



**MATHSPORT INTERNATIONAL 2017  
CONFERENCE**

**- Proceedings -**

*University of Padua*

*Botanical Garden 26-28 June 2017*



DIPARTIMENTO  
**MATEMATICA**



UNIVERSITÀ  
DEGLI STUDI  
DI PADOVA



Original title Proceedings of MathSport International 2017 Conference

Editors: Carla De Francesco, Luigi De Giovanni, Marco Ferrante, Giovanni Fonseca, Francesco Lisi, Silvia Pontarollo

© 2017 Padova University Press  
Università degli Studi di Padova  
via 8 Febbraio 2, Padova  
[www.padovauniversitypress.it](http://www.padovauniversitypress.it)

ISBN 978-88-6938-058-7

All rights reserved.

## Contents

<b>A. Adhikari, R. Saraf and R. Parma</b>	
Bowling strategy building in limited over cricket match: An application of statistics	1-10
<b>K. Andreeva</b>	
Manager's capacity and limit on foreign players: what influences variation of players in the field?	11-21
<b>M. Angonese, F. Cardin and P. Ferrandi</b>	
Performance Evaluation of Volleyball Serve using Dynamical Models of Flight Trajectory	22-31
<b>R. Arboretti, E. Carrozzo and L. Salmaso</b>	
Ranking ski courses using permutation methods for multivariate populations	32-38
<b>M. Asif and I. G. McHale</b>	
Estimating Margin of Victory in Twenty-20 International Cricket	39-45
<b>F. Bortolon, C. Castiglione, L. Parolini and L. Schiavon</b>	
A Markovian approach to darts	46-60
<b>P. J. Bracewell, A. K. Patel and J. D. Wells</b>	
Real time measurement of individual influence in T20 cricket	61-70
<b>P. Brown, P. J. Bracewell and A. K. Patel</b>	
Optimising a Batting Order in Limited Overs Cricket using Survival Analysis	71-80
<b>G. Budak, İ. Kara, Y. T. İç and R. Kasimbeyli</b>	
Optimization of Harmony in Team Formation Problem for Sports Clubs: A real life volleyball team application	81-86
<b>C. Casella and P. Vidoni</b>	
Formula 1 lap time modeling using generalized additive models	87-96
<b>P. Coletti and A. Pilkington</b>	
A comparison of the Performance of the generalized Google PageRank model (GeM) to other Ranking Systems on the NCAA Basketball Tournament	97-111
<b>L. Egidi and J. S. Gabry</b>	
Bayesian hierarchical models for predicting individual performance in football (soccer)	112-121
<b>M. Ferrante, G. Fonseca and S. Pontarollo</b>	
How long does a tennis game last?	122-130
<b>M. Fichman and J. O'Brien</b>	
Optimal Shot Selection Strategies for the NBA	131-140

<b>C. Frankland and G. Hunter</b>	
A Statistical Investigation of Factors Influencing the Results of One-Day Internationals in Cricket	141-150
<b>Y. Gerchak and E. Khmelnitsky</b>	
On Reducing Sequence Effects in Competitions	151-160
<b>A. Groll, T. Kneib, A. Mayr and G. Schauburger</b>	
On the Dependency of Soccer Scores - A Sparse Bivariate Poisson Model for the UEFA European Football Championship 2016	161-175
<b>N. Hirotsu, K. Inoue and M. Yoshimura</b>	
An analysis of characteristics of soccer teams using a Markov process model considering the location of the ball on the pitch	176-183
<b>S. Hoffmeister and J. Rambau</b>	
Sport Strategy Optimization in Beach Volleyball - How to bound direct point probabilities dependent on individual skills	184-193
<b>K. Jackson</b>	
Using Player Quality and Injury Profiles to Simulate Results in Australian Football	194-203
<b>E. Konaka</b>	
A statistical rating method for team ball games and its application to prediction in the Rio Olympic Games	204-216
<b>S. Kovalchik and M. Ingram</b>	
Estimating the Duration of Professional Tennis Matches with Varying Formats	217-228
<b>C. Liti, V. Piccialli and M. Sciandrone</b>	
Predicting soccer match outcome using machine learning algorithms	229-237
<b>M. de Lorenzo, S. Stylianou, I. Grundy and B. O'Bree</b>	
Draw importance in football	238-243
<b>P. Marek and F. Vávra</b>	
Home Team Advantage in English Premier League	244-254
<b>K. Marshall</b>	
The effect of leadership on AFL team performance	255-264
<b>R. Metulini, M. Manisera and P. Zuccolotto</b>	
Sensor Analytics in Basketball	265-274
<b>D. Meyer, S. Muir, M. Weerasinghe Jayawardana, D. Ho and O. Sackett</b>	
Evaluation of a corporate physical activity program using mixed methods	275-289

<b>D. Mon and A. Díaz</b>	
Which one has more influence in female air pistol performance: experience or training?	290-294
<b>A. Owen</b>	
The Application of Hurdle Models to Accurately Model 0-0 Draws in Predictive Models of Football Match Outcomes	295-304
<b>D. A. Perdomo Meza</b>	
Flow Network Motifs Applied to Soccer Passing Data	305-319
<b>F. Salassa and F. Della Croce</b>	
The Social Doubles Tournament Problem	320-329
<b>J. Schönberger</b>	
The Championship Timetabling Problem - Construction and Justification of Test Cases	330-339
<b>R. Stefani</b>	
What Performance Data Tells Us about PEDs in Olympic Athletics and Swimming	340-346
<b>R. A. Syme</b>	
The Nappy Factor in Golf: The Effect of Children on the Sporting Performance of Professional Golfers	347-356
<b>T. Toriumi</b>	
Compare the superiority of Japanese Collegiate Baseball Leagues	357-362
<b>J. A. Trono</b>	
Applying Occam's Razor to the Prediction of the Final NCAA Men's Basketball Poll	363-369

# **Bowling strategy building in limited over cricket match: An application of statistics**

Akash Adhikari, Rishabh Saraf and Rishikesh Parma

Indian Institute of Technology (ISM), Dhanbad, Jharkhand, India.

rajaadhikari23@gmail.com, rishabh.15je1745@am.ism.ac.in, rishikesh.parma9@gmail.com

## **Abstract**

Cricket has been one of the oldest form of sports played around the globe. Though not as popular as Football it has its' own charm and spectators do get frenzy with every ball to the bat. This bat and ball game has always attracted Mathematicians and Statisticians for the enormous scope of research to improve the game. Our idea involves implementation of statistical operations and data analysis, to deduce a Dominance factor, that will facilitate to select the most efficient bowler in a limited overs cricket match. The factor will ultimately simplify strategy building for the corresponding team. The research includes analysis of data of all bowlers in their past matches. The data used span the years 2007 to 2016. Assessing the statistics of the data, two important parameters: Runs conceded in between fall of consecutive wickets (RBW) and number of deliveries bowled in between fall of consecutive wickets (BBW) in a spell of the respective bowler, are evaluated. We have used these parameters to calculate a factor (Dominance factor) for the bowlers which are among the top 30 rankings (ICC ODI rankings) as of 2016. The Dominance factor will sort the bowlers in terms of priority for a particular over span. This factor will be of great help for respective teams to decide their bowling strategy. The result can be of immense use to bid for bowlers during cricket league auctions, to make an optimum bowling unit for the team. In addition to this, people who frequently play bets and are very much enthusiastic about gambling can make good use of the Dominance Factor to decide the bowler to bet on. Apart from the Dominance factor, we have done detail study about the variation pattern in bowling of the bowlers. We compared the characteristics of closely ranked bowlers in part, and discussed contradictions and supports to these rankings, on the basis of the parameters (BBW and RBW) and other statistical operations.

## **1 Introduction**

Cricket has always been an amusing game for public in general and statisticians in particular. Given its popularity, its slow pace, its length of play time, each game of cricket throws up a huge amount of performance related statistics and provide enormous data set to analyze on.

Cricket has been one of those sports which has been evolving with time. With an intention to increase excitement and entertainment this game has attracted huge crowds in following years. The formats of the game include Test cricket which can go as long as five days, One day International (ODI) comprising of 50 overs and T20 which limits to 20 overs. Moreover the introduction of new rules for example batting and bowling powerplay (a field restriction where only limited number of fielders are allowed outside the

30 yards circle around the batsman) have made the game an interesting one. Producing statistical and/or computational models which would enable reliable predictions of match results or scores, and probabilities relating to these, could be of interest to the cricketing, gambling and legal authorities.

In this paper, we have focused on the bowler's performance in the matches and attempted to build a strategy based on their performance. The two parameters BBW and RBW are used to formulate a factor called Dominance factor. A separate analysis of a bowler's performance in 1-30 overs and 31-50 overs is done in order to build a strategy to use a more efficient bowler in that very period where his Dominance factor is numerically higher among the other bowlers under consideration.

## 2 Need for Strategy Building

In a limited overs cricket match, a bowler is constrained to bowl maximum of 10 overs (ODI) or 4 overs (T20). These are the limitations on deliveries for a bowler. In the early hours of the game it is very important for the fielding team to get early breakthroughs to avoid any long partnerships from the batting side. Whereas in the death overs bunch of dot balls are quite satisfying. The present statistic mainly used to decide the bowling approach is the *bowling average* which is defined as total number of runs conceded by the bowler divided by the number of wickets obtained. The second is the *economy rate* which is defined as the total number of runs conceded by the bowler divided by the number of overs bowled. The third is the bowler's *strike rate* which is defined as the total number of balls bowled divided by the number of wickets obtained. However, these statistics are individually deficient as they do not adequately account for overs, wickets and runs respectively.

Previous works done for better precision in calculating a bowler's performance, includes calculating a performance indicator by Attanayake and Hunter (2015), in which the performance of the bowler is calculated by the formula  $\text{Performance} = (\text{Runs conceded})^\alpha (\text{Balls bowled})^\beta (\text{Wickets taken})^\gamma$  where  $\alpha, \beta, \gamma$  are some constants for each model. The traditional *bowling average* has  $\alpha=1, \beta=0, \text{ and } \gamma=-1$ , and the *bowling economy rate* have  $\alpha=1, \beta=-1, \text{ and } \gamma=0$ . Another paper to evaluate batting performance by Damodaran (2006) uses the concept of Stochastic Dominance to calculate a better batting average by finding the conditional average for the "not out" scores of the batsman. However, very less exploration has been done in detailing the bowling statistics. When the name cricket strikes the first thing which anyone asks is who is your favorite batsman? delving into something out of the box, we decided to focus on bowling statistics.

## 3 Data and Method

The performance of a bowler in context of one-day cricket relates differently when he bowls in the first 30 overs, Set 1 (1-30) and in the last 20 overs, Set 2 (31-50). We have proposed a new basis to judge a bowler's performance, to evaluate it distinctly in this two set of overs. The probability (P) to take wickets for any bowler is calculated distinctly in these two sets. We have plotted scatter graph for these calculated probabilities. Cumulative distributive function (cdf) is used to generate the formula for probability (P). The data for five best rated Slow-arm spinners and Fast-arm seamers (as per ICC ODI rankings 2016) who

have played ODI cricket at some point of time during 2007-2016 have been used in the analysis. For each bowler every match played by him during the period have been included in the analysis.

Say,  $w_i$  = total number of wickets taken by the bowler,  $w_i = i$ th wicket ( $1 \leq i \leq w_i$ ),  $b_i$  = number of balls delivered between wickets  $w_i$  and  $w_{i+1}$ ,  $r_i$  = number of runs conceded between wickets  $w_i$  and  $w_{i+1}$ . The procedure that is adopted in the paper is to first compute the value of BBW( $b_i$ ) by calculating the number of balls delivered between two consecutive wickets and RBW( $r_i$ ) by calculating the runs conceded between two consecutive wickets, these computation is continued till the last wicket of the bowler during the considered span ('Run-Out' is not considered). Also,  $b_k = k$  ( $1 \leq k \leq \max(b_i)$ ), and  $r_n = n$  ( $0 \leq n \leq \max(r_i)$ ). The next computation includes to calculate the probability to take wicket given the number of balls, which is defined as  $P(b_k)$  and the probability to take wicket given the number of runs, which is defined as  $P(r_n)$ . The following formulas have been defined for calculating these probabilities

$$P(r_n) = F(r_n) \div w_i + P(r_{n-1}). \tag{1}$$

$$P(b_k) = F(b_k) \div w_i + P(b_{k-1}). \tag{2}$$

where  $P(b_0) = 0$ , Frequency of  $b_n$  [ $F(b_n)$ ] = number of occurrences of  $b_n$  among all the values  $b_i$ , Frequency of  $r_n$  [ $F(r_n)$ ] = number of occurrences of  $r_n$  among all the values  $r_i$  ( $1 \leq i \leq w_i$ ).

These probability values are calculated independently, considering the effect of only one parameter at a time. The data that has been used corresponds to ODI cricket statistics, since the maximum number of permitted balls for a bowler in ODI cricket is 60, so the value of  $P(b_k)$  is calculated for the domain [0,60], Assuming that the maximum runs a bowler may concede in a match is 80, the value of  $P(r_n)$  is calculated for the domain [0,80]. The value of  $P(b_k)$  is calculated with respect to the variation in  $b_k$  neglecting the effect of  $r_n$  and vice-versa. To study these observations further, we plotted these values, to understand the dominance of a bowler provided any one parameter ( $b_k$  or  $r_n$ ) is considered irrespective of the effect of the other parameter. Considering two different types of bowlers – Slow-arm spinners and Fast-arm seamers, we have divided our study in these two categories. The above formula is implemented to calculate the probabilities of the bowlers in these two categories. As mentioned before, the overs under consideration are the two sets Set 1 (1-30) and Set 2 (31-50).

### 3.1 Slow-arm spinners

The bowlers which have been included in the study are as follows :





Bowlers	Indicators
Imran Tahir	
SP Narine	
Shakib Al Hasan	
R Ashwin	
RA Jadeja	



Table 1.a Values of  $P(b_k)$

$b_k$	SET 1 (1-30 overs)					SET 2 (31-50 overs)				
	Imran Tahir	SP Narine	S Hasan	R Ashwin	RA Jadeja	Imran Tahir	SP Narine	S Hasan	R Ashwin	RA Jadeja
0	0	0	0	0	0	0	0	0	0	0
10	0.1694	0.2244	0.2043	0.1194	0.2000	0.4130	0.3265	0.2500	0.2666	0.3164
20	0.4745	0.4081	0.3978	0.2537	0.2923	0.6304	0.4693	0.4642	0.4800	0.4810
30	0.6779	0.5102	0.4838	0.3582	0.4153	0.6956	0.6734	0.5476	0.6400	0.6329
40	0.7796	0.6122	0.6344	0.5074	0.5076	0.8043	0.7551	0.6190	0.8266	0.7468
50	0.8135	0.6938	0.6774	0.6119	0.6307	0.8913	0.7755	0.7619	0.8800	0.7848
60	0.8474	0.7551	0.7419	0.6865	0.6769	0.9347	0.8367	0.8452	0.9600	0.8860

Table 1.b Values of  $P(r_n)$

$r_n$	SET 1 (1-30 overs)					SET 2 (31-50 overs)				
	Imran	SP Narine	S Hasan	R Ashwin	RA Jadeja	Imran Tahir	SP Narine	S Hasan	R Ashwin	RA Jadeja
0	0.0338	0.0816	0.1075	0.0298	0.0153	0.1956	0.1020	0.0595	0.0533	0.1012
10	0.3559	0.4081	0.3440	0.2089	0.2307	0.4565	0.4285	0.3214	0.3600	0.4050
20	0.6610	0.5918	0.5591	0.4179	0.4153	0.6521	0.6938	0.5000	0.5733	0.5949
30	0.7796	0.6734	0.6666	0.5522	0.5384	0.7826	0.7346	0.5714	0.6666	0.7088
40	0.8474	0.7551	0.7634	0.5970	0.6307	0.8913	0.7959	0.7261	0.8000	0.7594
50	0.8644	0.8571	0.8064	0.7611	0.6923	0.9347	0.8979	0.8333	0.9066	0.8354
60	0.9152	0.8775	0.8709	0.8208	0.7534	0.9347	0.9387	0.8809	0.9600	0.8987
70	0.9152	0.9387	0.9247	0.8656	0.7846	0.9565	0.9795	0.9047	0.9733	0.9493
80	0.9322	0.9591	0.9354	0.8805	0.8307	0.9565	0.9795	0.9642	0.9999	0.9873

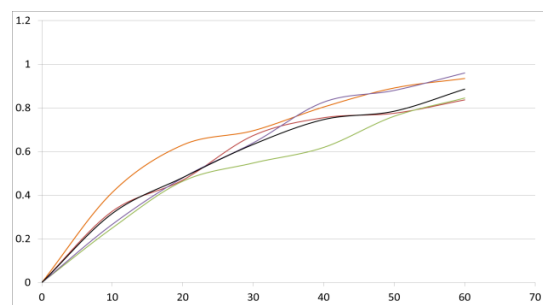
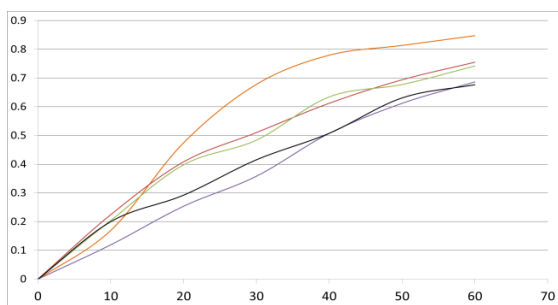


Figure 1. and 2. Dominance curve for different bowlers for SET 1 and SET 2 with  $P(b_k)$  as ordinate and  $b_k$  as abscissa.

Bowling strategy building in limited over cricket match

Adhikari, Saraf and Parma

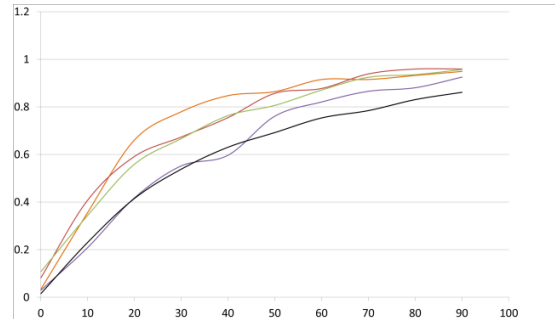
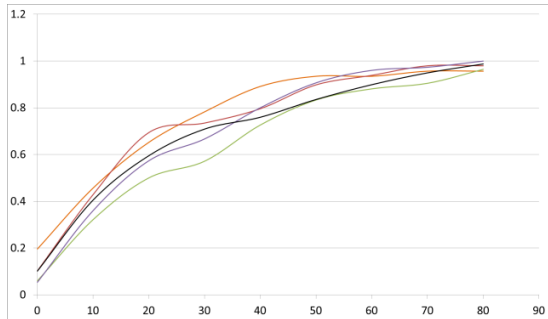


Figure 3. and 4. Dominance curve for different bowlers for SET 1 and SET 2 with  $P(r_n)$  as ordinate and  $r_n$  as abscissa.

