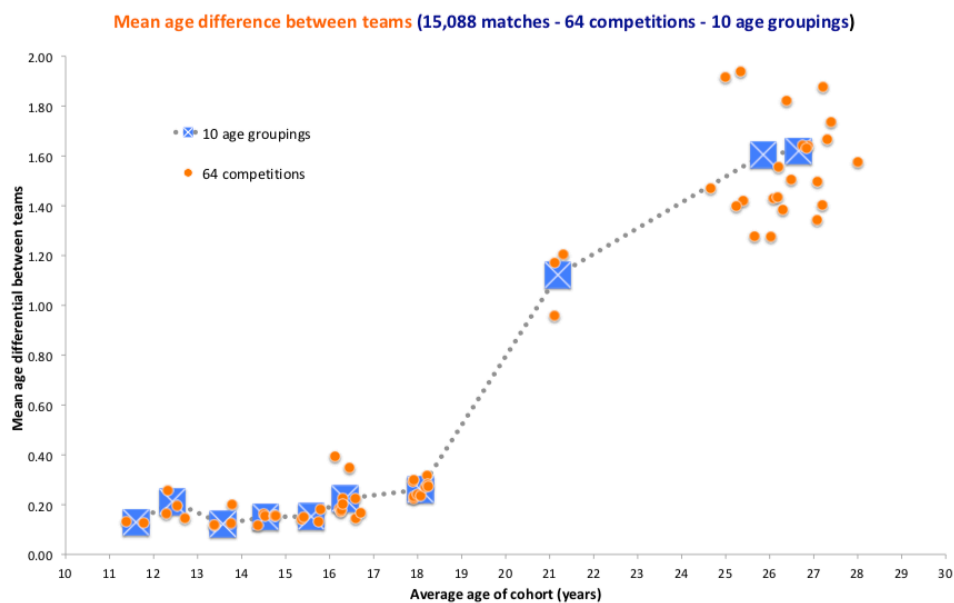


Figure 7. Graph of increasing mean age difference between teams as the average cohort age rises.

4 Conclusion

Our data shows that at youth ages, older football teams experience a competitive advantage when playing against younger teams (Figure 3). The data also shows that the advantage diminishes, as competition cohorts get older. This happens in the context of rigid competition eligibility age rules defined by a cut-off date. In order for any given team to be older than its competitor, it must necessarily consist of more players whose ages are closer to the cut-off date than its competitor. If that is the case the data must show a corresponding competitive advantage accruing to teams exhibiting a higher RAEi (Figure 4). This is precisely what the data shows. The advantage when measured as PPG is similar (as we would expect) for both higher average age and more biased relative age and both are similar to the known home team advantage which allows us to intuit its severity. Furthermore the evident relative age bias in competition steadily decreases as the average age of competition cohorts increases towards parity around the age of 27 when the average age advantage disappears (Figure 6).

The desire for competitive advantage in youth football drives up the average team age, which in turn, within eligibility cut-off date silos, causes relative age bias. The eligibility cut-off date rule is therefore causal in relation to the relative age effect and conversely it follows that removal of the cut-off date rule and replacement with an eligibility rule which disallows the possibility an older team being on the field would be causal in removing that effect.

Such an average team age (ATA) rule can be devised as follows: ‘A competing squad shall consist of no more than ‘X’ players whose average age on the competition start date shall be no more than ‘Y’. The average age of the starting team in any competition match shall be no more than ‘Y’. No player in the squad shall be more than ‘Z’ years older than the youngest player in the squad. The mean and the range of ages are thus defined on a team eligibility basis rather than on an individual eligibility basis allowing any individual player to participate across a spectrum of eligible age groups.

Survival Modelling of Goal Arrival Times in Champions League

I. Leriou, I. Ntzoufras, D. Karlis

Department of Statistics, Athens University of Economics and Business Athens, Greece
ntzoufras@aueb.gr, eleriou@aueb.gr, karlis@aueb.gr

Abstract

We consider possible survival modelling of goal arrival times using the bivariate Weibull distribution under a competing risks framework. The proposed approach takes into account the competitiveness of each team's goal arrival time by considering that the arrival goal time for the scoring team is taken as the censoring time for the opponent. Estimation of the parameters was made possible using MCMC methodology. Finally, a final model is selected using Gibbs Variable Selection and is compared in terms of its goodness of fit and prediction with a null model without covariates. The proposed methodology will be presented on data concerning the Champions League 2017-2018.

1 Introduction

The statistical modelling of soccer outcomes has been of enormous interest for statisticians especially during the recent decades. In particular, the modelling and prediction of the score outcomes has been thoroughly addressed by Karlis and Ntzoufras (2008), Karlis and Ntzoufras (2003) where the Bivariate Poisson is assessed and compared with existing at the time models. However, modelling the actual times of goals for each team is of equivalent importance in terms of research. Early attempts of tackling such problems include Nevo and Ritov (2013) where the effect of the first goal is examined in terms of its impact to the second goal by using a Cox model that includes time dependent covariates.

Our approach concerns modelling survival times regardless of whether we are focusing on the first goal, by accounting for a broader survival perspective. In particular, we propose a Marshall Olkin Bivariate Weibull distribution for the scoring times between two competing teams under a parametric survival analysis framework. Among the assumptions underlying our model is the fact that each goal time for the scoring team is considered as censoring time for the opponent. An additional aspect that is taken into account in our model, is the attacking and defensive ability of each team but since we are considering Champions' League data, the game effect and the round effect is also incorporated under a Bayesian framework.

The rest of the paper is organized as follows. In Section 2 we formally present the Marshall Olkin Bivariate Weibull distribution followed by some of its defects with regards to possible modelling such events. After that, we present our approach to avoiding those defects. In Section 3 we give a brief mention on the prior distributions concerning the parameters in our final model while in Section 4 we compare a variety of models including the one that was derived after we applied Gibbs Variable Selection. Before concluding this paper, we use our final model to describe the final match in Champions League 2017-2018 while presenting survival curves illustrating the probability that each competing team will score at each minute of the game. We conclude this paper in Section 6 where we are presenting the limitations of our approach and suggest possible solutions.

2 Bivariate Models and Data Layout

Let t_{1im} and t_{2im} be the event times for team 1 and team 2 respectively with $i = 1, 2, \dots, n$ and $m = 1, 2, \dots, M$ the game indicator. To be more precise, part of the data layout in our case is presented as follows:

Game	t_1	t_2	Home Team	Away Team
1	50	NA	Benfica	PFC CSKA Moskva
1	NA	13	Benfica	PFC CSKA Moskva
1	NA	8	Benfica	PFC CSKA Moskva
1	NA	NA	Benfica	PFC CSKA Moskva
...				
...				

Table 1: Data layout for survival modelling of Champions' League Data

From Table 1, it is clear that the survival times are presented in an unusual manner. This is because we are assuming that after a team has scored a goal, the time in that specific game resets. To be more precise, $t_{221} = 13$ means that 13 minutes have passed since the previous event has happened. At the last line concerning the first game, the two missing values (t_{141} and t_{242}) represents our inability to observe at what time would a team have scored from the time that the last team scored until the end of the game.

The first assumption that we are making at this point is that we are considering each pair of scoring times to be independent (each line in our data layout to be independent of one another) and hence $(t_{1im}, t_{2im}) = (t_{1i}, t_{2i})$. However, the game effect is not neglected since we are considering that it is indeed a random effect, similarly with the attacking and defensive ability of each team.

The essence of assuming a bivariate model, is to assume some kind of dependence between that two random variables under consideration. In survival analysis there exists a wide variety of bivariate models that seem to capture specific kind of dependencies. One of those models is called the Marshall Olkin bivariate Weibull model and is presented formally in the next subsection.

2.1 Marshall Olkin Bivariate Weibull

Let U_0 , U_1 and U_2 be independent Weibull random variables with the same shape parameter γ and scale parameters λ_0 , λ_1 and λ_2 respectively.

Define

$$T_1 = U_0 \wedge U_1 \quad T_2 = U_0 \wedge U_2.$$

Then

$$(T_1, T_2) \sim MOBW(\gamma, \lambda_0, \lambda_1, \lambda_2)$$

The Joint Probability Density Function of the Marshall Olkin Bivariate Weibull distribution is given by

$$f_{T_1, T_2}(t_1, t_2) = \begin{cases} f_W(t_1; \gamma, \lambda_1) f_W(t_2; \gamma, \lambda_0 + \lambda_2) & \text{if } 0 < t_1 < t_2 \\ f_W(t_1; \gamma, \lambda_0 + \lambda_1) f_W(t_2; \gamma, \lambda_2) & \text{if } 0 < t_2 < t_1 \\ \frac{\lambda_0}{\lambda_0 + \lambda_1 + \lambda_2} f_W(t; \gamma, \lambda_0 + \lambda_1 + \lambda_2) & \text{if } 0 < t_1 = t_2 = t \end{cases}$$

That means that we have a closed form expression of the joint pdf of the Marshall Olkin distribution as following:

$$f_{T_1, T_2}(t_1, t_2) = f_W(t_1; \gamma, \lambda_1) f_W(t_2; \gamma, \lambda_2) f_W(t_1 \vee t_2, \gamma, \lambda_0)$$

where $t_1, t_2, \lambda_0, \lambda_1, \lambda_2, \gamma > 0$.

Note that the parametrization of the above mentioned Weibull distribution is given by:

$$f_W(x; \gamma, \lambda) = \gamma \lambda x^{\gamma-1} e^{-\lambda x^\gamma}$$

with mean

$$E(X) = \lambda^{-1/\gamma} \Gamma(1 + 1/\gamma)$$

The joint survivor function has the following form:

$$S_{T_1, T_2}(t_1, t_2) = S_W(t_1; \gamma, \lambda_1) S_W(t_2; \gamma, \lambda_2) S_W(t_1 \wedge t_2, \gamma, \lambda_0), \quad \forall \lambda_0, \lambda_1, \lambda_2, \gamma, t_1, t_2 > 0$$

At this point it is worth noting that λ_0 plays the role of the dependence parameter between T_1 and T_2 . In particular, if $\lambda_0 = 0$ then T_1 and T_2 are independent.

The basic defect of this distribution it is not straightforward how it can be fitted under a bayesian survival framework using standard software like R and WinBUGS. Another defect is the fact that this particular bivariate distribution cannot be straightforwardly represented by the use of latent variables like the Bivariate Poisson model. To avoid these defects we assumed that the two arrival scoring times are coming for two independent Weibull distributions truncated at the censoring times of each team.

2.2 Independent Weibull Model

Let t_{i1} and t_{i2} be the goal arrival times (in the sense that was presented above) by home (HT) and away teams (AT) $i = 1, 2, \dots, n$. Then the "independent Weibull" model can be expressed by

$$T_{ij} \sim \text{Weibull}(\gamma, \lambda_{ij}) \quad \text{for } j = 1, 2, i = 1, 2, \dots, n$$

with

$$\begin{aligned} \log(\lambda_{i1}) = & \mu + \text{home} + a_{HT_i} + d_{AT_i} + ge_{GD_{descr_i}} + re_{GD_{descr_i}} \\ & + \beta_1 g d 1_i + \beta_2 (h f_i - 1) + \beta_3 g d 2_i + \beta_4 r t_i + \beta_5 g s_i \end{aligned}$$

$$\begin{aligned} \log(\lambda_{i2}) = & \mu + a_{AT_i} + d_{HT_i} + ge_{GD_{descr_i}} + re_{GD_{descr_i}} \\ & - \beta_1 g d 1_i + \beta_2 (h f_i - 1) - \beta_3 g d 2_i + \beta_4 r t_i + \beta_5 g s_i \end{aligned}$$

In this model, μ is the intercept, that exists commonly in both independent Weibull distributions, home depicts the home effect, a_k and d_k are the attacking and defensive parameters of team k with $k = 1, 2, \dots, K$,

$ge_{GD_{descr_i}}$ and $re_{GD_{descr_i}}$ are the game effects and the round effects. In terms of online covariates, we have considered including, an indicator for one goal difference (gd_1 for each team), different effect for goal difference that is higher than 2 (gd_2), a half time covariate indicating if we are goal was scored in the first or the second half of the game (takes values 1 or 2), the remaining time (rt_i) indicating what is the remaining time until the end of the game (90 minutes) and gs is the number of goals scored at each event time.

3 Prior Distributions

The prior distributions that were assigned to the parameters of our model, are weakly informative and are presented as follows:

$$a_k, d_k \sim Normal(0, 10^{-3})$$

$$\mu, home, ge_{GD_{descr_i}}, gs_{GD_{descr_i}} \sim Normal(0, 10^{-3})$$

The coefficients in our model are also assumed to have a weakly informative prior namely:

$$\beta_j \sim Normal(0, 10^{-3})$$

Finally, since the shape parameter γ is a positive parameter, a Gamma distribution as follows

$$\gamma \sim Gamma(10^{-3}, 10^{-3})$$

In order to make the model identifiable and make comparisons of the ability of each team with an overall level of attacking and defensive abilities we imposed Sum-To-Zero constraints on those parameters. In particular we assumed the following

$$\sum_{k=1}^K a_k = 0, \quad \sum_{k=1}^K d_k = 0$$

4 Bayesian Estimation and Model Fitting

Using the R2MultiBUGS package in R, we were able to use MCMC (Monte Carlo Markov Chains) to fit our model and to sample from the required posterior distributions. In terms of the data, we have used the Champions League 2017-2018 data, with 528 observations in total, 32 teams and 125 games.

Additionally to the actual model fitting, we conducted Gibbs Variable Selection (Dellaportas et al., 2002) to select a final model and make comparisons in terms of its DIC value. The DIC of each model fitted is presented in Table 2.

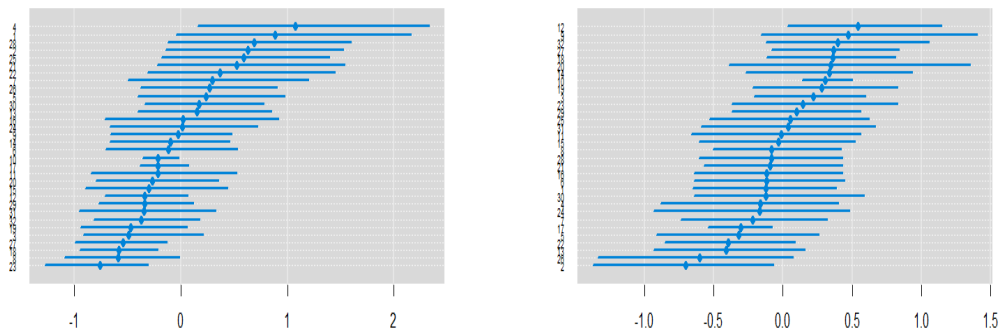
Concerning our final model, Figure 1 illustrates the 95% credible intervals of the attacking and defensive abilities of each team. Note that under this framework, the higher the attacking(defensive ability) the worst(better) that team is in terms on scoring frequency. Moreover, the posterior estimates of the parameters γ , $home$ and μ are also presented in Table 3 along with their standard errors and 95% credible intervals.

<i>Parameter</i>	<i>Mean</i>	<i>SD</i>	<i>2.5%</i>	<i>97.5%</i>
<i>home</i>	-0.197	0.071	-0.345	-0.050
μ	1.045	0.451	0.370	2.301
γ	1.357	0.057	1.247	1.468
<i>hf</i>	-0.533	0.112	-0.762	-0.314
<i>gd2</i>	-0.110	0.038	-0.179	-0.041
<i>rm</i>	-0.031	0.002	-0.035	-0.026

Model	DIC	Covariates
Null Model	4069.7	None
GVS Model	3826.7	hf + gd2 + rt

Table 2: Parameter Estimate of the final model and DIC comparisons.

From Table 2 we can infer that the GVS model is the better model since it seems to be fitting the data better according to the DIC (the lower the better).



(a) Credible intervals for the attacking abilities.

(b) Credible intervals for the defensive abilities.

Figure 1: Credible intervals for the abilities of each team.

5 Champions League Final Match Illustration using Survival curves

From Figure 2 we can infer that as the game time increases the probability that Real Madrid will not score decreases considerably faster than that of Liverpool until Real Madrid score at 51' (first dashed vertical green). Additionally, what is clear from this graph is the fact that as the Game time increases and tends to the end of the game, the difference in those survival curves tends to increase drastically. Therefore, the winner (Real Madrid) of the Champions League match can easily be inferred by looking at the differences between those survival curves that were derived from our final model. Those type of plots can be particularly useful for online prediction.

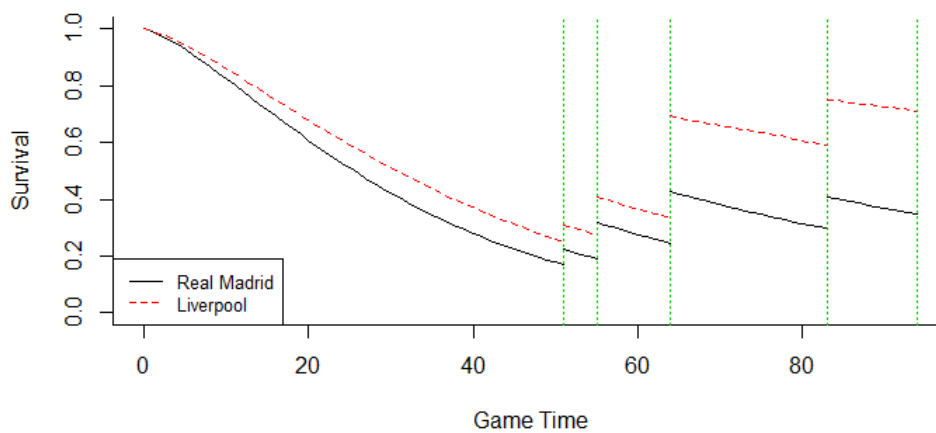


Figure 2: Survival Curves for the Champions League. The green vertical lines represent the goal times.

6 Discussion

In this paper we proposed an independent Weibull model, to model event times regarding soccer. We ended up choosing a model using Gibbs Variable Selection that tends to give reasonable results and to be able to describe a game using survival curves.

Further work includes, fitting the Marshall Olkin Bivariate Weibull model using a different software (Stan). Assessing those types of survival models' predictive ability, and reconstructing a League in order to check the general goodness of fit of each model chosen.

7 Acknowledgements

The research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT), under the HFRI PhD Fellowship grant (1809).

References

- Dellaportas, P., Forster, J. J. and Ntzoufras, I. (2002), 'On bayesian model and variable selection using mcmc', *Statistics and Computing* **12**(1), 27–36.
- 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.
- Karlis, D. and Ntzoufras, I. (2008), 'Bayesian modelling of football outcomes: using the skellam's distribution for the goal difference', *IMA Journal of Management Mathematics* **20**(2), 133–145.
- Nevo, D. and Ritov, Y. (2013), 'Around the goal: Examining the effect of the first goal on the second goal in soccer using survival analysis methods', *Journal of Quantitative Analysis in Sports* **9**(2), 165–177.

Fame and Fortune in Elite Tennis Revisited

Denny Meyer^{1*}, Minh Huynh¹, Kelly Marshall¹ and Geoff Pollard²

¹ Swinburne University of Technology, Melbourne, Australia

² Tennis Australia.

dmeyer@swin.edu.au, mh huynh@swin.edu.au, kmarshall@swin.edu.au,
gpollard@tennis.com.au

Abstract

Professional tennis is aging with older players able to prolong their careers thanks to science, technology and healthy pay cheques. However, change is in the air making this a good time to investigate the trends and patterns apparent in ATP statistics. Previous research, using data for only the top 128 ATP players in the 2011 ATP World Tour Media Guide, investigated the relationship between performance, prize money and rankings for the period 2004 to 2010. The results were contradictory, depending on what measure of performance were used. This paper again investigates the relationship between rankings, prize money and performance using On Court data for the top 200 ranked players for the years 2004 to 2018. This larger, more modern data finds more consistent results, with prize money consistently more important than rankings for predicting performance. Secondary analyses consider changes that generally occur through a player's career in terms of prize money, average rankings, percentage of matches and tie breaks won, average number of aces per match and several other important match parameters. Linear trends are detected, suggesting consistent changes over the careers of elite players. However, in the case of matches played, average rankings and prize money, the magnitude of the time trends are dependent on age in 2018, suggesting different trajectories for younger and older generation players. More rapid improvements are seen amongst younger generation than older generation players. The implications for younger players are exciting, suggesting that there will soon be a changing of the guard.

1 Introduction

For many years now, the aging of professional tennis has been noted. For example, Gallo-Salazar et al. (2015) reported that the mean age of the top 100 players increased from 24.6 to 27.6 between 1984 and 2013. Ramsey (2017), reporting on an interview with former British Number 1, Sam Smith, suggested that the reasons for this are numerous. In particular, he mentioned the greater power and

* Corresponding Author

control possible with modern racquets and strings, modern medical science and a better understanding of nutrition and diet. However, according to Sam Smith and others, it is prize money that keeps older players in the game, allowing them to travel with a team of experts to oversee their recovery and preparation. The effect of this aging cohort on the game itself is something that will be considered in this paper.

However, in recent times, a “changing of the guard” has also been predicted in ATP tennis. Just prior to the 2019 Australian Open McEnroe again repeated these sentiments, despite the fact that all the 2018 Grand Slams were shared out between Federer, Nadal and Djokovic, all aged in their 30’s. The defeat of Federer by Tsitsipas in the fourth round of the Australian Open left the crowd and Tsitsipas in some disbelief, but Djokovic still went on to win the tournament. This means that, since the 2003 Wimbledon championships, 52 of the 63 Grand Slams have been won by one of the so-called “big three”; 15 for Djokovic, 17 for Nadal and 20 for Federer. But what do the “stats” say about this “changing of the guard”? In particular, what is motivating our older and newer stars, fame or fortune? Building on the results of a previous study by Meyer and Pollard (2012) this paper considers the relative importance of fame and fortune for the motivation of older and younger generation players. Prize money is one part of the “fortune” and singles rankings are one part of the “fame”, however, with fame comes exhibition matches and advertising contracts that can have serious benefits in terms of “fortune” for some players. In this paper, we try to establish whether prize money is really a meaningful incentive for elite ATP players, or if it is “fame” factors that have more importance in the present environment.

Sunde (2009) has investigated whether higher prize money improves performance in terms of the proportion of matches won and the total number of matches played. Using the Association of Tennis Players (ATP) singles data from 156 Grand Slam and Masters tournaments, collected between 1990 and 2002, Sunde showed that the higher stage prizes, and in particular the substantially higher prizes won in the finals as compared to the semi-finals, significantly improved performance. However, rankings have also been linked to performance. Using Wimbledon singles data 1992-95, Klaassen and Magnus (2003) converted player rankings into the probability that a particular player would win a match between two players. They then recalculated this probability as the match progressed. For further probability predictions based on rankings see Bedford and Clarke (2000) and Clarke and Dyte (2000).

It seems therefore that both prize money and singles rankings contribute to the performance of elite tennis players. The previous small study by Meyer and Pollard (2012) suggested contradictory results depending on what performance measures are used. With an additional eight years of data, many more players and a new era in ATP tennis emerging, we will reinvestigate this issue, while also investigating the effect of our aging players on the game of tennis itself. In particular, we will determine how career trajectories are moderated by player generation.

2 Methods

In this paper statistical analyses have been performed using annual statistics for the Top 200 players in each of the years 2004 to 2018, as provided by <http://www.protennislive.com> and downloaded from the On Court database. The analysis considered an average of 4.7 years of data for each of 612 players, resulting in 2878 player years of data. The average age of these players was 33.2 years at the end of 2018 with a standard deviation of 6.6 years. However, as indicated in Figure 1 this age distribution includes some much older (sometimes currently retired) players. These players represent the “older generation” in the discussion below.

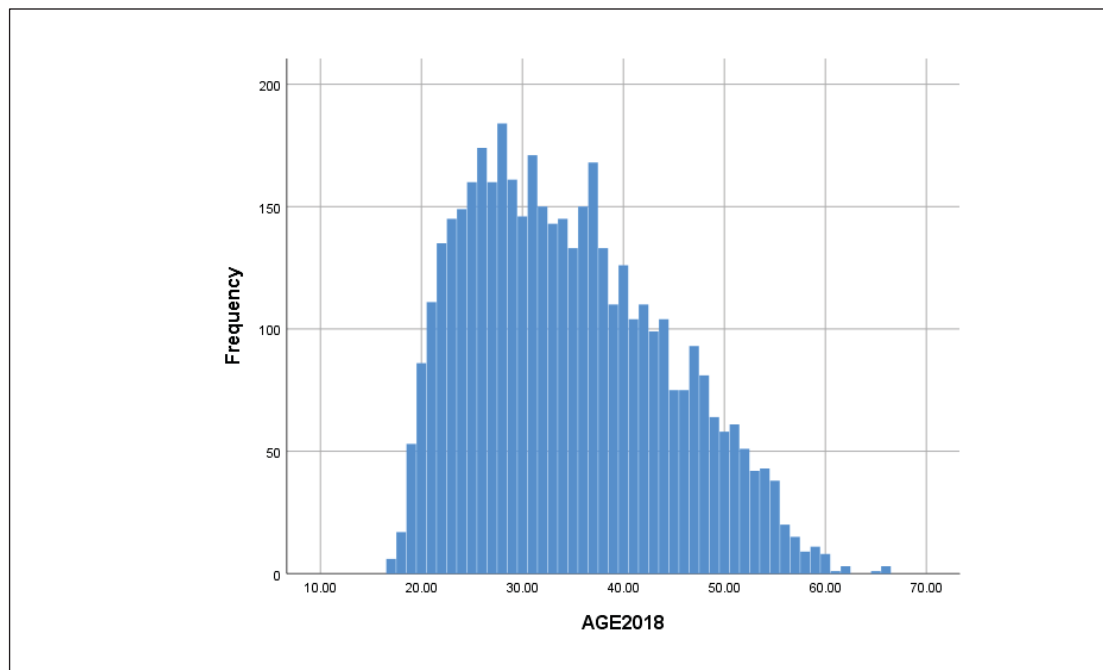


Figure 1: Age Distribution All Players in 2018

The average career prize money for these 612 players at the end of 2018 was \$3.8m, with a minimum value of \$95871 and a maximum value of \$125.8m for Djokovic, closely followed by Federer at \$120.5m. The average number of matches per player per year was 28.1 with a standard deviation of 22.1, and the average percentage of matches won per player per year was 42.8% with a standard deviation of 19.1%. It should be noted that during the period 2004-2018 there was a steady increase in prize money, especially for the bigger tournaments and for the tournament winners. However, since about 2012, players knocked out in earlier rounds have received a greater share of the prize money, perhaps providing a greater incentive to young up-and-coming players.

The data includes the number of singles matches won and lost for each player in each year, total prize money in each year as well as several other player performance statistics. This data has been supplemented with On Court (2018) data for weekly rankings, which have been averaged for each year for each player. The annual average rankings were transformed using a square root transformation and the annual prize money data were log transformed in order to reduce the effect of outliers and to create something closer to a normal distribution as shown in Figures 2 and 3. As recommended by Sunde (2009) in the initial analysis conducted in this paper the proportion of singles matches won in any year was used as the measure of performance, while in the second analysis the number of singles matches played in any year was used as the measure of performance. This second measure of performance acknowledges the knockout nature of most tournaments, resulting in more matches played in the case of more successful players. As shown in Figure 4 there is a strong linear relationship between these two performance measures (with $r = -0.90$).

Before commencing the analysis, a random intercept model was fitted for the two performance measures to find out what proportion of their variation could be attributed to players in any year. These models were fitted using HLM7 (SSI Central). A proportion in excess of 5% was obtained for both performance measures suggesting that multi-level analyses were required in this study in order to allow for the player dependence in the data across years (Raudenbush and Bryk, 2002).

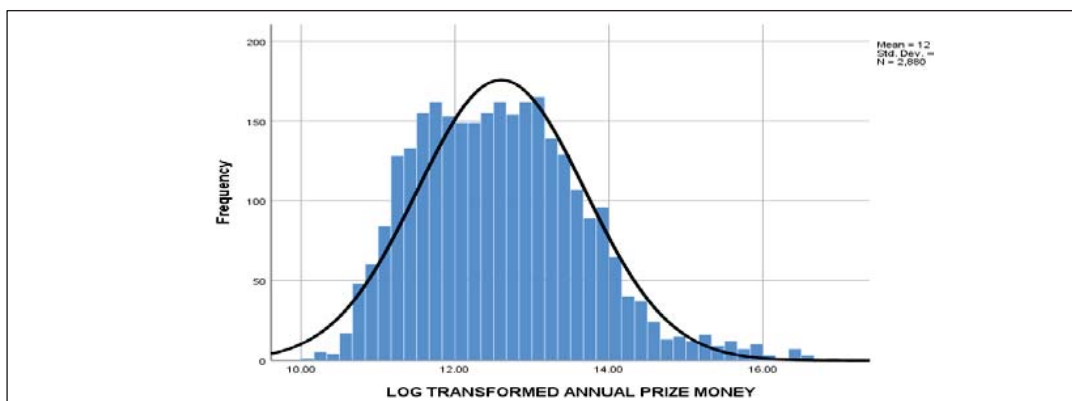


Figure 2: Log Transformed Prize Money Per Annum

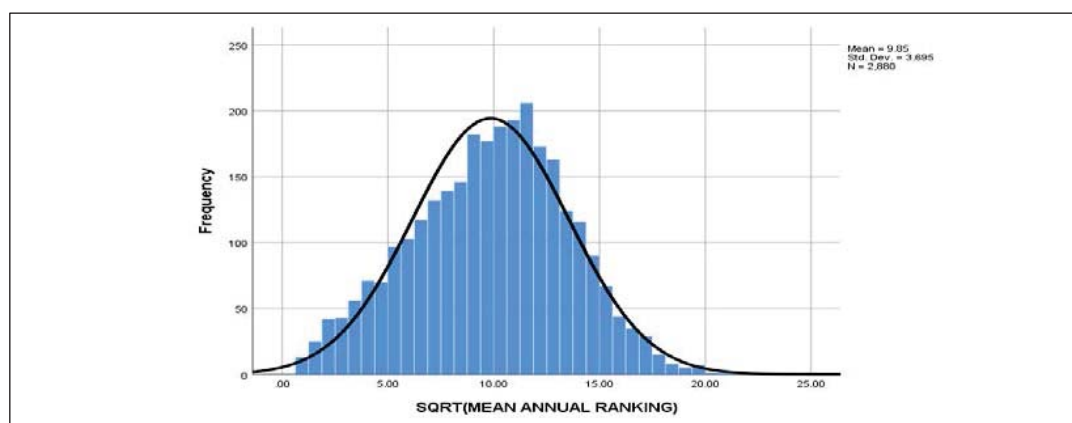


Figure 3: SQRT Transformed Average Ranking Per Annum

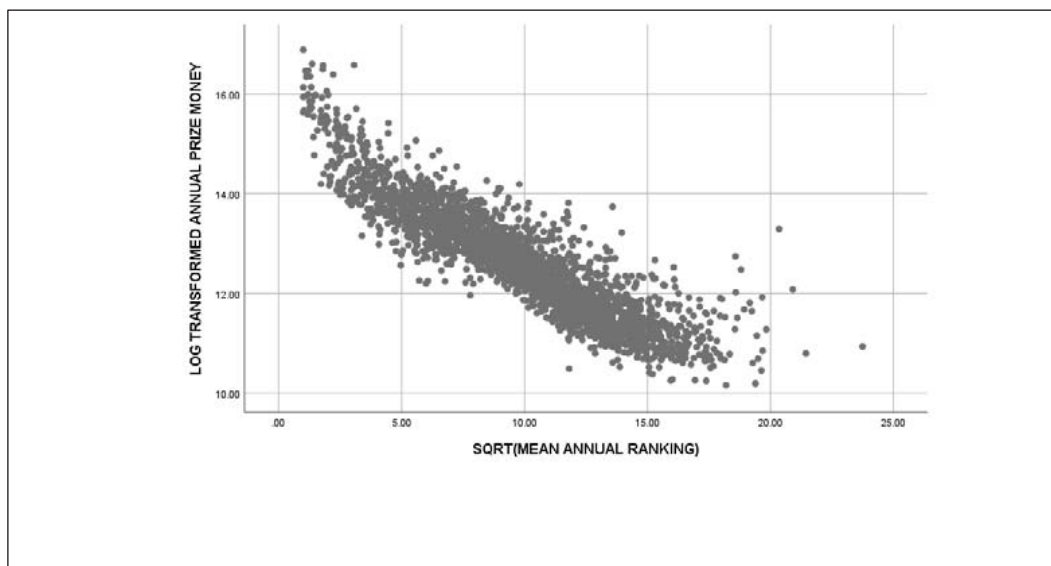


Figure 4: Relationship Between Annual Prize Money and Mean Annual Rankings After Transformation

A multi-level analysis was therefore conducted in order to determine for each player the effects of average ranking and prize money on performance in any year. These models allow for the creation of a population model based on the averaged regression equations of individual players. These models assume that on average the coefficients remained the same for each player over his career. This means that the incentive effects of rankings and prize money are assumed to remain the same over time for each player.

In Figure 5 ϕ_{ij} represents the expected proportion of matches won by player j in year i while in Figure 6 λ_{ij} represents the expected number of matches played by player j in year i . A binomial distribution is assumed in Figure 5 and a Poisson distribution with constant exposure is assumed in Figure 6, with an allowance for over-dispersion in both cases. In the level 1 models the SQRTRANK and LOGPRIZE variables are group centred by subtracting the mean values for each player. This allows the interpretation of the intercept in the final population models as an average performance statistic across players and careers.

LEVEL 1 MODEL (bold: group-mean centering; bold italic: grand-mean centering)

$$E(\text{MATCHWON}=1|\pi) = \phi \cdot \text{TOTALMAT}$$

$$\text{Log}[\phi/(1 - \phi)] = \eta$$

$$\eta = \pi_0 + \pi_1(\text{LOGPRIZE}) + \pi_2(\text{SQRTRANK}) + e$$

LEVEL 2 MODEL (bold italic: grand-mean centering)

$$\pi_0 = \beta_{00} + \beta_{01}(\text{AGE2018}) + \beta_{02}(\text{LNCPRIZE}) + r_0$$

$$\pi_1 = \beta_{10} + r_1$$

$$\pi_2 = \beta_{20} + r_2$$

Figure 5: Multi-Level Model for Expected Proportion of Matches Won

LEVEL 1 MODEL (bold: group-mean centering; bold italic: grand-mean centering)

$$E(\text{TOTALMAT}|\pi) = \lambda$$

$$\text{Log}[\lambda] = \eta$$

$$\eta = \pi_0 + \pi_1(\text{LOGPRIZE}) + \pi_2(\text{SQRTRANK}) + e$$

LEVEL 2 MODEL (bold italic: grand-mean centering)

$$\pi_0 = \beta_{00} + \beta_{01}(\text{AGE2018}) + \beta_{02}(\text{LNCPRIZE}) + r_0$$

$$\pi_1 = \beta_{10} + r_1$$

$$\pi_2 = \beta_{20} + r_2$$

Figure 6: Multi-Level Model for Expected Number of Matches Played

In both these models the effect of age in 2018 (AGE2018) as well as log transformed total career prize money to December 2018 (LNCPRIZE) were controlled. This effectively controlled for player generation and overall career success. These variables were grand centred across all players resulting in zero values for players with values equal to the grand mean. Moderation tests were also carried out in order to determine whether the importance of rankings and prize money as predictors for performance in any year differed, depending on player generation and overall career success. Next, multi-level models were used to test for changes over time in all the key player statistics, again controlling for the age of the player at the end of 2018 and their total career prize at this time. Trend moderation effects for these two player variables were tested in order to determine if player trajectories differ across player generations and/or across overall career success levels.

□

3 Results

Using random intercept models it was found that 15.2% of the variation in the proportion of matches won in any year and 5.8% of the variation in the number of matches played in any year could be attributed to player effects. This suggests that at least for these variables a multi-level analysis is needed in order to allow for the player dependence in the data.

3.1 Fame or Fortune Analyses

The multi-level analysis results in Table 1 suggest a significant positive relationship between prize money and percentage of matches won in any year, and a significant negative relationship between mean rankings and percentage matches won in any year. However, the t-ratios indicate a much stronger relationship in the case of prize money than rankings.

In addition, for any player the expected proportion of matches won is significantly lower for older generation players and significantly higher for players with greater career prize money. In particular, there was on average a 1% drop in the odds of winning a match for each additional year of age in 2018. However, there were no significant moderation effects for AGE2018 and LNCPRIZE, suggesting that the effects of prize money and rankings in any year are the same, regardless of generation or overall career success.

Table 1: Multi-Level Model for Proportion of Matches Won in Any Year

Effect	Coefficient	Odds Ratio	Confidence Interval	t-ratio	p-value
Intercept (β_{00})	-0.508	0.602	(0.581,0.624)	-27.945	<.001
AGE2018 (β_{01})	-0.009	0.991	(0.987,0.996)	-3.957	<.001
LNCPRIZE (β_{02})	0.336	1.400	(1.358,1.443)	21.601	<.001
Slope LOGPRIZE (β_{10})	0.247	1.280	(1.235,1.327)	13.528	<.001
Slope SQRTRANK (β_{20})	-0.022	0.978	(0.967,0.990)	-3.71	<.001

The same analysis was then performed using the number of matches played in any year as the measure of performance, assuming a Poisson distribution for this variable with a log link function. The Multi-level analysis shown in Table 2 suggests that prize money is a more important motivator for the number of matches played in any year than rankings.

Table 2: Multi-Level Model for Number of Matches Played in Any Year

Effect	Coefficient	Exp(B)	Confidence Interval	t-ratio	p-value
Intercept (β_{00})	2.588	13.308	(12.778,13.861)	124.92	<.001
AGE2018 (β_{01})	0.007	1.007	(1.000,1.014)	1.97	0.049
LNCPRIZE (β_{02})	0.553	1.738	(1.690,1.788)	38.36	<.001
Slope LOGPRIZE (β_{10})	0.541	1.717	(1.646,1.792)	24.95	<.001
Slope SQRTTRANK (β_{20})	-0.110	0.896	(0.884,0.907)	-16.9	<.001

However, although players with greater overall career success (higher LNCPRIZE) played more matches in any year on average than players with less success, it seems that older generation players also tended to play more matches than younger generation players on average. This result is barely significant ($p=.049$), but clearly of interest since it suggests that, when we control for level of career success, older generation players are on average playing more matches although not with as much success as younger generation players. Again, there were no significant moderation effects for AGE2018 and LNCPRIZE, suggesting that the effects of prize money and ranking on the number of matches played in any year are the same, regardless of generation or overall career success.