### 3.2 Fast-arm seamers

The bowlers which have been included in the study are as follows :






Bowlers	Indicators
MA Starc	
TA Boult	
DW Steyn	
M Morkel	
TG Southee	

Table 2.a Values of  $P(r_n)$

$r_n$	SET 1 (1-30 overs)					SET 2 (31-50 overs)				
	MA Starc	TA Boult	D Steyn	Morkel	TG Southee	MA Starc	TA Boult	D Steyn	Morkel	TG Southee
0	0.2156	0.1794	0.0405	0.0632	0.0933	0.1600	0.142857	0.125	0.1000	0.05556
10	0.4117	0.4102	0.2837	0.3544	0.3733	0.5800	0.5238	0.4167	0.4500	0.3889
20	0.5882	0.5384	0.5540	0.5189	0.6400	0.8400	0.7619	0.6458	0.6800	0.537
30	0.7058	0.6666	0.5945	0.6708	0.6933	0.9000	0.8571	0.7604	0.8300	0.7222
40	0.8431	0.8461	0.6351	0.7341	0.7866	0.9000	0.9048	0.9063	0.9000	0.8148
50	0.8823	0.9230	0.7702	0.8101	0.8400	0.9400	0.9048	0.9479	0.9600	0.9259
60	0.9411	0.9230	0.8513	0.8987	0.8533	0.9800	0.9524	0.9479	0.9800	0.9444
70	0.9607	0.9487	0.8918	0.8987	0.8800	0.9999	0.9999	0.9688	0.9900	0.9444
80	0.9607	0.9487	0.9189	0.9113	0.8933	0.9999	0.9999	0.9688	0.9990	0.9441

Table 2.b Values of  $P(b_k)$

$b_k$	SET 1 (1-30 overs)					SET 2 (31-50 overs)				
	MA Starc	TA Boult	D Steyn	Morkel	TG Southee	MA Starc	TA Boult	D Steyn	Morkel	TG Southee
0	0	0	0	0	0	0	0	0	0	0
10	0.2941	0.2820	0.1891	0.1645	0.2533	0.4200	0.4671	0.3958	0.3800	0.3148
20	0.4705	0.4871	0.3918	0.3797	0.4666	0.7200	0.7619	0.6458	0.6900	0.537
30	0.6078	0.6153	0.5270	0.4936	0.6133	0.9000	0.8095	0.7500	0.8300	0.7407
40	0.6862	0.7435	0.5810	0.5569	0.6800	0.9200	0.8571	0.8958	0.9200	0.8333
50	0.7450	0.8205	0.6621	0.7088	0.7733	0.9999	0.9047	0.9375	0.9400	0.9444
60	0.7843	0.8717	0.6891	0.7721	0.8133	0.9999	0.9999	0.9479	0.9800	0.9444

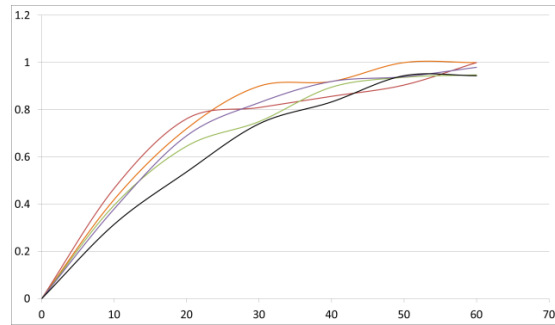
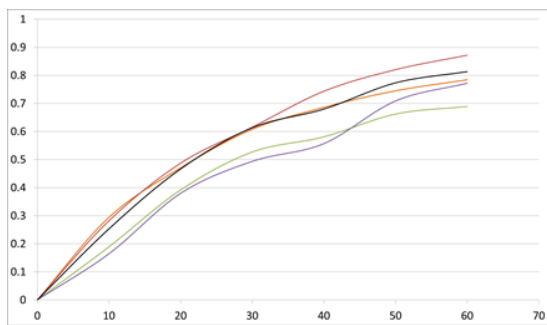


Figure 5. and 6. Dominance curve for different bowlers for SET 1 and SET 2 with  $P(b_k)$  as ordinate and  $b_k$  as abscissa.

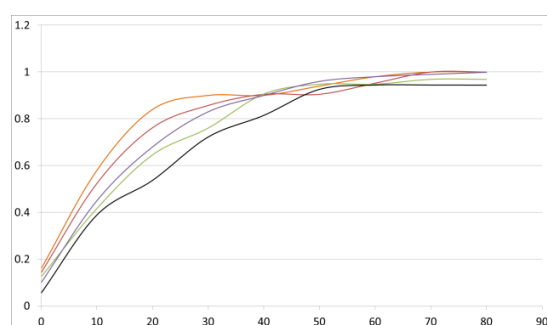
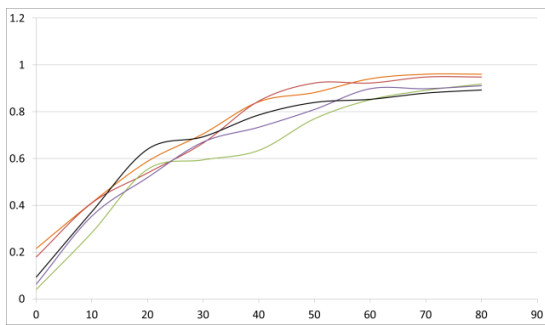


Figure 3. and 4. Dominance curve for different bowlers for SET 1 and SET 2 with  $P(r_n)$  as ordinate and  $r_n$  as abscissa.

## 4 Discussions

We infer from these computations that the probability for a bowler to take a wicket given the number of balls he can deliver  $P(b_k)$ , or the runs he can concede  $P(r_n)$  does not entirely support each other. In fact, we have two independent parameters ( $b_k$  or  $r_n$ ) to decide which is more required in accordance with the situation of a match. For instance, Team A is defending a target of 150 runs in a one-day cricket match,

and the opponent Team B needs just 20 runs in the last 5 overs with 6 wickets remaining, in such case, a bowler with higher probability ( $P$ ) to take wicket with respect to  $b_k$  will be preferred against  $r_n$  i.e.  $P(b_k)$  will be preferred over of  $P(r_n)$ . Similarly, there may be other instances where  $P(r_n)$  will be more optimum to select the bowler. However there may be instances where an exclusive preference for either of the parameters does not exists, and there is a need to consider the effect of both of the parameters. Due to this limitation, that the probability to take a wicket will depend partially on both the parameters, we defined Striking Score of a bowler which includes the effect of both the parameters  $b_k$  and  $r_n$  in line with the probability of taking wickets. Striking Score is defined in such a manner so that it includes the effect of  $b_k$ ,  $r_n$  and the probability to take a wicket ( $P$ ) after delivering  $b_k$  number of balls. Given  $b_k$ ,  $P$  is equal to  $P(b_k)$ , and  $r_n$  is equal to that value for which  $P(b_k) = P(r_n)$ . Thus striking score ( $Y$ ) as a function of  $b_k$  is,

$$Y = P + b_k \div r_n. \quad (3)$$

Where,  $P = P(b_k) = P(r_n)$ . The weighted average of striking score ( $Y$ ) with respect to  $b_k$  was calculated. This average is defined as the Dominance Factor.

$$D.F. = \{\sum(Y * b_k)\} \div \{\sum b_k\}. \quad (4)$$

Table 3.a Striking score and Dominance factor of Slow-arm spinners for Set 1 (1-30)

SET 1	STRIKING SCORE (Y)				
	$b_k$	IMRAN TAHIR	SP NARINE	S HASAN	R ASHWIN
10	2.6694	2.7244	2.7043	2.1194	1.3111
20	1.8079	2.6303	1.8264	1.7921	1.9589
30	1.9823	2.5102	2.2485	2.1229	1.9153
40	2.1589	2.5170	2.1158	1.8867	2.0461
50	2.3287	2.2563	2.2903	1.8314	1.9128
60	2.4264	2.4217	2.3208	1.8314	1.9535
D.F.	2.2414	2.4474	2.2353	1.9391	1.9259

Table 3.b Striking score and Dominance factor of Slow-arm spinners for Set 2 (31-50)

SET 2	STRIKING SCORE (Y)				
	$b_k$	IMRAN TAHIR	SP NARINE	S HASAN	R ASHWIN
10	2.0797	1.5765	1.9166	1.9333	1.5664
20	1.6304	2.1360	1.6407	1.8133	1.8143
30	2.0000	2.4381	1.7476	1.7511	1.8329
40	1.8854	1.9315	1.8311	1.7357	1.7224
50	2.1413	2.2460	1.9814	1.9216	1.9476
60	2.1592	2.3001	1.9773	1.9944	1.9575
D.F.	2.0259	2.1867	1.8826	1.8729	1.8603

Table 4.a Striking score and Dominance factor of Fast-arm seamers for Set 1 (1-30)

SET 1	STRIKING SCORE (Y)				
$b_k$	MA STARC	TA BOULT	DW STEYN	M MORKEL	TG SOUTHEE
10	5.2941	2.7820	2.1891	2.1645	2.2133
20	2.1372	2.0256	1.8204	1.9182	2.2315
30	2.0364	1.9790	2.1936	2.1603	2.1789
40	2.0655	1.8547	2.2477	2.3751	2.2666
50	2.2602	2.1025	1.8526	2.0977	2.1488
60	2.4985	2.0962	2.0528	2.0487	2.2634
D.F.	2.3920	2.0609	2.0467	2.1316	2.2192

Table 4.b Striking score and Dominance factor of Fast-arm seamers for Set 2 (31-50)

SET 2	STRIKING SCORE (Y)				
$b_k$	MA STARC	TA BOULT	DW STEYN	M MORKEL	TG SOUTHEE
10	2.0866	1.4761	1.5069	1.8085	2.8148
20	1.9700	1.8730	1.6984	1.6423	1.5711
30	1.9000	2.2380	1.8214	1.8300	1.7222
40	1.8723	2.2364	1.9484	1.7895	1.8674
50	1.8064	2.1868	1.9579	1.9604	1.9675
60	1.9677	1.6185	2.1479	2.1120	2.0158
D.F.	1.9073	1.9775	1.9447	1.9150	1.9298

Higher the Dominance factor higher will be the chance to take wicket for a bowler, given the number of balls (here,  $b_k$ ). From these observations it is found that SP Narine has highest Dominance factor among the considered Slow-arm spinners followed by Imran Tahir though, Imran Tahir is rated better than SP Narine in ICC ODI bowler rankings as per 2016. Similarly, MA Starc has the highest Dominance factor among the Fast-arm seamers. Coming to the concept of building of strategy in a particular match, a noticeable variation is seen in the Dominance factor of the bowlers in the two set of overs. For instance, TG Southee has a Dominance factor of 2.2192 in Set 1 in contrary to, 1.9298 in Set 2. Therefore his performance will be comparatively better in overs of Set 1, where he can get early breakthroughs, than in Set 2. Moreover it is observed that almost every high rated Fast-arm seamers have better Dominance factor in Set 1 (all values are  $\geq 2$ ).

## 5 Conclusion and Future work

The method that has been developed only provides an alternative approach to represent the bowling performance of cricket players. This alternative approach is visually and intuitively appealing. The attempt in this paper is not to arrive at a model to rank the utility of players. Factors like tactical skills, passive support to the partner bowler, etc. cannot be gauged by looking at the Striking score or the Dominance

factor. One of the limitations of this paper is that the difference interval in the values of  $b_k$  and  $r_n$  are assumed as 10. Lesser the value of this interval, more accurate results will be obtained. Furthermore, increasing the number of simulations to compute  $b_i$  and  $r_i$  will enhance the accuracy as well as the precision of the various end results like, Probability to take wicket, Striking score and consequently the Dominance factor.

In a cricket match, the projected score of a team is one of the most discussed topics. While complex techniques based on dynamic programming such as the Winning and Score Predictor (WASP) Technique have been used, in its simplest form, the projected score is the product of current run-rate and total number of overs. If  $R$  is the current run-rate, projected score  $P$  is given as,  $P = R \times 50$ . We found that the team scores more than the projected score. The rate of change of run rate is directly proportional to number of overs bowled. Hence we categorized bowling strategies in two parts, strategy for first 30 overs and strategy for final 20 overs. In first 30 overs Batsmen are comparatively defensive while in the later stage they turn aggressive. Hence perfect bowling strategies is required for these two parts.

Table 5. Rate of change of Run Rate

		No. Of over remaining →					
		30-25	25-20	20-15	15-10	10-5	5-0
No. Of wickets fallen →	0	0.062	0.078	0.089	0.103	0.120	0.140
	1	0.058	0.075	0.083	0.097	0.120	0.140
	2	0.043	0.062	0.072	0.084	0.110	0.129
	3	0.030	0.053	0.065	0.078	0.099	0.118
	4	0.021	0.042	0.051	0.066	0.086	0.107
	5	-0.024	0.007	0.033	0.053	0.074	0.093
	6	-0.037	-0.026	0.012	0.037	0.049	0.076
	7	-0.044	-0.034	-0.023	0.010	0.034	0.055
	8	-0.051	-0.046	-0.035	-0.021	-0.006	0.032
	9	-0.064	-0.585	-0.047	-0.036	-0.024	-0.02

Within the limits of this study, the paper seeks to highlight the tremendous scope that exists to improve and develop on the measures currently used to describe the performances of cricket players in general, and bowlers in particular. Dominance factor for a particular bowler is prone to change depending on his performance in career. The measures used today do not adequately capture the richness of the underlying data. Similar approach can be adopted to predict the performance of bowlers in T20 (20 over limit) matches also.

## References

- [1] D. Attanayake and G. Hunter Probabilistic Modelling of Twenty-Twenty (T20) Cricket : An Investigation into various Metrics of Player Performance and their Effects on the Resulting Match and Player Scores, *Proceedings of 4<sup>th</sup> International Conference on Mathematics in Sport 2015*.

Bowling strategy building in limited over cricket match

Adhikari, Saraf and Parma

[2] U. Damodaran. Stochastic dominance and analysis of ODI batting performance : The Indian cricket team 1989-2005. *Journal of Sports Science & Medicine*, 5:503-508, 2006.

# **Manager's capacity and limit on foreign players: what influences variation of players in the field?**

K.Andreeva\*

\*Russia, 119049, Moscow, 26 Shabolovka St.,ka.andreeva@hse.ru

## **Abstract**

Limit on foreign players is a common instrument to promote national players in different sports, including football. It is usually imposed in several forms, such as limit of players in an application for a season (Portugal, Turkey, Austria), limit of players on the field (Ukraine, South Korea), stronger requirements for foreigners (England), restrictions on import from a particular country (Spain, Finland), restrictions on transfer turnover (Italy), limit of young players (Netherlands, Norway, Denmark), etc. Sometimes even economic barriers may refer to regulative policy (the lower wage level is, the less attractive a league becomes for foreign mature players). The core issue about the limit are uncertain consequences of its implementation. The purpose of this research is to evaluate the effect of the limit on the intermediate indicators, such as number of combinations available to team's manager when making decisions on particular players on the field. Practical results of this research may be used in making decision process on regulation mechanisms of the labor market in developing countries.

## **1 Limit as a regulation mechanism on the football market**

In contrast to traditional microeconomic concepts of quotas reducing surplus on traditional markets, football is interesting for economists because of government as a specific market player with its own interest to increase the number of national talents. The main instrument for this type of regulation is limit on foreign players. However, it is usually a political discussion on the issue of limit, not the economic one. As a result, more and more ineffective measures are being implemented within the government policy in the recent years in countries like Russia with high level of social expectation of results of the national team.

Generally, there are several specific features of the labor market in football that may cause uncertainty of regulation. First of all, it is differentiated: there are a couple of positions on the field with particular requirements to players, it may even have sense to analyze markets of these types of players as separate sub-markets. Secondly, marginal product of players also depends on the field position: forwards are associated with scored goals, middle players – with goal attacks, goalkeepers – with «saves» (when the opponent's goal attack doesn't lead to the scored goal).

From theoretical point of view it makes an analysis more complicated. However, for simplicity these specifics are usually not considered and all players are treated as a uniform «product». For example, when using formal demand and supply models, football labor market is treated as an oligopoly market, because there are just a few clubs that buy players' skills, although this number may vary from 9 in Lithuania to 20 in the Spanish League. While using another approach, for example, maximization of the objective function, then players' salaries become clubs' costs that diminish the function.



Manager's capacity and limit on foreign players:  
what influences variation of players in the field?

K.Andreeva

The economic theory deals with the issues of discrimination and quotas on the labor market. Limit on foreign players is a common instrument to promote national players in different sports, including football. It is usually imposed in several forms: limit of players in an application for a season (Portugal, Turkey, Austria); limit of players on the field (Ukraine, South Korea); stronger requirements for foreigners, for example to the number of games for the national team (in England it is an obligatory to play not less than 75% of such games); restrictions on import from a particular country (in Spain there is a discrimination towards South African and Asian players, in Finland – towards Chinese and Japanese players); restrictions on transfer turnover (Italy), which means that clubs cannot buy more than a particular number of players at the same time (Italian clubs cannot buy more than two foreigners at the same transfer period); discrimination towards young foreigners: on their import (Netherlands), on their representativeness in the core team (according to the UEFA rules in European club tournaments from the season 2008/2009 there should be more than 8 players who had already played at least four seasons in the same league in the youth team).

Sometimes even economic barriers may refer to regulative policy (the lower wage level is, the less attractive a league becomes for foreign mature players). This example reflects the basic model of an open economy with labor as one of the major factor. There are also several approaches to the term «foreign players». Sometimes it reflects the fact that a player doesn't have a passport from a particular country or from a group of countries (like in Spain, France or Austria), sometimes it refers to the player's incapability to play for a certain national team. According to Russian legislation, a foreigner in football is a player, who doesn't have Russian passport and doesn't have Russian nationality. It is important to emphasize that it is a regulative measure and may vary due to the national goals. For example, in Russia in 2005 in order to improve the quality of the national league foreigners who played in their national team were not treated as foreign players.

We can evaluate the consequences of the limit by several ways. First of all, we may look at the effectiveness of the national team in international tournaments taking into account that limit is aimed at promoting national players. Secondly, we may look at mean index of time that national players spend on the field in national leagues (the hypothesis of increasing opportunities of game practice due to the limit on foreigners). Thirdly, it is possible to analyze sources of new players preferred by clubs (the hypothesis of a growing interest in young prospective players). Finally, we may simply assess the dynamic of number of foreigners in clubs before and after the limit's implementation or observe improving game indicators or players' characteristics of new foreigners.

Russian case shows that all these indicators have decreased since the limit was implemented in 2005. During the whole period the Russian national team has usually reached only group stages of international tournaments (European Championships 1992, 1996, 2004, 2012, 2016; World Championships 1994, 2002, 2014; bronze medals of European Championship 2008 was rather an exception). From the season 2002 Russian national team players have played from 21 to 24 matches in a season, the median number of games has not changed either - it varies from 22 to 25 (half of the national team players have played most of the games in seasons, regardless to the limit on the foreign players, and spent on the field from 75 to 80 minutes). [1]

Since 2002, the leading Russian clubs have not increased their investments in their children's and youth schools, foreign transfers continue to be the main source of talents in clubs. The transfer activity of clubs has decreased, teams have become more stable. It disproves the argument of limit's supporters that limit leads to the development of children's and youth football in the country. If it was true, then

Manager's capacity and limit on foreign players:  
what influences variation of players in the field?

K.Andreeva

clubs would buy less players from the market and pay more attention to transferring players from the youth team to the main ones (since it is binding investment for clubs).

Russian clubs even did not become more selective when buying foreign players. The level of game characteristics sufficient to buy a legionnaire of the attacking and middle lines has remained the same: on average Russian clubs are interested in forwards with the indicators for the previous season of 20-25 matches and the performance of 6-11 goals. The share of foreigners in the Russian Premier League has also varied only within 15%: from 50% in 2005 to 38% in 2017.

From methodological point of view all these indicators have one disadvantage as they do not allow to predict consequences of the decision on the level of limit. They give us a retrospective view, but many football regulators are more interested in prospective view when making decisions. The approach in this article refers to calculation of possibilities of managers when choosing players on the field which helps overcome the outlined disadvantage.

## 2 Combinatorial analysis in assessment of manager's capacity

Combinatorics helps us to calculate the number of options depending on the importance of order of the elements in these options. Professional football team managers have a number of players available during the season and should decide on the combination of players in the particular game. They have to follow a number of rules, for example that they should choose a number of players for the application for a particular game and then choose a number of players on the field who start the game - not less than 7 people and not more than 11 people. [2]

The number of combinations depends on several factors, including meaning of number of players permitted for a game, meaning of number of players permitted for a season and additional constraints on players like limit on foreign players. All these factors influence the basic formula of combinations without repetition and the rule of addition and multiplication [3]:

$$C_n^k = n! / (n - k)! k! . \quad (1)$$

How this methodology may be applied in assessing manager's capacity? Let's assume that a manager has several players in a club for a season (designated as  $a$ ) divided into four types: goalkeepers ( $a_1$ ), defenders ( $a_2$ ), midfielders ( $a_3$ ) and forwards ( $a_4$ ). A manager is able to include in the game list  $b$  players, including not less than  $b_1$  goalkeepers,  $b_2$  defenders,  $b_3$  midfielders and  $b_4$  forwards. It means that a number of possibilities to choose players for a game is equal to:

$$p_1 = C_{a_1}^{b_1} * C_{a_2}^{b_2} * C_{a_3}^{b_3} * C_{a_4}^{b_4} * C_{a-b_1-b_2-b_3-b_4}^{b-b_1-b_2-b_3-b_4} . \quad (2)$$

Then a manager should decide on the game circuit (the number of defenders ( $c_2$ ), midfielders ( $c_3$ ) and forwards on the field ( $c_4$ ). The number of goalkeepers  $c_1$  is always 1. If we assume that all players within one type are unified (there are no right and left forwards, for example), then the number of possibilities to choose lineup is equal to:

$$p_2 = C_{b_1}^{c_1} * C_{b_2}^{c_2} * C_{b_3}^{c_3} * C_{b_4}^{c_4} . \quad (3)$$

Omitting substitutions, we get  $p$  combinations for managers:

$$p = p_1 * p_2 . \quad (4)$$

Generally there are several possible meanings of  $a$ ,  $b$  and  $c$ . According to the statistics on the website <http://www.transfermarkt.com/>, the most popular game circuits are 1-6-3 ( $c_2 = 1$ ;  $c_3 = 6$ ;  $c_4 = 3$ ), 2-3-5 ( $c_2 = 2$ ;  $c_3 = 3$ ;  $c_4 = 5$ ), 3-3-4 ( $c_2 = 3$ ;  $c_3 = 3$ ;  $c_4 = 4$ ), 3-4-3 ( $c_2 = 3$ ;  $c_3 = 4$ ;  $c_4 = 3$ ), 3-5-2 ( $c_2 = 3$ ;  $c_3 = 5$ ;  $c_4 = 2$ ), 3-6-1 ( $c_2 = 3$ ;  $c_3 = 6$ ;  $c_4 = 1$ ), 4-2-4 ( $c_2 = 4$ ;  $c_3 = 2$ ;  $c_4 = 4$ ), 4-3-3

Manager's capacity and limit on foreign players:  
 what influences variation of players in the field?

K.Andreeva

( $c_2 = 4; c_3 = 3; c_4 = 3$ ), 4-4-2 ( $c_2 = 4; c_3 = 4; c_4 = 2$ ), 4-5-1 ( $c_2 = 4; c_3 = 5; c_4 = 1$ ), 4-6-0 ( $c_2 = 4; c_3 = 6; c_4 = 0$ ), 5-3-2 ( $c_2 = 5; c_3 = 3; c_4 = 2$ ), 5-4-1 ( $c_2 = 5; c_3 = 4; c_4 = 1$ ).

Key inequations for  $a$ ,  $b$  and  $c$  are listed below:

$$a = a_1 + a_2 + a_3 + a_4. \tag{5}$$

$$c \leq b \leq a. \tag{6}$$

$$b \geq b_1 + b_2 + b_3 + b_4. \tag{7}$$

$$c = c_1 + c_2 + c_3 + c_4 = 11. \tag{8}$$

$$c_n \leq b_n \leq a_n, 1 \leq n \leq 4. \tag{9}$$

For example, if a club has 3 goalkeepers, 7 defenders, 8 midfielders and 7 forwards, and has a right to apply 16 players for a particular game, then decisions on lineup and game circuit that are made by a team's manager lead to the following variation of possible combinations:

Table 1. Combinations for a club with 25 players

<b>a</b>	<b>a1</b>	<b>a2</b>	<b>a3</b>	<b>a4</b>	<b>b</b>	<b>b1</b>	<b>b2</b>	<b>b3</b>	<b>b4</b>	<b>c</b>	<b>c1</b>	<b>c2</b>	<b>c3</b>	<b>c4</b>	<b>p1</b>	<b>p2</b>	<b>p</b>
25	3	7	8	7	16	2	3	7	4	11	1	1	6	3	29400	168	4,94*10 <sup>6</sup>
25	3	7	8	7	16	2	3	4	7	11	1	2	3	5	7350	504	3,70*10 <sup>6</sup>
25	3	7	8	7	16	2	4	4	6	11	1	3	3	4	51450	480	2,47*10 <sup>7</sup>
25	3	7	8	7	16	2	4	5	5	11	1	3	4	3	123480	400	4,94*10 <sup>7</sup>
25	3	7	8	7	16	2	5	6	3	11	1	3	5	2	61740	360	2,22*10 <sup>7</sup>
25	3	7	8	7	16	2	7	6	1	11	1	3	6	1	588	70	4,12*10 <sup>4</sup>

Now let's add additional restrictions on foreign players  $l$ . Let's look over two possibilities: limit on the application for a season and limit on the number of players on the field.

First option doesn't cause any changes: managers just have one more restriction  $a_1^l + a_2^l + a_3^l + a_4^l \leq l \leq a$ , where  $a_i^l$  ( $i = 1 \dots 4$ ) refers to the number of foreign goalkeepers, defenders, midfielders and forwards. However, we do not have any changes in a formula:

$$p = p_1 * p_2 = (C_{a_1}^{b_1} * C_{a_2}^{b_2} * C_{a_3}^{b_3} * C_{a_4}^{b_4} * C_{a-b_1-b_2-b_3-b_4}^{b-b_1-b_2-b_3-b_4}) * (C_{b_1}^{c_1} * C_{b_2}^{c_2} * C_{b_3}^{c_3} * C_{b_4}^{c_4}). \tag{10}$$

In the second case ( $l$  refers to a number of foreigners on the field) the situation is more complicated and may be presented in the game tree view where  $ck_i^l$  ( $1 \leq k \leq c_i, 1 \leq i \leq 4$ ) means that there are  $k$  foreigners of the  $i^{th}$  position on the field ( $1^{st}$  – goalkeepers,  $2^{nd}$  – defenders,  $3^{rd}$  – midfielders,  $4^{th}$  – forwards):

Manager's capacity and limit on foreign players:  
what influences variation of players in the field?

K.Andreeva

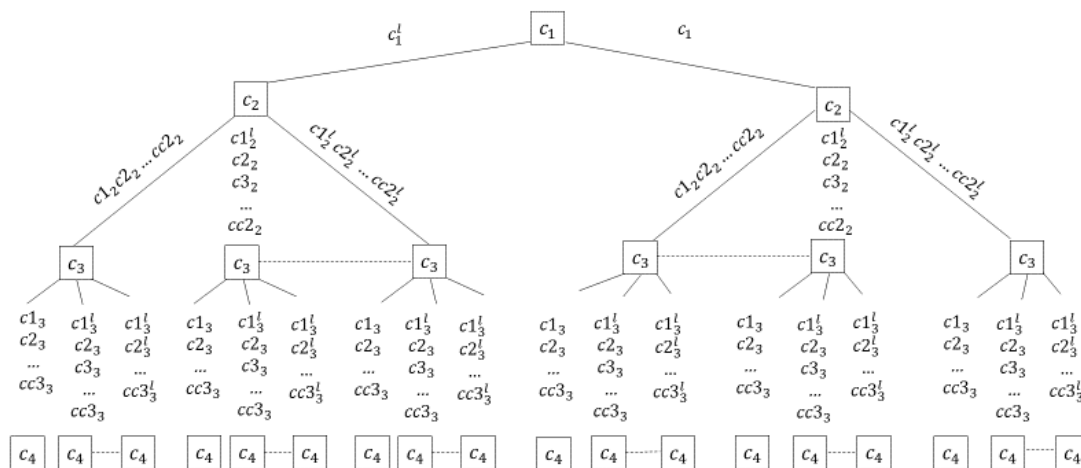


Figure 1a. Tree of combinations of foreigners on the field, without forwards

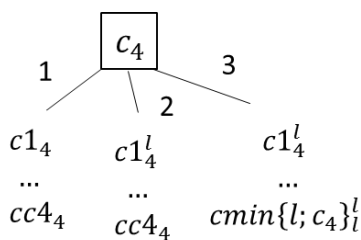


Figure 1b. Tree of combinations of foreigners on the field, choice of forwards

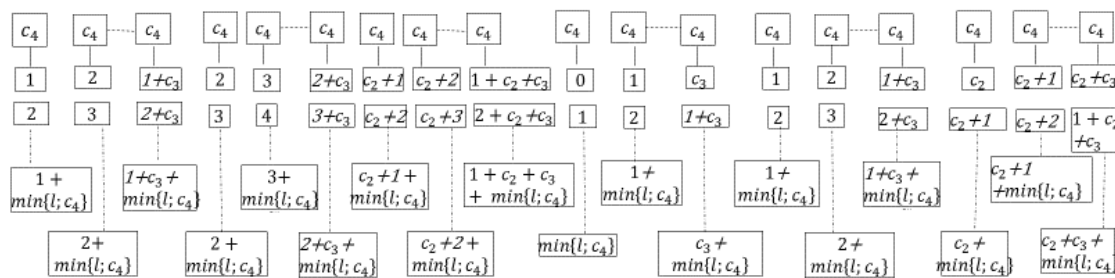


Figure 1c. Total number of foreigners on the field

The key change caused by the limit refers to one more decision that a manager should make, a decision on the number of foreigners on the field. Manager has several variations to put foreigners on the field depending on the game circuit (formula of combinations with repetitions, where  $k$  refers to an allowed number of foreigners on the field,  $n$  means possible positions for foreigners on the field; basically  $n = 4$  (goalkeepers, defenders, midfielders, forwards)):

Manager's capacity and limit on foreign players:  
 what influences variation of players in the field?

K.Andreeva

$$p_0 = \widetilde{C}_n^k = (n + k - 1)! / (n - 1)! k! . \tag{11}$$

Table 2. Combinations of foreigners on the field

$l$	Formula for $p_0$	Result for $p_0$	$l$	Formula for $p_0$	Result for $p_0$
1	$\widetilde{C}_4^1$	1+4=5	6	$\widetilde{C}_3^5 + \widetilde{C}_3^6$	1+5+15+37+83+177+49=367
2	$\widetilde{C}_3^1 + \widetilde{C}_3^2$	1+5+9=15	7	$\widetilde{C}_3^6 + \widetilde{C}_3^7$	1+5+15+37+83+177+367+64=749
3	$\widetilde{C}_3^2 + \widetilde{C}_3^3$	1+5+15+16=37	8	$\widetilde{C}_3^7 + \widetilde{C}_3^8$	1+5+15+37+83+177+367+749+81=1515
4	$\widetilde{C}_3^3 + \widetilde{C}_3^4$	1+5+15+37+25=83	9	$\widetilde{C}_3^8 + \widetilde{C}_3^9$	1+5+15+37+83+177+367+749+1515+100=3049
5	$\widetilde{C}_3^4 + \widetilde{C}_3^5$	1+5+15+37+83+36=177	10	$\widetilde{C}_3^9 + \widetilde{C}_3^{10}$	1+5+15+37+83+177+367+749+1515+3049+121=6119

There is a sum in a formula because there cannot be two goalkeepers on the field, so combinations should be divided into two parts – when a goalkeeper is a foreigner and when he is not a foreigner. For example, when  $l = 2$ , a manager may put a foreign goalkeeper and then put a second foreigner on one of the three left positions (defender, midfielder or forward),  $\widetilde{C}_3^1$ , or he may put two foreigners on one of these three positions,  $\widetilde{C}_3^2$ . Moreover, the limit doesn't make an obligation for managers to choose the maximum number of foreigners – he may simply choose less number of foreigners and does not overcome the limit.

It is quite obvious that the stricter the limit is, the less opportunities for coaches are – because without the limit a coach has more «net» opportunities  $\widetilde{C}_3^{10} = 66$  and total  $1+5+15+37+83+177+367+749+1515+3049+6119+66=12183$ , which means ten times more combinations than in case of limit of 8 foreigners, for example.

After a manager chooses a number of  $c_i^l$  (a number of foreigners in each line), he makes a decision on a lineup. Let's distinguish foreigners and national players with index  $l$  (for example, there are  $a_1^l$  foreign goalkeepers and manager chooses  $b_1^l$  of them to be in a lineup). It should be considered that limit influences not only  $p_2$ , but also  $p_1$  because managers may change their tactics depending on the possibilities of substitutions and application itself). For example, at  $l = 1$  the result for  $p_2$  and  $p_1$  will be

$$p_2 = C_{b1l}^{c1l} * C_{b2}^{c2} * C_{b3}^{c3} * C_{b4}^{c4} + C_{b1}^{c1} * C_{b2l}^{c2l} * C_{b2}^{c2} * C_{b3}^{c3} * C_{b4}^{c4} + C_{b1}^{c1} * C_{b2}^{c2} * C_{b3l}^{c3l} * C_{b3}^{c3} * C_{b4}^{c4} + C_{b1}^{c1} * C_{b2}^{c2} * C_{b3}^{c3} * C_{b4l}^{c4l} * C_{b4}^{c4} . \tag{12}$$

$$p_1 = C_{a1l}^{b1l} * C_{a1}^{b1} * C_{a2l}^{b2l} * C_{a2}^{b2} * C_{a3l}^{b3l} * C_{a3}^{b3} * C_{a4l}^{b4l} * C_{a4}^{b4} * C_{a-b1-b2-b3-b4}^{b-b1-b2-b3-b4} . \tag{13}$$

Manager's capacity and limit on foreign players:  
what influences variation of players in the field?