3.2 Trend Analyses

Table 3 and 4 show the results of trend analyses (expected change per annum) for various performance measures when controlling for career prize money (LNCPRIZE) and age at the end of 2018 (AGE). Normal distributions are assumed for all these variables except the proportion of matches won per annum for which a binomial distribution is assumed, and number of matches per annum for which a Poisson distribution is assumed. The TREND variable is group centred. This means that the intercept relates to the middle year of a player's career, whenever that may occur. As before the AGE2018 and LNCPRIZE variables are grand centred.

In Table 3 no significant moderation effects were detected for the trend, suggesting that career trajectories were similar regardless of generation or overall career success in terms of prize money. However, there are some significant trends that tell us about the average playing trajectory that can be expected for all elite ATP players. The results in Table 3 suggest a significant improvement in service over the careers of the average elite player but no significant change in terms of receiving performance or the percentage of matches won or the percentage of tie breaks or breakpoints won. Interestingly the trend for double faults is significantly positive, suggesting that the mean number of double faults per match increases over the career of an average player.

As shown in Table 3 and 4 the proportion of matches won is significantly lower for older generation players and so is their prize money, however, these players tend to get more of their first serves in and they tend to win more points on first serve and more breakpoints. The effect of career prize money is

significant and positive for all these outcome measures as well as those in Table 4, except for the average number of double faults per match. This suggests that all these outcome measures except double faults are rewarded in terms of career prize money.

Table 3: Trend Analysis Controlling for Age in 2018 and Career Prize Money (LNCPRIZE)

Outcome Measure	Trend (years)	AGE 2018	LNCPRIZE
Proportion matches won	-0.01	-0.02***	0.36***
% first serves in	0.19**	0.12**	0.92***
% points won on first serve	0.19***	0.08	2.38***
% points won on second serve	0.16***	0.05	1.82***
% service games won	0.33***	0.02	3.61***
% break points saved	0.31***	0.05	.19+2.27***
% points won on return 1 st serve	-0.06	0.06*	1.27***
% points won on return 2 nd serve	0.06	0.06	1.56***
% breakpoints won	-0.07	0.09*	1.05***
% receiving games won	-0.07	0.00	1.77***
Average aces per match	0.07***	-0.01	0.51***
Average double faults per match	0.02*	0.01	-0.06
% tie breaks won	-0.34	-0.08	4.50***

* p<.05, ** p<.01, *** p<.001

Table 4: Trend Analysis moderated by AGE2018

Outcome Measure	Trend (years)	AGE 2018	LNCPRIZE
Number matches played	0.03-0.02Age2018 ***	.00	.54***
LN(Prize money)	0.12-0.03Age2018 ***	-.03***	.66***
SQRT(Mean Ranking)	-0.25+0.09Age2018 ***	-.01	-2.10***

* p<.05, ** p<.01, *** p<.001

Table 4 shows the outcomes where it was found that the trend (expected change per annum) was significantly moderated by age at the end of 2018. The results suggest that, as their career progressed

older generation players saw less of an increase in the number of matches played per annum than younger generation players. Similarly, the increase in prize money per annum appears to have been smaller for older generation players than younger generation players. Finally, the rate of decline (improvement) in rankings appears to have been lower for older generation than younger generation players.

These results seem to suggest that younger generation players are experiencing a more meteoric improvement in terms of matches played, prize money and ranking than was the case for older generation players. However, career prize money failed to significantly moderate the trends in these outcome variables or any of the outcome variables in Table 3.

4 Discussion

The previous small study by Meyer and Pollard (2012) suggested that the relative importance of rankings and prize money for predicting performance depends on what performance measures are used. With an additional eight years of data, many more players and a new era in ATP tennis emerging, we have revisited this issue, while also investigating the effect of our aging players on the game of tennis itself.

The previous analysis based on the data for only the top 108 players in 2010 suggested that, in the case of the proportion of matches won in any year, rankings was a better predictor of success, while, for the number of matches played in any year, prize money was a better predictor of success. Using a larger number of players (612) and considering matches played in the period 2004-2018 instead of 2004 to 2010, this study suggests that prize money is now a more important predictor of performance than average annual ranking for both the proportion of matches won and the number of matches played. Tournament prize money has increased markedly since 2010 and it may be this increase which has produced the new finding for the proportion of matches won.

A second important finding relates to the consistency of the coefficients for prize money and ratings across all generations of players and across levels of career success. It seems that both older and younger generations of players, regardless of their level of success, are better motivated by prize money than rankings. Rankings are tied to product promotion and competition earnings, and as such they certainly do act as an incentive for players. However, the immediate reward of tournament prize money is clearly more of an incentive to players.

The third important finding concerns the effects of generation on the proportion of matches won. It seems that on average the younger generation have won a greater proportion of their matches than the older generation of players. This is despite the fact that the older generation has a clear advantage on first serve and breakpoints. This may have something to do with the way the game is changing, relying more on longer rallies as new racquets mean that it is easier to return even the fastest of serves.

The fourth finding relates to the career trajectories that are apparent for all players. Clearly as players mature their service improves, resulting in more service games won, more aces and, perhaps not surprisingly, more double faults. Also, there is a tendency to do better with breakpoints. However, all these career trends, with three exceptions, are consistent across all generations of players and for players at all career success levels, suggesting that they are generic. The three exceptions are number of matches played per annum, prize money per annum and mean annual ratings. In the case of these three measures it seems that the younger generation are at an advantage in that trajectories for these three variables are more rapid for younger than older generation players. It may be the increased prize money available to modern generation players that is fueling these more rapid trajectories. Players can now better afford to travel to tournaments, providing more opportunities to play matches and improve ratings.

Finally, although overall career success in terms of prize money does not influence career trajectories it is clear that a successful career in tennis requires high performance in all aspects of the

game. The only aspect of the game that seems to be a bit more forgiving is double faults. Double faults, it seems, are not a good discriminator for overall career prize money.

5 Conclusion

This paper has sought to better understand the effects of maturation in our ATP players and, at the same time, to look for evidence of a “changing of the guard” in the near future. The above analysis suggests that in recent years prize money has become more important than ranking as a driver of performance. In addition, the results have shown consistent improvement in service performance over the years. This has been associated with consistent improvement in annual mean rankings, prize money and number of matches played over player careers. However, the speed of this trajectory appears to be hotting up for younger players, suggesting that the hiatus that has typified ATP tennis in recent years, with the top three players dominating, is nearly over.

However, this study has one very serious limitation. Coupled with the maturation of our players there is continuous improvement in racquets, strings and the medical support that our players receive. This means that player generation effects are confounded with technological developments. This means that we cannot be sure if the changes in the game that we have observed are due to changes in technology rather than the aging of our elite players.

Nevertheless, the results do suggest that younger players are winning matches as never before, despite the dominance of the older generation in terms of service. It therefore seems that McEnroe is correct. The younger generation has the statistics on its side. The “changing of the guard” cannot be too far away.

Acknowledgements

We wish to thank the International Tennis Federation and KAN-soft for the use of their data.

References

- Bedford, A.B., & Clarke, S.R. (2000), A comparison of the ATP rating with a smoothing method for match prediction. Pp 43-51 in Cohen, G, Langtry, T eds. *Proc. 5th Australasian Conference Maths and Computers in Sport*.
- Clarke, S.R., & Dyte, D. (2000), Using official ratings to simulate major tennis tournaments, *International Transactions in Operational Research* 7, 585–594.
- Gallo-Salazar, C., Salinero, J.J., Sanz, D., Areces, F. & del Cos, L. (2015) *International Journal of Performance Analysis in Sport*, 15, 873-883.
- Klaassen, F.J.G.M., & Magnus, J.R. (2003), Forecasting the winner of a tennis match. *European Journal of Operations Research*; 148: 257-267.
- Meyer, D. & Pollard, G. (2012). Fame and fortune in elite tennis. *Proc. 11th Australasian Conference Maths and Computers in Sport*.
- On-Court (2018), www.oncourt.info, Copyright © 2001-2012 KAN-soft.
- Ramsay, A. Australian Open 2017: Why today’s tennis players are fitter, stronger – and older. January 21, Sydney Herald.
- Sunde, U. (2009). Heterogeneity and performance in tournaments: a test of incentive effects using professional tennis data. *Applied Economics*, 41, 3199-3208.

Lessons Learned in Scheduling the Finnish Major Ice Hockey League

Kimmo Nurmi, Jari Kyngäs and Nico Kyngäs

Satakunta University of Applied Sciences, Pori, Finland

cimmo.nurmi@samk.fi, jari.kyngas@samk.fi, nico.kyngas@samk.fi

Abstract

Ice hockey is the biggest sport in Finland both in terms of revenue and in number of spectators. The Finnish Major Ice Hockey League involves significant investments in players, broadcast rights and merchandising. The quality of the League schedules is extremely important, as the schedule has a direct impact on revenue for all involved parties. We have generated the League schedule for the last eleven years. During this time, the League has continuously looked for improvements in its schedule format and the schedule itself. We believe that scheduling the Finnish Major Ice Hockey League is one of the most difficult sports scheduling problems because it combines break minimization and traveling issues with dozens of other constraints that must be satisfied. We have used the PEAST algorithm and its predecessors to schedule the League since the 2008-2009 season. This paper summarizes the lessons learned in scheduling the League. We present the most important academic and practical findings that we believe will give new ideas to the sports scheduling community.

1 Sports Scheduling

This paper summarizes the lessons learned in scheduling the Finnish Major Ice Hockey League. The paper is organized as follows. This section introduces the sports scheduling problem. The next section describes the characteristics of the Finnish Major Ice Hockey League. In Section 3 we give a detailed discussion of the constraints and goals determining the generation of the League schedule. Sections 4 and 5 present our practical and academic lessons learned during the last eleven years.

Professional sports leagues are huge businesses. The quality of the schedules are extremely important; the schedule has a direct impact on revenue for all involved parties. TV networks that pay for broadcasting rights want the most attractive games to be scheduled at commercially interesting times. Furthermore, the schedule influences the number of spectators in the stadiums and the traveling costs for the teams. Thousands of journalists and millions of fans follow the leagues on an everyday basis. A good schedule makes a season more interesting for the media and the fans, and fairer for the

teams, so that all teams play under the same transparent conditions. The scheduler should find the best possible schedule where often colliding business issues and fairness issues are both optimized.

Most important applications of sports scheduling are found in football, soccer, basketball, baseball, cricket and hockey. The general sports scheduling problem involves scheduling the games of a tournament by determining the date and the venue in which each game will be played (see e.g. (Nurmi et al. 2010)). Furthermore, professional sports leagues face so many priorities, requests and requirements that it is impossible to generate a good schedule without efficient computer algorithms. Fortunately, a growing number of academic researchers have been interested in creating algorithms that find optimized schedules based on the league owners' preferences. The sports scheduling researchers have proved that they can generate high quality and profitable schedules for professional leagues. An extensive sports scheduling bibliography can be found in (Knust 2018) and a good overview in (Easton et al. 2004).

In a sports tournament, n teams play against each other in given rounds. Table 1 shows an example of a sports tournament with four teams and six rounds. In a round robin tournament each team plays against every other team a fixed number of times. The tournament in Table 1 is a double round robin tournament (2RR) where the teams meet twice, once at home and once away. A schedule is *compact* because each team plays in each round; otherwise, it is *relaxed*. If a team does not play in a round, it is said to have a *bye* in that round. If the league contains an odd number of teams, a dummy team is added to the league. Then in each round, the team playing against the dummy team has a bye.

Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
Reds	Whites	Blacks	Blacks	Blacks	Greens
vs.	vs.	vs.	vs.	vs.	vs.
Blacks	Blacks	Whites	Greens	Reds	Blacks
Whites	Greens	Reds	Whites	Greens	Reds
vs.	vs.	vs.	vs.	vs.	vs.
Greens	Reds	Greens	Reds	Whites	Whites

Table 1: A compact double round robin tournament with four teams

A team is said to have a *break* if it plays two consecutive home or two consecutive away games. Breaks are to be avoided, when the distances between the team venues are reasonably short allowing the teams to return home after each game. However, a team may prefer to have two or more consecutive away games if its venue is located far from the opponents' venues, and the venues of these opponents are reasonably close to each other. A series of consecutive away games is called an *away tour*.

Sports scheduling involves four basic academic problems: Finding a schedule

- 1) with the minimum number of breaks (Schreuder 1992)
- 2) which minimizes the travel distances, called the traveling tournament problem (Easton et al. 2001)
- 3) with the minimum number of breaks and, at the same time, take additional requirements and requests into account is known as the constrained minimum break problem (Rasmussen and Trick 2008)
- 4) which considers the break minimization and the travel issues simultaneously as well as many additional requirements, called the constrained sports scheduling problem (Nurmi et al. 2010).

The professional sports league scheduling mainly follows the fourth problem model. Consider the tournament shown in Table 1. The constrained sports scheduling problem could include the following criteria:

- 1) The same teams cannot play against each other in consecutive rounds
- 2) Team Greens cannot play at home in the first two rounds, because their stadium is reserved for another event
- 3) Team Reds and Team Whites cannot play at home in the same round, because they are located in the same region
- 4) The total number of home breaks should be minimized.

Table 2 shows one possible solution to the problem. The constraints 1, 2 and 3 hold. The schedule has three home breaks, one for Team Blacks and two for Team Greens. Team Greens has a three-game home break partly due to its home game restriction in the first two rounds. Team Reds and Team Whites have no home and no away breaks.

In the constrained sports scheduling problem the goal is to find a feasible solution that is most acceptable to both the league authorities and the teams. Hard constraints are requirements that the schedule has to fulfil. Soft constraints are requests presented by the league and the teams. They prefer to optimize many goals at the same time. The priorities between the goals are given by the soft constraint weights. It should be noted that hard and soft constraints vary quite substantially depending on the problem and the problem instance at hand. Nurmi et al. (2010) have presented a collection of constraints that occur frequently in the constrained sports scheduling problem.

Round 1	Round 2	Round 3	Round 4	Round 5	Round 6
Blacks	Blacks	Whites	Reds	Greens	Blacks
vs.	vs.	vs.	vs.	vs.	vs.
Reds	Greens	Blacks	Blacks	Blacks	Whites
Whites	Reds	Greens	Greens	Whites	Reds
vs.	vs.	vs.	vs.	vs.	vs.
Greens	Whites	Reds	Whites	Reds	Greens

Table 2: One possible solution to the constrained sports scheduling problem

2 Finnish Major Ice Hockey League (FMIHL)

Ice hockey is the biggest sport in Finland both in terms of revenue and in number of spectators. The Finnish Major Ice Hockey League involves (for Finnish standards) significant investments in players, broadcast rights and merchandising. The quality of the League schedules is extremely important, as the schedule has a direct impact on revenue for all involved parties. Finding the best schedule of games is a difficult task with multiple decision makers, constraints, and objectives involving logistics, economical and fairness issues.

Since the 2015-2016 season, the League has had 15 teams (for earlier seasons, see e.g. (Kyngäs and Nurmi 2009) and (Nurmi et al. 2014a)). Six of the teams are located in “big” cities (over 100,000 citizens) and the rest in smaller cities. Team Kärpät is quite a long way up north and teams Sport and KalPa slightly separate from the other teams (see Figure 1).

The format played in the League is unique. The competition starts with a regular season in September and ends with the playoffs from late March to early May. The League fixes the dates on

which the games should be played. The basis of the regular season is a quadruple round robin tournament (each team plays against each other twice at home and twice away) resulting in 56 games for each team. In addition, the teams are divided into five groups of three teams to get a few more games to play. The teams in the groups are selected based on the traveling time between the teams. These teams play a double round robin tournament (once at home and once away) resulting in 4 games for each team totaling 60 games for each team. Therefore, the teams play either four or six games against each other. The total number of games is 420 from the standard 4RR and 30 from the extra 2RR, totaling 450.

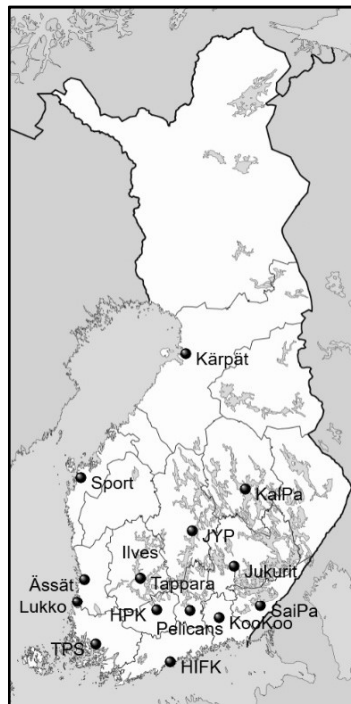


Figure 1: The fifteen teams on the map of Finland.

The schedule of the League is relaxed, that is, each team does not play in each round. The main reason for this is the complexity of the optimization problem, so it is theoretically not possible to generate a compact schedule. In particular, two factors complicate the scheduling. First, some of the teams cannot play on certain days because they also play in the European Champions Hockey League (CHL). Second, most of the teams cannot play at home on certain days because their venues are in use for some other events, mostly concerts. Other reasons include that a local football team having scheduled a game there or a world championship event of some sport being scheduled to the region.

The best six teams of the regular season proceed directly to the quarterfinals. Teams placed between 7th and 10th play preliminary playoffs best out of three. The two winners take the last two quarter-final slots. Teams are paired up for each playoff round according to the regular season standings, so that the highest-ranking team plays against the lowest-ranking, and so on. The playoffs are played best out of seven. The semi-final teams are also paired up according to the regular season standings. The winner of the playoffs receives the Canada Bowl, the championship trophy of the League. The League is closed, that is, no team is relegated.

The standard game days for FMIHL used to be Tuesday, Thursday and Saturday, but for the 2017-2018 season Tuesday was changed to Wednesday, and Thursday was changed to Friday to maximize the revenue. The schedule should maximize the number of games on Fridays and on Saturdays. However, due to the travel distances between some venues, certain combinations of a Friday home team playing a Saturday away game (and vice versa) are not allowed. The games that cannot be scheduled on Fridays due to these restrictions are played on Thursdays.

3 Generating the League Schedule

We have generated the League schedule for the last eleven years since the 2008-2009 season. During this time, the League has continuously looked for improvements in its schedule format and the schedule itself. We believe that scheduling the FMIHL is one of the most difficult sports scheduling problems because it combines break minimization and traveling issues with dozens of other constraints that must be satisfied. We have used the PEAST algorithm (Kynäas et al. 2013a) and its predecessors to schedule the League. The algorithm is dynamic in many senses. It is very important that the input of the games need not be round robins. The algorithm optimizes any number of any games based on any given constraints and goals.

The generation of the League schedule is a good example of the constrained sports scheduling problem (Nurmi et al. 2010). It includes the three basic sports scheduling problems introduced in Section 1:

- 1) break minimization,
- 2) traveling distance minimization and
- 3) a large number of additional requirements and requests.

There are four main reasons why we have to minimize the number of breaks: the fans do not like long periods without home games, consecutive home games reduce gate receipts, consecutive away games might cause a serious income gap for the away team and long sequences of home or away games might influence the team's current position in the tournament.

Mainly four issues can prevent the team playing a home game. First, the team can share a home venue with another team. Second, two teams cannot play at home on the same day because they share the same (businessmen) spectators. Third, the team might not want to play at home if another team in the same region plays at home in another league. Finally, the team cannot play at home if the venue is reserved for some other event, such as rock concerts. A venue can also be under renovation, which can delay the first of the team's home games. Venue restrictions have a major impact on the difficulty of generating the schedule. The usual number of venue restrictions in the League is close to one hundred.

Traveling issues must be considered based on two requirements. In most of the cases, where a team has two away games on consecutive days, it prefers to return home after the first game. However, some teams prefer to stay overnight between the games. Furthermore, in general, two away games on consecutive days are forbidden due to the distances between the away venues. Note, that in the traveling tournament problem (Easton et al. 2001) the teams do not return home after each away game but instead travel from one away game to the next as in most major leagues in USA where long away trips must be scheduled. This of course implies long home breaks as well.

In addition to the break minimization and the traveling distance minimization, the following business issues must be considered (in order of importance):

- B1. The schedule should include as many Friday and Saturday games as possible.

- B2. The teams have priorities on which weekdays they prefer their home games to be played. Most of the teams prefer weekends to maximize number of overall spectators, but some teams prefer working days to maximize number of business spectators.
- B3. The interest of media and fans must be increased by introducing special interest games, e.g. local rival games and back-to-back games.
- B4. Two games between the same opponents should not be played on close days. There must be at least three weeks between two games with the same opponents, and the games should be played on different venues.

Furthermore, the following fairness issues must be considered (in order of importance):

- F1. The further located a team is from the other teams the more away tours the team should be assigned.
- F2. For each team and at any point in the tournament, the difference between home and away games played should be at most three.
- F3. The difference in the number of games played between different teams should be at most two at any point in the tournament.

Finally, the League has two teams, Ilves and Tappara (see Figure 1), who share a venue, so they cannot play at home in the same round. It should also be noted, that because of the League format, a break cannot occur in the last round.

Table 3 shows the most important indicators for the last two seasons, 2017-2018 and 2018-2019. The schedules were generated using the PEAST algorithm (Kygäs et al. 2013a). The schedules were feasible, that is, they had no hard constraint violations. Home game restrictions refer to the cases where a team cannot play at home because its stadium is reserved for another event. Restricted travel combinations refer to the cases where a team has two away games on consecutive days even though it is forbidden due to the distances between the away venues.

#Rounds	#Full rounds	#B1	#B3	#F2 errors	#F3 errors
84	36 (43%)	261 (58%)	67	1	2
94	30 (32%)	288 (64%)	61	6	1
#Home game restrictions	#3-breaks at home	#3-breaks away	#Away tours	#Restr. travel combinations	#Traveling errors
54	11	11	29	52	11
66	8	22	30	52	11

Table 3: The indicators of the schedules generated for the 2017-2018 and 2018-2019 seasons

4 Practical Lessons Learned

Eleven years ago, we thought that to generate good-quality League schedules, the most important thing is to build the best possible (academic) optimization algorithm for the problem. However, we were very wrong. Generating the League schedule is a long process to ensure satisfaction of League owners, team owners, broadcast right holders, merchandisers and fans, while at the same time competing between them for their preferences. In this section, we give a discussion of important practical issues, our experiments and lessons learned.

Figure 2 shows the main screen of the software. The software includes the same criteria that are used to generate the optimized schedule. Furthermore, the software uses exactly the same violation and fitness calculations. The competition manager told us that he has time and again used the software to show a team manager in real time the cost of moving a requested game. Now he is able to rationalize the consequences and/or the impossibility of a move to himself and to the team managers involved. The number of actualized moves has substantially decreased after the pilot season 2017-2018. Furthermore, now that the team managers have personal experiences of the very complicated correlations between their criteria, they put much more effort of the pre-analysis of their requests, restrictions and priorities.

We are currently generating the schedule for the 2019-2020 season. The process now goes as follows. We first brainstorm the possible improvements to the format with the League's competition manager. The manager has a very deep understanding of both business and sporting issues, as well as what is theoretically possible or impossible. This is a highly important factor in generating high-quality schedules. Then, the format gets approval by the team CEOs and the main broadcasting company. Next, the competition manager tries to gather all the restrictions, requirements and requests from the teams and from the broadcasting company. Then the competition manager together with us decide the priorities of the gathered constraints. Next, we generate some test schedules to check whether it is possible to handle the given framework. Quite often, some changes or relaxations to the given constraints must be suggested to the competition manager. During the test scheduling, the team managers introduce new restrictions. The most common restriction is that the team cannot play at home in certain round because the venue has been sold for some other event, usually for a concert. Of course, we have to add these new restrictions to the framework.

The competition managers signals us to start generating the final schedule. We run the PEAST algorithm several days to find the best possible schedule, which every involved party should be satisfied with - in theory. The competition manager sends the optimized schedule to the team managers and our three-month contribution is over. For the last couple of weeks before the schedule goes public, the competition manager gets feedback from the teams. He uses the JSRoundRobin software to consider whether it is possible to move requested games to the new rounds. Most often this is not possible, because a request from a team would worsen the schedule for the other team involved in that game. The software is also used during the season when unavoidable changes must be processed.

During these eleven years, the League has continuously looked for improvements in its schedule format and the schedule itself. The improvements have focused on two categories:

- A. How to make a more interesting season for the broadcasting company, sponsors and fans
- B. How to cut down the expenses of the teams.

Table 4 shows the number of spectators in the League in the last thirteen seasons. Our contribution started from the 2008-2009 season. The number of spectators increased until the 2014-2015 season, when the biggest team (Jokerit) decided to move to the Kontinental Hockey League (KHL). Two new teams (KooKoo and Jukurit, see Figure 1) were promoted to the League in 2015. However, the League was not able to reach the number of spectators in the earlier seasons. Furthermore, the broadcasting company started to broadcast and stream all the games. Table 4 shows the most important League format innovations we have introduced during these years to make the competition more interesting to all the parties involved, and to secure the number of spectators.

Away tours were introduced in the 2009-2010 season. Western teams Sport, Ässät, Lukko and TPS (see Figure 1) would like to meet eastern teams KalPa, Jukurit, SaiPa and KooKoo in two consecutive rounds, and vice versa. Furthermore, the northern team Kärpät basically wishes to meet whatever southern teams in two consecutive rounds. This has a serious drawback. We cannot schedule an away tour to Friday and Saturday since that would imply two weekend home games in a row for another team. Recall, that one of our goals is to maximize the number of compact rounds. It is obvious that the more we introduce away tours the more we will have breaks in the final schedule. Actually, traveling issues very often compete with the smooth flow of the total schedule. Away tours may imply a decreased

sports-related success. The probability for the team to win its second away game has been 30% smaller than to win any away game.

2005-2006	2006-2007	2007-2008	← Seasons before our contribution		
1 958 843	1 943 312	1 964 626			
2008-2009	2009-2010	2010-2011	2011-2012	2012-2013	2013-2014
	<i>Away Tours</i>	<i>January leveling</i>		<i>Local rivals</i>	
1 997 114	2 015 080	2 036 915	2 145 462	2 189 350	2 088 690
2015-2016	2016-2017	2017-2018	2018-2019	← Seasons with 15 teams	
<i>Back-to-back games</i>		<i>Weekend games</i>			
1 912 200	1 946 790	1 914 360	1 903 500		

Table 4: The number of spectators in the League in the last thirteen seasons, and the timing of the League format innovations

During the past years, we have analyzed the general factors, that influence the revenue of a single game. The three most important factors in order are the opponent, the weekday and the current position in the standings. We also thought that the day of the month when the game is played might have some correlation to the revenue. We thought that the correlation factor could be the salary day. We found no evidence on this. Instead, “Dry January” and “Depressing November” are clearly the most unprofitable months. The former refers to the Finnish habit of having an alcohol free January and the latter to the dark and rainy weather in November. We tried to tackle this by introducing the January leveling games and the back-to-back games in November.

For the 2010-2011 season, we introduced the so-called January leveling to add two extra games for each team. In January, in the middle of the season, the last team on the current standings selects an opponent against which it plays once at home and once away on two consecutive days on Friday and on Saturday. The opponent selects the day for its home game. Then, the second last team (or the third last if the second last was selected by the last team) selects its opponent from the rest of the teams and so on. The teams can choose to select their opponents either by maximizing the winning possibilities or by maximizing the ticket sales. The January leveling was dropped when the League was extended to 15 teams for the 2015-2016 season.

For the 2012-2013 season, we increased the number of local rival games. As described in Section 2, some teams play against certain other teams more than four times. The local rivals play in the same group. The defined local rivals play as many games as possible in the first two rounds. Furthermore, the number of Friday and Saturday games between some local rivals is maximized.

From the 2015-2016 season, we have scheduled so-called back-to-back games where seven pairs of teams play against each other on consecutive rounds on Friday and on Saturday. These match-up games have been highly welcomed by the fans and by the media. However, they have one drawback. The gap between the teams’ other confrontations may be quite long. We tackle this by matching as many teams within the same group as possible (see Section 2).

To maximize the profit for the teams for the 2017-2018 season, the competition manager changed the standard game days to Wednesday, Friday and Saturday. The schedule should now maximize the number of games on Fridays and on Saturdays. This is a somewhat complicated task. Due to the travel distances between some venues, certain combinations of a home team playing the next day an away game against some opponents are not allowed. Similarly, certain combinations of a home team playing

on the previous day away against some opponents are not allowed. The number of forbidden pairs is as high as 52 (see Table 3). In case of existing traveling error in the final schedule, the teams can agree to move the Friday game to Thursday. Another complication is that each team requests to play at least one Friday or Saturday home game against the HIFK and Kärpät teams. These two teams yield the most revenue for the home team.

Most of the teams play two consecutive games on Fridays and Saturdays, but some teams have a rest day on Friday before the Saturday game. Sometimes the coaches have argued (but not after a victorious game) that the other team had the advantage of not playing on Friday. We have found no statistical evidence of that. In fact, the team who has played on Friday has a slightly better probability to win.

Saturday games are also important from the sporting point of view. The probability of a team to win a Saturday home game is more than 20% higher compared to any other weekday. The probability of a top-three-team to win a Saturday home game is almost 30% higher compared to any other weekday. It should be noted here, that we have found no evidence that the carry-over effect has influence on the final standings. Carry-over effect occurs when a team often plays against teams that play against strong teams in previous/next round. The experiences from some other major leagues suggest that there might be some effect, if a team plays against the top three and/or bottom three teams in the consecutive rounds. We have found no evidence of this. One interesting correlation exists. The team's rank group (best six, preliminary playoffs, eleven or worse) after 30 rounds almost certainly implies its rank group after the regular season, i.e. 60 rounds.

Based on the feedback from the League authorities and team CEOs, we have been able to generate schedules that have improved quality and are more valuable. They told us, that the last two schedules have been superior, even though the framework has been most challenging.

5 Academic Lessons Learned

As stated in the beginning of the last section, we thought that to generate good-quality League schedules, the most important thing is to have the best possible optimization algorithm for the professional sports scheduling problem. We were very wrong. However, we still need an academic insight to the problem solving.

During the eleven years, we have used the PEA algorithm to schedule the League. The evolution of the algorithm is closely related to our research in workforce optimization (see e.g. (Kyngäs et al. 2013b) and (Kyngäs et al. 2014)). The algorithm for staff rostering has been integrated into market-leading workforce management software (Visma Numeron WFM) in Finland. Furthermore, we have used the algorithm to solve somewhat more academic problems, such as balanced incomplete block design (Nurmi et al. 2014b), single round robin tournaments with balanced home-away assignments and pre-assignments (Nurmi et al. 2014b), days-off scheduling (Nurmi and Kyngäs 2011) and constraint minimum break problems (Nurmi et al. 2010).

The acronym PEA stems from the methods used: Population, Ejection, Annealing, Shuffling and Tabu. The heart of the algorithm is the local search called GHCM (greedy hill-climbing mutation). In the GHCM search, the basic hill-climbing step is extended to generate a sequence of moves in one step, leading from one solution candidate to another. The GHCM search is used to explore promising areas in the search space. Simulated annealing refinement and tabu search are used to avoid staying stuck in promising search areas too long. Shuffling operators assist in escaping from local optima. The algorithm uses population of candidate solutions. The least fit solution is replaced with a clone of the fittest individual after given number of iterations. A detailed description of the algorithm is given in (Kyngäs et al. 2013a).

The PEA algorithm has been criticized of using many different heuristic methods, and being nothing more than a collection of old ideas. A closer look at the algorithm reveals that the GHCM operator, the shuffling operators, simulated annealing refinement and the ADAGEN penalty method have been clear novelties. The most important observation during the lifespan of the algorithm is that the solution quality clearly decreases if we omit any of the five core components of the algorithm. This holds true for the real-world problems and for the academic problems described in the beginning of the section.

We have been able to improve the schedule of the League by using the same version of the PEA algorithm that we each year use in the workforce management optimization. The performance and the quality improvements of the algorithm implemented during the year are each tested in the sports scheduling framework before starting to generate the final schedule for that season. As an example, the last year's improvement was to randomly choose the version (inner or outer) of the simulated annealing refinement (see Kyngäs et al. 2013a) in each iteration of the algorithm. As another example, the new minimum and maximum limits of the dynamic weights in the adaptive genetic penalty method (ADAGEN) were implemented for this year's schedule generation. The ADAGEN method assigns dynamic weights to the hard constraints based on the constant weights assigned to the soft constraints. The soft constraints are assigned fixed weights according to their priority and significance.

An important finding concerns the generation of the initial solution. Majority of algorithms created to solve real-world problems use a fast greedy method to generate a good-quality initial solution. This significantly cuts down the running time of the algorithm. Furthermore, this is practically without exception argued to generate better final solutions. For the PEA algorithm, we have found no evidence that a greedy method or other sophisticated method outperforms a random initial solution. On the contrary, random initial solutions yield superior or at least as good results in all the real-world problem types and cases where we have applied the algorithm. Recall that we have plenty of time to run the algorithm to generate the schedule for the Finnish Major Ice Hockey League. For the real-world workforce optimization, we have only some minutes or hours to complete the job. However, the customers have preferred to run the algorithm longer to be able to get better final solutions.

The first time we generated the League schedule (for 2008-2009 season) we solved it as a quadruple round robin tournament (Kyngäs and Nurmi 2009). Our ambition level and confidence to the PEA algorithm was too high back then. The sports scheduling problem is not just NP-hard (Easton et al. 2004), but also extremely challenging practical problem. For example, scheduling the Australian Football League (AFL) has been claimed to be the most difficult mathematical problem in world sport (Herald Sun 2012). We have also experience in solving the AFL problem (Kyngäs et al. 2017). We believe that the FMIHL and AFL problems are the most difficult ones.

To cut down the huge search space of 4RR, we decided to schedule the 2009-2010 season by splitting it into two 2RRs. This requires us to predetermine which games and which away tours are played in which 2RR. In addition, a team might prefer home games against important opponents in the second half of the season, as these games are likely to be more attractive near the end of the competition. For the same reason, a game between local rivals might be preferred to be scheduled early in the season. Recall, that the input of the games for the PEA algorithm need not to be round robins. The algorithm optimizes any number of any games based on the given constraints and goals. However, we have to add a constraint to the framework to ensure that there are at least a given number of rounds between two games with the same opponents. The quality of the 2x2RR schedule has proven to be superior to 4RR schedule.

Finally, once more, we want to emphasize that being able to produce good-quality schedules is not about using the best possible algorithm or finding near-optimal solutions. Even more important is to understand the various parties involved, to gather all essential data and to realize the priorities of each party. We call the final generated schedule as the candidate for practical optimum. The final practical optimum is the one generated by the competition manager after he has considered the moves requested by the team CEOs using the JSRoundRobin software.

6 What next?

We presented the most important academic and practical findings that we believe will give new ideas to the sports scheduling community. We hope that we have been able to bridge the gap between theory and practice.

We continue on scheduling the Finnish Major Ice Hockey League and on improving the PEA algorithm. It would be very interesting to use the PEA algorithm to solve sports scheduling benchmark problems. To the best of our knowledge, the only test instances published are the somewhat outdated ones in (Nurmi 2009). We encourage sports scheduling community to create new ones.

References

- K. Easton, G. Nemhauser, and M. Trick “The traveling tournament problem: description and benchmarks” in Proc of the 7th. International Conference on Principles and Practice of Constraint Programming, Paphos, 2001, pp. 580–584.
- K. Easton, G. Nemhauser and M. Trick, “Sports scheduling” in Handbook of Scheduling: Algorithms, Models and Performance Analysis, J. T. Leung, Ed. CRC Press Inc, Florida, USA, 2004, pp 1–19.
- Herald Sun (M. Warner), “Supercomputers tackle toughest mathematics problem in world sport - the AFL draw”, [Online], Available: <http://www.news.com.au/sport/afl/supercomputers-tackle-toughest-mathematics-problem-in-world-sport-the-afl-draw/story-fne1ctok-1226500295493> (Published 22.10.2012).
- S. Knust, “Sports Scheduling Bibliography”, [Online], Available: http://ww2.informatik.uni-osnabrueck.de/knust/sportssched/sportlit_class/, (Last update 23.12.2018).
- J. Kyngäs and K. Nurmi, “Scheduling the Finnish Major Ice Hockey League”, in Proc of the IEEE Symposium on Computational Intelligence in Scheduling, Nashville, USA, 2009.
- J. Kyngäs, K. Nurmi, N. Kyngäs, G. Lilley, T. Salter and D. Goossens, “Scheduling the Australian Football League”, Journal of the Operational Research Society, 2017.
- N. Kyngäs, K. Nurmi and J. Kyngäs, “Crucial Components of the PEA Algorithm in Solving Real-World Scheduling Problems”, in Proc of the 2nd International Conference on Software and Computer Applications, Paris, France, 2013a.
- N. Kyngäs, K. Nurmi and J. Kyngäs, “Solving the person-based multitask shift generation problem with breaks”, Proc. of the 5th International Conference On Modeling, Simulation And Applied Optimization, Hammamet, Tunis, 2013b, pp. 1-8.
- N. Kyngäs, K. Nurmi and J. Kyngäs, “Workforce Scheduling Using the PEA algorithm”, in Ao, Sio-Long (ed.): IAENG Transactions on Engineering Technologies, Lecture Notes in Electrical Engineering Volume 275, Springer, USA, 2014, pp 359-372.
- K. Nurmi, “Sports Scheduling Problem”, [Online], Available: <http://www.computationalintelligence.fi/ssp.htm> (Last update 28.11.2009).
- K. Nurmi, D. Goossens, T. Bartsch, F. Bonomo, D. Briskorn, G. Duran, J. Kyngäs, J. Marengo, CC. Ribeiro, FCR. Spieksma, S. Urrutia and R. Wolf-Yadlin, “A Framework for Scheduling Professional Sports Leagues”, in Ao, Sio-Long (ed.): IAENG Transactions on Engineering Technologies Volume 5, Springer, USA, 2010.
- K. Nurmi and J. Kyngäs, “Days-off Scheduling for a Bus Transportation Company”, International Journal of Innovative Computing and Applications 3 (1), 2011, pp. 42-49.
- K. Nurmi, J. Kyngäs, D. Goossens and N. Kyngäs, “Scheduling a Professional Sports League using the PEA Algorithm”, Lecture Notes in Engineering and Computer Science: Proceedings of

The International MultiConference of Engineers and Computer Scientists, pp. 1176-1182, Hong Kong, 2014a.