K.Andreeva

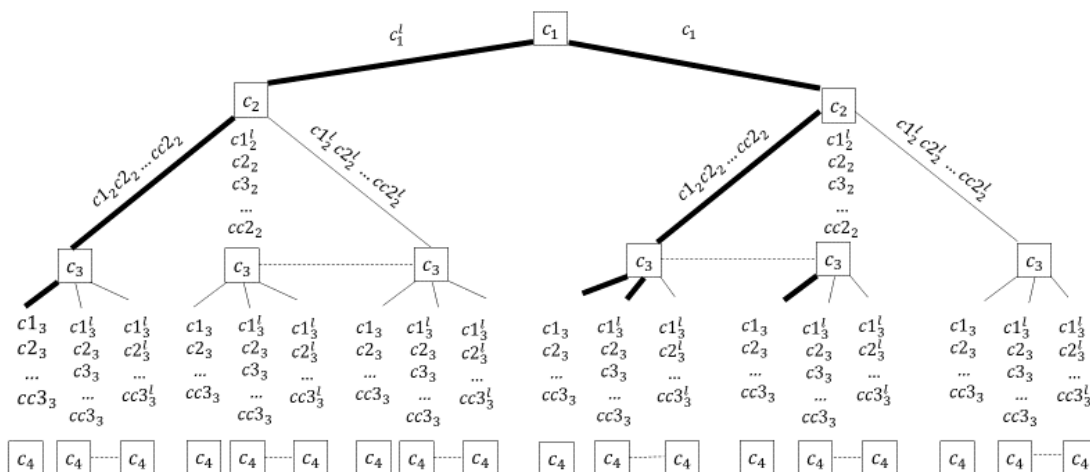


Figure 2a. Tree of combinations of foreigners on the field at the limit  $l = 1$ , without forwards

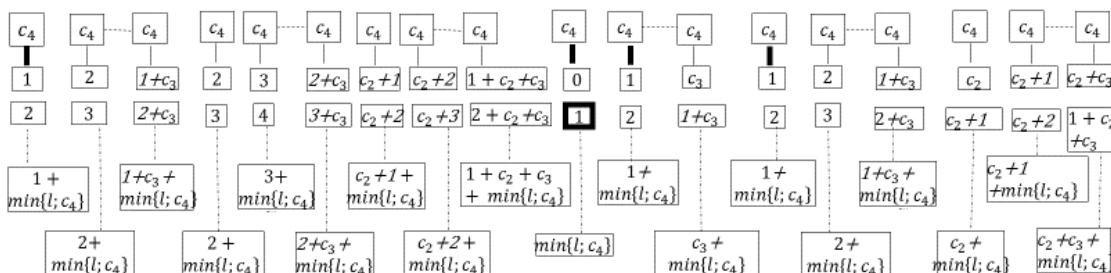


Figure 2b. Total number of foreigners on the field at the limit  $l = 1$

Remark 1.  $C_{n_1}^{k_1} * C_{n_2}^{k_2} \leq C_{n_1+n_2}^{k_1+k_2}$

$$C_{n_1}^{k_1} * C_{n_2}^{k_2} = n_1! n_2! / (n_1 - k_1)! k_1! (n_2 - k_2)! k_2! \quad (14)$$

$$C_{n_1+n_2}^{k_1+k_2} = (n_1 + n_2)! / (n_1 + n_2 - k_1 - k_2)! (k_1 + k_2)! \quad (15)$$

Let's implement a recursion approach to prove inequation in the Remark 1:

- if  $n_1 = n_2 = k_1 = k_2 = 1$  then  $C_{n_1}^{k_1} * C_{n_2}^{k_2} = C_{n_1+n_2}^{k_1+k_2}$
- if  $n_1 + n_2 = 3; k_1 = k_2 = 1$  then  $C_{n_1}^{k_1} * C_{n_2}^{k_2} < C_{n_1+n_2}^{k_1+k_2}$
- let  $C_{n_1}^{k_1} * C_{n_2}^{k_2} \leq C_{n_1+n_2}^{k_1+k_2}$ . Then for  $n_1 + 1; n_2 + 1; k_1 + 1$  and  $k_2 + 1$  we have

$$C_{n_1+1}^{k_1+1} * C_{n_2+1}^{k_2+1} = n_1! n_2! (n_1 + 1)(n_2 + 1) / (n_1 - k_1)! k_1! (k_1 + 1)(n_2 - k_2)! k_2! (k_2 + 1) =$$

$$C_{n_1}^{k_1} * C_{n_2}^{k_2} * (n_1 + 1)(n_2 + 1) / (k_1 + 1)(k_2 + 1) \leq C_{n_1+n_2}^{k_1+k_2} * (n_1 + 1)(n_2 + 1) / (k_1 + 1)(k_2 + 1) =$$

$$(n_1 + n_2)! (n_1 + 1)(n_2 + 1) / (n_1 + n_2 - k_1 - k_2)! (k_1 + k_2)! (k_1 + 1)(k_2 + 1) \leq$$

Manager's capacity and limit on foreign players:  
 what influences variation of players in the field?

K.Andreeva

$$\frac{(n_1 + n_2)! (n_1 + n_2 + 1)(n_1 + n_2 + 2)}{(n_1 + n_2 - k_1 - k_2)! (k_1 + k_2)! (k_1 + 1)(k_2 + 1)} \leq \frac{(n_1 + n_2)! (n_1 + n_2 + 1)(n_1 + n_2 + 2)}{(n_1 + n_2 - k_1 - k_2)! (k_1 + k_2)! (k_1 + k_2 + 1)(k_1 + k_2 + 2)} = C_{n_1+n_2+2}^{k_1+k_2+2} \quad (16)$$

Thus  $C_{n_1+1}^{k_1+1} * C_{n_2+1}^{k_2+1} \leq C_{n_1+n_2+2}^{k_1+k_2+2}$ .

Because a manager should make an extra decision on the number of foreigners on the field, the number of combinations, according to the Remark 1, decreases (as  $a_i^l \leq a_i, b_i^l \leq b_i, c_i^l \leq c_i$ ) as well. It means that implementation of the limit on foreign players worsen manager's capacity of players' choice.

### 3 Russian case: national and club manager's capacity with and without the limit on foreign players

What does this approach allow? It helps analyze three types of decision: decision on the level of limit, decision on the number of foreigners in a club and decision on the game circuit. Let's look over two seasons – 2003 (before limit's implementation) and 2007 (after limit's implementation), both years were one year before European Championships, and assume that national team managers are interested only in players who play in the highest division in one of the sixteen teams.

Table 3. Russian players in the Premier League in 2003 and 2007

	Number of Russian goalkeepers	Number of Russian defenders	Number of Russian midfielders	Number of Russian forwards	Total
2003	43	93	130	66	332
2007	41	93	182	80	396

Russian national team played 8 official games within UEFA Euro qualifying stages in both years. National team managers attracted for these games 33 national players in 2003 (2 goalkeepers, 11 defenders, 15 midfielders and 5 forwards) and 26 players in 2007 (3 goalkeepers, 6 defenders, 12 midfielders, 5 forwards). All the information is taken from the official site of the Russian national team: <http://www.russiateam.com/>.

The first decision that national team managers had made was a decision on players that would be included in the list of 33 and 26 players. The total number of combinations is too big to be presented as a total number of players with Russian passport is very high. Thus we may assess only the influence of chosen game circuits and lineups in both cases. In 2003 lineup was all time different from the previous games, the coincidence of all players in the start team was 0% (it means that managers have tried to differentiate a number of players in each particular game). Managers used 5 different game circuits in the official games in 2003, 3 of them were used just once, so a share of unique game circuits was 60%. In 2007 the situation with lineups was the same (0% of coincidence), while a share of unique game circuits decreased to 25%. As a result,  $p_2$  which reflected available combinations of players on the field, also decreased three times:

Table 4. Game circuits of the Russian national team in 2003 and 2007

	1	2	3	4	5	6	7	8
2003	4-5-1	3-7	3-6-1	4-5-1	3-5-2	4-4-2	4-4-2	4-5-1

Manager's capacity and limit on foreign players:  
what influences variation of players in the field?

K.Andreeva

2007	3-5-2	3-5-2	4-4-2	3-4-3	4-4-2	4-4-2	4-4-2	4-4-2	4-3-3
------	-------	-------	-------	-------	-------	-------	-------	-------	-------

Table 5. Possible combinations of Russian national team players in 2003 and 2007

year	a	a1	a2	a3	a4	b	b1	b2	b3	b4	c	c1	c2	c3	c4	p2
2003	332	43	93	130	66	33	2	11	15	5	11	1	4	5	1	9,91*10 <sup>6</sup>
	332	43	93	130	66	33	2	11	15	5	11	1	3	7	0	2,12*10 <sup>6</sup>
	332	43	93	130	66	33	2	11	15	5	11	1	3	6	1	8,26*10 <sup>6</sup>
	332	43	93	130	66	33	2	11	15	5	11	1	4	5	1	9,91*10 <sup>6</sup>
	332	43	93	130	66	33	2	11	15	5	11	1	3	5	2	9,91*10 <sup>6</sup>
	332	43	93	130	66	33	2	11	15	5	11	1	4	4	2	9,01*10 <sup>6</sup>
	332	43	93	130	66	33	2	11	15	5	11	1	4	4	2	9,01*10 <sup>6</sup>
	332	43	93	130	66	33	2	11	15	5	11	1	4	5	1	9,91*10 <sup>6</sup>
total																68,04*10 <sup>6</sup>
	396	41	93	182	80	26	3	6	12	5	11	1	3	5	2	4,75*10 <sup>5</sup>
	396	41	93	182	80	26	3	6	12	5	11	1	3	5	2	4,75*10 <sup>5</sup>
	396	41	93	182	80	26	3	6	12	5	11	1	4	4	2	2,23*10 <sup>5</sup>
	396	41	93	182	80	26	3	6	12	5	11	1	3	4	3	2,97*10 <sup>5</sup>
	396	41	93	182	80	26	3	6	12	5	11	1	4	4	2	2,23*10 <sup>5</sup>
	396	41	93	182	80	26	3	6	12	5	11	1	4	4	2	2,23*10 <sup>5</sup>
	396	41	93	182	80	26	3	6	12	5	11	1	4	4	2	2,23*10 <sup>5</sup>
	396	41	93	182	80	26	3	6		5	11	1	4	3	3	0,99*10 <sup>5</sup>
total									12							22,37*10 <sup>5</sup>

Now let's use the combinatorial approach to analyze manager's opportunities at club level. The main hypothesis that has been predicted in the theoretical part, is that limit on foreign players worse manager's capacity in preparation for games. One of the Russian clubs, FC Lokomotiv Moscow, played in both seasons, 2003 and 2007, 30 games in the Premier League and had 25 Russian players and 16 foreign players in the season 2003 and 28 Russian players and 17 foreign players in 2007. Three types of decision were made: on the application for a game, on the game circuit and on the particular lineup at the start. In 2003 application and lineup were mostly all times unique (96,6 % of uniqueness), in 2007 – 93 and 87% respectively. At the same time game circuits were more diversified in 2007 (44% share of uniqueness) than in 2003 (33% share of uniqueness).

Table 6. Number of foreigners and Russian players of different positions, FC Lokomotiv Moscow, 2003 and 2007

	Number of goalkeepers		Number of defenders		Number of midfielders		Number of forwards		Total	
	Russian	Foreign	Russian	Foreign	Russian	Foreign	Russian	Foreign	Russian	Foreign
2003	5	0	8	2	8	5	4	9	25	16
2007	3	3	6	8	14	3	5	3	28	17



Manager's capacity and limit on foreign players:  
what influences variation of players in the field?

K.Andreeva

Table 7. Number of foreigners and Russian players of different positions for each game, FC Lokomotiv Moscow, 2003 and 2007

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2003	application	6-7-3	6-7-3	4-8-3	5-8-3	6-7-3	5-8-3	5-8-3	5-8-3	6-6-4	5-7-4	6-6-4	6-6-4	6-6-4	6-6-4	6-6-4
	game circuit	4-5-1	5-3-2	4-5-1	3-5-2	3-5-2	4-4-2	4-6	4-5-1	4-4-2	3-6-1	4-5-1	5-3-2	5-3-2	5-4-1	4-4-2
2007	application	7-6-3	7-5-4	6-6-4	6-6-4	7-4-5	7-4-5	8-3-5	8-4-4	7-5-4	8-5-3	7-5-4	6-6-4	6-6-4	6-7-3	6-7-3
	game circuit	5-2-3	5-2-3	6-3-1	6-2-2	5-2-3	5-2-3	6-2-2	6-2-2	5-2-3	5-3-2	5-3-2	5-3-2	5-4-1	4-4-2	5-3-2
		16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
2003	application	5-6-5	4-8-4	4-7-4	5-7-4	5-6-5	5-6-5	5-6-5	6-5-5	4-8-4	5-7-4	6-7-3	6-6-4	6-5-5	6-5-5	5-6-5
	game circuit	4-4-2	3-6-1	3-5-2	4-4-2	4-4-2	4-4-2	3-4-3	5-2-3	3-5-2	4-4-2	5-4-1	4-5-1	3-5-2	4-4-2	3-5-2
2007	application	7-5-4	6-6-4	6-6-4	7-6-3	6-6-4	6-6-4	7-5-4	7-5-4	6-7-3	5-7-4	7-5-4	7-5-4	7-5-4	7-4-5	7-4-5
	game circuit	5-3-2	4-4-2	4-4-2	6-3-1	4-3-3	4-3-3	4-3-3	4-4-2	5-3-2	4-5-1	6-2-2	5-3-2	5-3-2	6-2-2	4-3-3

Calculation of the potential number of combination of players on the field for both seasons enables to disprove a hypothesis of the negative influence of the limit on the manager's capacity: in 2003 the total sum of possible combinations of players was  $1,25 \cdot 10^{14}$  and in 2007 –  $1,42 \cdot 10^{15}$ . Despite the predictions, we have ten times more opportunities with the limit than without. It could have happened because of difference in game circuits, difference in number of players at the start of the season (what influences  $p_1$ ), or even because of legislation gap (at club level a player with Russian passport isn't treated as a foreigner, even if he is not able to play for a Russian national team; it helps overcome difficulties caused by limit). All in all, it may be stated that phenomenon of the limit is still not very well known and requires further analysis.

## 4 Conclusions

Practical results of this research may be used in making decision process on regulation mechanisms of the labor market in developing countries. It becomes more and more obvious that the number of foreign players in the league should be regulated by the market, not government, because clubs have incentives to attract domestic players even without any limit. This conclusion is very important for those who protect idea of competitive football market in Russia. Freedom in labor market policy allows clubs to conduct a more meaningful and independent transfer policy instead of trying to cope with the effects of restrictions imposed by the Russian Football Union.

The methodology used in this article may be implemented to predict consequences of decision referred to the level of limit. It may be less sensible to make conclusions regarding to one particular game, but doubtless becomes more relevant for a long-time period. The main idea of this methodology is to analyze the potentiality of team managers (no matter club or national team managers) to choose 11 players to start a game depending on several decisions he should firstly make (on the game circuit, lineup or a number of foreigners).

Although theoretical conclusions of the worse conditions for managers when they cope with the limit have been proved only at the national level, we may still continue to gather data and check it not only

Manager's capacity and limit on foreign players:  
what influences variation of players in the field?

K.Andreeva

for the Russian case or not only for two-year period. What is more, Russian case gives us one more proof of the uncertainty of regulation and makes even more relevant researches on this issue.

## References

- [1] Andreeva K.A. (2017). *Regulation of transfer market: limit on foreign players in the Russian football*. Actual Problems of Economics and Law, №2.
- [2] The International Football Association Board. (2016). Laws of the game. [http://static-3eb8.kxcdn.com/documents/60/Laws%20of%20the%20Game\\_16-17\\_Digital\\_Eng.pdf](http://static-3eb8.kxcdn.com/documents/60/Laws%20of%20the%20Game_16-17_Digital_Eng.pdf)
- [3] Mazur, David R. (2010), *Combinatorics: A Guided Tour*, Mathematical Association of America, ISBN 978-0-88385-762-5

# Performance Evaluation of Volleyball Serve using Dynamical Models of Flight Trajectory

M. Angonese\*, F. Cardin\*\* and P. Ferrandi\*\*\*

\*MATHandSPORT srl - Moxoff Spa start up, via Durando 38/A, 20158 Milano.

E-mail: maria.angonese@mathandsport.com

\*\* Università degli Studi di Padova, Department of Mathematics “Tullio Levi-Civita”, via Trieste 63, 35121 Padova.

E-mail: cardin@math.unipd.it

\*\*\* Moxoff Spa - Politecnico di Milano spinoff, via Durando 38/A, 20158 Milano.

E-mail: paolo.ferrandi@moxoff.com

## Abstract

MOVIDA is a software platform focused on analyzing and optimizing movements by extracting data, trajectories and other measures automatically. The main goal is to evaluate the performance objectively, by assigning an index to each gesture and monitoring the consistency of the movement. Studying volleyball serves, restricting to ball flight trajectories, the observed data are the starting and ending point of the trajectory, the height of the ball over the net, the velocity during the flight and others. Three different aerodynamic models for the trajectory simulation are considered: the ball is idealized as spherical in the simplest one, while it has prolate ellipsoidal shape in the other two. The models simulate float serves, but also overhand serves in which the ball has an arbitrary spin direction. The most elaborate model takes into account the viscoelastic deformation that the ball may undergo, due to its compressibility, after the impact with the server's hand. After some considerations on the models' sensitivity analysis, MOVIDA technology is presented and it is highlighted its potential to support the athlete performance improvement.

## 1 Introduction

Coaches focus their time on developing players' skills that have the highest correlation to winning: this means that every player of the team should be guided and enabled to produce his best performance during the competition.

Methods to rate players performance are increasingly popular in sports analytics (i.e. [4] and [7]).

MOVIDA (Movement's Optimization through Video and Data Analysis) is a software platform that helps the coach to identify the optimal gesture, assuming that each athlete has his proper optimum. It provides an index in order to evaluate the performance objectively and to improve the gesture on the base of the specific physical and technical characteristics.

The main aim of this work is presenting a method to rate objectively and automatically the volleyball serve trajectory, both float and spin, regardless of the reception.

Before starting our work, we have to point out the importance of serving in modern volleyball, hence the

advantage of controlling and optimizing this athletic performance. The serve is the only skill that the player can totally control, moreover each rally begins with a serve. In the age of rally point system (since 1998 every action assigns a point) this skill takes greater importance, in fact a bad-executed serve may let the opposite team to score a direct point.

In the literature many models for volleyball serves suppose that angular velocity is zero, as in the case of floating serves, or they assume for simplicity that the angular velocity vector has constant direction to simulate a topspin/backspin serve (i.e. the top/back of the ball is spinning in the same direction as the ball translational motion). See [6], [10], [13].

The dynamical models of volleyball serve proposed in [1] are adopted in MOViDA. The innovative aspect is that the ball can take an arbitrary spin: in other words, it is reasonable to admit changes in magnitude, direction and sense of the angular velocity vector.

Some tests regarding the accuracy of the models and some numerical simulations results will be presented.

## 2 Dynamical Models of Volleyball Serves

In every application of a mathematical model we must take into account the trade-off between model accuracy and computation time required to perform the simulations. With respect to this, the main aim of this chapter is to expose three models of different complexity to simulate volleyball serves.

The first model is the simplest: it idealizes the ball as a sphere; the particular symmetry of this solid facilitates the form of the equations of motion, in particular the expression of the inertia tensor.

In the second model the ball is supposed to be a prolate ellipsoid in the attempt to take into account the deformation of the ball, due to its compressibility, after the impact with the server's hand. No data has been found about the compressibility of volleyballs; for this reason it is defined a coefficient equal to the ratio of the initial lengths of the semi-axes, called aspect ratio, which describes the deformation of the ball during the flying trajectory. To study this second problem it is introduced the non-inertial frame, whose reference axes always coincide with the semi-axes of the ellipsoid.