K. Nurmi, D. Goossens and J. Kyngäs, “Scheduling a Triple Round Robin Tournament with Minitournaments for the Finnish National Youth Ice Hockey League”, Journal of the Operational Research Society, Vol. 65(11), 2014b, pp. 1770-1779.

R. Rasmussen and M. Trick, “Round robin scheduling - A survey”, European Journal of Operational Research 188, 2008, pp. 617–636.

J. A. M. Schreuder, “Combinatorial aspects of construction of competition Dutch Professional Football Leagues”, Discrete Applied Mathematics 35, 1992, pp. 301–312.

Statistical Models of Horse Racing Outcomes Using R

Alun Owen

Coventry University, U.K.
aa5845@coventry.ac.uk

Abstract

The published literature on statistical modelling of horse racing outcomes is sparse relative to many other sports, apart from some key seminal works that are now very dated. In addition, there appears to have been little, if any, reference to modelling in horse racing at the MathSport series of conferences since it began as the “IMA International Conference on Mathematics in Sport” in 2007. This lack of published work is not surprising given the potential commercial value of such knowledge and hence the potential unwillingness for authors to share their work.

This paper therefore aims to fill this gap by presenting details of how a potentially profitable model could be developed, using a discrete choice modelling approach within the statistical computing environment R (R Core Team, 2018). Suitable data sources are considered, along with the process of model development and implementation in R, as well assessments of model fit and the potential profitability of the model when used for betting on future races.

1 Introduction

The published literature relating to the development of predictive models in horse racing is relatively sparse, despite it being one of the oldest sports still in common existence today and its long historical relationship with betting. There are many variations of the sport, with two of the most common forms being flat racing (no jumps) and hurdle or steeplechase races over jumps. In addition, races may or may not involve handicapping horses, where the “better” horses are required to carry heavier weights in their saddles to attempt to “even out” the field.

Amongst the seminal published papers in the context of the development of predictive models in horse racing, are Bolton and Chapman (1986), Benter (1994) and Chapman (1994). The first of these, Bolton and Chapman (1986), was the first to outline the use of a multinomial logistic regression model (see Section 3 for details of this model) in the context of 200 handicap races in the USA and utilizing 10 predictor variables. Chapman (1994) extended this work with the multinomial logistic model with a similar much larger database of 2,000 handicap races in Hong Kong, utilizing 20 predictor variables. Both argue that it was possible to derive positive returns from a unit stake betting strategy using a multinomial logistic regression modelling approach. Benter (1994) provides an excellent insight into the potential predictor variables that might be incorporated into an empirical model, and reports on his own successes in developing a profitable system. All three of these papers

appear in the highly recommended publication by Haush et. al. (2008), which is essential reading for anyone interested in developing a predictive model in horse racing, particularly in the context of betting.

A more recent publication is that by Wong (2011), which also includes some reference to the implementation of a multinomial regression model for horse racing in the context of the R programming language (R Core Team, 2018). However, implementing such a model using R presents a number of challenges which are not fully address in that reference. Therefore, this paper aims to provide an outline of how to implement such a model in R, discussing some of the potential problems, as well as possible data sources and the process of developing suitable predictor variables from the data. In addition, suitable model assessments are considered both in relation to model fit using in-sample data and also betting performance when applied to out of sample data.

Section 2 of the paper discusses potential data sources and the data used in the modelling work here, whilst Section 3 then describes the multinomial logistic regression model in the context of horse racing and develops a predictive model using our data. The predictive performance when deployed to new unseen (out of sample) data is then considered in Section 3, with Section 4 providing a summary of the work presented.

2 Data

There is a wealth of available data related to horse racing allows for a rich variety of predictive models to be developed. There are also many data sources that could be considered but many of these are not very helpful. Amongst the more useful data sources are *Proform* (<https://www.proformracing.com/>), *Timeform* (<https://www.timeform.com/horse-racing/>) and *Raceform* (<http://www.raceforminteractive.com/>). These include data on the horses running in each race, the jockeys and the horses' trainers.

The data source we use was compiled from a number of these sources and comes from the Flat Turf Handicaps in the UK, with 16,685 horses taking part in 1,693 races. This data set is obtainable from the author on request and is contained in a .csv file named *horse.data*. This includes for each horse in each race the following variables:

- race.id - unique reference number for each race;
- horse.ref - unique reference number for each horse in each race;
- age - age of the horse (years);
- sireSR - win percentage by offspring of the horse's sire prior to this race;
- trainerSR - win percentage achieved by the horse's trainer prior to this race;
- daysLTO - days since last race (days since Last Time Out);
- position1 - finishing position in the previous race (1, 2, 3 or 4, 0 = anywhere else);
- position2 - finishing position two races ago (1, 2, 3 or 4, 0 = anywhere else);
- position3 - finishing position three races ago (1, 2, 3 or 4, 0 = anywhere else);
- finpos - finishing position in the current race;
- win - indicator of whether each horse won (yes) or not (no);
- sp - starting price obtained from Betfair;
- various other indicators (1=yes, 0=no) of whether the horse was an entire (male horse that has not been castrated), gelding (male horse that has been castrated), and if it was wearing blinkers, a visor, cheekpieces or a tongue tie. Note that a horse that is neither a gelding nor an entire was female.

The variable sp is the starting price obtained from the Betfair betting exchange (<https://www.betfair.com>). This is used here as it represents a kind of consensus of the market just before the event begins, and avoids having to account for the fact that odds change over time. Appendix 1 provides a sample of the data for illustrative purposes. Note that this data set excludes many other variables, such as jockey statistics, but we restrict the number of variables both for brevity, and to maintain the commercial value of the models developed with a fuller version of this data set.

2.1 Data Management and Potential Pitfalls

We split our data (16,685 horses from 1,693 races) into a training sample (70% of races) in order to develop a model and to propose a future possible betting strategy. We retain the remaining 30% of races for out-of-sample assessments of betting performance when the model is applied to new unseen race data. The training set thus consists of 11,710 horses taking part in 1,181 races, with the remaining 4,975 horses from 512 races making up the test set.

Section 3 describes a multinomial logistic regression model developed here, which requires careful consideration of two key variables included in our training data. One of these is `horse.ref`, which provides a means of identifying the horses (or choices) within each race. These only need to be unique within a race as the model provides a means of estimating the probability (more precisely the odds) of winning conditional on the horses it is competing against in that race. The `horse.ref` does not need to be unique to each horse across the data set and in our modelling approach we do not make any connection with horse 1 in race 1 and horse 1 on race 2, for example. The other key variable is `race.id`, which indicates the “choice” groups of horses in each race from which “nature” chooses a winner. The `race.id` must be unique to each race and experience suggests that not checking that this is the case is a common cause of problems, which we consider here.

The model in Section 3 is fitted using the function `mlogit()` from the `mlogit` package of Croissant (2019) within the R statistical computing environment. This requires the data to be in a particular format such that the following function variables need to be specified with care:

- `choice` indicator of which horse won each race
(in our data set this is the variable called `win`);
- `chid.var` variable that defines the choice sets (races) from which the choice of winner is made
(in our data set this is `race.id`);
- `alt.var` variable that defines the choice alternatives (horses) in each set (race)
(in our data set this is `horse.ref`)

The above function variables can be defined in the required format using the `mlogit.data()` function available within the `mlogit` package (where our data set is contained in a data frame in R called `model.data`) as follows:

```
h.dat<- mlogit.data(data=model.data,choice="win",chid.var="race.id",
alt.var="horse.ref",shape="long")
```

The shape is “long” since in our data set we have one row for each horse. We could have included all the horses from one race together on one “wider” row in which case our data shape would have been wide. A common cause of error is with not ensuring the `race.id` (or other variable used to specify the `chid.var` function variable) are unique for each race. Our data set has been checked and “cleaned” to ensure no problems will occur. Resolving issues with this command for other data sets may not always be easy and experience suggests trying running the command on just the first few races in the data set, and increasing the amount of races until the error occurs. This will at least identify which set of race data is the first row to cause problems. Another issue to be wary of is that the `mlogit` package cannot accommodate ties and so any races where two or more horses are tied as winners would need to be removed from the data set.

Within our data set, we also need to consider the issue of the market over-round in the starting prices or odds held in the variable `sp`. These odds do not include the backer’s stake, so for example odds (`sp`) of 1.5 would mean that betting 1 unit on that horse would win 1.5 units and so would see 2.5 units returned to the backer if the horse won. The implied probability (of the horse winning the race) from these odds = $1/(sp+1) = 1/2.5 = 0.4$. Totaling these implied probabilities across all horses in the race should result in a total = 1. However, this total is often greater than 1 and the amount by which this total exceeds 1 is often referred to as the over-round. The histogram

in Figure 1 below provides a summary of the over-rounds evident in our training data set and shows that this is mostly around the 1.1 or 1.2 mark, but can be as much as 1.5.

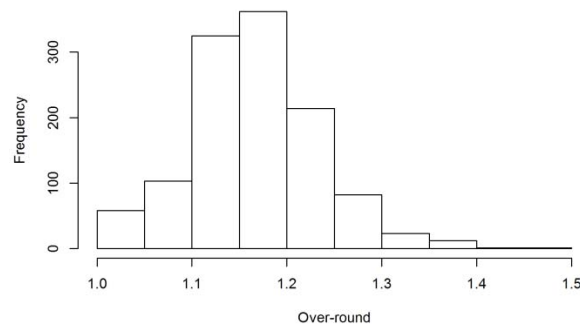


Figure 1: Histogram summary of the over-rounds within the training data set

In order to make more informed assessments of the model versus the market odds, and also to make better judgements with respect to betting strategies, we derive to additional variables to add to our data set. The first of these we name `market.prob`, which contains the market implied probabilities of each horse winning any particular race. The other we call `market.prob.adjusted`, which contains the market implied probabilities of each horse winning any particular race but adjusted by dividing by the over-round for that race so that those probabilities do sum to 1.

`SireSR` and `trainerSR` are, respectively, the strike rate (percentage of races won) by the offspring of the horse's Sire (father) prior to this race and the win percentage achieved by the horse's trainer prior to this race. The boxplots in Figure 2 summarise the data for both `sireSR` and `trainerSR` in the training data set. In both cases, the vast majority have `sireSR` and `trainerSR` values of less than 20% and only a handful of horses have values above this. A value above 20% would also be considered to be very unusual. Hence for both variables, we censor the values so that any above 20% instead take the value of 20%.

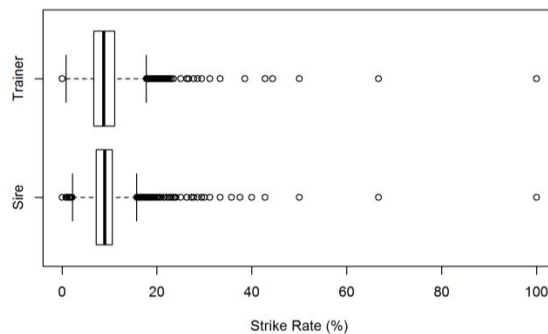


Figure 2: Boxplot summaries of `sireSR` and `trainerSR` within the training data set

Examining `daysLTO`, which is the number of days since the horse's last race (days since Last Time Out), lead us to decide to also censor this variable so that values of `daysLTO` above 60 days are capped at 60 days. For brevity, the data summaries used to support this decision are not included here.

2.2 Exploratory Analysis

This section explores the training data set, and aims to provide an illustration of some of the types of initial assessments that need to be undertaken before any attempt is made to develop a model. One of the key ingredients in a successful model, is the development of new more meaningful variables from those that are collected within the original data set. A knowledge of horse racing and how various factors could affect a horse's performance is therefore crucial in this work.

We first consider age. Gramm and Marksteiner (2011) offers evidence that a horse's peak age for racing is around 4.5 years. These findings are supported by the results in Figure 3, which summarises the proportion of winning horses in each age category (in years) in our training data set.

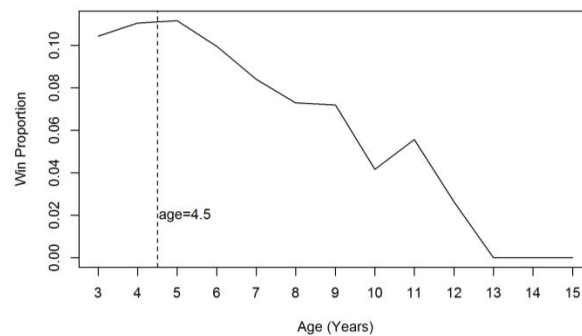


Figure 3: Win proportion versus Age

It would therefore seem sensible to define a revised version of the variable age that might serve as a better predictor of the odds of winning. Here, the absolute difference in age from 4.5 years is utilized instead and referred to as `age.diff`. Note that this assumes the rate of improvement from age 3 to 4.5 is the same as the rate of decline after age 4.5, which would seem to be supported in Figure 3. However, Gramm and Marksteiner (2011) found that the improvement from age 3 to 4.5 is greater than the rate of decline after age 4.5.

The conditional density plots in Figure 4 summarise the win proportion against `sireSR` and `trainerSR`. Conditional densities are used to smooth out random fluctuations in win proportion observed with increasing strike rate. Both plots show an increase in win proportion with higher values of `sireSR` and `trainerSR` and hence evidence their values as potential predictors in the model.

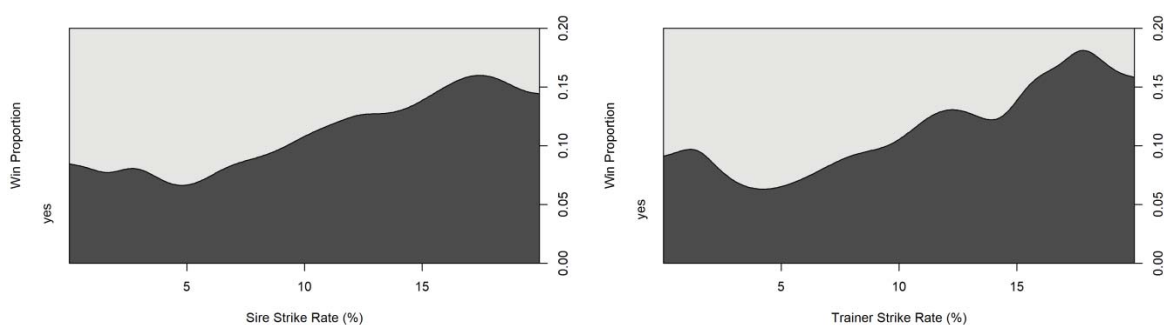


Figure 4: Win proportion versus Sire Strike Rate and Trainer Strike Rate

The conditional density plot in Figure 5, similarly summarises the win proportion against daysLTO, and illustrates how win proportion decreases with higher values of daysLTO.

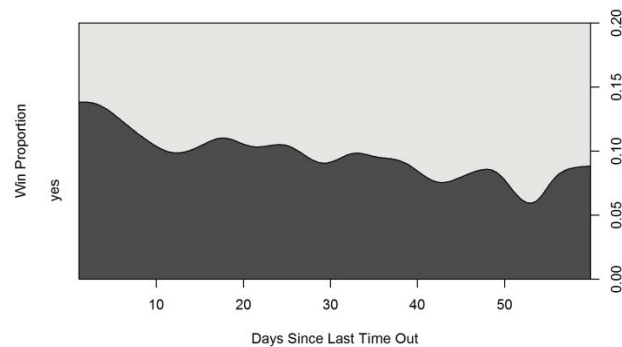


Figure 5: Win proportion versus Days Since Last Time Out

Figure 6 below summarises how win proportion varies with a horse's finishing position in its previous race, two races previous and three races previous. This suggests that higher finishing positions in the previous race and two races previous (variables position1 and position 2 in our data set) are associated with a greater win proportion in the subsequent race. The result three races previous does not seem to have much impact on the current race outcome.

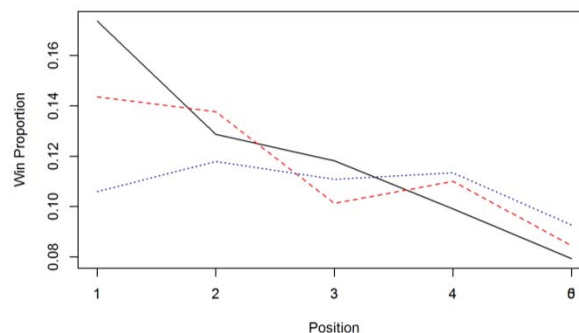


Figure 6: Win proportion versus Position in the horse's previous race (—), two races previous (---) and three races previous (.....)

Finally, Table 1 summarises the remaining categorical variables in relation to win proportion. The first row indicates the win percentages for horses that were entire, geldings, or wore blinkers, visor, cheekpieces or a tongue tie. The second row indicates the win proportions for horses that were not entire, or not geldings or did not wear the item listed. This suggests that greater win proportions in the sample are evident with horses that were entire and geldings (i.e. male horses) compared to female horses, and also with horses that were wearing blinkers or not wearing cheek pieces or a tongue tie. The impact of a visor across the sample seems minimal in terms of win proportion.

	Entire	Gelding	Blinkers	Visor	Cheek Pieces	Tongue Tie
Yes	0.115	0.106	0.111	0.103	0.069	0.084
No	0.099	0.091	0.100	0.101	0.103	0.102

Table 1: Win proportion versus the remaining categorical predictor variables

3 Model and Results

3.1 Model Definition and Fitting Using R

The model we utilise is the multinomial logistic regression model. There are actually a number of different forms of the multinomial logistic regression model and they are often collectively referred to as discrete choice models. In our context we treat each horse in a race as the set of choices from which nature will choose a winner when the race is run. The multinomial model is usually framed in the context of people making choices and so we often see references to individual-specific variables and alternative-specific variables. In our case the race is making the choice of horse and so individual-specific variables are the race-specific variables and the alternative-specific variables are the horse-specific variables. This is often the source of confusion that prevents many implementing the multinomial logistic regression model in the context of horse racing.

For the multinomial logistic regression model we define linear predictor terms V_i for horse i as:

$$V_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}$$

where x_1, x_2, \dots, x_p are the values of the horse-specific variables (age, sireSR, trainerSR etc.) for horse i and $\beta_1, \beta_2, \dots, \beta_p$ are the associated model parameters to be estimated in the model. We could also extend this to include other variables that are race-specific if required, such as the going (the condition of the course) or the value (standard) of the race. However, our data set does not include such variables and so we ignore this aspect of the model here. The model can then be stated in terms of the probability P_i that horse i will win the race as follows:

$$P_i = \frac{\exp(V_i)}{\sum_{i=1}^n \exp(V_i)},$$

where there are n horses in the race. Note that we have no intercepts (β_0 terms) in the model, since all we are interested in are the relative chances of each horse winning relative to all the other horses in the same race. The above model also has the useful property that for any given race, $\sum_{i=1}^n P_i = 1$.

Fitting the model with all of our predictor variables in the model is then undertaken using the `mlogit()` function as follows:

```
mlogit(win~age.diff+sireSR+trainerSR+daysLTO+position1+position2+position3+entire
+gelding+blinkers+visor+cheekpieces+tonguetie|0|0,data=h.dat)
```

The two vertical lines after `tonguetie` are used to separate the predictor variables into three parts. The first part (before the first of the two vertical lines) contains the predictor variables that are horse-specific, which in our case is the list of all of our predictor variables. The second part (between the two vertical lines) which currently has just a zero listed, would contain the predictor variables that are race-specific, but we have none of these and so we state this using 0. The third part (after the last of the two vertical lines) which also currently has just a zero listed, can be used to list the predictor variables that are again horse-specific (like the first part), but in this case the way in which these variables affect the win probability would be allowed to vary from horse to horse. Again, since we have none of these we simply state this by using 0.

3.2 Model Results

The results of fitting an initial model with all predictor variables in the model are shown in Table 2. The reference categories for the three "Position" variables was "0" (i.e. unplaced in the previous races).

Parameter	Estimate	Std. Error	p
age.diff	-0.153	0.0314	<0.001
sireSR	0.048	0.0093	<0.001
trainerSR	0.051	0.0093	<0.001
daysLTO	-0.004	0.0018	0.020
Position1	1	0.602	<0.001
	2	0.324	
	3	0.312	
	4	0.159	
Position2	1	0.368	<0.001
	2	0.363	
	3	0.066	
	4	0.213	
Position3	1	-0.046	0.43
	2	0.109	
	3	0.117	
	4	0.130	
entire	0.499	0.1297	<0.001
gelding	0.557	0.0948	<0.001
blinkers	0.016	0.1125	0.89
visor	0.027	0.1443	0.85
cheekpieces	-0.504	0.1470	0.001
tonguetie	-0.297	0.1632	0.069

Table 2: Parameter estimates for an initial model including all predictor variables

This suggests the following are all highly statistically significant ($p < 0.001$) predictors of the odds of winning: age.diff, sireSR, trainerSR, daysLTO, position1, position2, entire, gelding and cheekpieces. In addition, tonguetie is borderline not significant ($p = 0.069$) but we chose to exclude this. There was no evidence that the following should be retained in the model: position3 ($p = 0.43$), blinkers ($p = 0.89$) and visor ($p = 0.85$).

A process of developing a model by removing the most non-significant predictor variable above (i.e. starting with the removal of blinkers) was undertaken until only significant predictors remained. The resulting model is summarised in Table 3. Note that the exclusion of visor was a borderline decision and the decision to exclude this was argued on the basis that the more parsimonious model is preferable. The results in Table 3 indicate that the odds of winning are increased for male horses (entires and geldings) compared to female horses. Increased odds are also associated with horses closer to the optimal age (4.5 years) and who have higher values for sireSR and trainerSR, lower values for daysLTO and a higher placing (1st being highest) in the previous two races. In addition, the wearing of cheek pieces is also seen to decrease the odds of winning.

Parameter		Estimate	Std. Error	P
age.diff		-0.152	0.0313	<0.001
sireSR		0.048	0.0093	<0.001
trainerSR		0.051	0.0093	<0.001
daysLTO		-0.004	0.0018	0.021
Position1	1	0.607	0.0918	<0.001
	2	0.329	0.1003	
	3	0.313	0.1026	
	4	0.157	0.1081	
Position2	1	0.377	0.0970	<0.001
	2	0.373	0.0975	
	3	0.075	0.1070	
	4	0.220	0.1048	
entire		0.489	0.1294	<0.001
gelding		0.548	0.0946	<0.001
cheekpieces		-0.503	0.1454	0.001

Table 3: Parameter estimates for a "final" model

3.3 Model Fit Assessments

One key assessment of model fit is that the model should be well calibrated. That is, for example, on average, horses with a model win probability of 0.2 should win 20% of races they take part in. The horses in the training set were therefore grouped into 9 sets (bins) based on whether their model fitted win probabilities fell into the following intervals: $[0,0.05]$, $(0.05,0.1]$, ..., $(0.35,0.40]$, $(0.4,1.0]$. Figure 7 plots the mean value for the model fitted win probabilities for each of these intervals and the associated win proportions that were actually observed (plotted using the symbol "o"). The plot also shows the same results for the win probabilities implied from the market odds adjusted for over-round (+). The closer the points are to the reference line shown the more accurate the fitted probabilities are on average, and hence are better calibrated. This suggests the model is well calibrated although less so for the higher win probabilities, but the sample size for these higher win probabilities is much smaller being subject to more random fluctuation. Figure 7 would also seem to suggest that the model performs very well compared to the market implied probabilities.

Another potentially useful measure of model performance are those related to scoring rules used for prediction, although in this case we are considering in-sample predictions. Two such measures we consider here are $P1$ and $P2$ as follows:

$$P1 = \exp \left\{ \frac{1}{N} \sum_k \log(P_{jk}) \right\}$$

where P_{jk} represents the model fitted win probability (or win probability implied by the odds adjusted for over-round) for horse j that was observed to win race k , for races $k=1, \dots, N$, and

$$P2 = \exp \left\{ \frac{1}{N} \sum_k \left[\log(1 - P_{jk})^2 + \sum_{i \neq j}^{n_k} \log(P_{ik})^2 \right] \right\}$$

where P_{jk} represents the model fitted win probability (or win probability implied by the odds adjusted for over-round) for horse j that won race k . Here, P_{ik} represents the model fitted win probabilities (or win

probability implied by the odds adjusted for over-round) for all other horses $i \neq j$, that did not win race k , for horses $i=1, \dots, n_k$ in race k , for $k=1, \dots, N$.

In effect, $P1$ is equivalent to the geometric mean of the model fitted win probabilities for the horses that won the races in the training set. Hence, higher values of $P1$ are preferable. In contrast, $P2$ represents a form of quadratic loss or brier scoring rule, and as such measures the squared error in the model, and so smaller values are preferable.

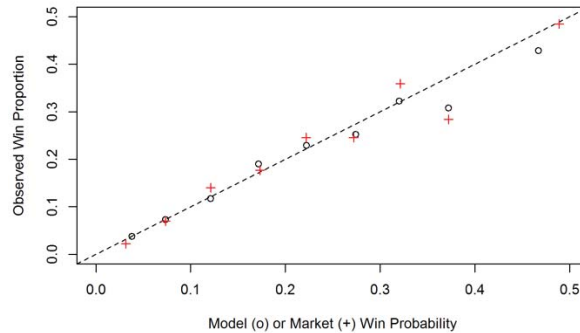


Figure 7: Calibration plot for the model (o) and market implied win probabilities (+)

Calculating $P1$ and $P2$ for the training set, for both the model fitted win probabilities and the win probabilities implied by the market odds adjusted for over-round, gave $P1_{\text{model}} = 0.120$ and $P1_{\text{market}} = 0.136$, and $P2_{\text{model}} = 0.860$ and $P2_{\text{market}} = 0.831$. These suggest the market outperforms the model overall. However, this is not the full story as shown in Figure 8. This plots the values of $P1_{\text{model}}$ compared to $P1_{\text{market}}$ and $P2_{\text{model}}$ compared to $P2_{\text{market}}$, but based only on horses with win probabilities above the values shown on the x-axis, which ranged from 0 to 0.5. For example, based only on horses where the model fitted win probabilities exceed 0.13, we see that $P1_{\text{model}} = P1_{\text{market}}$ and $P2_{\text{model}} = P2_{\text{market}}$. Therefore, if for example we focus only on horses where the model fitted win probabilities exceed, say, 0.15, we see that $P1_{\text{model}} > P1_{\text{market}}$ and $P2_{\text{model}} < P2_{\text{market}}$ and so in that case the model outperforms the market. One possible reason for this could be due to the well-established favourite-long-shot bias. In this case, the odds of long-shots are lower than those that might reflect "fair" odds and so the implied probabilities are higher than would be expected and hence biased. In contrast, the odds for favorites (or horses with higher win probabilities) are actually relatively higher than expected, resulting in lower implied win probabilities. These results suggest there is scope using this model, betting only in horses with a model win probability above some threshold such as 0.15.

A naïve betting strategy to deploy to the so far unseen test data, might therefore to simply choose to bet on horses which has a model win probability greater than 0.15 (established above). In addition, if we consider the model win probability implying "fair" odds (often referred to as an odds-line), then we might wish to also restricts bets to where the model win probability is greater than that implied from the market odds (sp). However, experience with developing betting strategies of this nature, suggests that this may not represent an optimal strategy. A more informed strategy would be to bet where the ratio of model win probability / market implied win probability is greater than some threshold (greater than 1). Further analyses based on the training data set again only (not shown here for brevity) suggested that betting when the ratio of the win probabilities from the model/market is above 1.3 appeared optimal.

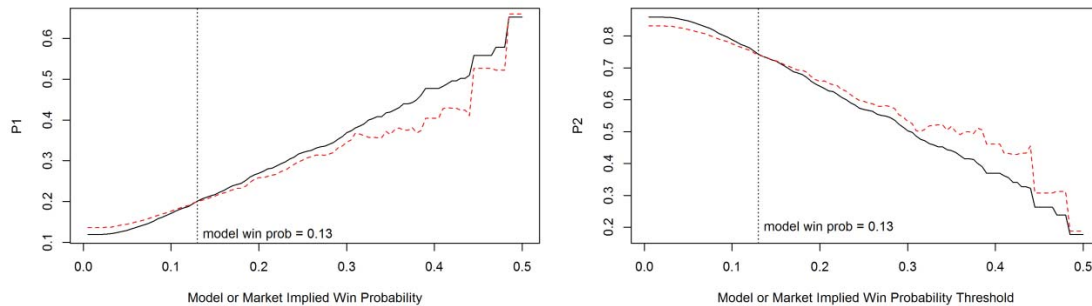


Figure 8: Plots of $P1$ and $P2$ based on win probabilities above the values shown on x-axis for Model (—) and the Market implied win probabilities (---)

3.4 Future Predictive Performance

The model summarised in Table 3 was deployed to the so far unseen test data set (4,975 horses from 512 races) to derive estimated win probabilities for each horse in those races. Unit bets were placed "virtually" on those horses where the model win probability was greater than 0.15 and the ratio of the win probabilities from the model/market was above the threshold of 1.3. This returned a profit of 54 units betting on 264 horses, which represents a positive return on investment (ROI) of around 20% (return of 54 units from 264 units staked). Clearly this might be due to chance and so in Figure 9 we summarise the potential variation in this return using bootstrap estimates of the standard error. This shows the potential ROI when deploying the model using a betting strategy adopting a range of thresholds for the ratio of the win probabilities from the model/market (from 0.9 to 2.0), and includes the 90% confidence interval for this estimate. This provides evidence to suggest that whilst positive returns are not guaranteed, they would appear to have been quite likely using our model and betting strategy.

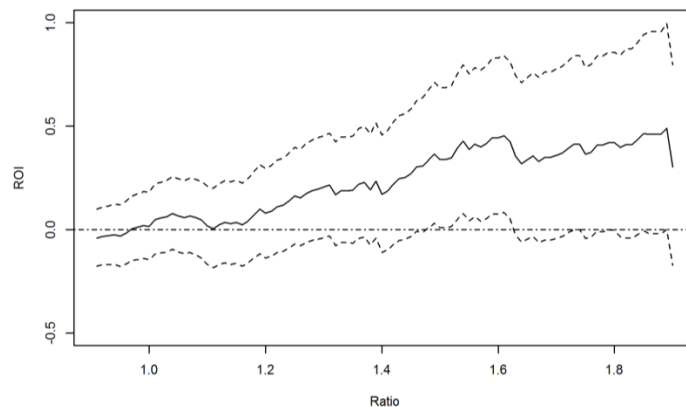


Figure 9: ROI when betting above a range of thresholds of model win probability/market implied win probability (—) and 90% Confidence Intervals (---)

4 Summary

This paper has shown how a potentially profitable model could be developed, using a discrete choice modelling approach within the statistical computing environment R. A large data set with 16,685 horses taking part in 1,693 races from the Flat Turf Handicaps in the UK was utilised. A model was developed from a training set representing 70% of the races (11,710 horses taking part in 1,181 races). with the remaining 4,975 horses from 512 races making up a test set. A multinomial logistic regression model was developed which was shown to be well-calibrated and competitive in relation to the market prices contained within the Betfair starting prices. A betting strategy, again developed using only the training set, was deployed to with the model using the test data. The results suggest that positive returns are possible using such an approach. The data set and R code used in the analysis are both available from the author on request.

References

- Benter, W. (1994), Computer-based Horse Race Handicapping and Wagering Systems, in *Efficiency of Race Track Betting Markets*, eds. Haush, Lo and Ziemba, pp 183-198. Re-printed 2008.
- Bolton, R. N. and Chapman, R., G. (1986), Searching for Positive Returns at the Track: A Multinomial Logit Model for Handicapping Horse Races, *Management Science*, 32 (8), pp1040-1060. <http://www.ruthnbolton.com/Publications/Track.pdf>
- Chapman, R., G. (1994), Still Searching for Positive Returns at the Track: Empirical Results from 2,000 Hong Kong Races, in *Efficiency of Race Track Betting Markets*, eds. Haush, Lo and Ziemba, pp 173-181. Re-printed 2008.
- Croissant, Y. (2019). mlogit: Multinomial Logit Models. R package version 0.4-1. <https://CRAN.R-project.org/package=mlogit>.
- Gramm, M. and Marksteiner, R. (2011). The effect of Age on Thoroughbred Racing Performance. *Journal of Equine Science*, 21(4), pp 73-78.
- Haush, D. B., Lo, V. S. and Ziemba, W. T. (2008). *Efficiency of Race Track Betting Markets*. World Scientific, Singapore.
- R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- Wong, C. X. (2011). *Precision: Statistical and Mathematical Methods in Horse Racing*. Outskirts Press Inc., Colorado.

Appendix 1: Data Sample

race.id	horse.ref	age	sireSR	trainerSR	daysLTO	position1	position2	position3	finpos	win	sp	entire	gelding	blinkers	visor	cheekpieces	tonguetie
1	1	7	6.2	5.4	96	0	0	0	5	no	18	0	1	0	0	0	0
1	2	7	10	9.7	4	3	1	2	9	no	3.5	0	1	0	0	0	0
1	3	4	8	11.1	23	0	4	1	6	no	8	0	0	0	0	0	0
1	4	6	8.8	11.4	40	4	1	0	3	no	3.5	0	0	0	1	0	0
1	5	8	4.7	11.9	14	0	1	3	4	no	11	0	1	0	0	1	0
1	6	9	2.5	2.8	16	3	0	0	1	yes	6	0	1	1	0	0	0
1	7	5	9.5	8.7	16	0	0	0	2	no	4.5	0	1	0	0	0	0
1	8	6	8.1	9	2	0	2	0	7	no	9	0	1	0	1	0	0
1	9	7	8.3	9	23	0	0	0	8	no	20	0	0	0	1	0	0
2	1	9	8.1	5.2	16	3	0	3	3	no	4	0	1	0	1	0	0
2	2	6	7.4	8.8	159	0	0	2	7	no	8	0	1	0	0	0	0
2	3	10	0	0	5	0	0	0	8	no	16	0	1	0	0	0	0
2	4	6	8.8	14	5	0	0	1	5	no	9	0	1	0	0	0	0
2	5	5	9	13.6	23	4	0	1	2	no	2.25	0	0	0	0	0	0
2	6	9	8.3	8.7	19	4	1	2	1	yes	7	0	1	0	0	0	0
2	7	8	7.3	11.4	31	0	0	0	6	no	12	0	1	0	0	0	0
2	8	7	7.1	10	14	2	0	0	4	no	5	0	1	0	0	0	0
3	1	5	13.8	11.9	40	3	1	0	2	no	2	1	0	0	0	0	0
3	2	7	7.3	9	13	1	3	0	1	yes	6	0	1	0	0	0	1
3	3	4	12.4	15.5	12	4	0	0	3	no	8	0	0	0	0	0	0
3	4	7	7.1	7	19	4	2	4	5	no	9	0	1	0	0	0	0
3	5	6	8.8	11.4	12	3	1	1	6	no	10	0	1	1	0	0	0
3	6	5	8.4	6	14	1	1	1	4	no	2	0	0	0	0	0	0
3	7	5	8.1	5.2	19	0	4	0	7	no	33	0	1	0	0	0	0

Multi-Criteria Solutions for Optimizing Lineup in Baseball

Marina V. Polyashuk

Northeastern Illinois University

Chicago, IL 60625, USA

m-polyashuk@neiu.edu

***Abstract:** This paper focuses on the task of creating an optimal lineup for a baseball game, which is one of the most important tasks facing a Major League Baseball team manager. The lineup problem is decomposed into two sequential parts: first, selection of an optimal 9-player set of position players, and second, choosing the best way to arrange this set of nine players in the batting order. The problem of choosing a set of starting players for a game is formulated as a bi-criteria problem with such criteria as the offensive potential of the set and the defensive quality of the set, with optimal solutions defined as non-dominated (Pareto-optimal) solutions with respect to both criteria. The problem of the batting order is viewed as a combinatorial problem with two lexicographic-ordered criteria, the first of them being the expected number of runs scored by one rotation of the lineup. Throughout the paper, the proposed methodology is applied to the example of the 2012 Chicago White Sox team; actual managing decisions are compared to the solutions proposed by our model.*

1. Introduction

If you decided to use the key-word “baseball” in your search of a scientific database, you would be buried under thousands upon thousands of titles. These titles cover all possible aspects of the game, including but not limited to identifying winning parameters or combinations of factors, improving the chances of winning the game; or having an edge over the competition in the long run. After all, effectiveness and success of a team has a great influence on this multi-million-dollar business. And then there is sabermetrics (derived from SABR, the Society for American Baseball Research represented by publications like Baseball Info Solutions and James B., 2016), which became known to the general audience of movie-goers through the motion picture “Moneyball”. Most mathematicians in the US, let alone statisticians, are aware that baseball is an ultimate breeding ground for statistical studies and analyses. In short, baseball is not just “America’s favorite pastime”, but also an inspiration for numerous research and business ideas.

Besides the United States, baseball is a favorite sport in such countries as (in alphabetical order) Canada, Cuba, Dominican Republic, Israel, Italy, Japan, Netherlands, South Korea, and Spain. It was a part of summer Olympics from 1992 through 2008 and was voted out in 2012, one of the reasons being that Major League Baseball (MLB) decided against taking a two-week break during the Games so that the players could participate. Baseball’s long history of international tournaments (since 1938) suggests that this sport is popular in many countries and still has a potential to grow in popularity in the future. One of the signs of such potential is recent report by the Associated Press that two MLB teams, namely New York Yankees and Boston Red Sox will be playing a game series at London Olympic Stadium in June of 2019 (Guardian, May 3, 2018).

The goal of this paper is to suggest solutions for optimizing the starting lineup in a baseball game. What is the lineup? It is a pre-arranged order in which nine players of the team will come to bat when the team is in offense (batting) mode. The same nine players (with one exception, which will be clarified later) are assigned to their specific defensive positions when the team operates in defense (fielding) mode. The starting lineup is decided by the team manager before every one of the 162 games of the regular baseball MLB season; an optimal lineup would aim at maximizing the chance of winning a single game by scoring more runs than the opposite team. All baseball experts agree that different lineups bring different results in the game, whether it has to do with a single game, the entire season, or championship playoff series. There is constant and justified interest towards this topic among professionals and fans of baseball, and there does not exist a single approach to solving this problem that would be universally recognized, thus prompting this research and the attempt to apply multi-criteria solutions to optimizing lineups in baseball.

We will break the lineup problem into two parts: one, selection of the 9-player set for a starting fielding position assignment, and two, deciding on the order in which this set of players will come to bat (batting order). The problem of choosing a set of starter players for a game will be formulated as a bi-criteria problem with such criteria as the offensive potential of the set and the defensive quality of the set. The problem of the batting order will be viewed as a combinatorial problem with two lexicographic-ordered criteria, the first of them being the expected number of runs scored by one rotation of the lineup. We will then compare results of the model applications with the results of actual managerial decisions. In the paper, we will also address situational choices based on players' "handedness" (throwing/batting from right or left side) when considering possible applications of the proposed model selecting a starting position assignment.

Majority of the examples that are used in this paper are of the 2012 Chicago White Sox (CWS) season. The main reason for this choice is that the author is a long-time White Sox fan; year 2012 was when the author collected statistical data for this research. It turns out that not only 2012 was the first and the best year for Robin Ventura as the team manager (2012 – 2015), but that 2012 is the last season with a winning record for the team up until now.

As was mentioned earlier, we limited ourselves to two criteria in choosing the starting position assignment. Because of this, vectors of criteria values for all feasible assignments can be viewed as points in a two-dimensional graph where optimal points can be easily identified, and the corresponding assignments defined. In all practicality, with a little training a manager (or his assistant) could use this graph as a tool before each game, given that all necessary offensive and defensive statistics of the players are kept up to date. Limiting the multi-criteria model to two objectives also allows the manager to use his own determination to make final selection of the starting position assignment, since there very likely will be more than one optimal solution provided by our model. For example, in the case of 2012 CWS demonstrated in this paper (with statistics measured in mid-August), there were just 3 optimal points, which were obtained from the pool of 720 feasible options. Further, the manager as the decision maker may apply his knowledge of "intangibles" and hard to quantify factors such as current physical condition of the players, their psychological evaluation, "chemistry" between the players, etc. to make the ultimate choice. In other words, we may use the "drawback" of the model having too few

criteria to manager's advantage. This last feature could make a decision-support system based on the ideas of our research attractive to MLB management; however, such a system is a matter of future work and is out of the scope of the current research.

Some of the ideas presented here could potentially provide support in managing sports other than baseball (e.g. softball), where the players are fulfilling distinctive responsibilities as both fielders (in defense) and batters (in offense) with a prearranged offensive order determined before the game. In addition, the part of the model that chooses the set of starters based on defensive quality of the set and offensive potential of the set can be applied to many other popular sports such as basketball, hockey, and soccer. Indeed, in all these games the team players (which even includes goalkeepers in soccer and hockey), regardless of their assigned positions, must fulfill both offensive and defensive functions. Therefore, it makes sense to use offensive and defensive qualities when selecting the set of starters in a soccer game or choosing the set of players with most time played on the court in basketball.

The paper is structured as follows. Section 2 contains a literature review on the lineup selection problem; in section 3 we present the formulation and proposed solutions for the problem of selecting an optimal starting position assignment. In section 4, we formulate and present solutions for the problem of choosing an optimal batting order, once the set of starting players has been determined. In both sections 3 and 4, we will use the actual data from the 2012 season of the CWS team to demonstrate how the models work in practice and to compare the results with the actual managing solutions. The overall results of the current research, its merits and limitations, and future directions are discussed in section 5, which concludes the paper. Appendix contains some of the rules of baseball, technical terms and notations that the reader might not be familiar with and that are necessary for reading the paper.

2. Lineup selection problem: a literature review

In sports sections of newspapers and online, one may find many experts' opinions and suggestions on forming an "optimal" lineup, especially if it relates to a specific team and/or game (such as an All-Star game between American League (AL) and National League (NL), which happens every year in July, when the teams of the most popular players from each league are selected to compete). There are websites (e. g. www.baseballmusings.com), where you may input a set of nine players and receive "the best" batting order with this set.

The above examples do not help to build a scientific foundation for solving a lineup problem and, unfortunately, one can find very few items in scientific literature on theoretical and/or practical approaches to doing it. This does not mean that there are no efforts of creating such methodology, however they are most likely limited to each organization's (team's) use.

As was mentioned in the introduction, there is an abundance of statistical data on baseball performance, which was collected over many years, since 1870s. Since 1920s, the log of each game was recorded, therefore allowing an analysis of historical baseball data and a comparison of players' performance across generations of players. Without such careful preservation of factual data and the categorizing of various occurrences during games, without the aid of the official scoring and statistical analyses of baseball games it would be impossible to approach any

research goal involving evaluations of players' and team's performance. The statistical data may be utilized and combined in various ways (see Albert and Bennett, 2003; Baseball Info Solutions and James, 2016; Gleeman and Sayre, 2017; Tanko, Lichtman, and Dolphin, 2015) while providing both professionals and baseball fans with foundational resources, whether it is scientific research or playing fantasy baseball.

While tremendous progress has been made over the years by baseball organizations and various sabermetrics entities in developing new, more advanced and refined, measurements evaluating players' and teams' performance (see section 2), there are barely any references on optimizing the set of starters in a baseball game. In Graham, 2012 the lineup optimization problem is bisected into two parts: the player selection model and the batting order optimization model. For the player selection model, the model aims at maximizing the number of runs scored while considering player-pitcher matchups, the playing field, and the set of referees (umpires). A genetic programming algorithm is used to find an optimal selection of the set of starting players, which seems to produce good results in an example provided in the paper. The entire approach to the player-pitcher matchup is based on a run production formula with the empirically obtained parameters. However, the model does not consider defensive qualities of the set of starting players and focuses exclusively on the probability of winning a game as a function of the number of runs scored by a team. While it is true that the probability of a team winning a game increases as the number of runs scored increases, it also depends on the number of runs allowed to be scored by the opposite team, which is entirely overlooked by the author.

In Polyashuk, 2005 selection of the set of starters was presented as an example of portfolio selection problem. Portfolio is defined as a set of objects that contribute to one or more common goals; it is then argued that quantitative criteria evaluating offensive and defensive characteristics of a team are more important than qualitative criteria such as presence of leadership, psychological health or compatibility, so these quantitative criteria must be considered for the initial selection of an optimal set of players.

There is a larger number of publications addressing optimization of the batting order, once the set of players has been determined (e. g. Clayton, 2012; Pankin, 2018; Sugrue and Mehrotra, 2007; Thaker, 2011; Tanko et al., 2015). Tanko et al. for instance, assert that the number of times per game a player comes to bat is decreasing as we go down in the batting order, so best offensive players should be set earlier in the batting order; on the other hand, the fourth player is coming to bat with the largest (on average) number of runners on bases, so putting the best hitter in the fourth position would seem to be beneficial to the run production. Therefore, the authors assert that the best players must bat at 1st (lead-off player), second, and fourth. Some of these and similar conclusions are contradictory, and there is no clear agreement on how the manager should approach the problem.

Pankin (2018) uses stochastic modeling and creates a Markov chain with 28 states representing the number of outs plus the set of bases occupied by a runner. When the next player in the batting order goes to bat, it is possible to consider probabilities of the Markov chain transitioning from the current state to other 27 states, thus creating the transition matrix for the chain. There is a dilemma on whether to consider a generic transition matrix for an average player or to create such matrices for each player in the batting order. The first one of these two approaches is easier

but falsely assumes that all players are identical (Thaker, 2011), while the second approach would jeopardize tractability of the model. A Markov chain approach is also explored in Bukiet et al. (1997) where a transition matrix contains probabilities of transitioning among 25 possible states. However, in this paper Markov chain is not used for determining an optimal batting order but to approximate probability of a given team winning a game against another team; the batting order itself is determined based on the ranking provided by the players' scoring index (D'Esopo and Lefkowitz, 1960), a version of expected number of runs produced by the player in one inning.

In Sugrue and Mehrotra (2007), the problem of an optimal batting order is stated as the longest cycle problem in a complete weighted digraph on 9 vertices with the weight of the branch from node i to node j indicating the probability l_{ij} of player i scoring a run if followed in the batting order by player j . The total weight of a Hamiltonian cycle represents the expected number of runs produced by a single rotation of the corresponding batting order. This approach allows the use of a simple and efficient way to find an optimal rotation of the batting order. In this paper, we capitalize on the results presented by Sugrue and Mehrotra to improve the formula for P_{ij} and use an integer programming model which finds optimal solutions to the batting order problem.

In this paper, we propose a multi-criteria approach to evaluating and comparing players and sets of players playing on the field in any given game. We will be using some of the measurements listed in the Appendix but will not limit ourselves to a single refined measure such as WAR (wins above replacement). Reasons for a multi-criteria approach are several. First, it is the desire to avoid reliance on empirical algorithms that are modified and adjusted every year to be more consistent with the players' actual impact on winning games for their teams. For instance, WAR formulas involve weighting coefficients, which vary from year to year and from source to source. Another reason for separating different aspects of players' contributions to winning games into criteria is that their defensive successes often depend on the position they play, so that an overall characteristic would not help the manager to make decisions on specific position assignments. Finally, such measure as WAR will not help the manager to evaluate potential strength of his starting player assignment separately in offensive and defensive aspects because, depending on the opposing team and its pitcher for a given game, the manager might wish to strengthen either offense or defense if it is not possible to strengthen both.

3. Position player selection as a bi-criteria problem

3.1. Preliminaries

As was discussed in the introduction, we have set the goals of finding solutions for optimizing the starting lineup in a baseball game. The team manager writes his starting lineup in the lineup card and hands it to the umpire before the game. It contains the batting order of nine players. These are the same players who will be also working in their defensive positions during the defensive portion of each inning. The only exception is the designated hitter (DH) in AL teams where the pitcher can (and practically always is) replaced by a DH. Therefore, when setting up a batting order, the manager must know which player will take each defensive position in the field. It is only logical that we will consider two parts of the lineup problem: one, selection of the 9-

player set for a starting defensive (fielding) position assignment, and two, deciding on the batting order.

The active roster of any MLB team consists of 25 players; usually, 12 or 13 of them are batters and the rest are pitchers. Pitching assignments are usually decided well ahead of a game because ordinarily a team has 5 “starting” pitchers who have rotating assignments and pitch every 5th game; usually such a pitcher plays at least 5 innings of the game. The remaining pitchers on the roster are “relief” pitchers and usually work during the late innings of the game, once the starting pitcher needs to be replaced. In other words, short of injuries, emergency situations, or scheduling changes, the pitching assignment is determined in advance. At this point, we will assume that we are solving the lineup selection problem for a manager of an AL team where the designated hitter rule is adopted; therefore, we will consider the player selection problem as the task of choosing 9 players from N non-pitching members of the team’s active roster that will be both in the starting defensive positions (except the DH) and in the starting lineup. There would be no principal difference between this case and that of a NL team, except that for a NL team, we would seek to optimize an 8-player selection instead of 9.

Once the problem has been specified, let us clarify it further. The manager has to select an assignment of 9 players (see Figure 2): one catcher (C), one first baseman (1B), one second baseman (2B), one shortstop (SS), one third baseman (3B), one left fielder (LF), one center fielder (CF), one right fielder (RF), and one designated hitter (DH). How many possibilities are there? Since these players must be selected from the set of non-pitching members on the team, and a team usually has 12 or 13 such players, it seems that the total number of selections is between $P(12,9) = 79,833,600$ and $P(13,9) = 259,459,200$, which is a difficult choice. These numbers represent the total number of alternatives in the decision space; however, the number of feasible alternatives is much smaller because the above calculations involving permutations assume that any player can play any position. It is not the case. The fact is, an overwhelming majority of players can play three or fewer defensive field positions, and, to understand this, one only must glance at the “depth chart” of any given team.

Table 1: The depth chart for the 2012 Chicago White Sox team

CATCHER	1ST BASE	2ND BASE	SHORTSTOP
1. Pierzynski	1. Konerko	1. Beckham	1. Ramirez
2. Flowers	2. Dunn	2. Olmedo	2. Olmedo
	3. Youkilis		
3RD BASE	LEFT FIELD	CENTER FIELD	RIGHT FIELD
1. Youkilis	1. Vicedo	1. Wise	1. Rios
2. Olmedo	2. Wise	2. Danks	2. Wise
	3. Danks	3. Rios	3. Danks

A depth chart is used to show the placements of the starting players and the secondary players for various positions. Usually, a starting player will be listed first while back-up players will be listed below. For example, based on the depth chart for the Chicago White Sox (CWS) as of August 31, 2012 (see Table 1), we can determine in how many ways it is possible to assign

infielders and outfielders to their starting positions. In section 2.3, it will be shown that the 2012 CWS had 720 feasible starting position assignments.

While the number of feasible starting assignments measures in hundreds and not in hundreds of thousands, assigning the starting players is not an easy choice, and therefore a “typical” manager usually goes with the first player on the list for each position. Those players are the first-choice selections who are considered regular starters and play every game, unless the manager decides that they need rest. Most of the time, the team fans perceive the choice as a good one. However, the approach based on predetermined order on the depth chart does not consider recent player performance, since such a chart in many cases is constructed even before the season begins.

Let us go over some examples of choices made by Chicago White Sox managers⁷. In 2011, CWS manager Ozzie Guillen used Alex Rios as the center fielder (he was number one on the depth chart for that position). In 2012, the current manager Robin Ventura moved Rios to the right field, and both the team’s defense and offense benefited from this player’s success. (for example, Rios’ *DRS* (defensive runs saved) statistic improved from -9 to 7, and his *WRC* (weighted runs created) improved from 42 to 87). Does this mean that 2011 depth chart was faulty, which led to a continual misuse of a valuable player? Is it entirely possible that among hundreds of possibilities there are several that can compete or be better than the general rule of following the depth chart? To find out answers to these questions, it is necessary to formulate the problem of the starting lineup selection as a multiple criteria problem.

3.2. Problem formulation

When formulating a multiple criteria problem, we must first determine the decision space and the criteria space. A 9-player starting assignment can be presented as a $9 \times p$ matrix where a row represents a selected starter player with his characteristics, and the length of a row, p , is the number of the characteristics of a single player that fully describe him as a team member. Such a row has fields that may be labeled by the player’s name, weight, height, age, salary, batting/throwing sides, the number of games in which the player was a starter and innings played in each position, as well as his offensive and defensive numbers in the positions played. This approach is consistent with the formulation of portfolio selection problem with several criteria in Polyashuk (2005). Let us assume that the first row corresponds to the starting catcher (C), the second – to the first baseman (1B), the third – to the second baseman (2B), and so on, according to the standard defensive position numbers chart; let the ninth row correspond to the selected starting designated hitter (DH). The set of all possible starting selections, which forms decision space, is the set of all possible $9 \times p$ matrices.

To form criteria space for the starting assignment problem, let us consider what the manager is trying to achieve when solving it. The goal, of course, is to win the game, and to reach this goal the manager has to put together a set of players that has the potential to score most runs while allowing the least runs to the opposing team. At the same time, since a regular baseball season lasts 6 months and consists of 162 games, the manager must use his team’s resources wisely, which means maintaining the team’s overall physical and psychological health. The goal of the current research effort is to produce a set of optimal starting assignments which would maximize both offensive and defensive potential of the starting set of players while assuming that the team

manager can make the final selection of the starting assignment based on the overall players' physical/psychological condition and other conditions.

The next step is to use various characteristics of individual players in order to form two criteria, which would adequately represent offensive potential and defensive quality of a set of players. Philosophically, we will rely on Pythagorean Expectation formula, a composite feature of team performance that has been extensively tested by baseball statisticians, which allows using offensive and defensive team numbers to approximate winning percentage of a team and, ultimately, predicting the outcome of the regular season very accurately (for instance, in 2017 only 3 out of 30 teams were outside a 4% error in prediction of the number of wins by this formula). Pythagorean Expectation Formula allows approximating the winning percentage for a

team in a season as follows: $\frac{RS^2}{RS^2 + RA^2} = \frac{1}{1 + (RA/RS)^2}$. More recently, the exponent used in

the formula was adjusted and is currently set as 1.83 (Baseball Reference.com), which gives a better fit with the empirical data: $\frac{1}{1 + (RA/RS)^{1.83}}$. As it is clear from the formula, the winning

percentage depends on the ratio of RA (the total number of runs allowed or runs scored by opposite teams) and RS (the total number of runs scored by the team).

The formula above is a key to our approach to forming offensive and defensive criteria for a set of starters in any given game, which is effectively based on composite characteristics of individual players reflecting their offensive potential and defensive value. More specifically, since winning a baseball game depends on the number of runs scored versus the number of runs allowed, let us use runs created (RC) by a player to quantify the player's overall offensive performance, and defensive runs saved (DRS) by a player to quantify the player's defensive performance in a given fielding position. Therefore, we will assume that offensive potential of a single player in a lineup is defined by the number of runs created by a player per game,

$RC / G = \frac{RC \cdot 27}{AB - H + SH + SF + CS + GDP}$. Further, let us assume that the defensive quality of

single player in each fielding position is characterized by defensive runs saved above average, DRS , as defined by Baseball Info Solutions (BIS). Note that in 2012 the most recent values of both RC and DRS were readily available on many websites keeping statistical data on baseball players' and teams' performance, such as BaseballReference.com where we took the data for our examples.

Definition 3.1. Let the set $T = \{p_1, p_2, \dots, p_{25}\}$ be the team roster. Suppose a permutation of players $S = (p_{j_1}, p_{j_2}, \dots, p_{j_9})$ is a starting assignment of players consisting of a catcher, a first baseman, a second baseman, a shortstop, a third baseman, a left fielder, a center fielder, a right fielder, and a designated hitter, in this order. Suppose the i -th player in the assignment S , p_{j_i} , has offensive potential (RC/G) equal to r_i and defensive quality (DRS) equal to d_i , $i = 1, \dots, 9$. Then the overall offensive quality of the 9-player starting assignment S is defined as

$R_s = \frac{1}{9} \sum_{i=1}^9 r_i$. The overall defensive quality of the starting assignment S is defined as $D_s = \sum_{i=1}^9 d_i$.

The intent of the above definition is to provide the criteria mapping, which assigns to every alternative in the decision space (in our case, to every feasible starting assignment) the corresponding vector of criteria values in the criteria space. Further, we will consider a starting player assignment optimal if it provides maximum simultaneously for R_s and D_s . The concept of simultaneous maximization of two criteria values will be interpreted in this paper in the context of the classical definition of Pareto binary relation, which defines a preference binary relation on the set of possible starting assignments.

Definition 3.2. Starting assignment S_A is less preferred than starting assignment S_B , which is expressed as $S_A \prec S_B$, if and only if either $R_{S_A} \leq R_{S_B}$ and $D_{S_A} < D_{S_B}$, or $R_{S_A} < R_{S_B}$ and $D_{S_A} \leq D_{S_B}$.

Definition 3.3. Let A be the set of all feasible 9-player starting assignments and $X = \{S_1, S_2, \dots, S_N\}$, $X \subseteq A$, is the set of the feasible starting assignments for a given game. A starting position assignment is optimal with respect to both offensive potential and defensive quality criteria if and only if this assignment is non-dominated with respect to the binary relation \prec ; the set of all such assignments is identified as $MAX(X)$:

$$S^* \in MAX(X) \Leftrightarrow \neg \exists S \in X \text{ such that } S^* \prec S.$$

In other words, we are using a traditional bi-criteria Pareto model (Sen, 1970) to choose the set of optimal starting player assignments as the set of efficient points in our bi-criteria space. Isolating the set of Pareto-optimal (efficient) points should substantially reduce the total number of possible alternatives which should allow the manager to make the final choice confidently. In any case, the greatest benefit from applying this model would be the elimination of dominated selections, which could become quite an eye opener. Let us now follow up with the example of starting 9-player selection for the 2012 Chicago White Sox.

3.3. Finding optimal starting player assignments: 2012 Chicago White Sox

Figure 3 below shows the decision tree for selecting the starters for infield positions (1B, 2B, SS, and 3B) starting with the choice of a first baseman (1B), to a second baseman (2B), to a shortstop (SS), and a third baseman (3B) while Figure 4 represents the decision tree for choosing outfielders: left fielder (LF), center fielder (CF), and right fielder (RF). These charts confirm that CWS can assign infield starters 9 different ways and outfield starters – 10 ways. Because there were 2 catchers in the team and any remaining player can be assigned as the designated hitter (DH) for a game (in the case of 2012 CWS team, four players would be available for the DH position), the number of CWS manager's choices of the starting position assignment is limited by the following number: $n = 2 \cdot 9 \cdot 10 \cdot 4 = 720$. This is the number of feasible solutions in the problem of finding optimal starting player assignments for the 2012 CWS team.

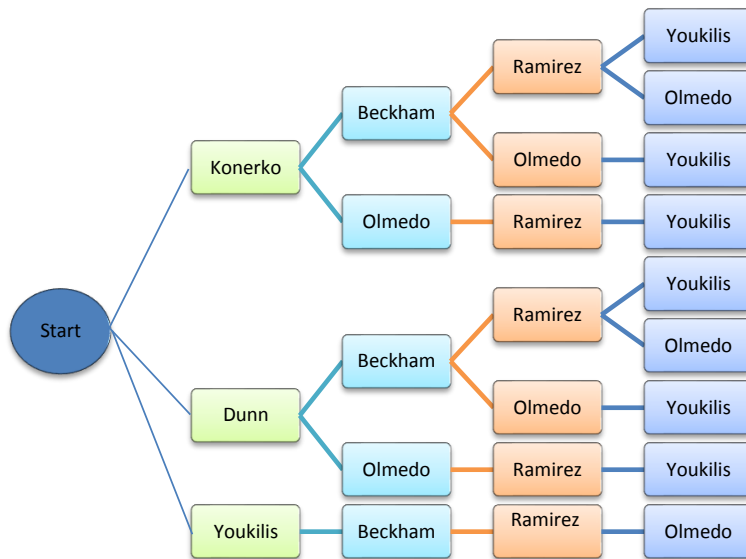


Figure 3

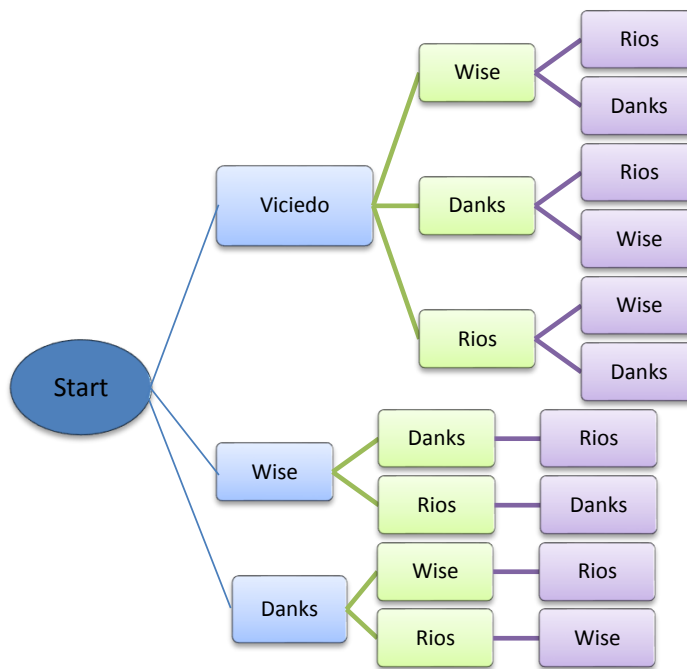


Figure 4

After the set of feasible position assignments has been determined, we proceed to create criteria values by measuring offensive potential and defensive quality for each position player (Table 2).

The values in the table (*RC* and *DRS*) were taken from BaseballReference.com website in August of 2012. Note that INF stands for infielder, OF for outfielder, and the last column contains different *DRS* values in case the player took on different positions on the field during 2012 season.

Once we have individual players' criteria values, we combine them to form the values for each infield and outfield assignments of starting players; these numbers are shown in Table 3 (for all 10 infield assignments), and Table 4 (for all 9 outfield assignments). Please notice that we included players' handedness in the Table 2 as well as combined numbers for the total number of right-, left- and switch batters (who can bat from either side) in Tables 3 and 4, since players' handedness is an important consideration when addressing the so called "platoon advantage": playing more righties against left-handed pitchers and more lefties against right-handers (see section 3.4).

Table 2: Individual players' stats

#	Players	Position	Offense: <i>RC/G</i>	Bats	Defense: <i>DRS</i>
1	Beckam	INF	3.3	Right	2B -4
2	Danks	OF	3.0	Left	LF 1, CF 1, RF 0
3	Dunn	1B	5.8	Left	1B -4
4	Flowers	C	4.0	Right	2
5	Konerko	1B	6.9	Right	-6
6	Olmedo	INF	2.1	Switch	2B 1, SS 0, 3B 0
7	Pierzynski	C	6.3	Left	-9
8	Ramirez	INF	3.5	Right	SS 10
9	Rios	OF	5.6	Right	RF 5, CF -9
10	Viciedo	OF	3.8	Right	LF -1
11	Wise	OF	5.3	Left	LF -1, CF -2
12	Youkilis	INF	6.0	Right	1B 0, 3B 2

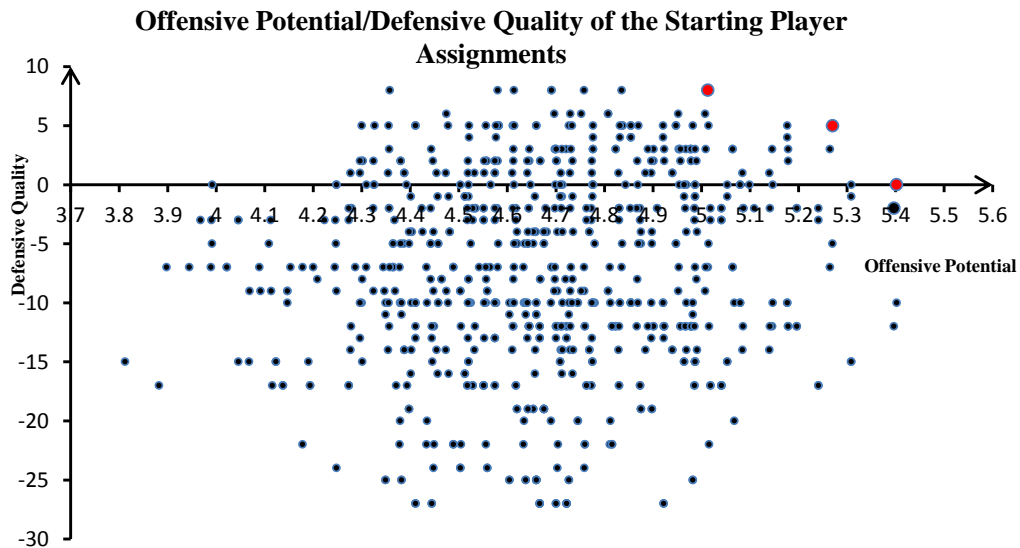
Table 3: Stats for all possible infield assignments

#	Infield player assignment 1B, 2B, SS, 3B	Offense: Average <i>RC/G</i>	The number of Right/Left/Switch Batters	Defense: total <i>DRS</i>
1	Konerko, Beckham, Ramirez, Youkilis	5.45	4/0/0	2
2	Konerko, Beckham, Ramirez, Olmedo	4.48	3/0/1	0
3	Konerko, Beckham, Olmedo, Youkilis	4.58	3/0/1	-8
4	Konerko, Olmedo, Ramirez, Youkilis	5.15	3/0/1	7
5	Dunn, Beckham, Ramirez, Youkilis	5.18	3/1/0	4
6	Dunn, Beckham, Ramirez, Olmedo	4.20	2/1/1	2
7	Dunn, Beckham, Olmedo, Youkilis	4.43	2/1/1	-6
8	Dunn, Olmedo, Ramirez, Youkilis	4.88	2/1/1	9
9	Youkilis, Beckham, Ramirez, Olmedo	4.25	3/0/1	6

Table 4: Stats for all possible outfield assignments

#	Outfield player assignment LF, CF, RF	Offense: average RC/G	The number of Right/Left/Switch Batters	Defense: total DRS
1	Viciedo, Wise, Rios	4.90	2/1/0	5
2	Viciedo, Wise, Danks	3.03	1/2/0	0
3	Viciedo, Danks, Rios	4.13	2/1/0	8
4	Viciedo, Danks, Wise	4.03	1/2/0	5
5	Viciedo, Rios, Wise	4.90	2/1/0	-5
6	Viciedo, Rios, Danks	4.13	2/1/0	-7
7	Wise, Danks, Rios	4.63	1/2/0	5
8	Wise, Rios, Danks	4.63	1/2/0	-10
9	Danks, Wise, Rios	4.63	1/2/0	4
10	Danks, Rios, Wise	4.63	1/2/0	-10

After the statistical data have been organized, we can proceed with analyzing bi-criteria evaluations for all 720 possible starting player assignments. We organized the 720-line table by a simple Excel program. The same program allows producing a graph of this two-dimensional data as presented in Figure 5 below, which is the scatter diagram of these evaluations in criteria space. As can be seen from Figure 5, our example produces just three optimal (efficient) points in the bi-criteria space (shown in red color), and the same Excel file allows to easily identify the exact three starting player assignments that correspond to them.



It appears that the optimal points and the corresponding assignments are (left to right on the graph): **Optimal solution 1**, (5.01,8): infield assignment #8, outfield assignment #3, catcher

Pierzynski, and DH Konerko; **Optimal solution 2**, (5.27,5): infield assignment #8, outfield assignment #1, catcher Pierzynski, and DH Konerko; and **Optimal solution 3**, (5.40,0): infield assignment #5, outfield assignment #1, catcher Pierzynski, and DH Konerko. These optimal solutions would be presented to the manager before the game in a form like the table below (Table 5), together with the offensive and defensive criteria values, so that the manager could decide on the most suitable starting assignment.

Table 5: Optimal Solutions

Options Presented to the Manager	Starting Position Assignment: 1B, 2B, SS, 3B, LF, CF, RF, C, DH	OFFENSE (RC)	DEFENSE (DRS)
Optimal Solution 1	Dunn, Olmedo, Ramirez, Youkilis, Vicedo, Danks, Rios, Pierzynski, Konerko	5.01	8
Optimal Solution 2	Dunn, Olmedo, Ramirez, Youkilis, Vicedo, Wise, Rios, Pierzynski, Konerko	5.27	5
Optimal Solution 3	Dunn, Beckham, Ramirez, Youkilis, Vicedo, Wise, Rios, Pierzynski, Konerko	5.40	0

It is interesting to check how often these starting assignments were used by Robin Ventura, the 2012 CWS manager. Per Baseball-Reference.com website, out of 90 games since having acquired Youkilis, the first and the second sets were never played while the third set of starters was played exactly eight times. Certainly, the team roster changes many times during a season and several players exit and enter the active roster; however, this does not fully explain why the 2012 CWS manager did not use what we have found to be optimal starting assignments because the players considered in this paper were all available over a prolonged period. Notably, there existed another set of starters which was similar to one of the optimal solutions and which was used almost equally often (6 times), for which our model produced the same offensive number but a lower defensive number. This assignment is identical to our Optimal solution 3, except that two players have traded their positions: Adam Dunn was in the DH position while Paul Konerko was in the 1B position, rather than the opposite assignment in our optimal solution (the corresponding point is shown in black on our scatter diagram). Although this point is located close to the efficient set, from the defense perspective Dunn playing first base should have been preferable to Konerko. Nevertheless, CWS won 9 out these 14 games (or 64%), which is a much higher percentage of wins than in the overall season (85 wins out of 162 games, or 52%).

We can also observe that two of three optimal points correspond to starting assignments in which Olmedo, rather than the regular starter Beckham, is assigned to play the second base. This is also true for several close points: (5.27,3), (5.18,5), and (5.18,3). In retrospect, 2012 was the best year for Robin Ventura as the manager of Chicago White Sox, and we suggest it is because he used many assignments that were either optimal or close to optimal. Since the players' statistics change from day to day, without making far gone conclusions about effectiveness of any manager, we can still say that had the knowledge of the optimal solutions been available to a manager before a game, he would be able to make better informed decisions on the starting player assignments.

3.4. Applying players' handedness when choosing an optimal starting assignment

It turns out that baseball statistical data (see, for example, Baseball Info Solutions and James B., 2016) overwhelmingly show that a right-handed pitcher (the player who throws the ball to the batter) generally has a disadvantage against a left-handed batter, and a left-handed pitcher – against a right-handed batter. Since the name of the starting pitcher for the opposite team is announced in advance, it usually calls for the manager to utilize so called “platoon advantage” (Baseball Info Solutions, 2016) and use more righties against left-handed pitchers and more lefties against right-handers in the lineup.

Two figures below display scatter diagrams with points corresponding to 2012 CWS starting player assignments with either maximum possible number of players batting right (Figure 6) or with maximum possible number of players batting left (Figure 7). For the 2012 CWS team, these numbers were eight and five, respectively. Platoon advantage percentage (Baseball Info Solutions (BIS) and Bill James, 2011) is the percentage of players in the lineup who can either bat right against a left-handed starting pitcher of the opposing team or bat left against a right-handed pitcher. Ideally, a manager should try to maximize platoon advantage percentage.

If we apply this concept to the 2012 CWS team, we should select only such starting assignments that maximize the platoon advantage: due to the team roster at the time, the manager could use at most 8 hitters batting from the right and at most 5 hitters batting left-handed. It appears that in 2012 CWS there were just 20 player assignments with 8 right-handers, while there were 132 options for a starting assignment with 5 left-handed hitters. Applying our bi-criteria model allows us to identify five optimal solutions in the case of a right-handed assignment and three optimal solutions among left-handed assignments. (The corresponding points on the scatter plots are marked red.)

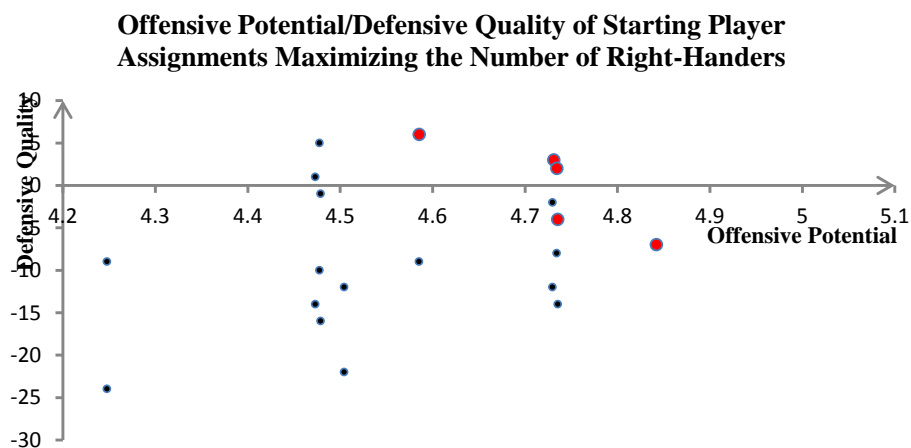


Figure 6

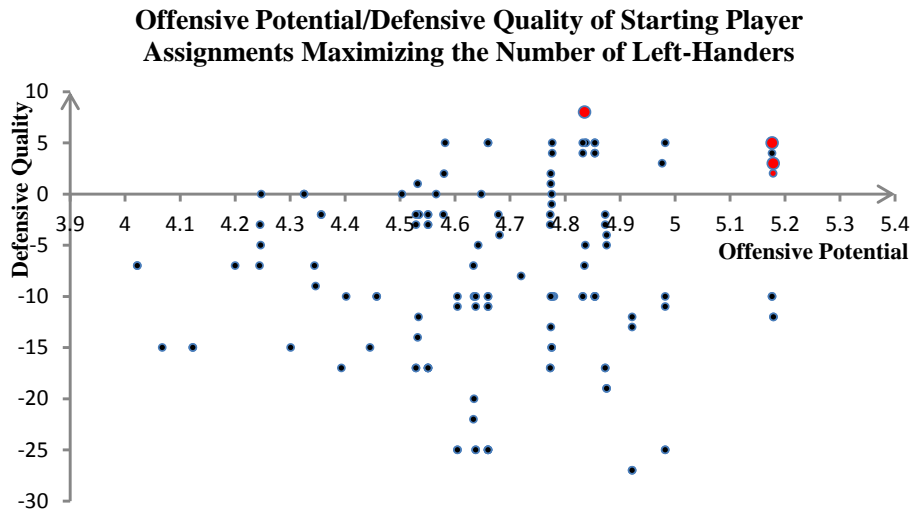


Figure 7

When comparing actual decisions with those suggested by our model, we can observe mixed results: in some respects, our model validates decisions made by the CWS manager but in other respects the results it provides are surprisingly different from the actual practice of choosing the set of starters. To analyze these similarities and differences, let us first list optimal solutions in each case. In the case the team is playing against a left-handed starting pitcher with the largest possible number of right-handed batters, we have 5 optimal solutions; in the case when the team tries to maximize the number of left-handers (when playing against a right-handed pitcher), our model detects 3 optimal solutions. We summarize these cases in tables 6 and 7 below.