Finally the third model, very similar to the second, takes into account the ball compressibility in flight, making use of a relaxing viscoelastic deformation function depending on time, on initial lengths of the semi-axes and initial angular velocity of the ball.

In order to validate the models, some serves of professional players are simulated. In particular the first and the third model show promising compliance. Starting from this point, in the next chapter, some considerations on this subject will be made and the results of other tests will be showed.

### 2.1 Framing the Issue

Boundary-layer phenomena provide explanations for the lift and drag characteristics of bodies of various shapes in high Reynolds numbers flows, including turbulent flows. To deduce the fluid mechanics of sports-ball trajectories is necessary to make some considerations about this subject.

Let the position of the center of mass, the velocity function and the angular velocity function be denoted by:

$$\mathbf{r} = (x, y, z)^T \in \mathbb{R}^3, \quad \mathbf{v} = (v_x, v_y, v_z)^T \in \mathbb{R}^3, \quad \boldsymbol{\Omega} = (\Omega_x, \Omega_y, \Omega_z)^T \in \mathbb{R}^3.$$

They all depend on time. It follows a survey of the parameters and boundary values to consider [5]:

Physical parameters		Boundary values	
acceleration of gravity	$g$	starting position	$x(0), y(0), z(0)$
density of air	$\rho$	starting translational velocity	$v_x(0), v_y(0), v_z(0)$
viscosity of air	$\nu$	starting angular velocity	$\Omega_x(0), \Omega_y(0), \Omega_z(0)$
width of the boundary layer	$\delta_b$		
specific weight of the ball	$m$		
radius of the spherical ball	$R$		
semi-axes of the ellipsoidal ball	$a, b$		
ellipsoid's aspect ratio	$\beta := a/b$		
ellipsoid's angle of incidence	$\alpha$		

*Remark 1.* The properties of air depend on altitude and they influence the calculation of the Reynolds number, thus the drag and lift coefficients (below they will be denoted by  $C_D$  and  $C_L$ ).

Let  $CM$  be the center of mass of the body,  $\mathbf{F}^{tot}$  the total force acting on the system,  $\mathbf{P}$  the total linear momentum,  $\mathbf{T}_{CM}^{tot}$  the total torque (moment of external forces),  $\mathbf{L}_{CM}$  the angular momentum with respect to  $CM$ . For a rigid body, that is a system with holonomic smooth constraints moving in any inertial frame of reference, all the dynamically possible motions are described by the balance laws of linear and angular momentum:

$$\frac{d\mathbf{P}}{dt} = \mathbf{F}^{tot}, \quad (1)$$

$$\frac{d\mathbf{L}_{CM}}{dt} = \mathbf{T}_{CM}^{tot}. \quad (2)$$

The first three scalar equations hold for the translational velocity of the orbit of the center of mass, while the second three govern the rotation of the body and can be written in the more convenient form of Euler Gyroscopic Equations.

## 2.2 Aerodynamic Forces

To calculate  $\mathbf{F}^{tot}$  the gravitational force  $\mathbf{F}_g = m\mathbf{g}$  and the aerodynamic forces of drag  $\mathbf{D}$  and lift  $\mathbf{L}$  must be taken into account: the first slows down the velocity, the latter causes the curvature of the ball orbit. The Earth's Coriolis force will be neglected.

The magnitude of the force can be determined by integrating the local pressure times the surface area around the entire body. The component of the aerodynamic force that is opposed to the motion is the drag, while the component perpendicular to the motion is the lift. Drag can be thought as aerodynamic *friction drag*, which depends on velocity of the main stream  $u_0$ , fluid density of the air and kinematic viscosity of the air, and *form drag*, which depends on the body shape. By dimensional analysis, using Buckingham theorem:

$$\mathbf{D}(\mathbf{u}_0) = -\frac{1}{2} C_D \rho S u_0 \mathbf{u}_0. \quad (3)$$

where:

$$S_{sphere} = \pi R^2, \quad S_{ellipsoid}(\alpha) = \pi ab \sqrt{\cos^2 \alpha + \left(\frac{4}{\pi\beta}\right) \sin^2 \alpha}.$$

On the other hand, the dependence of *lift* force on the angular velocity, due to spin, has to be considered. For this reason, in order to provide a formula for the lift force, more sophisticated tools and fluid dynamics results are needed.

### 2.2.1 Lift Force

Many studies have been conducted on the aerodynamics of sports spinning balls including soccer, tennis, golf and rugby balls [2], [12], [14], [15]. Some consideration about the *Magnus and Reverse Magnus Effect* are reported in [1].

The derivation of the lift force on a ball dates back to 1956-1957, 1987 in Auton's and Lighthill's works [3], [8], [9], [11]. These articles allow to calculate the lift force on a sphere in a rotational flow. Thus applying reciprocity principle, it is found the formula to determine the lift force on spinning spherical volleyballs.

Furthermore, in [1], it is studied the fluid dynamics underlying the models in order to take advantage of the more recent (1998) paper of Warsi [16]. This work gives the tools to generalize the formula for the lift force on a sphere to the case of a prolate ellipsoid in a rotational flow, retracing the reasoning made by Auton. With the same working assumptions on the flow of [3], the modellistic aspect is privileged to derive a computationally convenient formula for the lift force when the ball is deformed as a prolate ellipsoid of semi-axes  $a$  and  $b$  s.t.  $a \geq b$ .

**Theorem 1.** Let  $\mathbf{u}_0$  be the velocity far upstream on the stagnation streamline, let  $\boldsymbol{\omega}_0$  be the uniform oncoming vorticity and let  $C_L$  be the lift coefficient. Thus the lift force in an arbitrary rotational straining flow is:

$$\mathbf{L} = \frac{1}{2} C_L \rho S a \mathbf{u}_0 \times \boldsymbol{\omega}_0, \quad (4)$$

where:

$$S(\alpha) = \pi ab \sqrt{\sin^2 \alpha + \left(\frac{4}{\pi\beta}\right) \cos^2 \alpha}.$$

## 2.3 Spherical Ball

*Assumption.* In the first model the ball is a rigid sphere of radius  $R$ , not subjected to deformation.

First of all, a model to calculate the orbit of a ball, idealized as a sphere flying with generic spin in the air from a point to another one, is stated. This requires to know the forces and torques operating on the ball, due to friction with air and gravity. The balance of angular momentum is immediately written, using the following expression for the angular momentum with respect to  $CM$ :

$$\mathbf{L}_{CM}(t) = \text{diag}(\iota, \iota, \iota) \boldsymbol{\Omega}(t) \quad \text{where} \quad \iota = \frac{2}{3} m R^2. \quad (5)$$

On the other hand, taking a standard base of unit vectors in spherical coordinates, where  $\hat{\mathbf{r}}$  is the normal unit vector of the surface  $S$  and  $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}$  generate the tangential vector space at the point  $(r, \theta, \phi)$  on the surface, the resistance torque is:

$$\begin{aligned} d\mathbf{T}_r(\theta, \phi) &= -\frac{\rho v}{\delta_b} [R \hat{\mathbf{r}}(\theta, \phi)] \times [R \boldsymbol{\Omega} \times \hat{\mathbf{r}}(\theta, \phi) dS(\theta, \phi)] \\ \implies \mathbf{T}_r(\boldsymbol{\Omega}) &= \int_0^{2\pi} \int_0^\pi d\mathbf{T}_r(\theta, \phi) = -\frac{8}{3} \frac{\rho v}{\delta_b} \pi R^4 \boldsymbol{\Omega}. \end{aligned}$$

## 2.4 Ellipsoidal Ball

*Assumption.* In the second model the ball is a rigid prolate ellipsoid of semi-axes  $a$  and  $b$  s.t.  $a \geq b$ .

For a non-spherical body it is also necessary to account for the offset of the center of pressure in relation to the center of geometry (i.e. center of mass). The pressure distribution on the surface of a body, inclined to the flow direction, does not follow the symmetry of the body any longer. This gives rise to an additional torque due to the displacement of the *center of pressure*  $x_{cp}$ . Once the co-rotating frame  $(x'', y'', z'')$  is introduced, the torques on the ball are calculated.

- Aerodynamic torque:

$$\mathbf{T}_a(\mathbf{v}(t), \boldsymbol{\Omega}(t)) = x_{cp} \mathbf{x}''(t) \times [ \mathbf{D}(\mathbf{v}(t)) + \mathbf{L}(\mathbf{v}(t), \boldsymbol{\Omega}(t)) ].$$

- Resistance torque:

$$\begin{aligned} d\mathbf{T}_r(\theta, \phi) &= -\frac{\rho v}{\delta_b} \left( \frac{a+b}{2} \right)^2 \hat{\mathbf{r}}(\theta, \phi) \times [ (\boldsymbol{\Omega} \times \hat{\mathbf{r}}(\theta, \phi)) dS(\theta, \phi) ] \\ \implies \mathbf{T}_r(\boldsymbol{\Omega}) &= -\frac{\rho v}{\delta_b} \left( \frac{a+b}{2} \right)^2 \pi a^2 \begin{bmatrix} 1 \\ 1 + 2\frac{b^2}{a^2} \\ \frac{1}{2} \end{bmatrix} \boldsymbol{\Omega}. \end{aligned}$$

## 2.5 Ellipsoidal Deforming Ball

*Assumption.* In the third model the ball is a deforming prolate ellipsoid of semi-axes  $a(t)$  and  $b(t)$  s.t.  $\forall t a(t) \geq b(t)$ .

The motion equations are the same of the previous model, it is rather pointed out the construction of a proper deformation function.

- It is supposed that, at the moment of the impact with the server's hand, the volleyball is a prolate ellipsoid of semi-axes  $a_0 = a(0)$  and  $b_0 = b(0)$  ( $a_0 > b_0$ ); during the flight phase its shape relaxes and tends to a sphere of radius  $R$  (for simplicity it is assumed that there are not oscillations during this action). The change in the length of the semi-axes implies a variation of the principal moments of inertia. Since no specific data regarding the compressibility of volleyballs have been found, a negative exponential function to simulate the viscoelastic deformation is used, which obviously depends firstly on time. The following requires are made:

$$a(t), b(t) \rightarrow R \quad \text{as } t \rightarrow +\infty.$$

- Besides, it seems reasonable that the deformation depends also on the initial length of semi-axes of the ball and on the absolute value of the initial angular velocity, associated with the spin given to the ball. In particular, if the ball has no spin, for every  $t$  it remains a sphere:

$$\boldsymbol{\Omega}(0) = 0 \quad \implies \quad a(t) = R = b(t) \quad \forall t \quad \implies \quad D_f = 0.$$

- It can be verified that a consistent *relaxing viscoelastic deformation function* is the following:

$$D_f = D_f(t, a_0, b_0, |\mathbf{\Omega}(0)|) = |\mathbf{\Omega}(0)| (a_0 - b_0)^2 e^{-2t \frac{a_0}{b_0}}. \quad (6)$$

Then, in the co-rotating frame, the values of the semi-axes  $a(t)$ ,  $b(t)$  and the principal moments of inertia  $A(t)$ ,  $B(t)$  of the prolate ellipsoidal deforming ball at time  $t$  become:

$$\begin{aligned} a(t) &= R(1 + k_1 D_f), & A(t) &= \iota(1 - k_2 D_f), \\ b(t) &= R(1 - k_1 D_f), & B(t) &= \iota(1 + k_2 D_f), \end{aligned} \quad (7)$$

where  $\iota$  is the inertia tensor of a prolate ellipsoid of semi-axes  $a, b$  and  $k_1, k_2 \in \mathbb{R}^+$  are chosen in a proper way: during the flight, the sphere of angular momentum and the volume of the ball change characteristics, from ellipsoidal to spherical ones.

## 2.6 Models' Sensitivity Analysis

Various simulations to test the models by numerical analysis has been performed. The consistency of the models is verified: firstly, some well-known results of rigid body mechanics are reproduced (i.e. the tenacity of the gyroscope axis for the models in which the ball has gyroscopic structure); secondly, it is verified that the first two models are special cases of the third. Then, the models are compared among them, taking into account trajectories and energy balances. Finally, the paths of some recorded serves, performed by professional athletes, are used to assess the validity of the models. The results of these tests show promising compliance, especially for the first and third models. So, from now on, our analysis will be restricted to *spherical* and *ellipsoidal deforming* models.

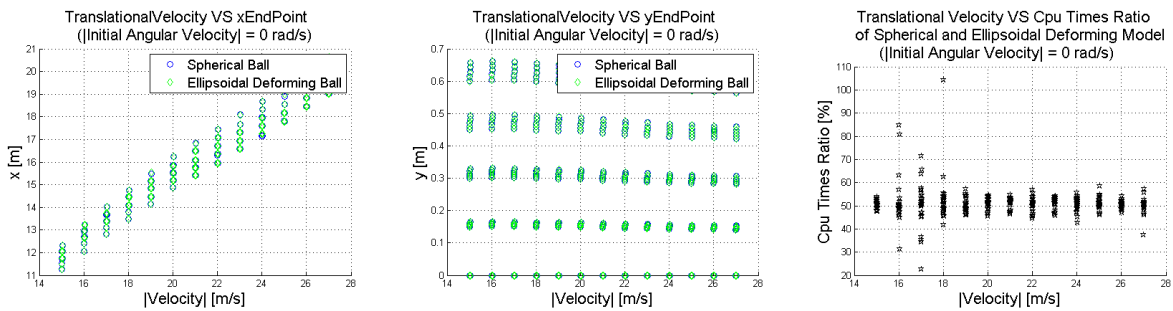


Figure 1: Comparison between spherical and ellipsoidal deforming models.

A couple of tests of models' sensitivity analysis is done in order to better understand the relationships between input and output variables, in particular between initial velocities and trajectory's end point.

Let  $(x, y, z) = (0, 0, 0)$  be the origin of the inertial frame of reference; it is put in the right corner of the volleyball field. The long side of the field grows along the  $x$ -axis, while the short side grows along the  $y$ -axis. For simplicity the starting position of the ball is set to  $(x, y, z) = (0, 0, 3)$ .

In the following the spherical and ellipsoidal deforming models are compared changing the value of the initial velocities vector, both translational and angular.

On the one hand, if no spin is supposed and the initial translational velocity of the ball is changed, no



considerable variation of the end point coordinate is found, as Figure 1 shows. It is more interesting the plot on the right side, which points out the percentage ratio between cpu timings of spherical and ellipsoidal deforming model: the spherical model shows evident advantages in terms of calculation times, when the involved velocities grow; in fact it halves the cpu calculation effort.

On the other hand, if the initial angular velocity is changed, there is a variation of the end point coordinate, as Figure 2 and 3 show. Comparing the deviation of coordinate values at the landing point of the ball in flight near and over the critical Reynolds regime, the spherical ball tends to have a smaller deviation than ellipsoidal ball.

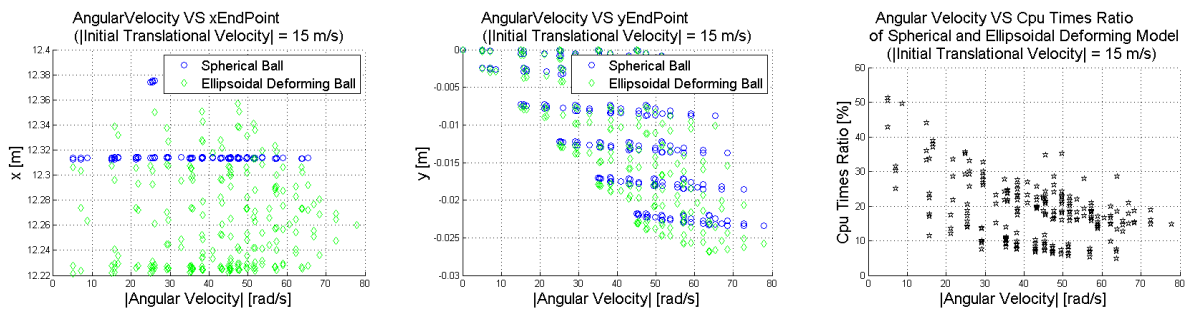


Figure 2: Comparison between spherical and ellipsoidal deforming models, when initial translational velocity is 15m/s.

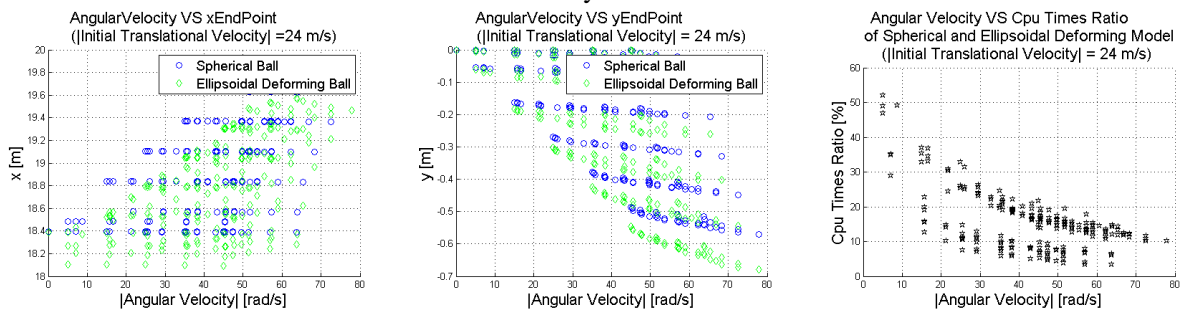


Figure 3: Comparison between spherical and ellipsoidal deforming models, when initial translational velocity is 24m/s.

The landing points with an initial velocity of 15 m/s admit the horizontal range ( $y$ -axis) of approximately 3cm and the vertical range ( $x$ -axis) of approximately 8cm for a spherical ball and 20cm for an ellipsoidal ball. The landing  $x$  points when the initial velocity is 24 m/s have the horizontal range of approximately 70cm and the vertical range of approximately 140cm for a spherical ball and 180cm for an ellipsoidal ball. It is evident that, as the initial velocity increases and the aerodynamic regime becomes turbulent, the two models simulate more and more different trajectories: in particular, the aerodynamics of the spherical model causes the ball to remain in flight for a longer time. In addition, the end point coordinates of the spherical model are more correlated despite the changing velocities. On the contrary, the end points of the ellipsoidal balls are more scattered. It is due to the intuitive fact that the aerodynamics of the ellipsoidal ball is more sensitive to disturbances, being less stable. This implies that, known the trajectory's starting and ending

points and the flight time, the third model can be more adaptable in the simulation; in fact, varying the velocities, it covers the codomain (i.e. the opponents' side of the volleyball court) in a more wide way.