Table 6: Maximizing the number of right-handed batters

Options Presented to the Manager	Starting Position Assignment: 1B, 2B, SS, 3B, LF, CF, RF, C, DH	OFFENSE (RC)	DEFENSE (DRS)
Optimal Solution 1	Konerko, Olmedo, Ramirez, Youkilis, Vicedo, Rios, Danks, Flowers, Beckham	4.47	6
Optimal Solution 2	Youkilis, Beckham, Ramirez, Olmedo, Vicedo, Rios, Danks, Flowers, Konerko	4.48	5
Optimal Solution 3	Konerko, Olmedo, Ramirez, Youkilis, Vicedo, Wise, Rios, Flowers, Konerko	4.731	3
Optimal Solution 4	Youkilis, Beckham, Ramirez, Olmedo, Vicedo, Wise, Rios, Flowers, Konerko	4.734	2
Optimal Solution 5	Konerko, Beckham, Ramirez, Olmedo, Vicedo, Wise, Rios, Flowers, Youkilis	4.74	-4

Table 7: Maximizing the number of left-handed batters

Options Presented to the Manager	Starting Position Assignment: 1B, 2B, SS, 3B, LF, CF, RF, C, DH	OFFENSE (RC)	DEFENSE (DRS)
Optimal Solution 1	Dunn, Olmedo, Ramirez, Youkilis, Vicedo, Danks, Rios, Pierzynski, Wise	4.84	8
Optimal Solution 2	Dunn, Olmedo, Ramirez, Youkilis, Wise, Danks, Rios, Pierzynski, Konerko	5.18	5
Optimal Solution 3	Konerko, Olmedo, Ramirez, Youkilis, Wise, Danks, Rios, Pierzynski, Dunn	5.19	3

Let us first go over several striking similarities between solutions proposed by our model and actual starting assignments for 2012 CWS. According to our model, Alexei Ramirez should have played shortstop in every one of the eight (combined for both right- and left-handed starting selections) optimal solutions, and indeed, he was a regular shortstop starter on the CWS team. Our model's optimal solutions suggest that Adam Dunn and A. J. Pierzynski, should have played in all games against right-handed pitchers; Dayan Vicedo should have started in the left field against left-handers, Alex Rios should have always started in the right field (with one exception of our model's suggestion); all starting assignments should have use Kevin Youkilis, and Paul Konerko should have played almost always (in seven out of eight optimal selections). All the above were regular practices in 2012 CWS team, which indicates that our model supports these decisions.

On the other hand, our model differs from other typical practices of the CWS starting selections. For example, infielder Ray Olmedo appears to be a member of every one of the eight optimal starting assignments. This makes sense due to his versatility as a position player (he could play any infield position but first base) and as a hitter (he was the only switch hitter on the team which means that he could bat from the right or left side). However, Olmedo mostly served as a substitute player and did not get a lot of starts for the team. Another observation: the team's right-handed catcher Tyler Flowers, according to our model's choices, should have been used in every game against left-handed pitchers, and second baseman Gordon Beckham should not have been played against right-handers. In contrast, Flowers was on the sitting out in many such games, and Beckham played second base in all but 9 out of first 131 games of the 2012 season.

Is it possible to conclude that Robin Ventura, the Chicago White Sox manager, had made some unjustified decisions in selecting starters? In the opinion of one who believes in strictly following the "law of large numbers" of statistical data and who can shake off prejudices of conventional wisdom, the answer is "yes". This indicates that the manager would have had been better off had he followed our model's recommendations.

4. Choosing the best batting order for a given set of starting hitters

4.1. Preliminaries

Once the manager has determined the set of players for the game, he must decide on the batting order. Unlike the case of player selection problem, this decision is solely based on the offensive

qualities of the lineup. Clearly the batting order is important, and every fan can observe its multiple variations over the course of a single season. Some managers like to change their batting order often; others prefer to keep the same order across different games for as long as possible. The maximum number of different lineups that a manager can use in a season is 162, the total number of games in a season. For example, Tony LaRussa used 153 when managing 2008 Cardinals while Charlie Manuel used 68 for the Phillies in 2009. But no manager can try all possible lineups because even if we assume that the same set of players is selected for every game, there are $9! = 362,880$ different batting orders. So how can managers come up with their choices when the number of possibilities is that large?

As was discussed in the literature review, this issue may be approached several ways: using empirical approach and utilizing statistical data to make recommendations on the batting order as it relates to players' hitting and running abilities; applying a stochastic model which would simulate an actual game based on empirical probabilities of transitioning from one state of the stochastic process to another; or using graph theory approach to represent a batting rotation. The latter was, as we have discussed in the literature review, the basis for Sugrue and Mehrotra (2007) paper, which we have adopted in principle for our approach to building an optimal batting order.

In the problem of finding an optimal batting order, unlike the problem of choosing the set of starters, we are not concerned with the defensive quality of the players; rather we focus exclusively on the offense. The idea is to create an ordering of the given set of nine players, which would produce the largest possible number of runs in the course of the game. Therefore, problem formulation depends on how run production is viewed and how a criterion for the number of runs produced is developed.

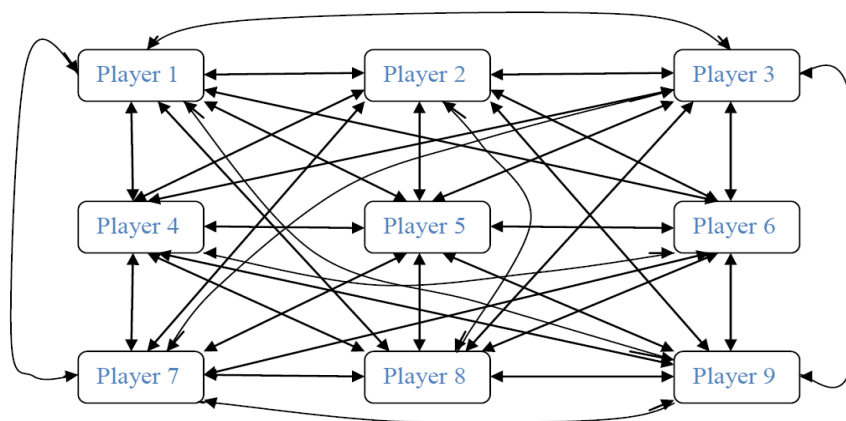


Figure 8

Let us consider a graph on nine vertices where each vertex represents a player participating in the batting rotation and each directed edge from player i to player j represents the immediate precedence of player i to player j in a batting order. In the game of baseball, during the offensive parts of the innings, the players come to bat in the order fixed by the team manager in advance;

once all nine players have taken their turns, the first player comes to bat, then the second and so on, until the game ends. This process justifies the term of batting order rotation, or simply rotation, in a baseball game. On the corresponding graph, such a rotation is represented by a Hamiltonian cycle, since each vertex is traversed exactly once.

Next, suppose we have a batting order (lineup) $L = (i_1, i_2, \dots, i_9)$ where $i_k \in \{1, 2, \dots, 9\}$, $\forall k$. Let

us now introduce the set of random variables $X_{i_k} = \begin{cases} 1, & \text{if player } i_k \text{ scores a run} \\ 0, & \text{otherwise} \end{cases}$,

$\forall k \in \{1, 2, \dots, 9\}$; then $X = \sum_{k=1}^9 X_{i_k}$ represents the total number of runs scored by one rotation of the batting order. The expected number of runs produced by one rotation of lineup L is

$E_L = E(X) = \sum_{k=1}^9 E(X_{i_k})$. Suppose that probability $P(\text{Player } i_k \text{ scores a run})$ does not depend

on the previous chain of events in the game; then the expected value

$E(X_{i_k}) = P_{i_k i_{k+1}} = P(\text{player } i_k \text{ scores a run} \mid \text{he is followed by player } i_{k+1})$; it follows that the

expected number of runs from one rotation of the batting order is the sum of these conditional

probabilities: $E(X) = \sum_{k=1}^9 P_{i_k i_{k+1}}$. Based on this approach, we may define the weight for the

directed edge from Player i to Player j in the graph in Figure 9 equal to P_{ij} , the probability that Player i scores a run if he is followed by Player j ; then the problem of maximizing the expected number of runs produced by one rotation of the batting order becomes the problem of finding a Hamiltonian cycle with the highest weight.

To develop a formula for P_{ij} , which would estimate of the probability that Player i scores a run if he is followed by Player j in the lineup, let us consider a tree of possibilities for Player i scoring a run. The tree in Figure 9 contains 16 paths resulting in a run; these paths end in a green-colored node. For example, Player i can hit a home run (HR), or he can get a walk or is hit by a pitch (HBP) followed by Player j hitting a triple; and so on. Note that the tree in Figure 9 contains possibilities resulting from the action of the “next player” (the player who follows Player j in the lineup) as well as running abilities of Player i . Since we want to make P_{ij} a function of the

ordered pair of players (i, j) and it should not depend on the identity of the next player, we will use assume that this player is a generic player from the team. Clearly, the tree somewhat simplifies and reduces the set of real-life possibilities. One omission is related to the fact that in a baseball game a player may get on base and/or advance because of errors committed by the opposite team; another omission has to do with “the player after the next player” who, in real life, also affects chances of Player i scoring a run. However, these omissions are justified by desire to not overcomplicate computations while adding negligible amounts to probabilities P_{ij} .



Figure 9

Let us agree on the following notations: for all $1 \leq k \leq 9$,
 $h(k)$ denotes the probability that Player k hits a home run (HR);
 $t(k)$ denotes the probability that Player k hits a triple;
 $d(k)$ denotes the probability that Player k hits a double;
 $s(k)$ denotes the probability that Player k hits a single;
 $b(k)$ denotes the probability that Player k gets a walk (BB) or is hit by a pitch (HBP);
 $r1(k)$ denotes the probability that Player k scores a run from the second base on a single;

$r2(k)$ denotes the probability that Player k scores a run from the first base on a double;

p_h denotes the probability that a generic (average) team player has a hit;

p_e denotes the probability that a generic (average) team player hits for extra bases (a home run, a triple or a double).

Based on the tree diagram above (Figure 9), we may introduce the following definition.

Definition 4.1. We will define P_{ij} , the probability that Player i scores a run if followed by Player j (where $1 \leq i \leq 9$, $1 \leq j \leq 9$, and $i \neq j$) as follows:

$$\begin{aligned} P_{ij} = & h(i) + t(i) \left(h(j) + t(j) + d(j) + s(j) + (1 - h(j) - t(j) - d(j) - s(j)) p_h \right) + \\ & + d(i) \left(h(j) + t(j) + d(j) + s(j) \left(r2(i) + (1 - r2(i)) p_e \right) \right) + \\ & + (s(i) + b(i)) \left(h(j) + t(j) + d(j) \left(r1(i) + (1 - r1(i)) p_h \right) + (s(j) + b(j)) p_e \right). \end{aligned}$$

Definition 4.2. We define the probability that Player i scores a run if followed by himself $P_{ii} = 0$ for all $1 \leq i \leq 9$, since no player can follow himself in a batting order.

Sugrue and Mehrota admitted that the formula developed by them is a lower bound for the actual probability that Player i scores a run if followed by Player j because they “do not consider the offensive capability from a chain of players preceding” Player i (Sugrue and Mehrota, 2007). In fact, these authors also did not consider the chain of players succeeding Player j , nor the running capabilities of Player i . By taking into account six additional possibilities of scoring (compared to Sugrue and Mehrotra’s formula), our definition of P_{ij} improves the estimate, while still being a lower bound for the actual probability of scoring a run.

4.2. Problem formulation

Let us assume that the problem of selecting a optimal set of players in their starting positions has been solved, and such a set of nine players has been determined. Then the problem of finding an optimal batting order has the decision space which consists of the set Y of all $9! = 362,880$

permutations of the given set of players: $Y = \{L_1, L_2, \dots, L_{9!}\}$. The criteria space must be defined by criteria which evaluate offensive capabilities of this set of players as functions of the order in which they are arranged in the lineup. The expected number of runs scored by a single rotation of the batting order, E_L , is such a criterion. Along with E_L , we will also consider

$E_{LF} = \sum_{k=1}^3 P_{i_k i_{k+1}}$, the expected number of runs scored by the first three players in the lineup. The

reasons for adding this criterion are multiple. The first three players have a higher impact on the team’s offense because they get more plate appearances than players at the end of the batting order (number one on the average gets roughly an extra at-bat compared to number 9). Further, statistically the first inning is the most productive inning of the game (Peterson, 2011), and the first three players in the lineup always play in the first inning. Finally, scoring runs earlier in the game has a positive psychological effect on the team. Since the overall output of the entire

lineup is more important than that of the first three players, E_{LF} will be adopted as a secondary criterion in our bi-criteria model.

Definition 4.3. For any $L \in Y$, consider the expected number of runs scored by a single rotation of lineup $L = (i_1, i_2, \dots, i_9)$, $E_L = \sum_{k=1}^8 P_{i_k i_{k+1}} + P_{i_9 i_1}$, and the expected number of runs scored by the

first three hitters in the rotation, $E_{LF} = \sum_{k=1}^3 P_{i_k i_{k+1}}$. Let $E^* = \text{Max}\{E_{L_1}, E_{L_2}, \dots, E_{L_{9!}}\}$ and

$E_F^* = \text{Max}\{E_{LF} | E_L = E^*\}$. Then a lineup $L \in Y$ is optimal if and only if $E_L = E^*$ and $E_{LF} = E_F^*$.

Definition 4.3 essentially describes a bi-criteria model with lexicographic-ordered criteria, where the first (and most important) criterion is the expected number of runs generated by one rotation of the batting order and the second criterion is the expected number of runs generated by the first three players in the rotation. The second criterion is also important because, by definition, E_L is invariant with respect to any circular permutation of the lineup L , so that there are at least nine lineups with the same optimal value of E^* . Finding $E^* = \text{Max}\{E_{L_1}, E_{L_2}, \dots, E_{L_{9!}}\}$ can be achieved by solving an integer programming problem resulting from the problem of finding the longest Hamiltonian cycle in a directed graph on nine vertices with the vertices corresponding to 9 players in the starting lineup (Figure 8) and the weight of a directed edge (i, j) equal to P_{ij} .

Binary variables x_{ij} are defined as follows:

$$x_{ij} = \begin{cases} 1, & \text{if player } i \text{ is followed by player } j \\ 0, & \text{otherwise} \end{cases}.$$

Then the total weight of a Hamiltonian cycle is given by $E = \sum_{i=1}^9 \sum_{j=1}^9 x_{ij} P_{ij}$, and we come up with the integer programming problem below.

$$\text{Maximize } E = \sum_{i=1}^9 \sum_{j=1}^9 x_{ij} P_{ij}$$

$$\text{Subject to: } \begin{cases} \sum_{i=1}^9 \sum_{j=1}^9 x_{ij} = 9 & (1) \\ \sum_{i=1}^9 x_{ij} = 1, \text{ for all } j = 1, 2, \dots, 9 & (2) \\ \sum_{j=1}^9 x_{ij} = 1, \text{ for all } i = 1, 2, \dots, 9 & (3) \\ \sum_{i \in S} \sum_{j \in S} x_{ij} \leq |S| - 1, \forall S \subset \{1, 2, \dots, 9\} & (4) \\ x_{ij} \text{ binary} & (5) \end{cases}$$

Constraints (1), (2), and (3) ensure that the cycle contains all nine vertices; constraint (5) ensures that there are no shorter cycles included.

This problem can be solved by enumeration, which is efficient for not very large sets of feasible solutions, which is our case. Further, for each solution with $E = E^*$ we can select the optimal batting orders as those maximizing the expected output from the first three batters in the rotation.

4.3. Finding an optimal batting order: 2012 Chicago White Sox

Let us return to the example of 2012 CWS team and consider one of the most frequently used starting assignments. The following batting order was used by CWS over 30 times during 2012 season: Wise – Youkilis – Dunn – Konerko – Rios – Pierzynski – Vicedo – Ramirez – Beckham (when Wise was on placed on disabled list and could not play, he was replaced by Alejandro De Aza, whose numbers were extremely similar, while keeping the rest of the order the same). Since the statistical data on the number of plate appearances, home runs, triples, doubles, etc. by each player is always readily available on websites such as BaseballReference.com or Fangraphics.com, all the probabilities that we use in formula for P_{ij} (Definition 5.1) can be obtained as relative frequencies of these events. For example, in the case of the ordered pair Konerko-Rios, we can use the following Table 8:

Table 8: Konerko-Rios P_{ij} computation

Player	$h(i)$	$t(i)$	$d(i)$	$s(i)$	$b(i)$	$r(i)$
Konerko	0.043	0	0.0379	0.1842	0.105	0.17
Rios	0.0398	0.0127	0.0589	0.1736	0.0462	0.42

Table 9 contains probabilities P_{ij} for the set of players used in the following lineup: Wise – Youkilis – Dunn – Konerko – Rios – Pierzynski – Vicedo – Ramirez – Beckham; this batting order generates the values of $E_L = 0.6813$ and $E_{LF} = 0.2533$. The situation improves if, while keeping the same cycle of players, we change the first (lead-off) player. Clearly, $E_L = 0.6813$ remains the same, however the value of E_{LF} will depend who is the lead-off batter. For example, leading with Youkilis (which means that the lineup is Youkilis – Dunn – Konerko – Rios – Pierzynski – Vicedo – Ramirez – Beckham – Wise) would produce $E_{LF} = 0.2737$, and leading with Dunn (in other words, if the lineup is as follows: Dunn – Konerko – Rios – Pierzynski – Vicedo – Ramirez – Beckham – Wise – Youkilis) would give us $E_{LF} = 0.2705$. It is easy to check all nine possible lead-off batters for this cycle to see that starting with Youkilis corresponds to the best E_{LF} while starting with Dunn – the second best E_{LF} for the cycle. Since the first three players in the lineup have roughly an extra one plate appearance in a game than those at the end of the lineup, the change in the lead-off spot could affect the total number of runs scored over the course of a season, which is a major factor in the number of wins in a given season.

Regarding various batting orders with the same starting assignment, it is possible to do a quick check of the expected number of runs E_L and compare it with the batting order mentioned

above. A much better batting order with the same set of starters would be as follows: Dunn – Pierzynski – Ramirez – Konerko – Rios – De Aza – Vicedo – Beckham – Youkilis. This order

Table 9: Probability that Player i scores a run if followed by Player j in the case of one starting assignment

P_{ij}	<i>Followed by Player j</i>								
<i>Player i</i>	Beckham	Dunn	Konerko	Pierzynski	Ramirez	Rios	Vicedo	Wise	Youkilis
Beckham	0	0.0641	0.0646	0.0650	0.0562	0.0664	0.0598	0.0593	0.0621
Dunn	0.0930	0	0.1011	0.1016	0.0923	0.1021	0.0963	0.0956	0.0992
Konerko	0.0762	0.0810	0	0.0869	0.0757	0.0866	0.0806	0.0794	0.0842
Pierzynski	0.0820	0.0892	0.0903	0	0.0819	0.0914	0.0854	0.0848	0.0877
Ramirez	0.0455	0.0527	0.0892	0.0903	0	0.0558	0.0484	0.0480	0.0504
Rios	0.0735	0.0798	0.0824	0.0828	0.0738	0	0.0770	0.0767	0.0784
Vicedo	0.0687	0.0759	0.0765	0.0769	0.0684	0.0773	0	0.0713	0.0745
Wise	0.0617	0.0683	0.0690	0.0694	0.0611	0.0714	0.0643	0	0.0662
Youkilis	0.0764	0.0860	0.0856	0.0861	0.0755	0.0864	0.0802	0.0792	0

generates $E_L = 0.7053$ and $E_{LF} = 0.2676$. The order Dunn – Rios – Pierzynski – Ramirez – Konerko – Vicedo– Youkilis – De Aza – Beckham gives $E_L = 0.7084$ and $E_{LF} = 0.2668$.

Replacing the batting order used by the CWS by this one would add 0.04 to the value of E_L , which would mean at least 0.18 extra runs in an average game and at least 30 extra runs in a season. (This analysis is based on the average of 4.5 rotations of a batting order in a game.) Extra 30 runs, based on the Pythagorean formula for the winning percentage, would lead to extra three wins, which would tie CWS with Detroit Tigers with 88 wins and could mean getting into the playoffs in 2012 instead of taking the second place in the Central division of AL!

5. Discussion and conclusions

This paper is an attempt at using a multi-criteria approach to solve the problem of selecting an optimal starting lineup, which is one of the most important and consequential tasks facing a Major League Baseball team manager 162 times in a single season. If approached in a straightforward way, the number of options available is very large for any person to consider without the help of technology. For an example of the 2012 Chicago White Sox (CWS) team, which was used as a source of data for the current research, the total number of feasible lineups was equal to $720 \cdot 9! = 261,673,600$. Without a universally adopted methodology of finding the

best solution for the lineup problem, a typical manager chooses the top player on the depth chart for each defensive position whenever these players are available; then uses conventional wisdom in arranging these players in a batting order. The problem with this process is that not every “top” player is always in his top form, and that managers often change their batting orders chaotically when the team has a long losing streak. In other words, there is a need for an unbiased method for lineup optimization, based on a solid foundation, both empirical and theoretical.

A lineup is essentially a set of nine players who play both offense and defense, together with the order in which these players will bat during the offensive part of each inning. Therefore, we have proposed to view the selection a lineup as a two-stage process: first, selecting the set of nine players who play in their defensive positions and in offense, and second, choosing the order in which these nine players come to bat. Such approach allows treating the problem of optimizing lineup in baseball efficiently while assisting the manager in making intelligent and informed decisions

During the first stage of the process, we are selecting an optimal 9-player starting assignment with one catcher, one 1B, one 2B, and so on, in the order of the standard defensive position numbers chart (see Figure 2), and one DH. The set of all feasible starting selections, which is a subset of the decision space for this problem, is viewed as the set of all possible 9-row matrices (in the case of the 2012 CWS team, 720 starting selections were possible). Further, we evaluate each starting assignment with respect to two criteria: the offensive potential of the set, and the defensive quality of the set. The bi-criteria nature of this problem allows for displaying the criteria space with the images of vectors of criteria values for all feasible starting assignments on a two-dimensional scatter diagram with easily identifiable Pareto-optimal points. In section 4.3, we demonstrate optimal solutions for the 2012 CWS team. As a result, a manager can have the set of optimal starting position assignments available before a game, which would allow him, as a secondary decision step, to consider the physical/psychological condition of the players, “chemistry” between the players, etc. to select the most appropriate starting assignment for the given game. This feature of the optimization model could make a decision-support system based on the ideas of our research attractive to the MLB management.

If the manager wishes to use “platoon advantage” by maximizing the number of right- or left-handed hitters in the lineup, he can consider optimal solutions with a corresponding reduced set of feasible starting assignments addressing the needed handedness as was demonstrated in Section 3.4.

Next, we turn to forming an optimal order in which the players come to bat, i.e. the batting order. This problem is viewed as a combinatorial problem with two lexicographic-ordered criteria, with the first criterion as the expected number of runs scored by one rotation of the lineup and the second criterion as the expected number of runs scored by the first three players in the batting order. An optimal batting order is obtained by finding a longest Hamiltonian cycle in a complete directed graph on 9 vertices with the weight of a directed edge (i, j) equal to P_{ij} , the probability of player i scoring a run if followed by player j . A computer program with the input as the set of starters and probabilities P_{ij} (where $1 \leq i \leq 9$, $1 \leq j \leq 9$) should give a recommendation on the batting order, which can be executed by the manager.

Clearly, the proposed two-stage approach requires maintaining all necessary offensive and defensive data for each player on the team roster. However, creating the scatter diagram for solving the starting assignment problem, as well as establishing the current values of probabilities P_{ij} may be achieved with simple Excel programs. We used an Excel program to organize data and compute criteria values for solving the starting assignment problem in the case of 2012 CWS (see sections 3.3 and 3.4); we also used Excel for computing probabilities P_{ij} for all possible ordered pairs of CWS players (see section 4.3). The problem of an optimal batting order is an integer programming problem with 81 variables, which can be solved using many open-source sites such as SolverStudio (at solverstudio.org). As a result, the entire two-stage process of lineup selection is highly tractable and manageable.

Certainly, it is necessary to note shortcomings and limitations of the proposed methodology of solving the lineup optimization problem. First, our desire for an easier visual representation of the criteria space leading to a fast identification of optimal starting player assignments had limited us to two criteria. While these criteria are the most important ones, our model could have been more thorough with the inclusion of other formally defined criteria. Another shortcoming is related to the fact that the suggested formula for the expected number of runs scored by a single rotation of a batting order, while an improvement over the previously published version (Sugrue and Mehrotra, 2007), is an underestimate of the actual expected number of runs because it does not consider some scoring possibilities for Player i in the lineup. This does not impact conclusions of the current research since the computation of the expected number of runs is used for comparison purposes only. However, this approach, in its current form, cannot be used for predicting the number of runs scored by a given lineup.

These shortcomings may be viewed as directions for future research, both theoretical and empirical, in baseball applications. For example, a decision support system for the manager to follow the proposed two-stage process of optimizing lineup for a given game is another important direction for future practical applications of the results presented in this paper. If such a system could be implemented, there would be a way to assess how our methodology improves team performance in the long run. Additional directions for future research and applications can be identified by a long list of unsolved problems that a baseball manager faces during a game, such as choosing a pitching substitution, a running substitution (pinch-runner), and a hitting substitution (pinch-hitter); making decisions on sacrifice flies/bunts, base stealing, etc.

Finally, the ideas presented here could potentially provide support in managing sports other than baseball. In various sports where the team players must fulfill both offensive and defensive functions (such as basketball, hockey, and soccer) it makes sense to use offensive and defensive qualities when selecting the set of starters before a given game or choosing the set of players with most time played.

References

- Albert J. and Bennett J. 2003, *Curve Ball: Baseball, Statistics, and the Role of Chance in the Game*, revised edition, Copernicus Books: New York
- Baseball Info Solutions and James B. 2011, *The Bill James Handbook 2012*, ACTA Publications

Baseball Info Solutions and James B. 2016, *The Bill James Handbook 2017*, ACTA Publications

Baseball-Almanac.com

BaseballReference.com

Bukiet B, Harold E. R., and Palasios J. 1997, A Markov Chain Approach to Baseball, *Operations Research* **45**: 14-23

D'Esopo D. A. and Lefkowitz B. 1960, The Distribution of Runs in the Game of Baseball, *SRI Internal Report*

Dummies.com

Duncan Petersen Publishing Ltd. and Palich, M., ed. 1998, *The Book of Rules*, Checkmark Books

Fangraphs.com

Gleeman A. and Sayre B., eds. 2017, *Baseball Prospectus 2017*. Turner Publishing Company: Nashville, Tennessee

Graham C. J. 2012, Baseball Enigma: The Optimal Batting Order, *MIT Sloan Sports Analytics Conference, 2012*

Official Rules of Major League Baseball, 2013 Edition

Pankin, M., 2018, Markov Chain Models: Theoretical Background, Mark Pankin Baseball Page, www.pankin.com/markov/theory.htm

Peterson J. 2011, All Innings Are Not Created Equal: How Run-Scoring Varies by Inning, <https://beyondtheboxscore.com>

Polyashuk M. 2005, A Formulation of Portfolio Selection Problem with Multiple Criteria, *Journal of Multi-Criteria Decision Analysis* **13**: 135-145

Sen A.K. 1970, *Collective Choice and Social Welfare*, Holden-Day: San Francisco

Sugrue P.K. and Mehrotra A. 2007, An Optimization Model to Determine Batting Order in Baseball, *International Journal of Operational Research* **2**: 39-46

Thaker, R. 2011, An Analysis of Lineup Optimization in Baseball, *Math 20: Discrete Probability*

wikipedia.org/wiki/Baseball

- A baseball game consists of 9 parts (innings); each inning is divided into two halves during which the teams switch roles from offense (batting and base running) to defense (pitching and fielding). Visiting team bats in the “top” (first) half of an inning while home team bats in the “bottom” (second) half. If the 9th inning results in a tie, the game continues until one of the teams wins; the bottom half of the last inning is not played if the home team has scored more runs and the outcome of the game is decided.
- The players on the team at bat attempt to score runs by circling or completing a tour of the four bases, starting and ending at the home plate in the counterclockwise direction. The team in the field attempts to prevent runs from scoring and record outs, which remove opposing players from offensive action. When three outs are recorded, the teams switch roles for the next half-inning. If the score of the game is tied after nine innings, extra innings are played to resolve the contest.
- The baseball is around 9 inches (23 centimeters) in circumference and has a rubber or cork center covered in white cowhide with red stitching; the bat is a hitting tool, traditionally made of a single, solid piece of wood. Bats are typically around 34 inches (86 centimeters) long; the players also use a glove as fielding tool, made of padded leather with webbing between the fingers (gloves may have shapes depending on the fielding positions) and protective helmets when at bat.
- There are usually four umpires (referees) who watch the game on the field and enforce the rules: the first, behind the catcher at the home plate, and three more behind each of the other three bases. The official scorer is the person appointed to observe from the press box and record the outcome of everything that happens during a game, and to make judgment calls that affect the official record of the game. The official scorer files a report after each game for documentation purposes.
- At the beginning of each half-inning, the nine players on the fielding (defending) team arrange themselves around the field. One of them, the pitcher (P), stands on the pitcher's mound. The pitcher begins the pitching delivery with one foot on the rubber, pushing off it to gain velocity when throwing toward home plate. Another player, the catcher (C), squats on the far side of home plate, facing the pitcher. The rest of the defending team faces home plate, typically arranged as four infielders who set up along or within a few yards outside the imaginary lines between first, second, and third base, and three outfielders. A first baseman (1B) is positioned several steps to the left of first base, a second baseman (2B) to the right of second base, a shortstop (SS) to the left of second base, and a third baseman (3B) to the right of third base. The basic outfield positions are left fielder (LF), center fielder (CF), and right fielder (RF).
- The nine players of the batting team are arranged in the batting order. In Major League Baseball (MLB), the batting order is set by the manager before each game begins and presented to the home plate umpire with two copies of his team's lineup card, a card on which a team's starting batting order is recorded.
- The Designated Hitter (DH) rule allows the use of a substitute offensive player for a pitcher when it is his turn to bat. The designated hitter does not participate when the team is fielding.
- MLB consists of American League (AL) and National League (NL), each with 30 teams. American League has adopted the DH rule while National League has not. When there is a game between teams from the different leagues, the game is played based on the home team's rule.
- The batter is positioned to the right or to the left of the home plate (depending on their handedness) facing the pitcher. The strike zone is the space above the home plate between the kneecap and the mid-chest of the batter. If the pitcher throws the baseball within the strike zone, it is called a strike. If the ball is thrown by the pitcher outside the strike zone and the batter did not swing at it, it means that the pitcher has thrown a ball.
- Once the batter hits the baseball, he becomes a runner. If the pitcher throws four balls, then a “base on balls” (or walk) is issued, and the batter becomes a runner by moving to the first base. If a player advances on a walk and a base is occupied, it causes another player (or players) to advance, since no more than one runner is allowed on a base. A walk is also issued if a player is hit by a pitch.
- A batter who hits the ball into the field of play must drop the bat and begin running toward first base. A batter-runner who reaches first base without being put out is said to be safe and is on base. A batter-runner may choose to remain at first base or attempt to advance to second base or even beyond, however far the player believes can be reached safely. A player who reaches base despite proper play by the fielders has recorded a hit. A player who reaches first base safely on a hit is credited with a single. If a player makes it to second base safely as a

direct result of a hit, it is a double; third base, a triple. If the ball is hit in the air within the foul lines over the entire outfield (and outfield fence, if there is one), or otherwise safely circles all the bases, it is a home run: the batter and any runners on base may all freely circle the bases, each scoring a run.

- Any runners already on base may attempt to advance on batted balls that land, or contact the ground, in fair territory, before or after the ball lands. A runner on first base must attempt to advance if a ball lands in play. If a ball hit into play rolls foul before passing through the infield, it becomes dead and any runners must return to the base they occupied when the play began. A runner intending to "steal a base" runs for the next base the moment the pitcher commits to pitch to home plate. If successful, a "stolen base" is recorder; otherwise, the runner is "caught stealing".
- Sometimes a batter is out intentionally, while a runner of his team scores a run. A sacrifice fly occurs when a batter hits a fly ball out to the outfield that allows a runner to score.
- The team in the field (defense) is attempting to record outs. In addition to the strikeout, a member of the batting team may be put out through the flyout, ground out, force out, and tag out. It is possible to record two outs during the same play. This is called a double play. Three outs in one play, a triple play.
- The batter is out if:
 - (a) after he strikes the ball, it flies in the air and is caught by a fielder before it lands, or
 - (b) a third strike is caught by the catcher.
- The runner is out if:
 - (a) he runs more than 3 ft away from the imaginary line connecting the bases, or
 - (b) he is tagged while off his base, or
 - (c) the first baseman secures the ball before he reaches the first base, or
 - (d) he is tagged before reaching a next base.
- A fielding error is an act, in the judgment of the official scorer, of a fielder misplaying a ball in a manner that allows a batter or baserunner to advance one or more bases or allows an at bat to continue after the batter should have been put out.
- Once three outs are recorded, the current half-inning ends, and the teams switch places.

After we have presented summarized rules of baseball (only the most basic rules have been included), let us introduce several technical terms and notations, which we will need when discussing statistical evaluations of players, sets of players, and teams. The list below contains some traditional evaluation terms, which are widely accepted in baseball related literature (items 1 through 25), along with and more recently introduced, advanced, refined measures (items 26 through 32). There are slight differences in how some of the terms are defined by different authors; in those cases, we provide a reference to the source we used.

1. *PA*: the number of plate appearances;
2. *BB*: the number of walks issued to the batter;
3. *1B*: the number of singles performed by the batter;
4. *2B*: the number of doubles performed by the batter;
5. *3B*: the number of triples performed by the batter;
6. *HR*: the number of home runs performed by the batter;
7. *H*: the total number of hits performed by the batter, $H = 1B + 2B + 3B + HR$;
8. *HBP*: the number of times the batter was hit by a pitch;
9. *SF*: the number of sacrifice flies;
10. *AB*: the number of at-bats by the player, $AB = PA - BB - HBP - SF$;
11. *RBI*: the number of runs batted in by the batter (the number of runs scored as a result of a batter's plate appearance, not counting situations in which an error caused the run to score or the batter hit into a double play);
12. *SB*: the number of stolen bases performed by the player;
13. *CS*: the number of times the player was caught stealing;
14. *GDP*: the number of times of getting into a double play by the batter;

15. *BA*: the player's batting average, $BA = \frac{H}{AB}$;
16. *TB*: the player's total bases, $TB = H + 2B + 2 \cdot 3B + 3 \cdot HR = 1 \cdot 1B + 2 \cdot 2B + 3 \cdot 3B + 4 \cdot HR$;
17. *SLG*: the player's slugging average, $SLG = \frac{TB}{AB}$;
18. *IP*: isolated power, $IP = SLG - BA = 2B + 2 \cdot 3B + 3 \cdot HR$;
19. *OBP*: the player's on-base percentage, $OBP = \frac{H + BB + HBP}{PA}$;
20. *OPS*: the player's on-base percentage plus slugging average, $OPS = OBP + SLG$;
21. *PO*: the number of defensive outs performed by the fielder (putouts);
22. *A*: the number of defensive plays assisting an out performed by the fielder (assists);
23. *DP*: the number of double plays turned by the fielder;
24. *E*: the number of fielding errors performed by the fielder;
25. *FP*: the player's fielding percentage, $FP = \frac{PO + A}{PO + A + E}$;
26. *RC* (or *WRC*): the (weighted) number of runs created by the player, $RC = \frac{(H + BB) \cdot TB}{AB + BB}$ (Albert and Bennett, 2003);
27. *RC/G*: the number of runs created by the player per game, $RC / G = \frac{RC \cdot 27}{AB - H + SH + SF + CS + GDP}$ (Albert and Bennett, 2003);
28. *RAR* (*runs above replacement*): the number of runs above or below the league average (determined by an algorithm), a characteristic measuring the number of runs created by the player above the league average (BaseballReference.com);
29. *DRS*: the number of defensive runs saved by the player above or below the league average (determined by an algorithm) (Baseball Info Solutions (BIS) and Bill James, 2011);
30. *DEF*: the number of runs above or below average a player has been worth on defense, combining Fielding Runs and the positional adjustment.
31. *OFF*:
32. the number of runs above or below average a player has been worth offensively, combining batting runs and baserunning runs.
33. *WAR*: the number of wins above or below replacement, a composite characteristic indicating the number of wins added to the team record (or lost by the team) by the player, compared to an average replacement player (determined by an algorithm) (BaseballReference.com).

Note that measures 28 through 32 are relative and compare with an average player, so positive and negative values suggest an above or below average performance, respectively. For instance, if *WAR* for a given player is equal to -10, it means that his performance cost the team 10 wins while *WAR* = 10 means that the player helped to earn extra 10 wins for the team (in a 162-game season).

We have presented an incomplete list of measurements, which are used to evaluate defensive, offensive, and overall performance of batters and teams and which are carefully accumulated, computed, and analyzed by MLB, SABR, and many other organizations and individuals. It is important to keep in mind that there are many other, more specified, measurements that are seriously considered when comparing/evaluating players and teams. For example, *BA* and *OBP* can be considered based on right- and left-handedness, on natural grass vs. artificial grass (turf), with or without runners in scoring position, at home and on the road; runs created may include park factor; defensive statistics are considered for each position, which any given player has taken on the field, etc.

Player impact measures for scoring in ice hockey

Carles Sans Fuentes, Niklas Carlsson, and Patrick Lambrix

Linköping University, Sweden

Abstract

A commonly used method to evaluate player performance is to attribute values to the different actions that players perform and sum up these values every time a player performs these actions. In ice hockey, such metrics include the number of goals, assists, points, plus-minus statistics and recently Corsi and Fenwick. However, these metrics do not capture the context of player actions and the impact they have on the outcome of later actions. Therefore, recent works have introduced more advanced metrics that take into account the context of the actions and perform look-ahead. The use of look-ahead is particularly valuable in low-scoring sports such as ice hockey. In this paper, we first extend a recent approach based on reinforcement learning for measuring a player's impact on a team's scoring. Second, using NHL play-by-play data for several regular seasons, we analyze and compare these and other traditional measures of player impact. Third, we introduce notions of streaks and show that these may provide information about good players, but do not provide a good predictor for the impact that a player will have the next game. Finally, streaks are compared for different player categories, highlighting differences between player positions and correlations with player salaries.

1 Introduction

In the field of sports analytics, many works focus on evaluating the performance of players. A commonly used method to do this is to attribute values to the different actions that players perform and sum up these values every time a player performs these actions. These summary statistics can be computed over, for instance, games or seasons. In ice hockey, common summary metrics include the number of goals, assists, points (assists + goals) and the plus-minus statistics (+/-), in which 1 is added when the player is on the ice when the player's team scores (during even strength play) and 1 is subtracted when the opposing team scores (during even strength). More advanced measures are, for instance, Corsi (sum of shots on goals, missed shots and blocked shots) and Fenwick (sum of shots on goals and missed shots)¹.

However, these metrics do not capture the context of player actions and the impact they have on the outcome of later actions. To address this shortcoming and to capture the ripple effect of actions (where one action increases/decreases the success of a later action, for example), recent works [9, 11, 4] have therefore introduced more advanced metrics that take into account the context of the actions and perform look-ahead. The use of look-ahead is particularly valuable in low-scoring sports such as ice hockey.

In this paper we use an existing approach for measuring player performance (based on actions performed by a player) as well as extend the approach (based on actions when a player is on the ice). Further, we introduce time-normalized versions of the approaches. The background to the existing approach is given in Sect. 3 and the performance metrics are defined in Sect. 4. In Sect. 5 we analyze these two metrics in different ways using data from the NHL 2007-2008 and 2008-2009 seasons. First, we look at the top 10 players for these metrics in the two seasons, and discuss performance distributions. Then, we compare these metrics with the traditional metrics goals, points and +/- and discuss the relation to salary. Third, we introduce two notions of streaks and show that information about these kinds of streaks may give indications on who is a good player but not for whether this particular player will contribute more or less in a game than on average over a season. Finally, we compare and contrast the streak durations observed for different player categories (e.g., based on player position and salary) and tie our findings to those in prior parts of the paper.

¹ See, e.g., [https://en.wikipedia.org/wiki/Analytics_\(ice_hockey\)](https://en.wikipedia.org/wiki/Analytics_(ice_hockey)).

2 Related work

Many of the models for evaluating player performance attribute a value to the actions the player performs and then compute a sum over all those actions. For instance, the goal measure attributes a value to goal-scoring actions, while the assists measure attributes a value to passes that lead to goals. This is also true for some newer performance measures such as Fenwick and Corsi that attribute value to shots. Several newer performance measures extend some of the traditional measures. For instance, several regression models have been proposed for dealing with the weaknesses of the +/- measure (e.g., [6, 7, 1]). Further, in [2] principal component analysis was performed based on 18 traditional measures and a performance measure based on the 4 most important components was proposed.

Some of the approaches take game context into account. Added goal value [8] is a measure that attributes value to goals, but the value of the goal is dependent on the situation in which it is scored, thereby taking some context into account. Another measure for player evaluation based on the events that happen when a player is on the ice is proposed in [10]. Event impacts are based on the probability that the event leads to a goal (for or against) in the next 20 seconds. Other works model the dynamics of an ice hockey game using Markov games where two opposing sides (e.g., the home team and the away team) try to reach states in which they are rewarded (e.g., scoring a goal). In [13] the scoring rate for each team is modeled as a semi-Markov process, with hazard functions for each process that depend on the players on the ice. A Markov win probability model given the goal and manpower differential state at any point in a hockey game is proposed in [3]. In [9, 11, 12, 4] action-value Q-functions are learned with respect to different targets. (See Section 3 for the model in [9].) Although the approaches use Markov-based approaches, the definitions of states and reward functions are different. The advantages of such approaches (e.g., [12]) are the ability to capture game context (goals have different values in a tie game than in a game where a team is leading with many goals), the ability to look ahead and thereby assigning values to all actions in the game, and the possibility to define a player's impact through the player's actions. In this paper we base our work on one of these approaches.

3 Background

We base our work on an initial model presented in [9], where action-value Q-functions are learned with respect to the next goal. The state space considers *action events* with three parameters: the action type (Faceoff, Shot, Missed Shot, Blocked Shot, Takeaway, Giveaway, Hit, Goal), the team that performs the action (home, away), and the zone (offensive, neutral, defensive). A *play sequence* is defined as the empty sequence or a sequence of events for which the first event is a start marker, the possible next events are action events, and the possible last event is an end event. If the play sequence ends with an end event, it is a *complete* play sequence. The start/end events are Period Start, Period End, Early Intermission Start, Penalty, Stoppage, Shootout Completed, Game End, Game Off, and Early Intermission End. Actions and play sequences occur in a context. In [9] a *context state* contains values for 3 context features. Goal Differential is the number of home goals minus the number of away goals. Manpower Differential is the number of home players on the ice minus the number of away players on the ice. Further, the Period of the game is recorded. A *state* is then a pair which contains a context state and a (not necessarily complete) play sequence.

Actions are performed in specific states. For action a and state $s = \langle c, ps \rangle$, where c is the context state and ps is the play sequence, the resulting state of performing a in state s is denoted by $s * a$ and is defined as $\langle c, ps * a \rangle$, where $ps * a$ is the play sequence obtained by appending action a to ps . For states with play sequences that are end events, the next state is a state of the form $\langle c', \emptyset \rangle$ where c' is defined by the end event. For instance, a goal will change the goal differential and update the context.

Table 1: Basic action sets.

A is the set of all state-action-pairs $\langle s, a \rangle$ where action a is performed in state s
$A_i(p_k)$ is the set of state-action-pairs when player p_k is on the ice
$A_p(p_k)$ is the set of state-action-pairs where the action is performed by player p_k $A_p(p_k) \subseteq A_i(p_k)$

Table 2: Player and player pair impact.

The direct goal-based impact of a player is the sum of the goal-based impact values of the actions performed by the player: $\text{DGB-impact}(p_k) = \sum_{\langle s, a \rangle \in A_p(p_k)} \text{impact}(s, a)$
The on-ice goal-based impact of a player is the sum of the goal-based impact values of the actions when the player is on the ice: $\text{OIGB-impact}(p_k) = \sum_{\langle s, a \rangle \in A_i(p_k)} \text{impact}(s, a)$

Transition probabilities between different states are based on play-by-play data. The transition probability $\text{TP}(s, s')$ for a transition from state s to state s' is defined as $\text{Occ}(s, s') / \text{Occ}(s)$ where $\text{Occ}(s)$ is the number of occurrences of s in the play-by-play data and $\text{Occ}(s, s')$ is the number of occurrences of s that are immediately followed by s' in the play-by-play data. Using a state transition graph with the computed transition probabilities, Q-values for states are learned using a value iteration algorithm. The **goal-based impact of an action a in a state s** , $\text{impact}(s, a)$, is then defined as $Q_T(s * a) - Q_T(s)$ where T is the team performing the action.

The performance of a player is computed as the sum of the goal-based impacts of the actions the player performs (over a game or a season). This is equivalent to comparing the actions taken by a specific player to the actions of an average player.

For our work, we re-implemented the code available from [9] using Python and R. The reward for goals for is +1 and goals against -1. (In the original implementation only +1 for goals is used.) The resulting goal-based impact values for the actions were used as a base for our work on player performance.

Recently, in [4] an updated model was introduced (which we have not used) where more events as well as more context features are taken into account. The Q-function represents the probability that a team scores the next goal. A neural net representing the Q-function was learned. It was shown that the updated player performance measure based on the updated goal-based impact measure is different from other measures such as +/-, expected goal, win-above-replacement and goal-above-replacement. It also correlates better than the other measures with many standard success measures such as goals, assists, points, shots, and face-off win percentage as well as with salary.

4 Player metrics

We introduce two basic measures for computing goal-based performance (similarly to [5]) as well as variants that normalize the measure with respect to time on ice.

First, we define different sets of actions² for players (Table 1). We differentiate between actions performed by a player and actions performed (by the player or another player) when a player is on the ice. In Table 2 we define the **direct goal-based impact of a player** (DGB-impact) based on the actions the player performs (and this is essentially the impact as defined in [9]). Further, we define the **on-ice goal-based impact of a player** (OIGB-impact) using the actions when the player is on the ice. This

²In the remainder we use action as a shorthand for action in a particular state.

Table 3: Top 10 players for 2007-2008 and 2008-2009 for the direct impact.

Player Name	Position	Age	Salary	GP	G	A	+/-	Points	Direct	Direct/h	On-ice	On-ice/h
2007-2008												
Alex Ovechkin	F	22	3.83	82	65	47	28	112	71.96	182.65	232.56	588.85
Dion Phaneuf	D	22	0.94	82	17	43	12	60	59.22	134.05	246.12	559.67
Rick Nash	F	23	5.50	80	38	31	3	69	59.01	181.80	158.82	485.99
Jarome Iginla	F	30	7.00	82	50	48	27	98	58.94	161.92	204.12	560.88
Dustin Brown	F	23	1.18	78	33	27	-13	60	53.78	156.41	171.40	501.48
Brenden Morrow	F	28	4.10	82	32	42	23	74	51.15	146.62	171.59	504.57
Zdeno Chara	D	30	7.50	77	17	34	14	51	50.74	117.69	203.78	468.89
Trent Hunter	F	27	1.55	82	12	29	-17	41	50.31	167.65	153.36	508.27
Mike Green	D	22	0.85	82	18	38	6	56	48.26	122.63	219.72	545.08
Pavel Datsyuk	F	29	6.70	82	31	66	41	97	48.22	134.68	198.44	559.41
2008-2009												
Alex Ovechkin	F	23	9.00	79	56	54	8	110	75.93	194.34	239.89	612.23
Dustin Brown	F	24	2.60	80	24	29	-15	53	59.76	177.60	178.34	540.84
Shea Weber	D	23	4.50	81	23	30	1	53	53.14	136.10	201.19	511.36
Evgeni Malkin	F	22	3.83	82	35	78	17	113	50.76	134.92	220.41	591.75
Dion Phaneuf	D	23	7.00	79	11	36	-11	47	50.34	122.64	240.57	532.49
Vincent Lecavalier	F	28	7.17	77	29	38	-9	67	49.46	143.99	188.17	549.37
Sheldon Souray	D	32	6.25	81	23	30	1	53	49.38	125.86	203.08	514.73
Jeff Carter	F	24	4.50	82	46	38	23	84	48.88	141.78	189.35	548.30
Rick Nash	F	24	6.50	78	40	39	11	79	48.88	145.11	171.59	498.26
Martin St. Louis	F	33	5.00	82	30	50	4	80	47.82	135.55	204.19	569.06

allows for a measure that includes indirect impact on the game by being on the ice. Even when players do not perform registered actions, they can still influence the game; e.g., by opening up a path for a teammate who may score. For both measures we also define variants normalized by time on ice (TOI). In this paper the normalization uses 1 hour of TOI.

5 Data-driven analysis

In this paper we use the play-by-play data for the NHL regular season games for the 2007-2008 and 2008-2009 seasons made available by [9]. Traditional performance metrics are gathered from www.nhl.com while salary information was taken from www.dropyourgloves.com. For the 2007-2008 and 2008-2009 seasons this resulted in information about 944 and 979 players, respectively.

5.1 Goal-based impact

In Tables 3 and 4 we show the top 10 players for the direct and on-ice goal-based impacts for the 2007-2008 and 2008-2009 seasons. For the on-ice impact we removed the goalkeepers as these are much longer on the ice than the other players and therefore collect more impact. For both the direct and on-ice impact measures we see that both defenders and forwards appear in the top-10 lists. This is similar to the +/- measure where 4 defenders were in the top 10 for each of the seasons. The goals and points measures, however, are heavily dominated by forwards. In 2007-2008 and 2008-2009 the best defenders regarding points held places 38 (Nicklas Lidström) and 32 (Mike Green), respectively, while the best defenders regarding goals held places 104 (Dustin Byfuglien) and 80 (Shea Weber), respectively. Similarly to the +/- measure, defenders show up in the top lists for the on-ice impact to a larger degree than for the direct impact. One reason is that the on-ice impact allows for indirect contributions. Another reason may be that, in general, defenders often play more than forwards.

Fig. 1 shows relative impact frequencies for 99.8 % of the values of the different player impact measures for the regular season games in seasons 2007-2008 and 2008-2009. For this figure we have

Table 4: Top 10 players for 2007-2008 and 2008-2009 for the on-ice impact (goalkeepers removed).

Player Name	Position	Age	Salary	GP	G	A	+/-	Points	Direct	Direct/h	On-ice	On-ice/h
2007												
Dion Phaneuf	D	22	0.94	82	17	43	12	60	59.22	134.05	246.12	559.67
Alex Ovechkin	F	22	3.83	82	65	47	28	112	71.96	182.65	232.56	588.85
Tomas Kaberle	D	29	4.25	82	8	45	-8	53	38.32	93.36	221.93	551.72
Mike Green	D	22	0.85	82	18	38	6	56	48.26	122.63	219.72	545.08
Andrei Markov	D	29	5.75	82	16	42	1	58	42.37	105.18	213.81	530.37
Nicklas Lidström	D	37	7.60	76	10	60	40	70	29.04	66.41	205.68	480.18
Jarome Iginla	F	30	7.00	82	50	48	27	98	58.94	161.92	204.12	560.88
Zdeno Chara	D	30	7.50	77	17	34	14	51	50.74	117.69	203.78	468.89
Lubomir Visnovsky	D	31	2.05	82	8	33	-18	41	32.64	83.52	201.34	523.00
Roman Hamrlik	D	33	5.50	77	5	21	7	26	37.79	93.89	201.29	509.39
2008												
Dion Phaneuf	D	23	7.00	79	11	36	-11	47	50.34	122.64	240.57	532.49
Alex Ovechkin	F	23	9.00	79	56	54	8	110	75.93	194.34	239.89	612.23
Evgeni Malkin	F	22	3.83	82	35	78	17	113	50.76	134.92	220.41	591.75
Dan Boyle	D	32	6.67	77	16	41	6	57	36.11	88.65	219.94	539.81
Chris Pronger	D	34	6.25	82	11	37	0	48	43.40	99.89	217.92	503.72
Mike Green	D	23	6.00	68	31	42	24	73	46.41	106.62	214.33	493.09
Nicklas Backström	F	21	2.40	82	22	66	16	88	37.12	111.83	214.19	630.43
Braydon Coburn	D	23	1.20	80	7	21	7	28	40.78	100.10	211.64	516.12
Andrei Markov	D	30	5.75	78	12	52	-2	64	38.03	96.17	209.18	527.62
Mark Streit	D	31	4.10	74	16	40	6	56	39.38	97.60	206.59	504.31

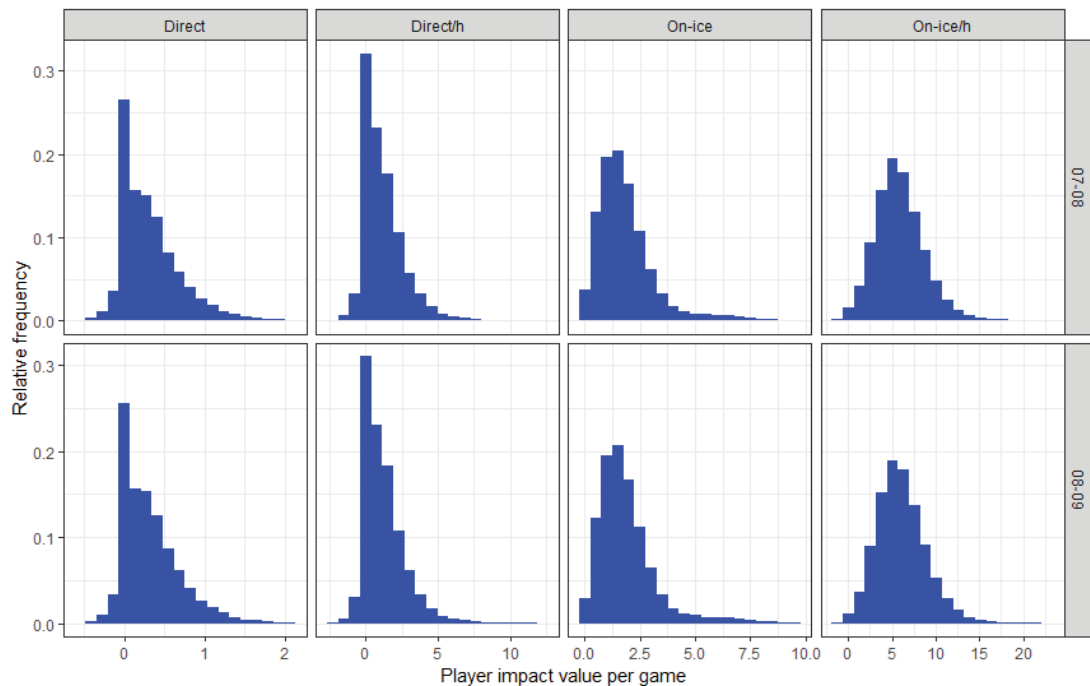


Figure 1: Player impact distributions for the 2007-2008 and 2008-2009 seasons.

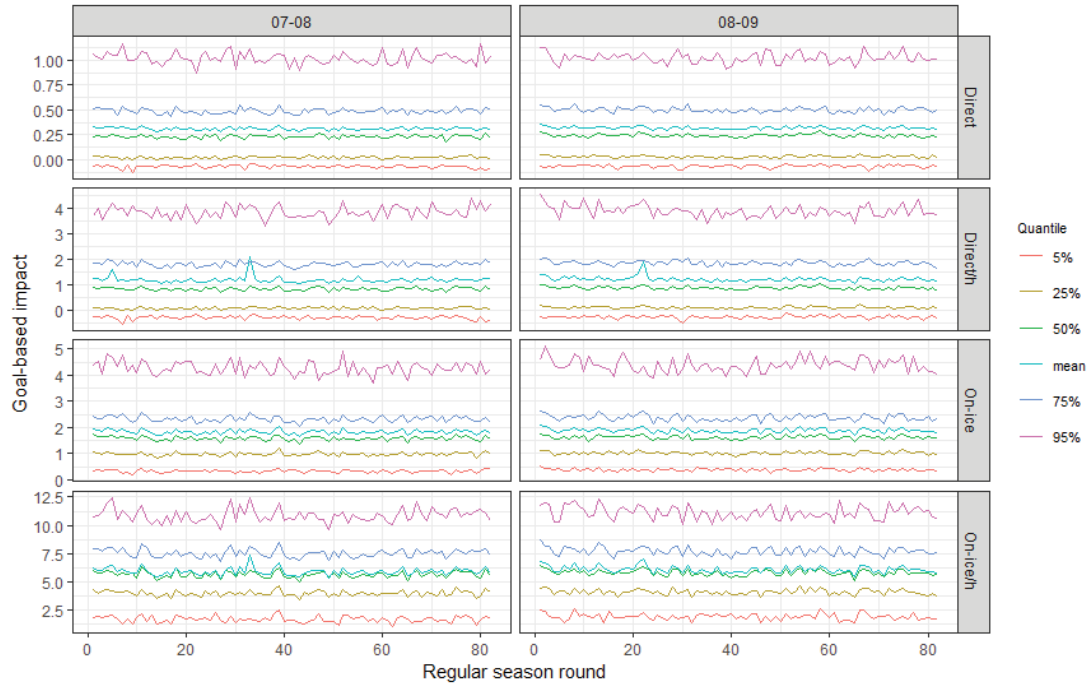


Figure 2: Player impact quantiles per game during 2007-2008 and 2008-2009 seasons.

excluded the two 0.01% tails. All measures are skewed towards the lower impacts. We also note that the distribution of the player impact is similar over the two seasons for each of the measures. In Fig. 2 we show the 5%, 25%, 75%, and 95% quantiles as well as the mean and median of all the goal-based player impacts per game for the regular season games in seasons 2007-2008 and 2008-2009. These are given for the direct and on-ice impacts and their normalized variants. The impact values for the quantile levels are rather stable during the season. Further, except for the normalized on-ice variant there is a clear separation between the mean and the median. We note that the levels of the player impact quantiles are similar over the two seasons for each of the measures. For the DGB-impact the values for the quantiles are about the following: 95% around 1, 75% around 0.5, mean around 0.36, 50% around 0.24, 25% around 0.03 and 5% around -0.06, while for the normalized DGB-impact the values for 95% around 3.9, 75% around 1.84, mean around 1.21, 50% around 0.86, 25% around 0.09 and 5% around -0.02. For the OIGB-impact these values are for 95% around 4.35, 75% around 0.23, mean around 1.8, 50% around 1.59, 25% around 1 and 5% around 0.34, while for the normalized OIGB-impact these values are for 95% around 11.95, 75% around 7.63, mean around 6, 50% around 0.57, 25% around 4 and 5% around 1.75. Further, we note that the gap between 95% and 75% is larger or much larger than between other quantiles of the same range. This may be interpreted as that top players contribute much more than good players.

5.2 Goal-based impact versus other performance measures versus salary

In this section we compare the impact measures to the goal, points and +/- measures (Figs. 3 and 4). The impact measures follow the goals and points for forwards and defenders, but are not correlated with the +/- measure. Further, they seem to allow for a more fine-grained measure than the points and the

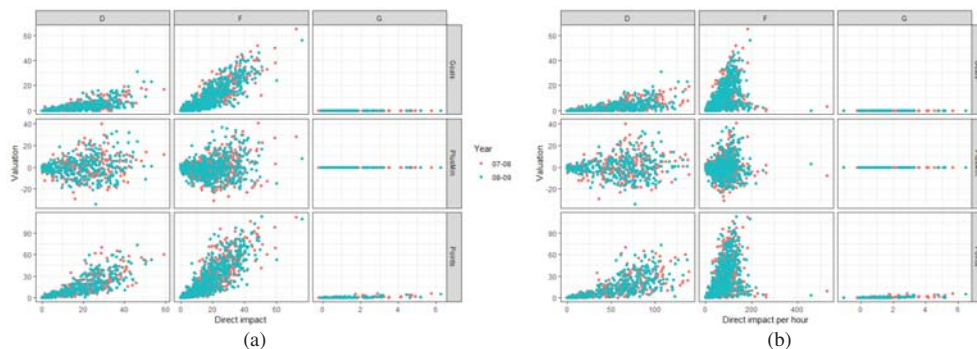


Figure 3: Direct impact (a) and Direct impact per hour (b) versus Goals, +/- and Points for defenders, forwards and goal-keepers for the whole 2007-2008 and the whole 2008-2009 seasons.

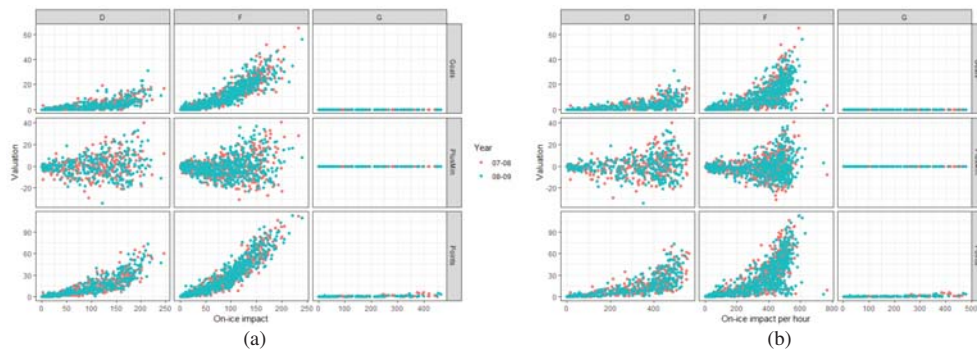


Figure 4: On-ice impact (a) and On-ice impact per hour (b) versus Goals, +/- and Points for defenders, forwards and goal-keepers for the whole 2007-2008 and the whole 2008-2009 seasons.

goals measures.

In Fig. 5 we plot the salary versus several performance measures. For defenders and forwards, the impact measures are similar to the goals and points in the sense that the higher the performance value, the higher the lowest salary for that value. However, the ranges for the salary for a particular performance value are quite large. For goalkeepers, as expected, the direct impact, goals and points have no correlation to the salary, but there is a similar trend as for forwards and defenders for the on-ice impact. We note that the on-ice impact for goalkeepers may be seen as a measure for the team when the goalkeeper is playing. The +/- measure does not seem to influence salary.

5.3 Streak durations

Over the duration of a season, player performance varies. Typically, point streaks (i.e., periods during which a player has points in consecutive games) are used to identify players that are “hot” or “cold”. Furthermore, the points over the last few games (e.g., five games) are often reported, providing some idea of how a player (or team) is currently playing. However, due to the low-scoring nature of the game, long point streaks are becoming rarer³, are seldom long lasting, and only assess offensive numbers. In this section we consider four alternative ways of identifying players currently on “hot streaks” and “cold streaks”. First, instead of using points, we use one of the two metrics: (i) direct impact, and (ii) on-ice

³Longest point streak in a season (NHL). <https://records.nhl.com/records/skater-records/scoring-streaks/longest-consecutive-point-streak>

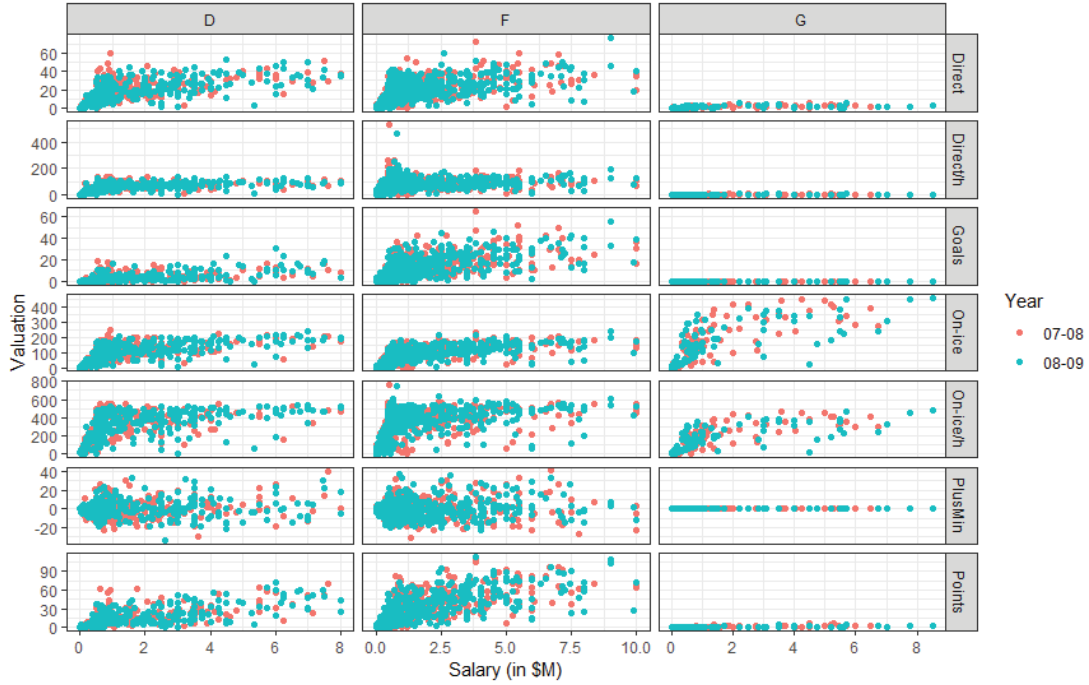


Figure 5: Salary versus different performance measures for defenders, forwards and goal-keepers for the whole 2007-2008 and the whole 2008-2009 seasons.

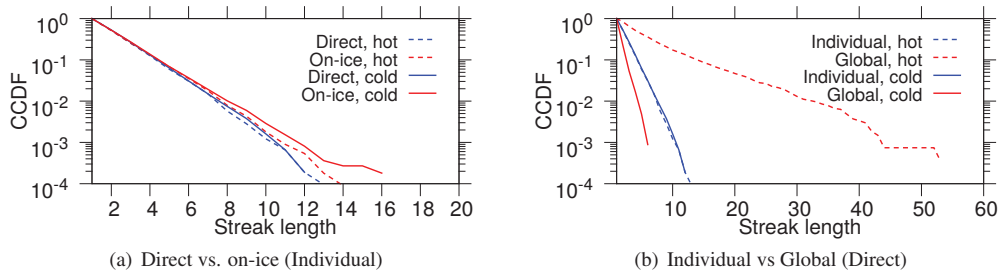


Figure 6: Streak durations shown as empirical Complementary Cumulative Distribution Functions (CCDFs) over all layers during the 2007-2008 season.

impact. Second, we define “streaks” either based on how the player performs relative to an individual threshold (i.e., its median impact) or a global threshold (i.e., whether it has positive or negative impact). For the individual metric, a streak is defined as a sequence of games over which the player has an impact above (or below) the player’s median score (over the games the player plays in the season). To avoid assigning “hot streaks” to players that currently have zero impact, we only consider players that have a median impact above $\epsilon = 10^{-5}$. For the global measure, a hot/cold streak is defined as a sequence of games over which the player consistently has an impact strictly above ϵ or strictly below $-\epsilon$ for some small $\epsilon = 10^{-5}$, respectively. Throughout this analysis we only consider games the player participate in, not games missed by injury, being benched, or that the player for some other reason misses.

Fig. 6 shows the empirical Complementary Cumulative Distribution Functions (CCDFs) of the

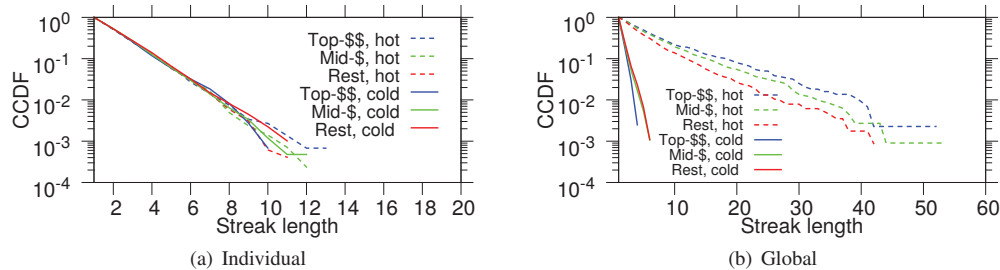


Figure 7: Impact of player's salary range on the streak distributions (CCDFs). Results for 2007-2008 season when using the direct impact measure.

streak durations observed with these four methods across all players. Fig. 6(a) compare the two metrics (i.e., direct vs. on-ice) when using individual per-player thresholds (i.e., their median impact). While the on-ice metric results in slightly longer streaks, all four curves show clear straight-line behavior on lin-log scale suggesting that hot-streak durations when using an individual threshold is exponentially distributed. This itself suggests that hot streaks, when assessed relative to the players' average performance over a season, may actually be memoryless and recent performance history (including longer streaks) may not add value compared to just reporting the average performance over the entire season.

In contrast, as shown in Fig. 6(b), the "hot streak" durations when defined relative a global baseline (i.e., strictly positive/negative impact), have a somewhat heavier tail than suggested by a straight line. The reduced hazard rates observed here suggest that hot streaks when defined relative to such global baseline in fact carry memory. However, we note that part of this simply is due to these streaks often being due to good players (and teams) being more likely to be associated with these streaks. In summary, these results suggests that providing information about who is on a hot streak may primarily help indicate who is a good player and that this player is likely to contribute positively in the next game; however, it does not seem to be a good indicator whether this particular player will contribute more/less than on average over a season.

5.4 Streak durations for player groups

The above observations also hold when looking at individual player groups; e.g., based on salary range (Fig. 7) and player position (Fig. 8). For the salary ranges, we split the players into three categories: (i) the top-10% (with the highest salary), (ii) the mid-range players (with salaries in the top-40%, but not in the top-10%), and (iii) the rest (with salaries below those in the top-40%). Again, we note the typical straight-line behavior of an exponential distribution for the individual measures (Figs. 8(a) and 7(a)) and slightly heavier tails for the global measure (Figs. 8(b) and 7(b)).

Given our prior observation that good players are more likely to be associated with longer hot streaks (when using a global baseline), it is perhaps not surprising that long hot streaks are more frequently among the best paid players. Similarly, these players typically have shorter cold streaks than the less paid players.

In general, we observe a strict ordering of the global CCDFs (shown Fig. 7(b)) based on salary range. This indicate that the better paid players in fact contribute more to the total impact of a team. To highlight these differences we also plotted CDFs of the per-game impact seen by players in the different salary categories (Fig. 9(a)) as a function of time (over the season) for different classes and example percentiles (Fig. 10(a)). And indeed, we did observed similar strict orderings here too.

Corresponding breakdowns based on player position highlight both differences and similarities. First, as shown in Fig. 9(b), although the direct impact is similar for the two player categories, de-

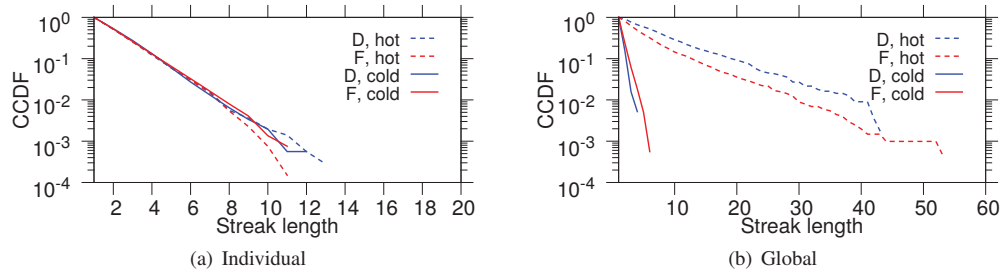


Figure 8: Impact of player position on the streak distributions (CCDFs). Results for 2007-2008 season when using the direct impact measure.

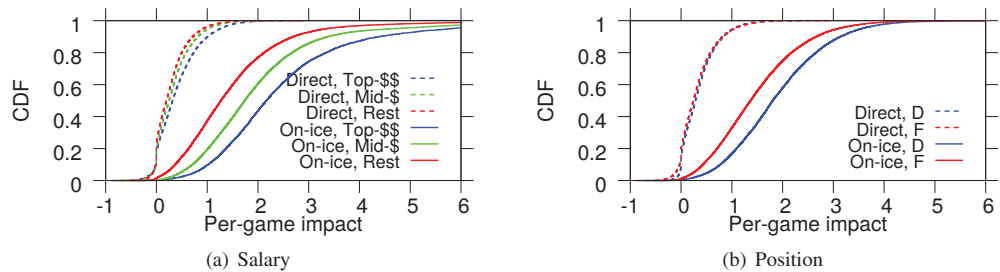


Figure 9: Empirical Cumulative Distribution Functions (CDFs) of the direct impact and on-ice impact for different player positions and salary ranges, as calculated on a per-game basis.

fenders typically have higher on-ice impact than forwards. Second, these observations are consistent across the season, as exemplified by the direct impact (median and 95%-ile scores) shown in Fig. 10(b). Third, although the longest hot streaks belong to forwards, the hot streaks in general are longer for defenders and shorter for forwards. Both the first and the last observation may in part be due to the top-defenders typically playing more minutes per game than forwards.

6 Conclusion

In this paper we introduced and analyzed approaches for measuring goal-based player impact in ice hockey. We showed that the measures, similar to the \pm measure, to a larger degree allow for defenders in the top rankings than goals and assists. There is a certain correlation (as expected) to goals and assists, but not to \pm . Further, we defined two notions of streaks that could be indicators of good players, but not for performance in the next game.

Regarding future work, one direction is to work with different reward functions in the Q-learning algorithm to investigate impact of player actions for different desirable outcomes (e.g., shots on goals, powerplays). Further, it would be interesting to extend the work in [5] and investigate alternative pair impact definitions.

References

- [1] Robert B. Gramacy, Shane T. Jensen, and Matt Taddy. Estimating player contribution in hockey with regularized logistic regression. *Journal of Quantitative Analysis in Sports*, 9:97–111, 2013. doi: 10.1515/jqas-2012-

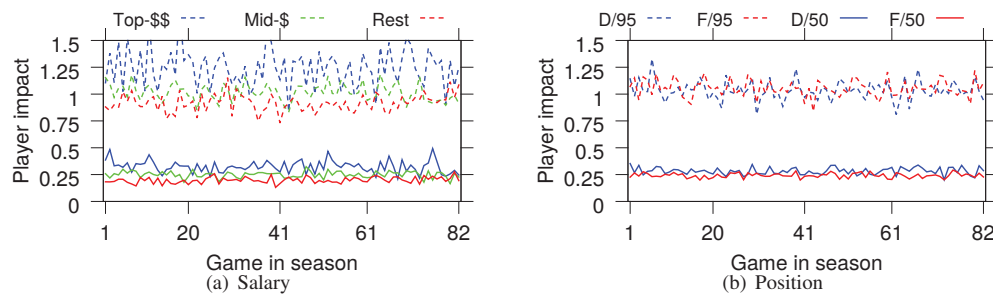


Figure 10: Comparison of the percentile values as seen for different player positions and salary ranges. Here, we show per-game impact values of the median and 95%-ile for each game during the 2007-2008 season.

0001.