The plots on the right side point out the percentage ratio between cpu timings of the spherical and the ellipsoidal deforming model: again, the first model shows evident advantages in terms of calculation times, when the involved velocities grow. The reason is pretty obvious: the spherical model is simpler in its construction. Furthermore, thanks to the symmetry of the sphere, the rotation matrix to pass from co-rotating to inertial frame of reference does not need to be calculated at every step of integration; in fact the inertia tensor does not depend on time, since every axis is principal.

### 3 Performance Evaluation with MOViDA

As a result of experiments conducted with the Italian National Men's Volleyball Team and with the CONI Institute of Medicine and Sports Science, Moxoff Spa has promoted research in sports science and sport technology and has developed several innovative tools for the world of sports, especially for volleyball. Mathematics and sport meet, thanks to Moxoff, both to improve the individual technique and to analyze the tactic. The growing interest in applied mathematics to sports analytics has led to MATHandSPORT: the Moxoff's startup devoted to the world of sport.

#### 3.1 MOViDA Technology

MOViDA (Movement's Optimization through Video and Data Analysis) is MATHandSPORT's software platform that applies innovative data-driven techniques to evaluate and optimize the individual technical gesture. In a standard MOViDA session, the same athletic movement is repeated several times and recorded by one or two cameras. Every gesture is classified based on objective criteria and some parts of the body are analyzed with particular attention, taking trajectories, measures, angles, velocities, relations between movements of various body segments. MOViDA algorithms automatically process a large amount of videos and extract smart data, in order to develop performance indeces, based on mathematical modeling and advanced statistical methods, as classification, clustering and functional data analysis. The final goal of MOViDA is to correlate the gesture to the performance in order to identify the player's optimal gesture, to describe athlete's current session efficiency and to monitor the performance over time. Furthermore, it can help the coach to evaluate the recovery of an athlete after a possible injury.

#### 3.2 Performance and Consistency Indeces

*MOViDA for volleyball serve* focuses on studying wrists and ankles movements during the run-up, analyzing the ball toss and, at last, correlating these with the flight trajectory. So, at the base of the performance evaluation, it is essential to simulate the ball trajectory in a trusted way. The dynamical models proposed above have been developed at this purpose. The characteristics of the optimal serve have to be preliminarily identified together with the coach: the initial velocity of the ball, the ball velocity and height over the net, the end point coordinates. For example, in the case of float serve, it is identified the initial velocity range that ensures the ball floats after the passage over the net (it depends on the drag coefficient, which is specific for each type of ball and is affected by altitude). Then MOViDA uses an appropriate metric to quantify the difference between the current gesture and the optimal target. Finally it weighs in a proper way all the

analyzed quantities and provides a single *performance index* for each gesture.



Figure 4: Trajectories evaluated with MOViDA indices and grouped by levels of performance.

Differently, the *consistency index* gives an indication of how much the gesture is regular, in particular it quantifies the variation range, along the curve, in a objective way; this information is very important for the coach because it gives a measure of the athlete ability to manage the skill, which is fundamental during the competition. The consistency index is developed using functional data analysis: it takes into account both the shape of the curve and the time warping function. As the performance index, it allows to monitor the athlete's gesture consistency over time and it underlines the gesture component where there is less regularity.

## 4 Summary and Conclusions

The purpose of this study is to present the way in which MATHandSPORT's software platform MOViDA evaluates a volleyball serve automatically and objectively. To reach this aim, two dynamical models of the flight trajectory are compared: the first idealizes the ball as a sphere, the second as a prolate ellipsoid which axes are subjected to a relaxing viscoelastic deformation. Thus, the trajectories' aerodynamic at the variation of the ball initial velocities can be examined.

The results may be summarized as follows. Both models have been tested at a preliminary stage and provide accurate outcomes. At low velocities, the two models simulate very similar trajectories and the only remarkable difference is the cpu calculation times: the spherical model halves the calculation effort. As the velocities increase, the aerodynamics of the ellipsoidal model causes the ball to remain in flight for a shorter time; moreover, the end point coordinates of the ellipsoidal balls are more scattered when the angular velocity changes; this is due to the less stable aerodynamics of the ellipsoidal ball. The spherical model calculation times remain significantly more efficient.

Often, in the industrial context, it is good to have the results as soon as possible; but it is also necessary to be aware of how much you lose in terms of accuracy of the output data. For every simulation, MOViDA's platform workflow captures from video, automatically, starting and ending points of the trajectory, then it requires to estimate translational and angular velocities. In this way an approximation of the data is introduced. In order to use the ellipsoidal model, it is also necessary to estimate the deformation of the ball; this risks to introduce a greater approximation than supposing the ball to be perfectly spherical. The previous consideration, added to the need to process large amounts of videos and thus to improve the cpu performance, makes

the spherical model more convenient in this situation.

It is worth pointing out that the ellipsoidal model is built to be more adaptable and it can be applied also to the simulation of soccer and rugby balls, in contexts where the ball undergoes greater deformation.

MOVIDA was initially developed and tested for volleyball, but today it is a versatile platform that can be used transversally in different sports, from tennis to football, from golf to rugby, to support the athlete performance improvement.

## Acknowledgements

A special thank you to the est. Prof. Edie Miglio and Dr. Matteo Pischiutta (Mathematics department of Politecnico di Milano), whose preliminary work inspired us and confirmed that it was a reasonable way to face new scientific challenges.

## References

- [1] Angonese, M., Cardin, F., Ferrandi, P. (2015) *Dynamical Models of Volleyball Serves*, Master's Degree, Università degli Studi di Padova, Department of Mathematics "Tullio Levi-Civita".
- [2] Asai, T., Seo, K., Kobayashi, O., Sakashita R. (2007) *Fundamental aerodynamics of the soccer ball*, Sports Engineering, 10: 101-110.
- [3] Auton, T.R. (1987) *The lift force on a spherical body in a rotational flow*, Journal of Fluid Mechanics 183: 199-218.
- [4] Corley, B., Brett, W. (2017) *Bump, Set, Spike: Using Analytics to Rate Volleyball Teams and Players*, MIT Sloan Sports Analytics Conference.
- [5] Fuchs, P.M. (1991) *Physical model, theoretical aspects and applications of the flight of a ball in the atmosphere. Part I*, Mathematical methods in the applied sciences 14.7: 447-460.
- [6] Kao, S.S., Sellens, R.W., Stevenson, J.M. (1994) *A mathematical model for the trajectory of a spiked volleyball and its coaching application*, Journal of applied biomechanics, 10.
- [7] Levin, A. (2017) *Ranking the Skills of Golfers on the PGA TOUR using Gradient Boosting Machines and Network Analysis*, MIT Sloan Sports Analytics Conference.
- [8] Lighthill, M.J. (1956) *The image system of a vortex element in a rigid sphere*, Mathematical Proceedings of the Cambridge Philosophical Society, Vol. 52. No. 02.
- [9] Lighthill, M.J. (1956) *Drift*, Journal of Fluid Mechanics 1.01: 31-53.
- [10] Lithio, D., Webb, E. (2006) *Optimizing a volleyball serve*, Rose Hulman Institute of Technology Undergraduate Math Journal, 7.
- [11] Lighthill, M.J. (1957) *Contributions to the theory of the Pitot tube displacement effect*, Journal of Fluid Mechanics 2.05: 493-512.
- [12] Mehta R.D., Pallis J.M. (2001) *The aerodynamics of a tennis ball*, Sports Engineering, 4: 177-189.
- [13] Ricardo, J. (2014) *Modeling the Motion of a Volleyball with Spin*, Journal of the Advanced Undergraduate Physics Laboratory Investigation.
- [14] Seo K., Kobayashi O., Murakami M. (2006) *Flight dynamics of the screw kick in rugby*, Sports Engineering, 9: 49-58.
- [15] Smits A.J., Ogg S. (2004) *Aerodynamics of the Golf Ball*, Biomedical Engineering Principles in Sports.
- [16] Warsi, Z.U.A. (1998) *Analysis and numerical evaluation of the drift function of Darwin and Lighthill for axisymmetric flows*, Theoretical and computational fluid dynamics 10.1-4: 439-448.

# Ranking ski courses using permutation methods for multivariate populations

R. Arboretti\*, E. Carrozzo\*\* and L. Salmaso\*\*

\*Department of Civil, Environmental and Architectural Engineering, University of Padova, Padova, Italy  
email address: rosa.arboretti@unipd.it

\*\* Department of Management and Engineering, University of Padova, Vicenza, Italy  
email address: annaeleonora.carrozzo@unipd.it  
luigi.salmaso@unipd.it

## Abstract

Monitoring perceived performance of sport trainers is usually a difficult task. The Board of Professional Ski Instructors of the Province of Bolzano and the Ski Schools in ‘Alto Adige’ were interested in investigating the satisfaction of people attending Ski courses in this area. A specific survey has been planned to investigate different aspects of satisfaction, such as on course organization, on teaching, on fun etc. The aim of the statistical analysis was multi-fold (1) to obtain a ranking of the ski courses from the ‘best’ to the ‘worst’, (2) to identify relevant drivers to differentiate ski instructors/schools and (3) to set up a suitable multivariate methodology able to test for multiple outcomes with stratified data. Usually such analysis is performed for each domain of satisfaction and they are kept separate. In the present work we provide a procedure to compare the performance of ski courses considering simultaneously all domains of satisfaction/quality and stratifying by covariates e.g. the nationality or the age of the attendees. This procedure is based on an extension of the Nonparametric Combination (NPC) of dependent permutation tests (Pesarin and Salmaso, 2010) and it allows us to take into account several aspects affecting the ski performances.

## 1 Introduction

Skiing activities are often linked with the winter tourism and so monitoring customer satisfaction about quality of provided services becomes more and more important for the economy of specific areas. Indeed many mountain cities in the world which provide skiing resorts, perform customer satisfaction surveys in order to understand how to become more attractive and competitive.

The Canadian Ski Council for example, thought that a way to continue to flourish was to attract new people into the sport by identifying potential skiers and converting them from non-skiers to skiers through specific marketing strategies (Williams and Fidgeon, 2000).

Some studies have shown that some factors specific of the customers, could influence the customer satisfaction itself and their consequent loyalty with ski resorts (Matzler et al. (2007), Matzler et al. (2008)). Several authors also studied how the perceived quality influences the tourist’s satisfaction (Alexandris et al., 2004, Kelley and Turley, 2001; Shonk and Chelladurai, 2008).

Thus, as Yoon and Uysal (2005) pointed out, in a competitive market environment, a careful analysis of tourist motivations, customer satisfaction and loyalty can make the difference.

Furthermore requirements for high quality service are also specified by ISO 9001 document (2008). The European regulation ISO 9001 states that organizations need to show its ability to regularly provide a product which satisfies customers' requirements and wishes to increase customers' satisfaction, the former related to monitoring of quality, the latter to improvement of quality.

In the present work we want to describe the statistical approach adopted to analyze data from a survey developed by the Board of Professional Ski Instructors of the Province of Bolzano and the Ski Schools in 'Alto Adige' to investigate the satisfaction of people attending Ski courses in this Italian area.

When checking quality of products or services, the research aim is often focused on evaluating the product/service performances from a multivariate point of view, i.e. investigating more than one aspect and/or under several conditions. From the statistical point of view, when the response variable of interest is multivariate in nature, the problem may become quite difficult to cope with, due to the high dimensionality of the parametric space. In this work we propose an appropriate statistical method based on an extension of the Nonparametric Combination (NPC) of dependent permutation tests (Pesarin and Salmaso, 2010), which allows us to compare multivariate performances by considering all aspects of interest separately albeit simultaneously and taking into account possible confounding factors.

The present paper is organized as follows. Next section introduces and describes the NPC-based method and the steps to be followed. In Section 3 the case of the ski courses evaluation is introduced and the application of the proposed method is explained and results presented. Finally in Section 4 we discuss the results and main advantages of the proposed method.

## 2 The NPC-based procedure

When developing new products or checking quality of products or services, complex problems of hypothesis testing arise. The complexity of the study is mainly referred to the presence of mixed variables (ordinal categorical, binary or continuous), missing values, stratification variables. Surveys performed to evaluate quality dimensions are often observational studies, where very little is known about the multivariate distribution underlying the observed variables and their possible dependence structure. In such cases conditional nonparametric methods can represent a reasonable approach. Multivariate permutation tests are conditional exact nonparametric procedures, where conditioning is on the pooled observed data as a set of sufficient statistics in the null hypothesis. In this contribution we consider permutation methods for multivariate stratified problems, allowing variables to be of a different nature (continuous, discrete, ordinal etc.). The proposed approach is based on the Nonparametric Combination (NPC) methodology (Pesarin and Salmaso, 2010) to reduce the dimensionality of the problem. Formalizing the problem, suppose to have two populations represented by two multivariate random variables, say  $\mathbf{X}_1$  and  $\mathbf{X}_2$  and we are interested in testing:

$$\begin{cases} H_0 : \mathbf{X}_1 \stackrel{d}{=} \mathbf{X}_2 \\ H_1 : \mathbf{X}_1 \stackrel{d}{>} \mathbf{X}_2 \end{cases} \quad (1)$$

Suppose to have two samples  $\mathbf{x}_1$  from  $\mathbf{X}_1$  and  $\mathbf{x}_2$  from  $\mathbf{X}_2$  respectively. In order to test (2) taking into account the possible effect of stratification factors, we break down the problem into sub-problems for each

combination of the levels of factors. Without loss of generality let us assume to have two stratification factors with  $S_1$  and  $S_2$  levels respectively, then we obtain  $S = S_1 \times S_2$  systems of hypotheses:

$$\begin{cases} H_{0(s_1, s_2)} : \mathbf{X}_{1(s_1, s_2)} \stackrel{d}{=} \mathbf{X}_{2(s_1, s_2)} \\ H_{1(s_1, s_2)} : \mathbf{X}_{1(s_1, s_2)} \stackrel{d}{>} \mathbf{X}_{2(s_1, s_2)} \end{cases} \quad (2)$$

where  $s_1 = 1, \dots, S_1$  and  $s_2 = 1, \dots, S_2$ . The idea of the NPC-based procedure for each comparison is to combine the p-values resulting from each stratum and then further combine the resulting p-values from different variables to build up the overall test. We briefly report the steps of the procedure:

1.  $\forall s_1 = 1, \dots, S_1$ 
  - 1.1.  $\forall s_2 = 1, \dots, S_2$ :
    - 1.1.1 Compute a suitable test statistic  $\mathbf{T}_{(s_1, s_2)}^O$  on the observed samples.
    - 1.1.2 Perform a random permutation of units between the two samples.
    - 1.1.3 Compute the test statistic on permuted samples obtaining  $\mathbf{T}_{(s_1, s_2)}^{*1}$ .
    - 1.1.4 Repeat steps 1.1.1-1.1.2 a number  $B$  of times obtaining  $\mathbf{T}_{(s_1, s_2)}^{*b}$ ,  $b = 1, \dots, B$ , the estimated null distribution of  $\mathbf{T}_{(s_1, s_2)}$ .

At the end of the previous step we obtained  $S_1 \times S_2$  test statistics and the related permutation distribution.

2. For each test statistic we can compute the related p-value as  $\lambda_{(s_1, s_2)} = \sum_{b=1}^B I(\mathbf{T}_{(s_1, s_2)}^{*b} \geq \mathbf{T}_{(s_1, s_2)}^O) / B$ ,  $\forall s_1 = 1, \dots, S_1$ , and  $s_2 = 1, \dots, S_2$ .
3. For each level of  $S_1$  consider a suitable combination function  $\Phi$  and combine the p-values related to different levels of stratum  $S_2$  obtaining  $\mathbf{T}_{(s_1, \bullet)}^O$  and the related permuted distribution  $\mathbf{T}_{(s_1, \bullet)}^*$ .
4. From step (3) compute the estimated p-values as  $\lambda_{(s_1, \bullet)} = \sum_{b=1}^B I(\mathbf{T}_{(s_1, \bullet)}^{*b} \geq \mathbf{T}_{(s_1, \bullet)}^O) / B$ ,  $\forall s_1 = 1, \dots, S_1$ ;
5. Combine the p-values related to different levels of stratum  $S_1$  obtaining  $\mathbf{T}_{\bullet}^O$  and the related permutation distribution  $\mathbf{T}_{\bullet}^*$ .
6. From step (5) compute the estimated p-values as  $\lambda_{\bullet} = \sum_{b=1}^B I(\mathbf{T}_{\bullet}^{*b} \geq \mathbf{T}_{\bullet}^O) / B$ ;
7. Finally combine the p-values related to different variables obtaining  $\mathbf{T}_{\bullet}''^O$  and the related permuted distribution  $\mathbf{T}_{\bullet}''^*$ .
8. From step (7) compute the estimated global p-value as  $\lambda_{\bullet}'' = \sum_{b=1}^B I(\mathbf{T}_{\bullet}''^{*b} \geq \mathbf{T}_{\bullet}''^O) / B$ ;
9. accept the null hypothesis if  $\lambda_{\bullet}'' \geq \alpha$ .

If we want to compare more than 2 population we can compute this algorithm for each pair of populations and adjust for multiplicity (Simes, 1986; Hommel, 1988; Shaffer, 1995, Benjamini and Hochberg, 1995).

### 3 Survey on ski courses

In the winter season of 2011 a large survey has been conducted in 38 ski schools of Alto Adige (an area of Italian Alps), in which customers and parents of children, who participated in a ski course, were asked to answer a questionnaire to express their level of satisfaction about some aspects of the experience. This study was innovative at a national level: it was the first systematic study performed on different schools, with quantitative evaluation, using a questionnaire specifically designed to measure satisfaction and quality perceived by customers. The questionnaire asked for opinions about different aspects of the service, related to:

- Improvement: progress in skiing skills;
- Courtesy: courtesy and availability of the instructor;
- Resort: satisfaction about ski resort;
- Safety: course has been performed in safety;
- Fun: fun during the course.