- [2] Wei Gu, Krista Foster, Jennifer Shang, and Lirong Wei. A game-predicting expert system using big data and machine learning. *Expert Systems with Applications*, 130:293–305, 2019. doi: 10.1016/j.eswa.2019.04.025.
- [3] Edward H Kaplan, Kevin Mongeon, and John T. Ryan. A Markov Model for Hockey: Manpower Differential and Win Probability Added. *INFOR Information Systems and Operational Research*, 52(2):39–50, 2014. doi: 10.3138/infor.52.2.39.
- [4] Guiliang Liu and Oliver Schulte. Deep reinforcement learning in ice hockey for context-aware player evaluation. In Jérôme Lang, editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pages 3442–3448, 2018. doi: 10.24963/ijcai.2018/478.
- [5] Dennis Ljung, Niklas Carlsson, and Patrick Lambrix. Player pairs valuation in ice hockey. In Ulf Brefeld, Jesse Davis, Jan Van Haaren, and Albrecht Zimmermann, editors, *Machine Learning and Data Mining for Sports Analytics. MLSA 2018*, volume 11330 of *Lecture Notes in Computer Science*, pages 82–92, 2019. doi: 10.1007/978-3-030-17274-9_7.
- [6] Brian Macdonald. A Regression-Based Adjusted Plus-Minus Statistic for NHL Players. *Journal of Quantitative Analysis in Sports*, 7(3), 2011. doi: 10.2202/1559-0410.1284.
- [7] Brian Macdonald. An Improved Adjusted Plus-Minus Statistic for NHL Players. In *MIT Sloan Sports Analytics Conference*, 2011.
- [8] Stephen Pettigrew. Assessing the offensive productivity of NHL players using in-game win probabilities. In *MIT Sloan Sports Analytics Conference*, 2015.
- [9] Kurt Routley and Oliver Schulte. A Markov Game Model for Valuing Player Actions in Ice Hockey. In Marina Meila and Tom Heskes, editors, *Uncertainty in Artificial Intelligence*, pages 782–791, 2015.
- [10] Michael Schuckers and James Curro. Total Hockey Rating (THoR): A comprehensive statistical rating of National Hockey League forwards and defensemen based upon all on-ice events. In *MIT Sloan Sports Analytics Conference*, 2013.
- [11] Oliver Schulte, Mahmoud Khademi, Sajjad Gholami, Zeyu Zhao, Mehrsan Javan, and Philippe Desaulniers. A Markov Game model for valuing actions, locations, and team performance in ice hockey. *Data Mining and Knowledge Discovery*, 31(6):1735–1757, 2017. doi: 10.1007/s10618-017-0496-z.
- [12] Oliver Schulte, Zeyu Zhao, Mehrsan Javan, and Philippe Desaulniers. Apples-to-apples: Clustering and Ranking NHL Players Using Location Information and Scoring Impact. In *MIT Sloan Sports Analytics Conference*, 2017.
- [13] A.C. Thomas, Samuel L. Ventura, Shane Jensen, and Stephen Ma. Competing Process Hazard Function Models for Player Ratings in Ice Hockey. *The Annals of Applied Statistics*, 7(3):1497–1524, 2013.

The Ancient Olympics: Events, Technology, Superstars, Women, Lessons for Them and for Us

Raymond Stefani^{1*}

¹ California State University, Long Beach, USA.
raystefani@aol

Abstract

A multidimensional analysis of the ancient Olympics is presented. Starting in 776 BC, the Olympic Games have emerged as the best known and documented of the four Panhellenic Games, compared to the Nemean, Isthmian and Pythian Games. A reliable list of 861 Olympic events and winners was analyzed. Among athletics events, were the stadion (a run of about 200m), the diaulos (about 400m), the diaulos in armor and the pentathlon (stadion, discus, javelin, long jump and wrestling). There were combat events (boxing, wrestling and the no-holds-barred pankration), chariot racing, equestrian racing and artistic performances (herald competition, trumpeting, lyre playing and acting). Significant technology included the clever use of ropes and levers to start the running events and the chariot races. A cord was wrapped around the javelin with a finger loop to create spin stabilization. Beginning with a standing start, long jumpers employed complex kinematics to extend their distances while carrying weights. The three greatest superstars were Leonides of Rhodes who won all three of the main running events four times in a row (12 wins), Herodoros of Megara who won the trumpeter competition nine times in a row and Astylos of Croton who won 7 athletics events. Five athletes won six times, including Nero (whose wins might have been somewhat contrived). Those superstars would have much to teach us as to training methods and techniques, while our video analysis and knowledge of nutrition could have helped them. Although the Olympic Games were only open to men, Kykniska of Sparta, a married woman, was a double Olympic champion, having twice owned and trained winning chariot horses. Women competed at Olympia in their own separate Heraean Games. The running distances were shortened from multiples of 600 Greek feet (for men) to multiples of 500 Greek feet (for women). In today's world, the Olympic flame is lit at Hera's shrine, providing women with a magnificent symbol of equality.

1 Introduction

It is well-known that the summer Olympic Games of today are an every-four-year event where the athletes of the world gather for a unique competition spanning a wide range of events. The ancient predecessor was first contested in 776 BC at the stadium in Olympia, Miller (2004). Unlike today, three more stadia were built for three other Games, the Nemean games, Isthmian Games and Pythian Games collectively called the Panhellenic Games, Thought Co (2019). Table 1 shows the dates when each Games were first held. A four-year pattern emerged after the last stadium was built. On year one of a cycle, the Games were at Olympia. On year two, the Nemean and Isthmian Games were held, followed on year three by the Pythian Games. Year four would again feature the Nemean and Isthmian Games.

The result was that six Games were held over a given four-year period, unlike our modern experience of one Games being held over each four-year period.

Games	First Held	Stadium Length (m)	Length of a Greek foot (ft)
Olympic	776 BC	192.28	1.050
Pythian (at Delphi)	582 BC	177.55	0.971
Isthmian	581 BC	181.20	0.991
Nemean	573 BC	178.00	0.972
Average			0.996

Table 1 Lengths of the Four Panhellenic Stadia with Opening Years

The primary event of the Panhellenic Games was a running event covering one length of what we would call the “stadium”. According to the Greek measurement system, one plethron was 100 Greek feet while six plethra was appropriately called a stadion, Romano (1993). The result was a running track, 600 Greek feet long. Thanks to existing stadia and the work of archaeologists, Table 1 shows the length of each stadium in meters, Romano (1993), which then gives us the length of a Greek foot in terms of the modern standard foot. The average for the five stadia is almost exactly equal to the modern standard foot. The Greek foot must have been reasonably standard by the time the Pythian, Isthmian and Nemean Stadia were built, given how similar the three lengths were, just a bit shorter than the modern standard foot. The Greek foot used two hundred years earlier at Olympia was 5% longer.

In the remainder of this paper, we will explore:

- the events held at Olympia and elsewhere, thanks to translated documents
- superstars and what they could teach us
- technology used in the games
- women’s sports in ancient Greece
- what our modern technology might tell those athletes of old

2 Events of the Ancient Olympic Games

Thanks to the Foundation of the Hellenic World (2015) and Wikipedia List of Ancient Olympic Victors (2019), a list of 902 events with dates and winners at Olympia (from 776 BC to AD 277) has been gleaned from ancient records. The writer or other source is shown for each contested event. We inserted those 902 dates, events and winners into an Excel spread sheet. Of those, 41 results were highly incomplete, creating a reliable list of 861 events. These were sorted alphabetically by event, to obtain the frequency of each event and thus of each sport. The list of 861 contested events were also sorted alphabetically by winner, to identify multiple winners. Of the 861 contested events, 49% were in athletics, 32% in combat events, 11% in chariot racing, 4% in equestrian racing and 4% in artistic performance. Table 2 lists the five sports, and their 30 events, eight of which were contested only once. The thousand years of competition at Olympia were narrowly focused on no more just 22 events, fewer at many Games.

The five sports in Table 2 include athletics (7 events), combat sports (6 events), chariot racing (10 events), equestrian racing (3 events) and artistic performance (4 events), Olympic-Legacy 1 (2015), Olympic-Legacy 2 (2015) and Perseus Project (2015).

In athletics, the most common running event was the stadion (nearly a 200 m run) contested 254 times, providing 30% of all Olympic winners over the 1000-year span. The other running events were

multiples of the stadion's length. The diaulos with and without armor was about a 400 m run, the dolichos varied from nearly 1500 m to nearly 5 km. The pentathlon was an elimination competition including the discus, javelin and long jump (only contested in the pentathlon) plus the stadion and wrestling. The stadion was likely contested first. The final competitors wrestled for the wreath called a "Stefani" my family name using Latin letters.

Event	Comments/Distance	First Year	Last Year	Times Held
Athletics (7 events, 419 competitions, 49% of all)				
Stadion	X1, 192 m	776 BC	AD 269	254
Stadion-Boys	X1, 192 m	632 BC	AD 133	31
Diaulos	X2, 384 m	724 BC	AD 153	43
Diaulos in Armor	X2, 384 m	520 BC	AD 185	28
Dolichos	X7-24, 1344-4608 m	720 BC	AD 221	30
Pentathlon	Discus, Javelin, Long Jump, Stadion, Wrestling	708 BC	AD 241	32
Pentathlon-Boys		628 BC	628 BC	1
Combat Sports (6 events, 279 competitions, 32% of all)				
Boxing		688 BC	AD 25	61
Boxing-Boys		540 BC	AD 89	40
Pankration		648 BC	AD 221	70
Pankration-Boys		200 BC	AD 117	7
Wrestling		708 BC	AD 213	68
Wrestling-Boys		632 BC	AD 97	33
Chariot Racing (10 events, 94 competitions, 11% of all)				
Apene	2 mules, x6, 7.2 km	500 BC	456 BC	4
Chariot Race		AD 65	AD 129	2
Chariot race-Foals		AD 65	AD 65	1
10 Horse Chariot		AD 65	AD 65	1
Synoris	2 horses, x6, 7.2 km	408 BC	60 BC	14
Synoris-Foals	2 foals, x6, 7.2 km	96 BC	AD 1	3
Synoris-Colts	2 colts, X6, 7.2 km	264 BC	264 BC	1
Tethrippon	4 horses, x12, 14.4 km	680 BC	AD 241	60
Tethrippon-Foals	4 foals, x12, 14.4 km	372 BC	AD 153	7
Tethrippon-Colts	4 colts, x12, 14.4 km	384 BC	384 BC	1
Equestrian Racing (3 events, 36 competitions, 4% of all)				
Foal Racing	X6, 7.2 KM	256 BC	72 BC	7
Horse Racing	X6, 7.2 KM	648 BC	AD 197	28
Mare Racing	X6, 7.2 KM	496 BC	496 BC	1
Artistic Performance (4 events, 33 competitions, 4% of all)				
Herald Competition	Gap of 420 years until AD 65	396 BC	AD 261	12
Lyre Playing		AD 65	AD 65	1
Tragedy Competition		AD 65	AD 65	1
Trumpeter Comp.		396 BC	AD 217	19

Table 1: Ancient Olympics by Sport (x6 means 6 lengths or circuits)

At either end of the running track still stand two long sets of double grooves carved in stone. Those were for the toes of runners starting each race. I verified the figure in Romano (1993), that 22 positions were separated by carved holes that used to hold posts. The starting mechanism will be discussed under technology. For the one-length stadion, 22 could run at once. For longer multiples of a stadion, every other lane was vacant. A runner would traverse the 200 m, run around a post and return in the adjacent

lane, looping as often as needed by the length of the race. It must have been quite a sight to see the diaulos when runners in full armor had to loop around a post and back down the adjacent narrow lane. Special skill and dexterity must have been demanded.

Combat events included boxing, wrestling and the no-holds-barred pankration, much like today's mixed martial arts. Chariot races with two animals (called a synoris) covered six circuits and with four animals (called a tethrippon) covered 12 circuits. Notice that two chariot races were only contested in 65 BC while another was contested in 65 BC and only once again. The three equestrian events covered six circuits of the hippodrome, with competition for foals, horses and mares. The hippodrome at Olympia contained a 1.2 km course. The winning chariot driver or equestrian jockey did not receive the winner's wreath as those competitors were paid and thus ineligible. The winner's wreath went to the owner and trainer of the winning animals.

Among the artistic performance events, the herald and trumpeter competitions reflected a then-practical skill. Since there were no loud speakers, heralds and trumpeters were employed to announce and control public events and to control military movements. They therefore had to be heard clearly over large distances. Other events included lyre playing and acting. Oddly, two events were held only in AD 65 and a herald competition was held in AD 65, after a long hiatus.

It is interesting that artistic performances comprised 4% of events contested. The Second World Mind Sport Games (chess, bridge, checkers, go and Chinese checkers) were held just after the 2012 Olympics, with 29 events, 10% as many as the 300 Olympic physical sports events. Thus, the modern world, as the ancient world, honors mental and physical prowess.

3 Superstars at Olympia and What They Could Teach Us

The 861 event results were sorted to determine the top event winners of antiquity. Two famous names emerged, one of which explained anomalies in the events list. Nero won six events in AD 65: the three chariot racing events run once or twice, the two artistic performance events contested only in AD 65 and the herald competition in 65 BC that was held after a 420-year hiatus. There is an old adage that "Nero fiddled while Rome burned". Actually, Nero won the lyre playing competition. One can only wonder how many ten-horse chariots can race at a time. Organizers obviously wanted Nero's patronage. It would also have been a bad career choice to defeat Nero. The other famous name was the father of Alexander the Great, King Philip II of Macedonia, who won the synoris (348 BC), the tethrippon (352 BC) and the horse race (356 BC). Defeating him was probably a bad career choice too. It is possible that neither King Phillip II and Nero actually competed in the chariot or horse races, as they only needed to own and train the winning horses to be the official winner.

Only three athletes won more events than Nero, although Nero's wins are highly questionable. In athletics, Leonides of Rhodes won 12 events: winning each of the stadion, diaulos and diaulos in armor for four consecutive Games starting in 164 BC, thus spanning a 12-year period. There is much Leonides could teach the athletes of today. He must have had superior training methods, superior running technique and agile turning ability for the diaulos races, given that all the youth of Greece wanted to win those same running events and that one slip in any one heat would have eliminated him, yet he reigned superior over a 12-year span. Leonides is arguably an Olympian for the ages. In our day, Carl Lewis won 10 Olympic gold medals in athletics from 1984 to 1996, including the 100 m run, 200 m run, long jump and 4x100 m relay. His four long jump wins covered a 12-year span, the same as Leonides' span.

Leonides outpaced an earlier athletics superstar who remains the third all-time winner, Astylos of Croton, who won seven times from 480 BC: including the stadion (three times), diaulos (once) and diaulos in armor (three times). Like Leonides, Astylos acquired skills that would be highly valuable to modern athletics.

The second most prolific winner in the Ancient Olympics was Herodoros of Megara, whose wins came in just one event. He won the trumpeter's competition a remarkable nine consecutive times starting in 328 BC, thus spanning 32 years, more than a generation. When he last won, he must have been in his 50s, about the life span of that era. Although today's trumpeters have perfected a technique of circular breathing, that enhances the length and quality of trumpet blowing, we could learn much from him. For Herodoros to have impressed a generation of judges and spectators implies that he had produced remarkable sound quality and a unique, absorbing repertoire that somehow wowed the judges over nine Games and 32 years. It is a shame that his performances are lost in time.

Four athletes won six events legitimately. Winners in wrestling were Hippsthene of Sparta (from 632 BC) and Milon of Croton (from 540 BC). Winners in athletics were Chionis of Sparta (from 664 BC) and Hermogenes of Xanthos (from AD 81).

One of the most irrepressible athletes was Sostratos of Sikyon, nicknamed "Mr. Fingertips". He won the no-holds-barred pankration at Olympia three times (from 264 BC), the Isthmian and Nemean Games 12 times and the Pythian Games two times for a total of 17 wins. His winning record can be traced to his unusual ability of breaking his opponent's finger tips, thus his nickname.

In 216 BC, Kleitomachos of Thebes won more championships in one day than most athletes could win in a lifetime. He won three combat events at the Isthmian Games (boxing, the pankration and wrestling), Wikipedia Isthmian Games (2019).

4 Technology at the Ancient Olympic Games

Given that as many as 22 at a time might start in the stadion and as many as 11 might start the other running races, it was necessary to ensure a fair start. Recall that runners were separated by posts. A clever rope mechanism, the *hysplex* was created as in Figure 1 from Archaeology Archive (2015), taken at Stephen Miller's re-enactment at Nemea in 1993. See also Miller (2004) for a thorough discussion of Ancient Greek Athletics. In the top left photo of Figure 1, starters in yellow at each end are cranking a narrow post from our left to our right, twisting a rope which would force the post to fall left were it not for restraining pins about to be put in place. The athletes are then restrained by two sets of connected ropes and the posts between them. In the top right, the head starter has yanked a rope, releasing the two restraining pins, causing the posts to fall to our left taking the ropes with them. The runners move to our left, where the outside runners have a small advantage, because the ropes fell from outside to inside. The result is a fair start since all runners have an equal chance to draw an outside position and they otherwise start together.

The hippodrome was a 1.2 km oval with very sharp corners. During a chariot race, as they moved counter-clockwise, the chariots nearest to the center divider had to slow down on each turn so as not to impact other chariots. The outermost chariots had to travel farthest around each curve. The chariots in the center of the track could travel at constant speed. An innovative and intricate starting mechanism was created to counter the disadvantage of being in inner or outer lanes. In the bottom photo of Figure 1, the chariots will move to our right after the start and have to navigate a sharp turn to our right. There are six pairs of chariots, with the first pair being farthest left and with pair six at the right center. A dolphin-shaped lever pulls on a rope that opens the starting gates for each pair of chariots, in order from pair one through pair six. Once pair one is cleared to run, pair two is not cleared until pair one is alongside, giving pair one a running start. Pairs three to six in sequence are not cleared to start until the previous pair moves alongside. By the time pair six is cleared to start, the chariots have formed the mirror image of the V-shaped starting alignment. Pairs one through six now have increasing advantage to compensate for the turns to follow, Starting Mechanism for Chariot Races (2012). That is a very clever and effective example of ancient technology.

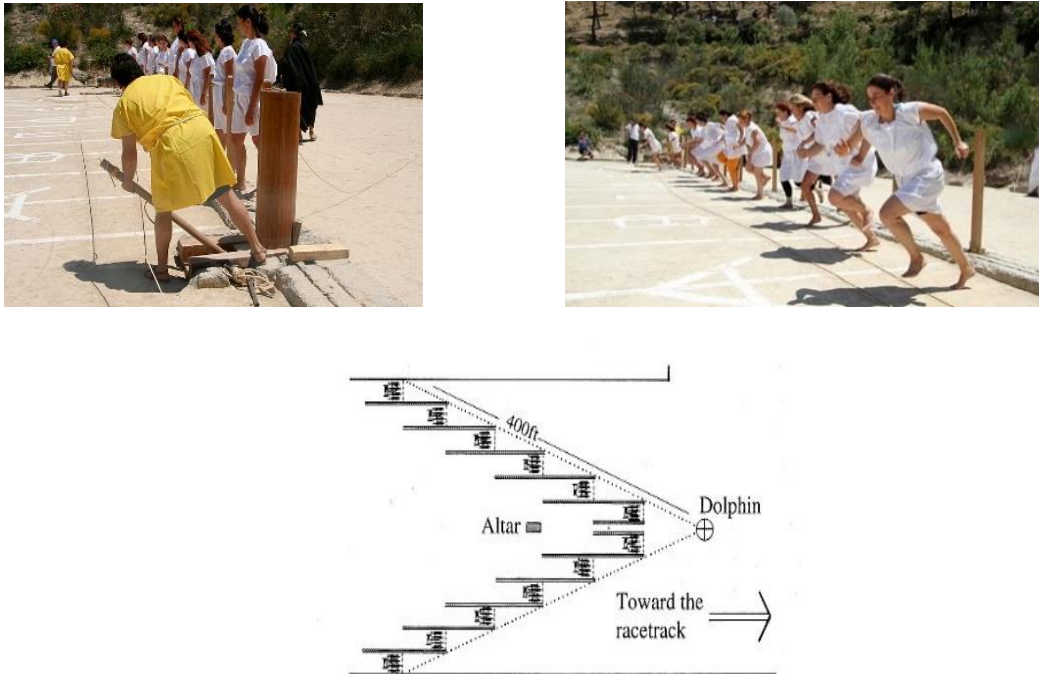


Figure 1: Starting Mechanism for Running Races (Top Left and Right and for Chariot Races (Bottom)

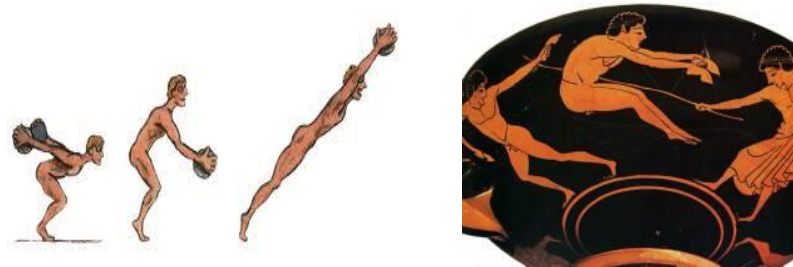


Figure 2: Technique for a Two Footed Long Jump with Weights

The long jump contestants in the pentathlon had to jump with halteres, smoothed weights, in each hand, weighing 1.5 to 2 kg. Minetti and Ardigo (2002) found that trained athletes could gain 5.7% carrying a 2 kg weight with the optimum weight being 5-6 kg. Huang et al. (2005) established a gain of 4.5% using 2 kg weighs with an optimum weight equal to 8% of body mass. Based on the average from these two studies, a 5% gain is possible with proper technique. The best weight is 8% of body mass.

An epigram indicated that Phayllos of Croton once jumped 55 feet (16.3 m), Harris (1960). He competed in the Pythian Games in 482 and 478 BC. Researchers from KU Leuven, The Ancient Long Jump and Phayllos (2012), indicated that after eight weeks of training, athletes jumped 15 m using five

two-footed jumps, the most likely staging for the ancient long jump since five is the number of events in the pentathlon and contemporary urns show two footed jumps. The left part of Figure 2 shows the resulting technique from wind up to push off through the forward thrust. The right photo from KU Leuven shows the landing, on a contemporary urn.

Phayllos competed in the pentathlon and thus also threw the discus. The exact wording on his epigram was that he jumped 5 feet more than 50 in the long jump and threw 5 feet less than 100 in the discus. This scenario suggests that measurement was from the nearest 50-yard marker in increments of five feet, followed by the elimination of those with the shortest throws.

Phayllos earned enough from his athletic career (and possible from family funds) to own and equip a battleship which he used to support Alexander the Great at the Battle of Salamis, Wikipedia Phayllos of Croton (2018). Alexander sent boatloads of loot from his next campaign to Croton, a Greek colony in southern Italy, to thank Phayllos. Winning at sport and battle paid handsomely.

Another insightful use of technology was twisting a cord around the javelin, ending in a loop for one or two of the thrower's upraised fingers. As the javelin was thrown, the loop provided leverage and the unwinding cord made the javelin spin, providing stable flight, Miller (2004, 69-70). Figure 3 shows my photos of a statue in the National Archaeological Museum in Athens. The museum information states that the statute is either Zeus throwing a thunderbolt or Poseidon hurling a trident. The upraised finger position strongly suggests that the sculptor's model had been a javelin thrower. Modern motion capture might suggest that the cord should have been wrapped in the opposite direction so the thumb might go through the loop making for a more efficient throw.



Figure 3 Zeus Throwing a Thunderbolt or Poseidon Throwing a Trident
(National Archaeological Museum)

5 Women in Ancient Greek Sports

This information is drawn from Wikipedia Heraean Games (2019), Were Women Allowed at the Olympics (2016), Miller (2014, pages 150-159) and Romano (1993). The religious practices of the day defined an athlete's role in ancient Greek sports, based on the gender of the god to whom a competition was dedicated. The ancient Olympic Games were dedicated to Zeus. Being a male god, only men were allowed to compete. Unmarried women could and did attend. The High Priestess of Demeter was an honored dignitary. Married women were not supposed to attend; however, Kyrniska of Sparta was a double Olympic champion in 396 BC and 392 BC, having owned and trained the winning chariot horses in the tethrippon. Recall that the chariot driver was not eligible to earn the wreath of victory as that driver was paid. She accepted her olive wreathes outside the stadium. Kallipateira of Rhodes trained

her son inside the stadium but was discovered. She was pardoned since her father, three brothers a nephew and her son were champions. Thereafter, trainers had to be naked.

Women competed in the Heraean Games, so named because these games were dedicated to Zeus' mythological wife Hera, contested in a different year from the Olympic Games; but in the same Olympic stadium. Unmarried women competed. Married women served as officials and trainers. We do not know if men were allowed to attend. The Greek government empowered the so-called Sixteen Women, all married, to coordinate female sports in all of Greece. Women competed in three age groups, over a stadion, reduced from 600 Greek feet to 500, over the two-length diaulos, over the four-length hippios and over the 18-24 length dolichos. The fact that women contested distances 5/6 of that of men implies that women were assumed to perform about 5/6 or 83% as fast as men. In fact, when women resumed Olympic athletics competition in 1928, the female champions performed 83% as well as their male counterparts, Stefani (2014). Today, female Olympic champions in athletics run about 90% as fast as the male champions, Stefani (2014).

Current tradition calls for the Olympic flame to be lit at Hera's shrine in Olympia. That lighting provides a symbolic equality for women, given that men and women now both compete at the same time, in the same place and in nearly equal numbers of events.

6 What We Could Teach Them

We could apply modern kinesiological analysis to the various events to improve technique. Starting reenactments show that the outer lanes gained at the start, as those athletes could see the outer poles begin to carry the restraining ropes forward and therefore time their start before the ropes hit the ground. The interior athletes also should have looked to the outer posts, not forward, to anticipate the fall of the restraining ropes, thus gaining a running start. Motion capture could create adept stopping-turning technique when rounding the posts for events lasting more than one length. Of course, athletes like Leonides and Astylos must have mastered techniques like those suggested.

These ancient athletes ate copious amounts of meat (to become strong), ate heart (to promote oxygen exchange) and ate animal testicles (to gain virility). They apparently believed they would gain the qualities of the animals ingested. Modern nutrition would teach them the biochemical bases. Protein can be gained from more than just meat. Aerobic exercise is needed to increased oxygen exchange. Hormones such as steroids secreted by testicular matter was what was needed.

Ironically, performance enhancing drugs were quite legal. They took hallucinogens, opium juice and chariot drivers took strychnine as a stimulant, Ancient Greeks Used Performance Enhancing Drugs (2015). We could have warned them of the side-affects. Thomas Hicks won the 1904 marathon while ingesting strychnine as a stimulant, Wallechinsky and Loucky (2012). He was almost awarded his gold medal posthumously.

7 Conclusions

Although the Olympic Games that started in 776 B comprised only one of the four Panhellenic Games, given that the Nemean, Isthmian and Pythian Games began about 200 years later; the Games at Olympia proved to be by far the best known and documented. Ten athletics events comprised 49% of the 861 completely documented events contested over the 1000-year span of the Olympic Games. A clever use of two posts with quick-release pins, connected to restraining ropes, allowed the nearly simultaneous starting of as many as 22 runners in the one length, nearly 200 m stadion, and as many as 11 runners in the two-length, nearly 400 m diaulos and diaulos with armor. Leonides of Rhodes was the

most successful athlete of antiquity and arguably all time, having won 12 olive wreaths for the stadion, diaulos and diaulos in armor over four consecutive Games starting in 164 BC. Leonides would have much to teach modern athletes. Another athlete with much to teach us was Astylos of Croton, who won seven times in athletics, the third most wins in the ancient Olympics.

The long jump, javelin and discus joined the stadion and wrestling in the five-part elimination athletics pentathlon. An epigram states that Phayllos of Croton once jumped 55 Greek feet. Long jumpers carried halteres weighing between 1.5 and 2 kg. Modern researchers have nearly duplicated that 55-foot jump as a sequence of five standing jumps, using a clever thrusting and retracting of the weights on each jump. The javelin throw was aided by a coiled cord and finger loop which provided leverage and spin stabilization.

The six combat events comprised 32% of all events contested, including boxing, the pankration and wrestling. Sostratos of Sikyon, won the no-holds-barred pankration a total of 17 times at the four Panhellenic Games, using a devious trick. He broke his opponent's finger tips.

The 10 chariot events (32% of all contested) were started by a clever use of restraining ropes and a so-called dolphin lever-activated release mechanism. The release sequence caused the innermost and outermost chariots to have the largest head starts to compensate for the sharp turn to come. Another 4% of all contested events were in three equestrian racing. The drivers and jockeys were not eligible to win the olive wreath, which was reserved for the owner-trainer. That is why a married woman, Kyrniska of Sparta, was a double Olympic champion in 396 BC and 392 BC, having owned and trained the winning chariot horses, in the men's-only ancient Olympics. King Phillip II of Macedon, father of Alexander the Great won three chariot races.

The four artistic performance events covered the remaining 4% of all events. We could learn much about trumpet technique and repertoire from the second most prolific winner in the Ancient Olympics, Herodoros of Megara, who won the trumpeter's competition nine consecutive times starting in 328 BC. Although a bit contrived, Nero won four times, including lyre playing.

Besides applying modern kinesiological analysis to the various events to improve technique, starting and turning, we could help the ancients with true nutrition, as their food intake seems based on a misguided idea that they would acquire the properties of ingested animals. We could also warn them about the side-effects of taking then-legal substances such as hallucinogens, opium juice and strychnine, much as today's athletes should understand the deleterious long-term effects of their misguided shortcuts.

Women competed in the Heraean Games, dedicated to Zeus' mythological wife, Hera. Women competed in three age groups, over a stadion, reduced from 600 Greek feet to 500 and longer multiples of a stadion. This implies that women were assumed to run about 5/6 or 83% as fast as men. Today, female Olympic champions in athletics run about 90% as fast as the male champions. The Olympic flame is now lit at Hera's shrine in Olympia, a magnificent symbol of equality for women.

References

- [1] Ancient Greeks Use Performance Enhancing Drugs. (2015). Retrieved 23 March 2015 from <http://sportsanddrugs.procon.org/view.timeline.php?timelineID=000017>.
- [2] Archaeology Archive. (2015). Starting Rope System for Running, Retrieved 28 December 2015 from <http://archive.archaeology.org/online/features/olympics/stadia.html>.
- [3] Foundation of the Hellenic World. (2015). Winners of the Ancient Olympics, Retrieved 24 October 2015 from <http://www.fhw.gr/olympics/ancient/en/db.html>.
- [4] Harris, H.A. (1960). An Olympic Epigram: The Athletic Feats of Phayllos. *Greece and Rome*, 7(1), March 1960, pp 3-8.
- [5] Huang, C. et al. (2005). The Effect of Hand Held Weights on Standing Long-Jump Performance, *ISBS Conference, Beijing*.

- [6] Miller, S.G. (2004). Ancient Greek Athletics, New Haven, CT: New Haven.
- [7] Minetti, A.E. and Ardigo, L.P. (2002). Biomechanics: halteres used in ancient Olympic long jump. *Nature* 420, 141-142.
- [8] Olympic-Legacy 1 (2015). General information about the Ancient Olympics, Retrieved 23 October 2015 from <http://olympic-legacy.com/>.
- [9] Olympic-Legacy 2 (2015). Ancient Olympic Events, Retrieved 24 October 2015 from http://www.pe04.com/olympic/olympia/events_o.php
- [10] Perseus Project (2015), Ancient Olympic Events and General Information, Retrieved September 2015 from <http://perseus.tufts.edu/Olympics/sports.html> and <http://www.perseus.tufts.edu/Olympics/index.html>.
- [11] Romano, D.G. (1993). Athletics and Mathematics in Archaic Corinth: The Origins of the Greek Stadion. *Memoirs of the American Philosophical Society*, v. 206.
- [12] Starting Mechanism for Chariot Racing (2012). Retrieved December 2015 from <http://ancientolympics.arts.kuleuven.be/eng/TA012aEN.html>.
- [13] Stefani, R.T (2014). Understanding the Velocity Ratio of Male and Female Olympic Champions in Running, Speed Skating, Rowing and Swimming, *Proceedings of ANZIAM Mathsport 2014, Darwin, Australia*.
- [14] The Ancient Long Jump and Phayllos (2012). Retrieved 19 September 2013 from <http://ancientolympics.arts.kuleuven.be/eng/TC003EN.html>.
- [15] Thought co. (2019). Panhellenic Games of Ancient Greece, Retrieved 6 January 2019 from <https://www.thoughtco.com/panhellenic-games-ancient-greece-116597>.
- [16] Wallechinsky, D. and Loucky, J. (2012) *The Complete Book of the Olympics 2012 Edition*. Aurum: London, 2012, pp 142-143.
- [17] Were Women Allowed at the Olympics? (2015). Retrieved December 2015 from <http://www.perseus.tufts.edu/Olympics/faq5.html>.
- [18] Wikipedia Heraean Games (2019). Retrieved 14 February 2019 from https://en.wikipedia.org/wiki/Heraean_Games.
- [19] Wikipedia Isthmian Games (2019). Retrieved 26 September 2019 from https://en.wikipedia.org/wiki/Isthmian_Games.
- [20] Wikipedia List of Ancient Greek Victors (2018). Retrieved 9 September 2018 from https://en.wikipedia.org/wiki/List_of_ancient_Olympic_victors.
- [21] Wikipedia Phayllos of Croton (2018). Retrieved 26 September 2018 from https://en.wikipedia.org/wiki/Phayllos_of_Croton.

Using a combination of the Generalized PageRank Model (GeM) with the Four Factors to Rank NBA Teams

Georgia Twersky^{1,2}, George Lyman^{1,3}, and Annette Pilkington^{1,4}

¹ University of Notre Dame, Notre Dame, Indiana, U.S.A.

² gtwersky@nd.edu

³ glyman@nd.edu

⁴ Pilkington.4@nd.edu

Abstract

In this study, we used the generalized version of Meyer's Google's PageRank (GeM) ranking system, combined with Oliver's Four Factors for basketball to make predictions for outcomes in the NBA playoffs. In particular, we compared the performance of this ranking system with the performance of GeM when calculated with the point differential and with the performance of Massey's ranking system with the point differential. We also created a ranking using Massey's system and the Four Factors for comparison. We ran all models on the NBA playoff data from 2007 to 2018. The results show no significant difference in the performance of the models using the four factors and the models using the point differential.

1 Introduction

The National Basketball Association (NBA) has thirty teams which play in two conferences of three divisions with five teams in each division. During the regular season, each team plays 82 games with each team playing all of the other twenty nine teams at least twice. Although teams play teams in their own conference much more often than the teams in the other conference, the interaction between the conferences makes it feasible to use ranking systems from linear algebra as a tool for creating rankings for the teams.

The NBA Playoffs begin in mid April, after the regular season, with the top eight teams in each conference, competing for the title. In the tournament, each team plays an opponent in a best-of-seven games series. The winning team advances to the next round and the losing team is eliminated from the tournament. The process continues until two teams remain, one from each conference. These two teams then compete in a best-of-seven series to determine the champion.

A popular challenge for spectators of most major sports in the United States is to fill out a bracket to predict the results of the playoffs prior to the start of the tournament, and see how the results match the predictions. Often pools are created and brackets are scored with the points scored for predicting a winner doubling with each round. The winner of the pool prize being the person who accumulates the most points. Clearly predicting the winner of the entire tournament correctly adds a large number of points, but one can also accumulate a large number of points by predicting a large number of winners in the rounds preceding the final.

After the playoff's have been completed and the winners of all matches have been determined, the bracket gives us a partial ordering of the teams. In this study, we used methods from linear algebra to create rankings for the teams from the regular season data, prior to the playoffs. Each ranking system gives us a total order of teams, which we can use to fill out a bracket, creating our own partial order of the teams. Our chosen measure of how effective a ranking system was in predicting wins was the number of points that one would accumulate on

a bracket, if it were filled out prior to the tournament in accordance with the ranking system. Obviously the weighting in the metric adds significantly to the value of a ranking system that can make predictions with a high degree of accuracy for the latter rounds.

In their 2008 paper, Meyer et al. [1] propose a generalization of Google's PageRank (GeM) to rank sports teams, using the point differential as weights for a weighted graph of games played. Towards the end of the paper, they suggest using a generalized version of the GeM model derived from such weighted graphs created with a variety of game statistics. Given that Oliver's Four Factors formula (calculated on game statistics), [3] and [4], gives a good prediction (up to a scalar multiple) of the point differential in basketball, we were interested to see if the generalized version of the GeM rankings using the Four Factors could produce better predictions for the playoff brackets than the regular model. It seemed that the generalized model might capture team strengths and weaknesses with greater subtlety by breaking the outcome down into components. Using data from bigdatataball.com for all NBA regular season games and playoff games for each year from 2007 to 2018, we created rankings and filled out brackets using the regular version of GeM, and the generalized version of GeM using the Four Factors. We also created rankings using Massey's method [5] and rankings using a combination of Massey's method and the Four Factors. The methods were implemented using R from the CRAN [6] website.

As it turned out, there was no significant improvement in the predictions when we incorporated the Four factors in the model. In fact, no significant differences were detected in the performance of the four different methods of ranking. Nevertheless there are some remaining avenues of inquiry to be pursued, that may yield an improvement on the effectiveness of the regular models using the point differential.

2 Measuring the Strength of a Ranking System

As mentioned in the introduction, our chosen measure for the strength of a given ranking system as a predictor of the results of the NBA playoffs was the number of points one would accumulate by filling out a bracket for the tournament using the ranking system. One scores 10 points for predicting a win in the first round, 20 for predicting a win in the second round, 40 for predicting a win in the third, and 80 for predicting the winner of the championship in the fourth round. Note that with this system of scoring, it is not enough to predict the winner of the seven game series for a particular match-up, you must also have predicted that the winning team won in all of the preceding rounds in order to collect the points.

For example, in Table 1 below, we show the ranking resulting from the generalized version of the GeM with the Four Factors in 2017. Alongside it, in Figure 1, we show a bracket which shows the points earned by the ranking for the games correctly predicted. It also shows the erroneous predictions in red. Note that although the ranking correctly predicted that Boston would beat Washington, points were not earned, since neither team was predicted to make it to round 2 of the tournament.

3 Generalized Page Rank (GeM)

The original PageRank algorithm (Brin and Page [2]) used by Google was based on the theory of Markov Chains. If a matrix G is stochastic (rows add to 1), is irreducible and has at least one positive diagonal entry, then the Perron Frobenius theorem guarantees that it has a unique eigenvector with eigenvalue 1 and norm 1. The algorithm creates the adjacency matrix of

Team	Rank
Golden State	1
Denver	2
San Antonio	3
Utah	4
Oklahoma City	5
Houston	6
LA Clippers	7
Cleveland	8
Chicago	9
Minnesota	10
Atlanta	11
Portland	12
Memphis	13
Boston	14
Brooklyn	15
LA Lakers	16
Sacramento	17
Milwaukee	18
New Orleans	19
New York	20
Miami	21
Washington	22
Indiana	23
Phoenix	24
Philadelphia	25
Detroit	26
Charlotte	27
Toronto	28
Orlando	29
Dallas	30

Table 1: GeM with Four Factors

the directed graph of web page links, and normalizes it to get a hyperlink matrix H . The matrix H is adjusted by adding a rank 1 matrix to deal with dangling nodes (nodes with no arrows pointing outwards, which represent pages with no links) and a small perturbation matrix to ensure irreducibility and positive diagonals. The associated rating vector for the webpages (the Perron vector) is an eigenvector of the transpose matrix, and can be computed by an algorithm for computing eigenvectors, or by repeatedly applying the matrix to an initial probability distribution vector (using the theory of Markov chains) to get the required vector as the stable vector of the system.