Each dimension was investigated with specific questions reporting the score on a scale 0-10 (0: not satisfied, 10: fully satisfied). Arboretti et al. (2014) analyzed the responses of children to this questionnaire for three aspects of satisfaction and investigating the covariates in a separate phase. NPC-Global ranking (Bonnini et al., 2006; Corain and Salmaso, 2007), based on the NonParametric Combination methodology (Pesarin and Salmaso, 2010), considers the problem of finding a global ranking of  $C$  populations with respect  $p$  variables, as formally represented in a testing-like framework where the hypotheses of interest are related to the stochastic inferiority or superiority when comparing  $C$  populations. The method considers first nonparametric tests for pairwise comparisons of  $C \times (C - 1)/2$  populations of interest for each variable, and then a combination of directional p-values (through a suitable combining function) in which all variables are simultaneously considered. On the basis of the NPC score a global ranking of the  $C$  populations is derived (see Arboretti Giancristofaro et al. (2014) for more methodological and computational details). In the present work we consider responses both from the adult and children questionnaire and considering age as a stratification factor. We also consider the nationality of the attendees as a further stratification factor. For illustrative purpose five of the 38 schools, selected for marketing reasons, have been codified as  $A, B, C, D, E$ . For each school we selected answers to the questionnaire from Italians and Germans, children and adults who attendee ski courses. People under 13 years old were considered children, otherwise adult. Thus two stratification factors, i.e. age class and nationality, have 2 levels respectively. Applying the NPC-procedure based on  $B = 10000$  permutations we obtain the final global results in Table 1 where numbers represent global multivariate p-value (adjusted for multiplicity) for each comparison. P-values in table have to be read as being related to comparison “school in row > school in column”. Thus applying an algorithm based on the number of significant comparisons, we obtained a ranking of the 5 schools (Table 2).

Note that this ranking is global, in the sense that schools are compared taking into account all aspects of satisfaction simultaneously stratifying for covariates. Furthermore it is possible to see the partial results, for example going back of one step in the algorithm we can see the comparisons for each aspect of satisfaction and also the results of a specific comparison separately for nationality.

Assume for example we are interested in investigate comparison  $C < D$  we can see from Table 3, that School  $D$  which is classified better than School  $C$  in the global ranking, presents significantly higher evaluations for *Improvement* and *Fun*, and also an evidence at 10% significance level for *Courtesy*.



Table 1: Pairwise multivariate comparisons between the five schools.

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
<i>A</i>	-	0.461	0.173	0.618	0.994
<i>B</i>	0.306	-	0.297	0.738	0.992
<i>C</i>	0.012	0.002	-	0.113	0.695
<i>D</i>	0.054	0.066	0.036	-	0.928
<i>E</i>	0.005	0.012	0.003	0.062	-

Table 2: Global ranking of the five schools.

<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
4°	4°	3°	2°	1°

Table 3: P-values for each aspect of satisfaction.

	<i>Improvement</i>	<i>Courtesy</i>	<i>Resort</i>	<i>Safety</i>	<i>Fun</i>
<i>A</i> < <i>B</i>	0.238	0.775	0.220	0.586	0.102
<i>A</i> < <i>C</i>	0.862	0.862	0.001	0.009	0.009
<i>A</i> < <i>D</i>	0.015	0.567	0.150	0.174	0.044
<i>A</i> < <i>E</i>	0.107	0.122	0.017	0.052	0.001
<i>B</i> < <i>C</i>	0.619	0.106	0.026	0.000	0.003
<i>B</i> < <i>D</i>	0.107	0.425	0.313	0.013	0.156
<i>B</i> < <i>E</i>	0.219	0.112	0.140	0.026	0.000
<i>C</i> < <i>D</i>	0.002	0.057	0.472	0.903	0.028
<i>C</i> < <i>E</i>	0.004	0.008	0.675	0.163	0.002
<i>D</i> < <i>E</i>	0.358	0.281	0.247	0.116	0.008

From Table 4 we can see that this difference in satisfaction between the two schools regarding *Improvement* is specific of Italian customers, whereas about *Fun*, Germans express a higher satisfaction in favour of school *D*. It is also possible to further investigate differences between age classes. Let us see for example the details for the comparison  $C < D$  in Table 5.

From these results we can see that, significant preference of the school *D* related to *Improvement* from Italian customers concerns both adults and children. Whereas preference about *Fun* from German customers looks to be more related to children experience.

It is also worth noting that evidence for the variable *Courtesy* derives mainly from a higher satisfaction of German children, in fact after combination with other results, related to the same comparison in different strata becomes non significant (at a  $\alpha$ -level of 5%).

Table 4: P-values of the comparison  $C < D$  for Italians and Germans for each aspect of satisfaction.

	<i>Improvement</i>	<i>Courtesy</i>	<i>Resort</i>	<i>Safety</i>	<i>Fun</i>
<i>Italian</i>	0.003	0.191	0.682	0.628	0.308
<i>German</i>	0.901	0.062	0.257	0.974	0.016

Table 5: P-values of the comparison  $C < D$  for each aspect of satisfaction for Italian and German adults and children.

		<i>Improvement</i>	<i>Courtesy</i>	<i>Resort</i>	<i>Safety</i>	<i>Fun</i>
<i>Italians</i>	<i>Adults</i>	0.001	0.491	0.402	0.382	0.134
	<i>Children</i>	0.018	0.174	0.938	0.852	0.968
<i>Germans</i>	<i>Adults</i>	0.974	1.000	0.999	0.973	1.000
	<i>Children</i>	0.645	0.034	0.109	0.897	0.007

## 4 Conclusions

With this work we aim at presenting a statistical procedure to analyze data from customer satisfaction survey related to sports teaching/services. The proposed method is based on an extension of the Nonparametric Combination (NPC) of dependent permutation tests (Pesarin and Salmaso, 2010), which allows to compare multivariate performances by considering all aspects of interest separately albeit simultaneously and taking into account possible confounding factors.

In particular we present a real application to a survey carried out for investigating the satisfaction of people attending ski courses in a specific Italian area. The objective was to compare ski schools on the base of satisfaction about different aspects (Improvement, Courtesy, Resort, Safety and Fun) considering that people attending ski courses have different age so that they could have different perception of quality.

We show that with the proposed method we are able to compare and rank different ski schools globally i.e. considering all domains of satisfaction of interest, taking into account possible confounding factors such as age classes and/or nationality. It is also possible to investigate each partial aspect related to each level of the confounding factors considered. We provided an easy algorithm which summarizes the steps to achieve this useful NPC-based permutation procedure.

## References

- [1] Alexandris, K., Zahariadis, C., Tsozbatzoudis, C. and Grouios, G. (2004) *An empirical investigation of the relationships among services quality, customer satisfaction and psychological commitment in a health club context*. European Sport Management Quarterly **4**, 36-52.
- [2] Arboretti Giancristofaro, R., Bonnini, S., Corain, L., and Salmaso, L. (2014). *A permutation approach for ranking of multivariate populations*. Journal of Multivariate Analysis, **132**, 39-57.
- [3] Arboretti, R., Bordignon, P., Carrozzo, E. (2014) *Two phase analysis of ski schools customer satisfaction: multivariate ranking and cub models*. Statistica **74**(2), 141-154.

- [4] Benjamini, Y., Hochberg, Y.(1995) *Controlling the false discovery rate: a practical and powerful approach to multiple testing*. Journal of the Royal Statistical Society Series B, **57**, 289- 300.
- [5] Bonnini, S., Corain, L. and Salmaso, L. (2006) *A new statistical procedure to support industrial research into new product development*. Quality and Reliability Engineering International, **22**(5), 555-566.
- [6] Corain, L., Salmaso, L. (2007) *A nonparametric method for defining a global preference ranking of industrial products*. Journal of Applied Statistics **34** (2) 203-216.
- [7] Hommel, G.(1988) *A stagewise rejective multiple test procedure based on a modified Bonferroni test*. Biometrika, **75**, 383-386.
- [8] International Organization for Standardization, International Standard ISO 9001 (2008). Quality management systems-Requirements.
- [9] Kelly, S. W. and Turley, L. W.(2001) *Consumer perceptions of services quality attributes at sporting events*. Journal of Business Research **54**, 161-166.
- [10] Matzler, K., Füller, J., Renzl, B., Herting, S., Späth, S. (2008) *Customer satisfaction with Alpine ski areas: the moderating effects of personal, situational, and product factors*. Journal of Travel Research **46**, 403-413.
- [11] Matzler, K., Füller, J., Faullant (2007) *Customer satisfaction and loyalty to Alpine ski resorts: the moderating effect of lifestyle, spending and customers' skiing skills*. International Journal of Tourism Research **9**(6), 409-421.
- [12] Pesarin, F. and Salmaso, L. (2010) *Permutation tests for complex data: theory, applications and software*, Wiley, Chichester.
- [13] Shaffer, J. P. (1995) *Multiple hypothesis testing*. Annual review of psychology, **46**, 561- 576.
- [14] Shonk, D. J. and Chelladurai, P. (2008) *Service quality, satisfaction, and intent to return in event sport tourism*. Journal of Sport Management **22**, 587-602.
- [15] Simes, R. J. (1986) *An improved Bonferroni procedure for multiple tests of significance*. Biometrika **73**, 751-754.
- [16] Williams, P. and Fridgeon, P. R. (2000) *Addressing participation constraint: A case study of potential skiers*. Tourism Management **21**, 379-393.
- [17] Yoon, Y. and Uysal, M. (2005) *An examination of the effects of motivation and satisfaction on destination loyalty: A structural model*. Tourism Management **26**, 45-56.

# Estimating Margin of Victory in Twenty-20 International Cricket

M. Asif\* and I. G. McHale\*\*

\*Department of Statistics, University of Malakand, Pakistan. email address: m.asif@uom.edu.pk

\*\* Salford Business School, University of Salford, UK. email address: I.McHale@salford.ac.uk

## Abstract

In this paper we propose a model of the runs remaining to be scored in the second innings of Twenty-20 International (T20I) Cricket. The proposed model takes account of overs left and wickets lost. Our model makes it possible to determine the runs differential of the two competing teams. The runs differential not only gauges the closeness of the game in terms of uncertainty of outcome, but also makes it possible to estimate ratings of the teams. Here we use the model to estimate the largest winning margins in T20I cricket history. Sri Lanka's 172 run victory over Kenya in 2007 appears to be the biggest margin of victory to date.

## 1 Introduction

In sports, the margin of victory is a useful statistic as it not only determines the closeness of a game but can also play an important role in rating teams, as it is a quantitative measure of the relative performance of the two teams in a game. The additional information given by margin of victory, compared with simply knowing which competitor (team, or individual) won, has been acknowledged in other sports. In football for example, cutting-edge forecasting models make use of score-lines (e.g. 3-1 or 0-2), rather than simply results (win, draw, loss). In tennis forecasting, leading models now make use of information on the points, games and sets won, rather than just the binary match result: won, or lost.

In sports like football, tennis, golf and basketball, the margin of victory is easily observable and is determined simply by the scores/points differential of the two competing players/teams. In contrast, measuring margin of victory in the game of cricket is not straightforward. This is because if the team batting second wins, the match is censored in that not all of the overs allotted to the second team are played: there is no point playing on once the winning target has been reached. As such, the metric for margin of victory depends on which team, the winner or loser, batted in the first/second innings.

For example, if a team batting in the first innings (team 1) wins a match then the margin of victory is simply determined by taking the difference of the two innings runs totals. However, if the winning team bats in the second innings (team 2) then the second innings is typically cut short so that not all of the allotted overs are played. In such circumstances it is traditional for the margin of victory to be described by how many wickets that team had remaining, *regardless of how fast the target was achieved*. Thus, it is difficult to compare the performances of sides as the margin of victory is measured using different units, depending on whether the victorious team batted first or second. It is noticeable that team 2's margin of victory can be considered as two dimensional, that is the team typically not only has a number of wickets in hand, but also a number of overs (or balls) remaining.

As a consequence of complexities in measuring margin of victory in T20I cricket, rating team performances and forecasting become more complicated. In this paper we present a Generalized Non-Linear Model (GNLM) that can be used to forecast the second innings total as a function of overs remaining and wickets lost, and hence we can calculate the margin of victory in T20I cricket. Although the model can be used for forecasting, here we consider its use when ranking all 556 T20I cricket matches by the margin of victory. To do so, we effectively convert all results into projected runs differentials.

Our work, to some extent, relates to the work that is done by Clarke and Allsopp (2001), de Silva et al. (2001) for One-Day International (ODI), and Allsopp and Clarke (2004) for ODI/Test cricket. Clarke and Allsopp (2001) used the Duckworth and Lewis (1998) resource table to project the second innings runs total of an ODI cricket match. Further, they fit a linear model to rate teams' performances in the ICC One-Day World Cup Championship held in the year 1999. Likewise, de Silva et al. (2001) used the same resource table in a different way, doing some ad hoc modification, to project the second innings runs total in One-Day International Cricket.

## 2 Projecting Second Innings Runs

Suppose,  $S_1$  and  $S_2$  are the total runs scored by team 1 and team 2 respectively. If team 1 wins the match then margin of victory can simply be determined by the runs differential,  $RD = S_1 - S_2$ . However, in case of team 2's victory  $S_2$  will be replaced by the projected runs,  $S_{2(proj)}$ . Let there be  $u$  overs left and  $w$  wickets lost when team 2 reached to the target, then the projected runs can be determined by the following relation.

$$S_{2(proj)} = S_{2(actual)} / \{1 - P(u, w)\} \quad (1)$$

where  $P(u, w)$  is the proportion of remaining runs when there are  $u$  overs left and when  $w$  wickets have already been lost. Herein and after, this proportion is referred as 'resources remaining'. Following Duckworth and Lewis (1998) the remaining resources can be estimated by  $P(u, w) = Z(u, w) / Z(N, 0)$ , where  $Z(u, w)$  is expected remaining runs in  $u$  overs when  $w$  wickets have been lost, and  $N$  is the total pre-allotted overs for each first and second innings. For example,  $N=20$  for T20I Cricket (unless the match has been shortened due to weather factors).

In regards to the functional form of  $Z(u, w)$ , various authors have proposed different models with a specific aim of revising targets for the team batting in the second innings in interrupted matches. For example, Duckworth and Lewis (1998) proposed an exponential type function, but due to commercial confidentiality the model fit results and estimation methods were kept hidden. Further, Duckworth and Lewis (2004) proposed some modification and provided an improved version of the model that can handle One-Day International Cricket matches in which the scoring rate is well above average. McHale and Asif (2013) proposed an arc-tangent based model for the expected remaining runs for ODI cricket. In this paper, a generalized form for  $Z$  is proposed. The model is based on certain properties, defined in the next section. These properties are needed to model the runs scoring patterns observed in all formats of the Limited Overs International (LOI) cricket.

## 3 The Generalized Non-Linear Model

The generalized model for the expected remaining runs,  $Z$ , as function of  $u$  overs remaining and  $w$  wickets lost, should have the following properties.

- i. For a given number of wickets lost, expected remaining runs should be non-decreasing with respect to  $u$ , overs left.
- ii. For a given number of wicket lost, the expected runs on the next ball should be non-increasing with respect to  $u$ , overs left.
- iii. For a given number of overs left, the expected remaining runs should be a non-increasing function of  $w$ , wickets lost.
- iv. For a given number of overs left, the expected runs on the next ball should be non-increasing function of  $w$ , wicket lost.

The above list of the properties can be used as a framework to build a model for expected remaining runs to be scored by a team with  $u$  overs remaining when  $w$  wickets have already been lost. A general form for  $Z$ , based on above standard properties, can be written as

$$Z(u, w) = Z_0 F(w) G(u | \sigma(w)) + \varepsilon \quad (2)$$

In order to satisfy the above properties, and to make the function more intuitive, some restrictions on  $Z_0$ ,  $F(w)$ , and  $G(u | \sigma(w))$  are made. For example,  $F(w)$  may be a non-increasing real valued function with domain  $[0,10]$  and range  $[0,1]$  such that  $F(0)=1$  and  $F(10)=0$ . The function  $G(u | \sigma(w))$  is also a real valued function defined on  $u>0$  such that the first order derivative is non-negative and second order derivative is non-positive for all  $u>0$ . Further,  $\sigma(w)>0$  is the parameter such that  $\sigma(w)=\sigma F(w)$ .  $Z_0$  is a constant and if the function  $G$  ranges from  $[0,1]$  such that  $G(0)=0$  and  $G(\infty)=1$ , then  $Z_0$  can be interpreted as the asymptotic runs obtainable with ten wickets in hands in an unlimited innings (infinite overs), but playing under the strategy of the specified format of the game, T20I for example. Finally,  $\varepsilon$  is an error term with zero mean.

If  $G(u | \sigma(w))$  takes the form of the exponential cumulative distribution function of  $u$  and  $F(w)$  is estimated in a non-parametric way for  $w=0,1,..,9$  under the constraint that  $F(0)=1$  and  $F(w)>F(w+1)$  then the model reduces to the Duckworth and Lewis (2004) model. However, if the function  $G(u | \sigma(w))$  is approximated by the Half-Cauchy cumulative distribution function, and  $F(w)$  is approximated by the truncated normal survival function with domain  $[0,10]$  and range  $[0,1]$  then the model becomes as proposed by the McHale and Asif (2013). Hence, Duckworth/Lewis and McHale/Asif versions are special cases of the proposed GNLM.

## 4 The Model Specification

In this section, we present a generalized functional form for  $Z$ . First, a function for  $F(w)$  is specified as

$$F(w) = \left\{ \exp\left(\frac{-w}{a}\right)^b - \exp\left(\frac{-10}{a}\right)^b \right\} / \left\{ 1 - \left(\frac{-10}{a}\right)^b \right\} \quad (3)$$

where  $a > 0$  and  $b > 0$  are the parameters to be estimated. The first order derivative of  $F(w)$  is negative in domain  $[0,10]$ , hence, the function is decreasing with range  $[0, 1]$ . Note that  $F(0)=1$  and  $F(10)=0$ . Hence, the above specified function for  $F(w)$  is appropriate as it satisfies the desirable properties described in the previous section.

Second, in regards to the function  $G(u | \sigma(w))$ , where  $\sigma(w)=\sigma F(w)$ , we adopt an arc-tangent type function as its first order derivative is non-negative and the second order derivative is non-positive, with respect to  $u$ , for all  $u>0$ . After specifying the functions for  $F(\cdot)$  and  $G(\cdot)$ , we have a model for  $Z$  as follows

$$Z(u, w) = Z_0 F(w) \tan^{-1}(u / \sigma F(w)) + \varepsilon \quad (4)$$

where  $Z_0 > 0$  and  $\sigma > 0$  are the parameters to be estimated, and  $F(w)$  is defined in Equation (3). Further, since  $\tan^{-1}(\cdot)$  ranges from  $[0, \pi/2]$ ,  $Z_0 \times (2/\pi)$  is the asymptotic runs to be scored in infinite overs under the rules and general strategy of the T20I cricket. The model in Equation (4) satisfies the four properties described above. It is a flexible model and can adapt to particular runs scoring patterns, and provides a superior fit to data than the Duckworth/Lewis original.