We apply an adaptation of Google's Page ranking system introduced by Meyer, Albright and Govan for sports leagues. The method (GeM) is described in detail in Meyer et al. [1]. A sport season is represented by a weighted directed graph with n nodes, where n is the number of sports teams involved. The teams correspond to the nodes, and each game is represented by an arrow from the loser to the winner, with weight w_{ij} equal to the absolute value of the point differential. The basic steps to constructing the stochastic matrix G are:



Figure 1: A page of a document created using the `debug` option

- Form the $n \times n$ adjacency matrix A of the graph of web pages :

$$A = \begin{cases} w_{ij} & \text{if team } i \text{ lost to team } j \\ 0 & \text{otherwise} \end{cases}$$

- Form the stochastic "hyperlink" matrix H where

$$H_{ij} = \begin{cases} A_{ij} / \sum_{k=1}^n A_{ik} & \text{if there is a link between } i \text{ and } j \\ 0 & \text{otherwise} \end{cases}$$

- Make an adjustment to H for the dangling nodes (rows of zeros corresponding to unbeaten teams) by adding $\frac{1}{n}ae^T$ to get $H + \frac{1}{n}ae^T$. Here a is an $n \times 1$ column matrix with 1's in the j position if j is unbeaten and 0's elsewhere, and e is an $n \times 1$ column matrix of 1's.
- Finally the adjustment to ensure irreducibility and primitivity is made to get the basic

version of the GeM (Generalized Markov Chain) matrix G , given by:

$$G = \alpha[H + \frac{1}{n}ae^T] + \frac{(1-\alpha)}{n}ee^T,$$

where α is a chosen scaling parameter which ensures that the resulting matrix is stochastic. It can be set at any value between 0 and 1 and allows us to adjust the size of the perturbation matrix $\frac{1}{n}ee^T$. Smaller values of α give a larger perturbation of the GeM matrix.

- We use the Perron eigenvector (v) of length one with $vG = v$ to give a rating and a ranking for the teams.

Example. As an example, we have chosen the games played in the first two rounds of the Six Nations Cup (Rugby) in 2015. The results of the games played up to that point are shown in Table 2. We also assign an index to each team, shown in the left hand column of Table 3. In the right hand column of Table 3, We show the point differential for each team.

	Ireland	England	Wales	Scotland	France	Italy
Ireland					18-11	26-3
England			21-16			47-17
Wales		16-21		26-23		
Scotland			23-26		8-15	
France	11-18			15-8		
Italy	3-26	17-47				

Table 2: Rounds 1 and 2 of Six Nations Cup 2015

Index	Team	Point Differential
1	Ireland	30
2	England	35
3	Wales	-2
4	Scotland	-10
5	France	0
6	Italy	-53

Table 3: Point Differential for Six Nations Example

For this example, with $\alpha = 0.85$, and teams ordered by the above index, the matrix H looks like:

$$H = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3/10 & 0 & 7/10 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 23/53 & 30/53 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Adjusting for unbeaten teams, by replacing rows of 0's by rows of $1/6$, and adding $(1-0.85)/6 = 1/40$ to each entry, we get;

$$G = \begin{pmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/40 & 7/8 & 1/40 & 1/40 & 1/40 & 1/40 \\ 1/40 & 1/40 & 7/25 & 1/40 & 31/50 & 1/40 \\ 7/8 & 1/40 & 1/40 & 1/40 & 1/40 & 1/40 \\ 167/424 & 1073/2120 & 1/40 & 1/40 & 1/40 & 1/40 \end{pmatrix}$$

The ratings and rankings produced are shown in Table 4.

Team	Rank	Rating
Ireland	1	0.269
England	2	0.252
Wales	4	0.124
Scotland	5	0.099
France	3	0.158
Italy	5	0.099

Table 4: Rankings for the Six Nations teams, 2015(after round 2), using the GeM model

3.1 The Generalized Version of the GeM Model

The generalized version of the GeM model proposed by Meyer et al. [1] using P game statistics, $\{g_1, g_2, \dots, g_P\}$, involves setting up a stochastic matrix S_i , for each game statistic in the set. One sets up a graph similar to that described for the point differential above, for each of the game statistics used. The weights are the absolute values of the difference between the stats for the teams, and the arrows point from the team with the smaller value to the team with the larger value. From the graph a stochastic matrix $S_i = H_i + \frac{1}{n}a_ie^T$ is created for each game statistic in a manner similar to the one described above. The matrix G which takes the place of G in the calculations above is

$$G = \alpha_1 S_1 + \alpha_2 S_2 + \dots + \alpha_P S_P,$$

where the parameters $\alpha_1, \alpha_2, \dots, \alpha_P$ have the properties $0 \leq \alpha_i \leq 1$ and $\sum_1^P \alpha_i = 1$.

4 Massey's Method

Massey's method of ranking is based on the idea that with a perfect set of ratings for the teams r_i , the difference in the ratings for two teams would equal the point differential for each game played between them. This gives us a system of equations $r_i - r_j = p_k$, one for each game. Obviously, with this approach the probability of getting a system of equations that is inconsistent is very high. The Massey algorithm takes the least squares solution to this system to derive a system of equations with infinitely many solutions. The last equation in the new system is replaced by the condition that the sum of the associated ratings adds to 0 to get a

unique solution, which gives us the Massey ratings. The result is the system of equations

$$\left\{ t_i r_i - \sum_{j \neq i} n_{ij} r_j = P_i, \sum_j r_j = 0 \right\}_{1 \leq i \leq n-1, 1 \leq j \leq n}$$

for the n teams in the tournament. For this system t_i is the number of games played by the row team, n_{ij} is minus the number of games played between team i and team j and P_i denotes the total point differential for team i . The solution to the system gives us Massey's ratings, from which we can derive a ranking for the teams.

Example. When applied to our working example, we get the following matrix equation $\mathbf{M}\mathbf{r} = \mathbf{P}$:

$$\begin{pmatrix} 2 & 0 & 0 & 0 & -1 & -1 \\ 0 & 2 & -1 & 0 & 0 & -1 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 \\ -1 & 0 & 0 & -1 & 2 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \\ r_6 \end{pmatrix} = \begin{pmatrix} 30 \\ 35 \\ -2 \\ -10 \\ 0 \\ 0 \end{pmatrix}$$

The solution to this equation gives Massey's ratings in \mathbf{R} . The resulting ratings, with corresponding rankings, are given in Table 5.

Team	Rank	Rating
Ireland	2	6.91667
England	1	9.58333
Wales	3	2.41667
Scotland	5	-2.75
France	4	2.08333
Italy	6	-18.25

Table 5: Rankings for the Six Nations teams, 2015(after round 2), using Massey's method.

5 The Four Factors

The Four Factors are game statistics that correlate very closely with the point differential in Basketball games. They give measures of the four important factors shooting, turnovers, rebounding and free throws and can be used to identify a team's strategic strengths and weaknesses. They can be calculated for each game, for both teams. They were introduced to basketball analysis by Dean Oliver. More background and details are given in [3]. They are defined as follows:

- **Effective Field Goal Percentage (eFG%)** ; this statistic adjusts for the fact that a 3-point field goal is worth one more point than a 2-point field goal. The formula is given by

$$eFG\% = [(FG + 0.5 * 3P) / FGA] \times 100\%.$$

- Turnover Percentage (TOV%) This is an estimate of Turnovers per 100 plays. The formula is given by

$$TOV\% = TOV/Poss \times 100\% \quad (Poss \approx FGA - OR + TO + 0.44 * FTA)$$

where *Poss* denotes possessions.

- Offensive Rebound Percentage (ORB%) ; an estimate of the percentage of available offensive rebounds grabbed by a player on the team. The formula is given by

$$ORB\% = ORB.P = ORB/(ORB + DRB)$$

where *DRB* is the opponoent'ss defensive rebounds.

- Free Throws Per Field Goal Attempt (FTPFGA) . This measures the team's ability to get to the foul line by dividing Free Throws Made by Field Goals Attempted

$$FTR = FTM/FGA.$$

The factors show a high degree of independence and do not carry equal weight in determining wins and losses. Oliver's theory assigns the the following linear weights to the factors [4]:

1. Shooting (40%),
2. Turnovers (25%),
3. Rebounding (20%),
4. Free Throws (15%).

Using this breakdown, theoretically, one could determine how a team could compensate for poor performance in one area by better performance in another. We were optimistic that breaking down the GeM model along these lines would also lead to a finer analysis of wins and losses and thus lead to better predictions.

6 Performance of Ranking systems on NBA tournament

We ran the regular GeM model on the data from the regular season for each tournament in the study. We set the parameter α to 0.85. We also ran Massey's algorithm as described above. We used the linear weights from Oliver's theory to incorporate the Four Factors into the GeM model, that is, we chose

$$\alpha_1 = 0.4, \alpha_2 = -0.25, \alpha_3 = 0.2, \alpha_4 = 0.15,$$

with the statistics ordered as above. Our hope was that by separating these factors in the generalized GeM model, the breakdown of the point differential would lead to better predictions than those obtained from the regular GeM model. We also adapted Massey's method to incorporate the Four Factors, by replacing the point differential by the linear combination of the difference of the four factors with the above coefficients, which is equivalent to solving for Massey's rankings for each factor individually and then taking a linear combination of the results.

Year Year	GeM	GeM with Four Factors	Massey	Massey with Four Factors
2007	210	80	210	210
2008	150	140	290	250
2009	160	180	140	110
2010	80	140	100	110
2011	110	90	130	110
2012	150	140	110	150
2013	130	80	130	60
2014	260	120	270	90
2015	260	300	240	140
2016	210	220	200	180
2017	310	260	250	260
2018	110	220	100	150

Table 6: Bracket scores by year and method (2016 marks end of 2015/2016 season).

The number of points earned on a bracket (out of a maximum of 320) for each of the four ranking systems described above, for the NBA tournaments from 2007 to 2018, is shown in Table 6.

Contrary to our expectations, neither an analysis of variance nor any of the pairwise t-tests showed any significant difference between the performance of the ranking systems using the four methods described. There are some significant differences on the performance of the models from year to year. Although the results are not promising, there are some further questions we wish to pursue here. As stated in Meyer et al. [1], the main challenge with using the Generalized GeM model is to decide which parameters are optimal. We chose the parameters from Oliver’s linear model, which may not be optimal, since we are dealing with eigenvectors and the eigenvector of a linear combination of matrices is not necessarily a linear combination of eigenvectors of the matrices. It might also be beneficial to pursue non-linear methods of combining eigenvectors derived for the individual factors to create a ranking. Although the adaptation of the Massey model here was intended to give us a benchmark for the effect of adding the Four Factors, some fine tuning of this model may also prove fruitful.

References

- [1] Carl D. Meyer Anjela Y. Govan and Russell Albright. Generalizing google’s pagerank to rank national football league teams. SAS Global Forum 2008, Paper 151, 2008.
- [2] Sergey Brin and Lawrence Page. The anatomy of a large scale hypertextual web search engine. Computer Networks and ISDN Systems, 33: 107-17, 1998. Available at <http://easychair.org/easychair.zip>.
- [3] K. Pelton J. Kubatko, D. Oliver and D. Rosenbaum. A starting point for analyzing basketball statistics. Journal of Quantitative Analysis in Sport, Vol 3, Issue 3, 2007, Article 1., 2007.
- [4] Konstantinos Kotzias. The four factors of basketball as a measure of success. statathlon.com: <https://statathlon.com/four-factors-basketball-success/>, 2018.
- [5] Kenneth Massey. Statistical models applied to the rating of sports teams. Thesis, Bluefield College, 1997.
- [6] R user groups. The comprehensive r archive network. <https://cran.r-project.org/>.

Sports betting strategies: an experimental review

Matej Uhrín¹, Ondřej Hubáček¹, Gustav Šourek¹, and Filip Železný¹

Czech Technical University, Prague

Abstract

We investigate the problem of optimal wealth allocation over predictive sports market's opportunities. We analyze the problem across diverse settings, utility targets, and the notion of optimality itself. We review existing literature to identify the most prominent approaches coming from the diverse sport and economic views on the problem, and provide some practical perspectives. Namely, we focus on the provably optimal geometric mean policy, typically referred to as the Kelly criterion, and Modern Portfolio Theory based approaches leveraging utility theory. From the joint perspective of decision theory, we discuss their unique properties, assumptions and, importantly, investigate effective heuristics and practical techniques to tackle their key common challenges, particularly the problem of uncertainty in the outcome probability estimates. Finally, we verify our findings on a large dataset of soccer records.

1 Introduction

In this work, we understand the game of betting as an investment decision making over probabilistic outcomes. We generally assume two sides acting in the game setting – a bookmaker (B) and a player model (M for “model”). Both the bookmaker and the player have their beliefs about the true probability distribution (P_R) over the outcomes. Prior to the outcome realization, the bookmaker sets up possible payoffs, representing the returns of investment for each possible outcome realization.

Example 1. *As a simplistic, artificial example of such a setting, imagine a game of fair coin tossing [7], where the payoffs are set up, for instance, as follows. If the coin ends up heads, our wealth grows by 50%, and if it is tails, our wealth shrinks by 40%.*

Generally, the goal of the player is to maximize the expected profits while minimizing expected losses (see Section 2 for specific target utility definitions), assuming repeated investment over a prolonged period of time.

Player's betting system typically consists of two major components. A model, estimating the probability of each outcome, and a betting strategy, combining the probabilities given by the predictive model with the bookmaker's payoffs to determine how to split the stakes. In this work, we focus solely on the second component – the betting strategy.

1.1 Betting Opportunities

The opportunities in the game represent simply the set of probabilistic outcomes currently available for betting. We can generally cover each betting game setting via definition of a “payoff matrix” \mathbf{R} , the columns of which represent different opportunities available, and rows represent different outcome realizations. Each element in \mathbf{R} then represents a single payoff from each opportunity realization. Additionally, we include a risk-free “cash” opportunity \mathbf{c} , which allows our strategy to put money “aside”. Also, \mathbf{c} allows to model situations where leaving money aside can cost small fraction of \mathbf{c} in every turn (“inflation”), or possibility to increase with some small interest rate (e.g. in a bank account).

The betting strategy will thus allocate the current wealth among n opportunities, $n - 1$ of which are the risky, probabilistic assets, with 1 opportunity being the added risk-free cash asset. Let us further assume that there are K possible worlds, i.e. K possible joint outcome realizations, in our probabilistic game. Each opportunity can then be fully specified using an asset vector \mathbf{a}_i with returns $r_{i,k}$ for each of the K probabilistic outcomes p_1, p_2, \dots, p_K [1].

$$\mathbf{R} = [\mathbf{a}_1 \quad \mathbf{a}_2 \quad \dots \quad \mathbf{a}_{n-1} \quad \mathbf{c}] , \text{ where } \mathbf{a}_i = \begin{bmatrix} r_{i,1} \\ r_{i,2} \\ \dots \\ r_{i,K} \end{bmatrix} , \mathbf{c} = \begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix} \quad (1)$$

1.2 Betting strategy

Betting strategy takes bookmaker's odds together with model's estimates, and outputs portions of wealth, i.e. the bets, to be waged on each of the betting opportunities. We can formally define the betting strategy f for n betting opportunities as follows:

$$f : P_M^n \times O^n \rightarrow B^n \quad (2)$$

where P_M is the probability distribution over the outcome realizations given by model, O are the bookmaker's odds, and B are the portions of wealth staked. We will further refer to the vector of wagers \mathbf{b} as "portfolio":

$$\mathbf{b} = [b_1, b_2, \dots, b_n]^T \quad (3)$$

where b_i stands for wealth portion allocated to i -th opportunity.

1.3 Betting dynamic

A betting dynamic is part of the game setting for the presented opportunities. The dynamic specifies the investment behavior as the opportunities are being repeated in time, determining the progression of wealth. For that goal, we will use the following definitions [7]:

- $W(t)$ is wealth in time t .
- δt is a regular time interval, e.g. 1 week.
- $r(t)$ is a return/payoff function in time t .

For example, the return function for the coin toss game from Example 1 can be defined as follows [7]:

$$r(t) = \begin{cases} 1.5 & \text{with probability } 1/2 \\ 0.6 & \text{with probability } 1/2 \end{cases} \quad (4)$$

We further assume that if we accept, we are obliged to play this game for a longer period of time (e.g. every week for 2 years). Generally, the investment behavior in this game can be either additive or multiplicative, determining the dynamic of our wealth progression.

1.3.1 Additive dynamic

Additive dynamic is simply a unit based investment, meaning that we invest a single unit (e.g. 1 dollar) at every single time step δt . Our wealth progression in time t is hence defined as:

$$W(t) = r(t) + W(t - \delta t) \quad (5)$$

Taking the above mentioned coin toss example, the “growth rate” of player’s wealth will approximately be the expected value $ev = \frac{1}{2} \cdot 1.5 + \frac{1}{2} \cdot 0.6 = 1.05$, in other words wealth growth rate under additive dynamic is ergodic [7].

1.3.2 Multiplicative dynamic

In this case, we are continuously reinvesting our current wealth over the set of the presented opportunities. Hence our wealth progression in time t is defined as follows:

$$W(t) = r(t) \cdot W(t - \delta t) \quad (6)$$

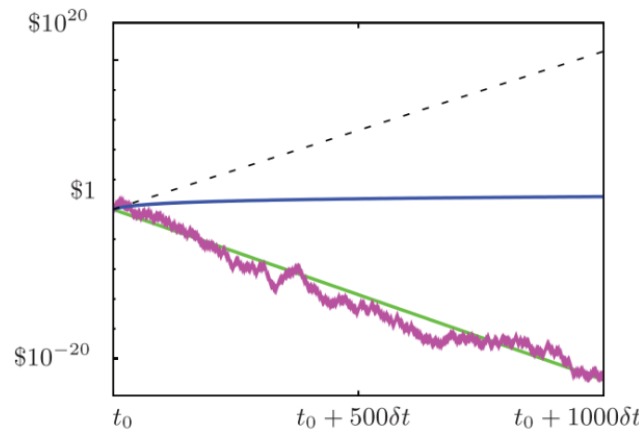


Figure 1: Multiplicative dynamic: Coin toss game. The magenta line represents median wealth trajectory in 1000 time steps of a coin toss game. The dashed line is the expected value [7].

In the Figure 1 we can see that, under multiplicative dynamic, the expected value does not indicate what happens to a single player in the “long run” [7]. Instead, the relevant quantity to look for in this case is the geometric mean $\bar{r} = (1.5 \cdot 0.6)^{\frac{1}{2}} \approx 0.95$. In other words, wealth growth rate under multiplicative dynamic is non-ergodic [7].

1.3.3 Ergodic property

Ergodic property of a dynamic process is often dubbed as the “equality of averages”, which stands for the time average [7] being equal to the ensemble average, i.e. the expected value. The growth rate of wealth is thus ergodic if its expected value is constant in time and its time average converges to this value with probability one [7].

2 Betting Strategies

In existing literature, the betting strategies range from informal “ad-hoc” approaches, such as betting according to the probability estimates, to the formal ones, represented mainly by Modern portfolio theory [6] and Kelly criterion [4].

2.1 Informal approaches

Here we list some of the informal betting strategies encountered in the literature for completeness of the review.

- Bet fixed amount on favorable odds (unif).
- Bet amount equal to the absolute discrepancy between probabilities predicted by the model and the bookmaker (abs disc bet).
- Bet amount equal to the relative discrepancy between probabilities predicted by the model and the bookmaker (rel disc bet).
- Bet amount equal to the estimated probability of winning (conf bet).
- Bet everything only on the maximum expected value opportunity (max ev).

While these informal betting methods are generally inferior to the formal approaches [2] and lack the necessary theoretical background, we do not investigate them experimentally much further in this work.

2.2 Modern Portfolio Theory

The idea behind Modern Portfolio Theory (MPT) is that portfolio \mathbf{b}_1 is superior to \mathbf{b}_2 if the expected gain $\mathbb{E}[g(\mathbf{b})]$ is at least as great.

$$\mathbb{E}[g(\mathbf{b}_1)] \geq \mathbb{E}[g(\mathbf{b}_2)] \quad (7)$$

and the risk, here generally denoted through a risk measure r , is no greater [6].

$$r(\mathbf{b}_1) \leq r(\mathbf{b}_2) \quad (8)$$

This creates a partial ordering on the set of all available portfolios. Taking the portfolios that no other portfolio is superior to gives us the set of efficient portfolios Θ . In simple terms we maximize the following:

$$\mathbb{E}[gain] - \gamma \cdot risk \quad (9)$$

In the most common setup, the risk of a portfolio is measured by its variance defined through a covariance matrix Σ . MPT can then be expressed as the following maximization problem:

$$\begin{aligned} & \underset{\mathbf{b}}{\text{maximize}} && \boldsymbol{\mu}^T \mathbf{b} - \gamma \mathbf{b}^T \Sigma \mathbf{b} \\ & \text{subject to} && \sum_{i=1}^K b_i = 1.0, \quad b_i \geq 0 \end{aligned}$$

where \mathbf{b} is portfolio, γ is risk aversion parameter and $\boldsymbol{\mu}$ is the expected gains vector of the offered opportunities.

2.2.1 Maximum Sharpe strategy

Sharpe ratio can be used as a criterion to choose from the set of efficient portfolios Θ . Sharpe ratio of a portfolio is defined as [2]:

$$\frac{r_p - r_f}{\sigma_p}$$

Where σ_p is standard deviation of portfolio return, r_p is expected return of the portfolio and r_f is a risk-free rate (we can neglect the risk free rate if there is no risk-free method how to appreciate the money). We can hence define a separate “MaxSharpe” strategy as follows:

$$\begin{aligned} & \underset{\mathbf{b}}{\text{maximize}} && \frac{\mu \mathbf{b}}{\sqrt{\mathbf{b}^T \Sigma \mathbf{b}}} \\ & \text{subject to} && \sum_{i=1}^K b_i = 1.0 \\ & && b_i \geq 0 \end{aligned}$$

2.2.2 Criticism

The MPT approach is often criticised for two main reasons, [7].

1. The growth rate of wealth under multiplicative dynamic is non-ergodic. In other words, the expected return does not tell us what will really happen in the long run, hence maximizing it does not produce a truly long term-optimal reinvestment strategy, which we showcased in the coin toss example 1.
2. The definition of risk as variance is often disputed. In many domains, risk is not easy to define.

2.3 Growth optimal strategy

A different approach to betting is to first make the growth rate of wealth ergodic, under specified dynamic using appropriate transformation, and then take the expected value. This approach is famously known as the “Kelly Criterion”, “Geometric Mean Policy” and under many other names.

For the multiplicative dynamic, the correct ergodicity transformation is the logarithm: $v(x) = \log(x)$ [7]. The growth optimal portfolio for a general case can hence be found by solving the following optimization problem:

$$\begin{aligned} & \underset{\mathbf{b}}{\text{maximize}} && \mathbb{E}[\log(\mathbf{R} \cdot \mathbf{b})] \\ & \text{subject to} && \sum_{i=1}^K b_i = 1.0, b_i \geq 0 \end{aligned}$$

The calculated portfolio is then growth-optimal under the following assumptions [4, 8, 7]:

1. True probability is known to the player.
2. There is a “long run” of “approximately similar” games.

3 Experiments

We showcase the above mentioned strategies on a dataset of 30000 soccer games from different leagues all over the world. The dataset consists of opening odds, closing odds, estimated outcome probabilities and results indicating the outcome realization in each game. The probabilities are modelled separately using a neural network model. Naturally, the true probabilities of the outcomes are unknown to both to the player and the bookmaker, which has a significant impact on how the strategies are applied, especially in the multiplicative dynamic scenario.

We have three outcomes per game and we always assume multiple simultaneous games happening simultaneously in each time step, which we will refer to as “round”. We investigate results from both the additive and multiplicative perspective.

In [6] the author proposes multiple measures of dispersion (risk measures) that can possibly be used to evaluate the investment strategy such as variance, standard deviation and coefficient of variation. In our evaluation framework we choose the standard deviation of portfolio return as the risk measure.

3.1 Additive dynamic

If our goal would be solely to maximize the expected profit, then the solution would be a trivial Bayesian strategy – bet the whole wealth on the opportunity with highest expected value of profit from presented opportunities. However, even under additive dynamic, we want some level of guarantee of not loosing too much money, hence our criteria for selecting the betting strategy are the expected profit and also the risk.

To demonstrate the differences between the betting strategies we randomly sampled data representing six illustrative betting opportunities with positive expected value [3]. Then we applied the betting strategies and analyzed the differences in the wealth allocations by the different strategies. Results are summarized in Table 1 (we note that the growth optimal approach only makes sense under the multiplicative dynamic, hence we do not analyse it in this sub-section). The strategies based on the absolute and relative discrepancies between the probability estimates always prefer higher expected profit of the opportunity regardless of the variance of the profit. On the other hand, the strategy based on the confidence of the probability estimate always prefers lower risk, regardless of the return. The MaxSharpe strategy is looking for a compromise between these two approaches.

#	p_M	p_B	σ	ev	MaxEV	Unif	Abs_disc	Rel_disc	Conf	MaxSharpe
1	0.30	0.26	1.80	1.19	1.0	0.17	0.20	0.35	0.08	0.09
2	0.59	0.52	0.94	1.12	0	0.17	0.27	0.24	0.16	0.23
3	0.75	0.70	0.62	1.07	0	0.17	0.21	0.15	0.21	0.30
4	0.60	0.57	0.86	1.06	0	0.17	0.13	0.12	0.17	0.12
5	0.74	0.71	0.62	1.04	0	0.17	0.11	0.08	0.20	0.17
6	0.64	0.62	0.77	1.03	0	0.17	0.07	0.06	0.18	0.08

Table 1: Comparison of betting strategies on simulated betting opportunities. The columns p_M and p_B represent the probability estimates of the model and the bookmaker, respectively, σ and ev refer to standard deviation and expected value of the opportunity [3].

The paper [3] provides a simulation of all the above mentioned informal strategies on a dataset of 5000 basketball games. The author comes to the conclusion that the only relevant

strategies are MaxEV and MaxSharpe. While MaxEV directly maximizes the expected profit, it completely ignores the risk and the possibility of ruin. Hence, we do not consider MaxEV to be a “reasonable” betting strategy, and we will further focus solely on the MaxSharpe strategy in our analysis of additive dynamic.

3.1.1 Simultaneous games

In the real world, there is almost always multiple soccer games occurring in any given moment. Hence it is reasonable to assume and conduct our simulation in such a way so as to test our strategies under this assumption. The number of simultaneous games i.e. “round size” clearly affects the number of presented opportunities. The more soccer games we can bet on, the more opportunities are presented to us. In this subsection we investigate whether round size significantly affects the overall cumulative profit.

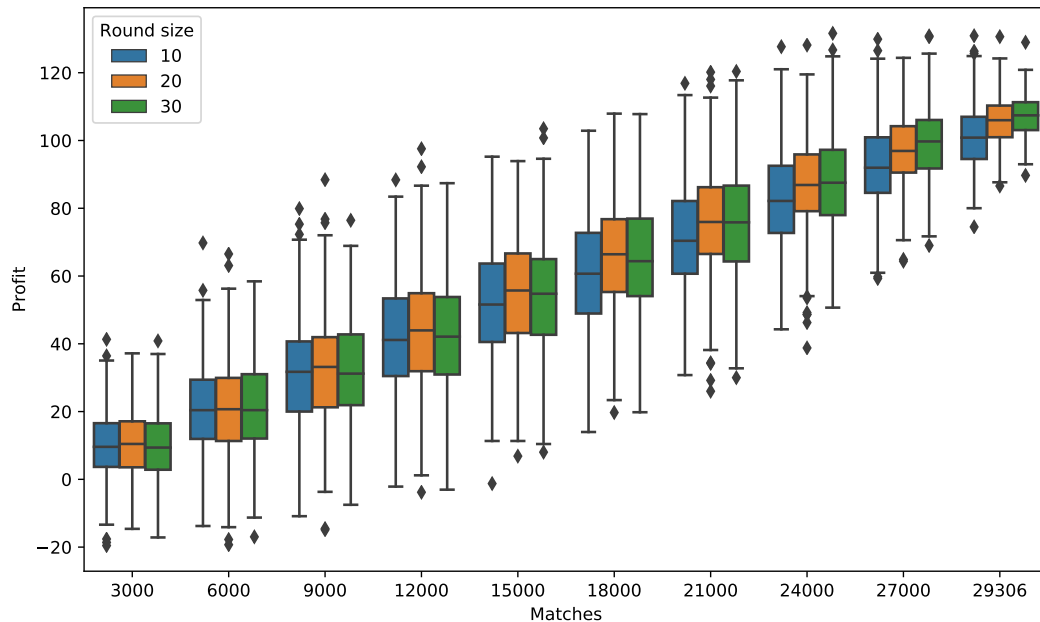


Figure 2: Simultaneous games: Effect of round size on profit under additive dynamic.

3.2 Multiplicative dynamic

In this section we focus on the reinvestment scenario. Our goal is hence to not only evaluate the presented opportunities, but to evaluate them in context of the resources available to us, i.e. our “bank”. We present two strategies that produced the best results in our experiments namely MaxSharpe strategy and the growth optimal Kelly criterion. We omit the informal approaches from the subsection 3.1 as they are insufficient for reinvestment scenario.

3.2.1 Fractional MaxSharpe

The main idea behind any “fractional” approach is to bet only a fraction ω of the calculated portfolio and leave the rest $1 - \omega$ in the cash asset for security. Here we investigate results of the previously mentioned MaxSharpe strategy. We define a trade-off index ω for a portfolio as:

$$\mathbf{b}_\omega = \omega \mathbf{b}_s + (1 - \omega) \mathbf{b}_c \quad (10)$$

where the \mathbf{b}_s stands for portfolio suggested by MaxSharpe strategy and \mathbf{b}_c is a portfolio where the only investment is a risk-free, “cash” investment.

Fraction, or a trade-off index, ω is now a lever between “growth” and “security” [5]. We hence fit the parameter ω on a training set and verify it on the testing set. We are looking for a reinvestment strategy that satisfies the following maximization problem across all the training dataset wealth trajectories.

$$\begin{aligned} & \text{maximize} && \text{median}(\mathbf{W}_F) \\ & \text{subject to} && Q_5 > 0.95 \end{aligned}$$

We are hence looking for a strategy that reaches the maximum median final wealth across all wealth trajectories, with the 5th percentile being the value below which 5% of all the wealth positions may be found. For the fractional MaxSharpe strategy this criterion yields a fraction $\omega = 0.11$

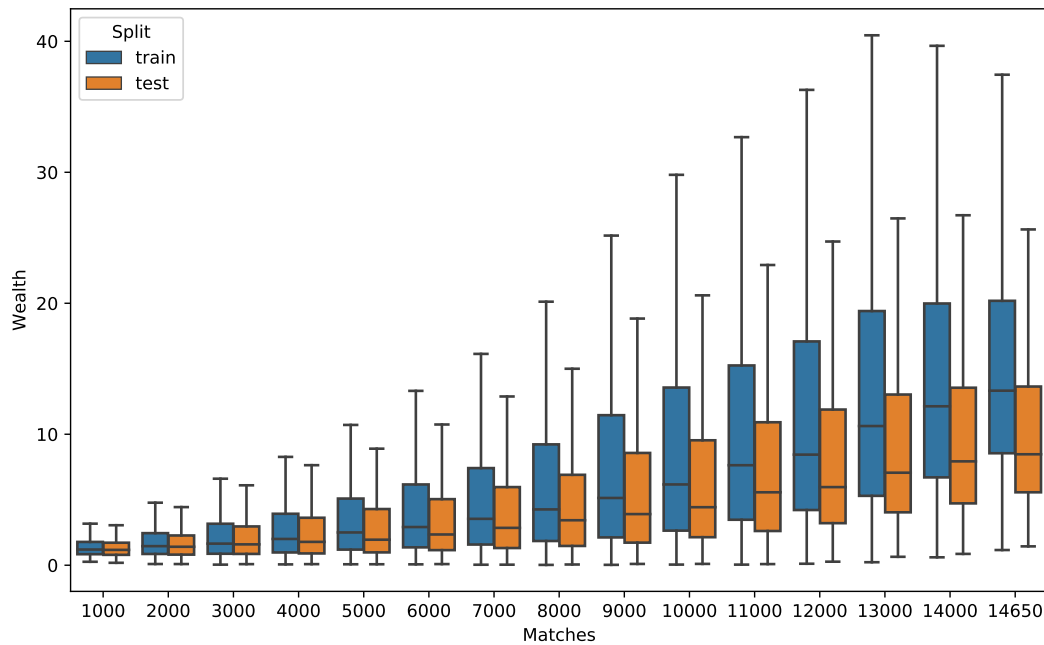


Figure 3: Results of the risk constrained growth optimal strategy on football dataset.

3.2.2 Risk constrained Kelly

In this case we adjust the growth optimal portfolio \mathbf{b} using the following maximum drawdown constraint.

$$P(W^{MIN} < \alpha) \leq \beta \quad (11)$$

representing the probability of our wealth falling below α being at most β . This constraint is approximately satisfied if the following is satisfied [1]:

$$\mathbb{E}[(\mathbf{R} \cdot \mathbf{b})^{-\lambda}] \leq 1 \text{ where } \lambda = \log(\beta) / \log(\alpha) \quad (12)$$

We hence fit the parameter λ on a training set and verify it on the test set using the same criterion as in the previous section. For the risk constrained Kelly the parameter is $\lambda = 9.4$.

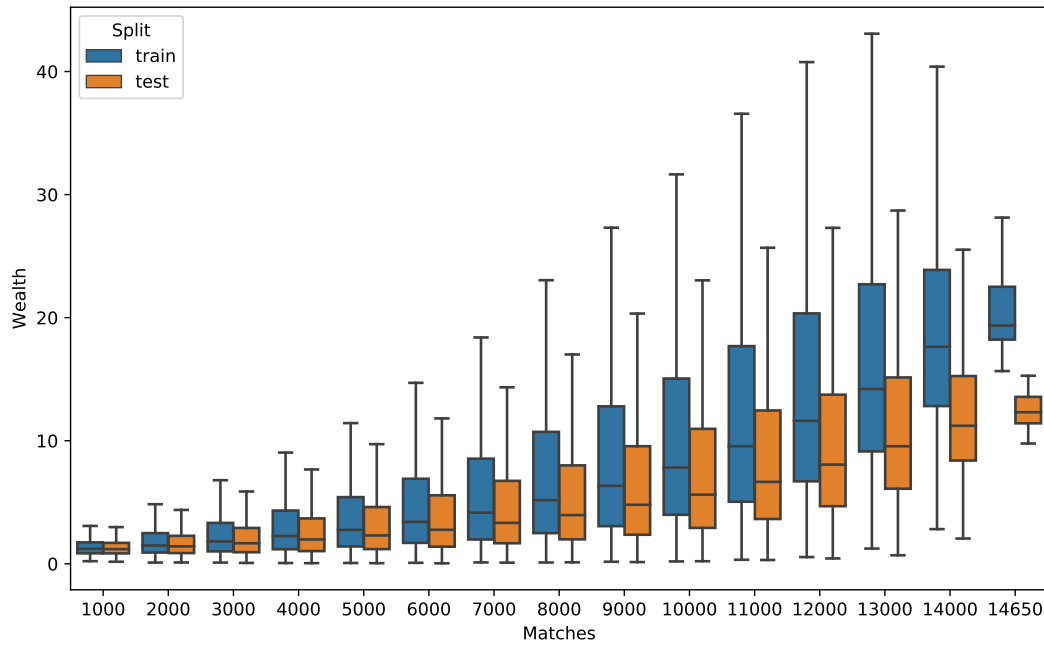


Figure 4: Results of the risk constrained growth optimal strategy on football dataset.

4 Results

We conducted experiments on a dataset of football games and our findings are as follows.

- The round size indeed does matter. Suppose that there are too few games taking place in a given day (e.g. 5). In such a case, the round is not representative enough and the strategy has too few opportunities to compare against each other, which leads to a lower

diversification of risk. On the other hand, the more games we are able to aggregate, the smaller the differences between the rounds, which as a result leads to a higher stability of the betting strategy.

- Both fractional MaxSharpe and risk constrained Kelly reach very similar results over the football dataset.

5 Conclusion

In this paper we review the most widely used betting strategies both theoretically and practically. From theoretical point of view we discuss the unique properties of the additive dynamic and the multiplicative dynamic, in other words the unit based and reinvestment scenario. Generally, we argue for the formal approaches as opposed to the informal ones often encountered in literature.

From the practical point of view, we showcase application of the approaches in a relevant sports domain, with several ideas on how to tackle one of its most difficult challenges – the uncertainty of probability estimates.

References

- [1] Enzo Busseti, Ernest K Ryu, and Stephen Boyd. Risk-constrained kelly gambling. *arXiv preprint arXiv:1603.06183*, 2016.
- [2] Ondrej Hubacek. Exploiting betting market inefficiencies with machine learning. Master’s thesis, Czech Technical University, Faculty of Electrical Engineering, 2017.
- [3] Ondřej Hubáček, Gustav Šourek, and Filip Železný. Exploiting sports-betting market using machine learning. *International Journal of Forecasting*, 35(2):783–796, 2019.
- [4] John L Kelly Jr. A new interpretation of information rate. In *The Kelly Capital Growth Investment Criterion: Theory and Practice*, pages 25–34. World Scientific, 2011.
- [5] LC MacLean, William T Ziemba, and George Blazenko. Growth versus security in dynamic investment analysis. In *The Kelly Capital Growth Investment Criterion: Theory and Practice*, pages 331–354. World Scientific, 2011.
- [6] Harry Markowitz. Portfolio selection. *The journal of finance*, 7(1):77–91, 1952.
- [7] Ole Peters and Murray Gell-Mann. Evaluating gambles using dynamics. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 26(2):023103, 2016.
- [8] Edward O Thorp. The kelly criterion in blackjack sports betting, and the stock market. In *Handbook of asset and liability management*, pages 385–428. Elsevier, 2008.

ISBN 978-618-5036-53-9