## 5 Results

We obtained data on all historical T20I Cricket matches, played from February 2005 to September 2016, from the [www.espncricinfo.com](http://www.espncricinfo.com) website in October 2016. During this time, a total of 570 matches were played though, 14 matches ended with ‘no result’ (due to weather interruptions) and were discarded from the sample. Non-linear weighted least squares was used to estimate the model parameters, using the Levenberg-Marquardt algorithm (LMA) provided in R package *minpak.lm* written by Timur et al. (2016). The observed and fitted curve is provided in Figure 1.

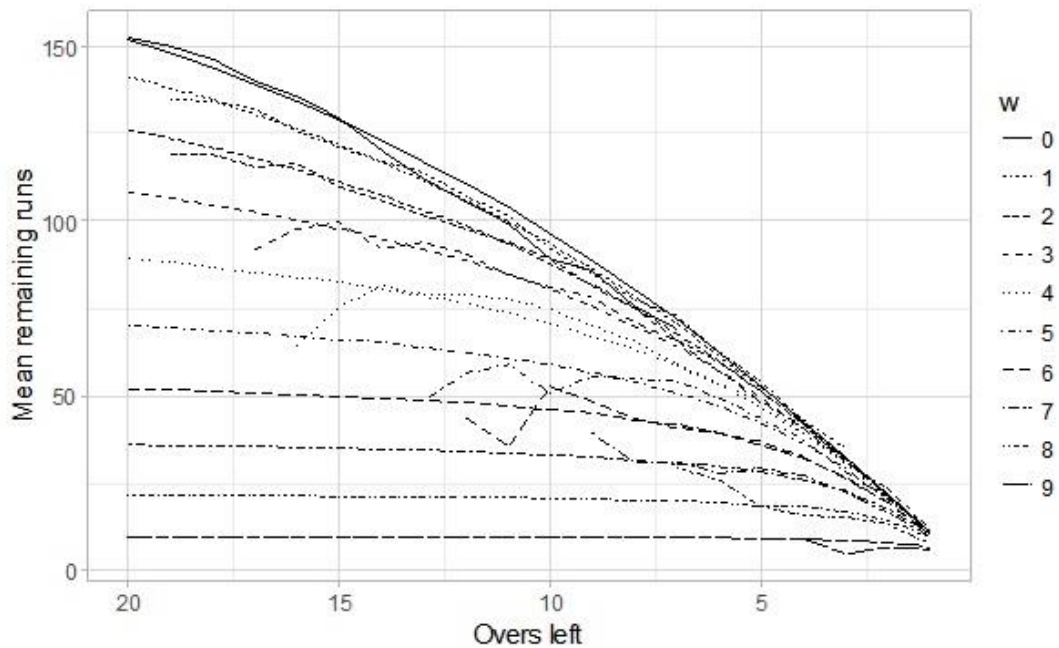


Figure 1: Plot of observed and fitted mean remaining runs versus overs left ( $u$ ) for T20I data. Top line is for  $w=0$  (no wicket lost), and the bottom line is for  $w=9$  (nine wickets lost).

The model in the equation (4), is suitable in matches when the pattern of scoring is “normal”. However, for well above average runs scoring matches the relationship between  $Z$  and  $u$  should tend to be more linear, in the range  $[0, 20]$ . This implies that the over-by-over runs scoring potential tends to uniformity as the run-rate tends to increase for any given number of wickets lost. Therefore, a new parameter,  $\lambda (\geq 1)$ , is introduced in equation (4) that allows the parameters  $\sigma$  and  $Z_0$  to be scaled up, in order to allow the relationship between  $Z$  and  $u$  to be more linear in range  $[0, 20]$ . Hence, introducing the parameter  $\lambda$ , equation (4) is rewritten as follows

$$Z(u, w, \lambda) = Z_0 \lambda^{n(w)+1} F(w) \tan^{-1}(u/\theta \lambda^{n(w)} F(w)) + \varepsilon \quad (5)$$

The parameter  $\lambda$  can be estimated based on the total runs scored by team 2 in  $20-u$  overs at the end of the T20I match. Therefore, the value of  $\lambda$  is dynamic and varies from match-to-match. For average or below average runs scoring matches the values of  $\lambda$  is equal to 1, otherwise its value will be greater than 1, depending upon how much team 2's runs are deviating from the average runs in  $20-u$  overs. In our data the average first innings total runs of the T20I cricket matches is equal to 152.2. Estimation of  $\lambda$ , is done in similar fashion as discussed in McHale and Asif (2013). In the following paragraph, using a real example, we explain the need of introduction of parameter  $\lambda$ .

On January 10, 2016 New Zealand (NZ) were set a target of 142 by Sri Lanka. They reached 147 runs in just 10 overs for the loss of just one wicket. Clearly, NZ scored with an exceptionally high run-rate (14.7 runs per over) which was well above the average for T20I cricket. If we use equation (4) to estimate the margin of victory (in runs), then NZ's expected remaining runs in the remaining 10 overs, with a loss of one wicket, is 233. That is clearly an unrealistic and over-inflated estimated number of runs to be scored in the remaining 10 overs. As a consequence, NZ's estimated runs margin of victory is 238 in a T20I cricket match. In contrast if we use equation (5) then our estimated value for  $\lambda$ , based on team 2's score ( $S_2=147$ ,  $w=1$ , and  $u=10$ ), is equal to 1.40. Hence, the expected remaining runs based on model in equation (5) is approximately 150. As a result, New Zealand's victory was by an estimated margin of 155 runs, and is the second greatest victory ever in T20I history.

Table 1 presents the top 20 largest winning victories in T20I history. Top of the table is Sri Lanka, who won against Kenya by a record runs margin of 172 runs at Johannesburg, South Africa. It is the greatest victory ever (by runs margin) by any team 1 in T20I cricket, and is indeed estimated as the biggest winning margin ever. In second place is New Zealand, who were victorious versus Sri Lanka. Since New Zealand batted second, the margin of victory in terms of runs is estimated from our model as 155.



MathSport International 2017 Conference Proceedings

Table 1: Largest margins of victory in T20 International history: February 2005 to September 2016.

Date	First Innings			Second Innings			Winner	Traditional Margin	Balls Left	Runs Differential
	Team 1	Score	Overs	Team 2	Score	Overs				
14/09/07	Sri Lanka	260/6	20	Kenya	88/10	19.3	<b>Sri Lanka</b>	172 runs	NA	172
10/01/16	Sri Lanka	142/8	20	New Zealand	147/1	10	<b>New Zealand</b>	9 wickets	60	-155
14/03/12	Kenya	71/10	19	Ireland	72/0	7.2	<b>Ireland</b>	10 wickets	76	-153
24/03/14	Netherlands	39/10	10.3	Sri Lanka	40/1	5	<b>Sri Lanka</b>	9 wickets	90	-142
22/03/12	Canada	106/8	20	Ireland	109/0	9.3	<b>Ireland</b>	10 wickets	63	-139
03/02/10	Bangladesh	78/10	17.3	New Zealand	79/0	8.2	<b>New Zealand</b>	10 wickets	70	-139
12/09/07	Kenya	73/10	16.5	New Zealand	74/1	7.4	<b>New Zealand</b>	9 wickets	74	-138
07/06/09	South Africa	211/5	20	Scotland	81/10	15.4	<b>South Africa</b>	130 runs	NA	130
09/07/15	U.A.E.	109/10	18.1	Scotland	110/1	10	<b>Scotland</b>	9 wickets	60	-124
20/09/07	Sri Lanka	101/10	19.3	Australia	102/0	10.2	<b>Australia</b>	10 wickets	58	-117
21/09/12	England	196/5	20	Afghanistan	80/10	17.2	<b>England</b>	116 runs	NA	116
02/02/07	Pakistan	129/8	20	South Africa	132/0	11.3	<b>South Africa</b>	10 wickets	51	-113
23/02/10	West Indies	138/7	20	Australia	142/2	11.4	<b>Australia</b>	8 wickets	50	-110
13/10/08	Zimbabwe	184/5	20	Canada	75/10	19.2	<b>Zimbabwe</b>	109 runs	NA	109
30/09/13	Afghanistan	162/6	20	Kenya	56/10	18.4	<b>Afghanistan</b>	106 runs	NA	106
03/03/16	U.A.E.	81/9	20	India	82/1	10.1	<b>India</b>	9 wickets	59	-105
30/12/10	Pakistan	183/6	20	New Zealand	80/10	15.5	<b>Pakistan</b>	103 runs	NA	103
01/07/15	Netherlands	172/4	20	Nepal	69/10	17.4	<b>Netherlands</b>	103 runs	NA	103
20/04/08	Pakistan	203/5	20	Bangladesh	101/10	16	<b>Pakistan</b>	102 runs	NA	102
13/06/05	England	179/8	20	Australia	79/10	14.3	<b>England</b>	100 runs	NA	100

## 6 Conclusion

The margin of victory is an important statistic in sport as it reflects not only the relative team performances in any one game but can also be used as the basis of forecasting models for future match results. In cricket the measuring margin of victory is not straightforward as the available information at the end of the match depends on whether the winning team batted in the first or in the second innings. However, the problem can be resolved if the number of runs team 2 would have achieved, had they continued batting are estimated. For this purpose, a Generalized Non-Linear forecasting model is proposed to project the expected runs to be scored in the remaining  $u$  overs such that  $w$  is lost. Some properties are associated to the model that are essential for the runs scoring in the Limited Overs International (LOI), for example T20I.

The proposed model does not directly estimate the runs margin of team 2's victory, but can be used to project the second innings runs total. In future work we plan to assess the goodness-of-fit of the model and develop a team ratings model which takes account of margin of victory. Here, we use the model to shed light on the largest margins of victory in T20I cricket history. To date, it appears that Sri Lanka's 172 run victory over Kenya in 2007 is indeed the biggest win ever.

## References

- [1] Clarke, S. R., and Allsopp, P. (2001) Fair Measures of Performance: The World Cup of Cricket, *The Journal of the Operational Research Society* 52, 471-479.
- [2] de Silva, B., Pond, G., and Swartz, T. (2001) Estimation Of the Magnitude of Victory in One-Day Cricket, *Australian & New Zealand Journal of Statistics* 43, 259.
- [3] Allsopp, P. E., and Clarke, S. R. (2004) Rating Teams and Analysing Outcomes in One-Day and Test Cricket, *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 167, 657-667.
- [4] Duckworth, F. C., and Lewis, A. J. (1998) A Fair Method for Resetting the Target in Interrupted One-Day Cricket Matches, *The Journal of the Operational Research Society* 49, 220-227.
- [5] Duckworth, F. C., and Lewis, A. J. (2004) A Successful Operational Research Intervention in One-Day Cricket, *The Journal of the Operational Research Society* 55, 749-759.
- [6] McHale, I. G., and Asif, M. (2013) A modified Duckworth–Lewis method for adjusting targets in interrupted limited overs cricket, *European Journal of Operational Research* 225, 353-362.
- [7] Timur, V., Elzhov, Katharine, M., Mullen, Spiess, A.-N., and Bolker, B. (2016) minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds. R package version 1.2-1.

# A Markovian approach to darts

Francesco Bortolon\*

bortolon.francesco@gmail.com

Cristian Castiglione\*

cristian.castiglione@studenti.unipd.it

Luca Parolini\*

luca.parolini93@gmail.com

Lorenzo Schiavon\*

lorenzo.schiavon@studenti.unipd.it

## Abstract

The aim of this paper is to study and simulate a classical Darts-501 match using Markov chains to describe score evolution. The Markovian approach, indeed, fits the problem since the probability of obtaining a certain score at each step depends only on the result of the last throw and on the score at the previous step. We first study the single dart throw, in order to determine the probability distribution on the dartboard and calculate the probability of hitting each score region, fixed an aiming point; these preliminary results have already been studied in other works. Then, we determine the best strategy the player would choose at each step and we construct the transition matrix of the Markovian process describing the score. We simulate the whole match obtaining results about the average number of steps needed to win it and about the chosen strategies at each step, that both depend on player skill. It is interesting to observe the influence the variance on the player throws has on these data.

## 1 Introduction and game rules

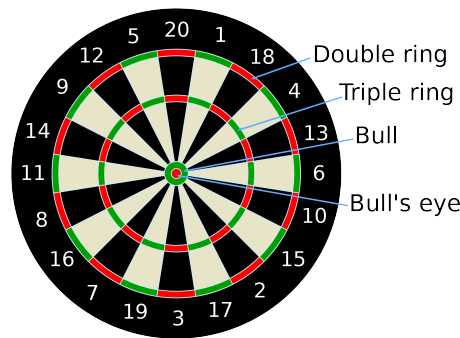
In this paper we construct a model that describes a whole dart match of a single player and then any chosen number of players. After a first part in which we describe the probabilistic model we will construct the transition matrix of the closing part of a match depending on the initial variance of the player studied, in order to simulate many matches and then study results such as average duration in terms of throws and strategies more suitable for each player. To solve this problem we have to make a choice of best strategies that takes in account all possible evolutions of the match in order to construct the transition matrix. This gives us, in the end of our simulation, knowledge of the best strategy given the initial variance on the player's throws.

Darts game consists of throwing little darts at a dartboard, scoring an amount of points that depends on where the player hits the target. The regular dartboard is circular, with a diameter of 453 mm and its center is 173 cm above the ground. The player stands at a distance of 237 cm from the dartboard plane. At each round of the match, the player throws a series of 3 darts and subtracts the scores obtained from the initial total of 501 points, then the match ends when someone gets to 0 points, with a double score at the last throw (see explanation later). The dartboard is divided in 20 slices of equal area, numbered from 1 to 20, as in the figure below, and the center has a central spot called inner bullseye and a small ring around it which is the outer bullseye. There are two more rings at a certain distance from the center that double or triple the score. Let's see how the player scores:

---

\*Università degli Studi di Padova

- hitting a slice, neither the center nor one of the additional rings, scores the slice number;
- hitting the double ring (outer ring), at a distance from the centre from 162 mm to 170 mm, scores twice the slice number;
- hitting the inner ring, or triple ring, at a distance from 99 mm to 107 mm from the center, scores three times the slice number. Notice that triple 20 is the highest possible score;
- the outer bullseye gives 25 points; and the inner bullseye gives 50 points counted as double (diameter 12,7 mm);
- no score is given if the dart misses the dartboard and throws that lead to negative total score, 0 points without a final double score and 1 point.



Using darts-501 rules all dartboard regions are always allowed and no maximum throws number is set. The match must end with a double score that leads one of the player to 0 points. At each round the player does not have to throw all the 3 darts in his hand, so that at the end he could win at the first or second round throw.

## 2 Distribution function of a single throw

Since our final aim is to study the evolution of a whole match, we firstly need to examine what happens during a single throw. The purpose of this section is to find a probability distribution on the dartboard, which is function of the target point chosen. This will be useful in order to describe our method in the next section and to obtain results from simulations. We can describe the dart throw with the laws of parabolic motion; we choose a Cartesian coordinate system such that the  $x$  axis is orthogonal to the dartboard plane and the  $y$  and  $z$  axes are parallel to the target. Let the initial point of the dart in player's hand be the origin of such coordinate system, so that the target stays at a distance  $x = d$ , where  $d = 237cm$ . Hence, the  $y$  and  $z$  coordinates represent the position on the dartboard and they describe what we will call from now on the target point  $(d, y, z)$ . Let  $\vec{v} = (v_x, v_y, v_z)$  be the starting velocity; then the intersection of dart's trajectory with the plane  $x = d$  gives the hit point on the target (or out of it if the player misses the throw). From Physics we have the following equations:

$$\begin{cases} d = v_x t \\ y = v_y t \\ z = v_z t - \frac{g}{2} t^2. \end{cases} \quad (1)$$

Hereafter we assume that the velocity component  $v_x$  is fixed; in our single throw simulation we choose the value  $v^* = 17882 \text{ mm/s}$  ( source: [1]), the average speed of a dart along the shooting line. From the system (1) we can find the hit point:

$$(x, y, z) = \left( d, \frac{v_y}{v_x} d, \frac{v_z}{v_x} d - \frac{g}{2} \frac{d^2}{v_x^2} \right).$$

To model the single throw we want to decide a target point, based on the score that the player wants to achieve. This information gives an associated target speed  $\vec{v}$ , i.e. the speed at which we should throw the dart to hit precisely the target point.

$$\vec{v} = (\bar{v}_x, \bar{v}_y, \bar{v}_z) = \left( v^*, \frac{v^* y}{d}, \frac{v^* z}{d} + \frac{g}{2} \frac{d}{v^*} \right)$$

We calculated also the same results using spherical coordinates; it is indeed useful a spherical approach during simulation to obtain an easy description of target regions.

Notice that the simulation uses two more assumptions: we decided that a player always shoot at the barycenter of the chosen score region and if there is not a unique region which gives a certain score the player will choose the one with greater area (easier to hit).

Now that we have a correspondence between initial velocity and achieved score we can construct the density as a function of the target point. We suppose that the components of the velocity, that will contain an error, have a normal distribution. Hence, we suppose that the shooting effective speed is  $\vec{v} = (v_x, v_y, v_z) \sim \mathcal{N}_3(\vec{v}, \Sigma)$ , where  $\Sigma = \text{diag}(\sigma_x^2, \sigma_y^2, \sigma_z^2)$  is the covariance matrix, as we can assume that velocity errors are independent.  $\Sigma$  depends on the ability of the player. To dramatically simplify the density function, i.e. to make the numerical integration much easier, we can assume, as mentioned previously,  $v_x$  to be fixed and condition on the speed  $v_x$ . Moreover, as the velocity components are independent we obtain:

$$f_{\vec{v}}(v_x, v_y, v_z) = f_{\bar{v}_x}(v_x) f_{\bar{v}_y}(v_y) f_{\bar{v}_z}(v_z).$$

Based on the assumptions we have made and taking advantage of the previous equality we want to write a density function of  $(y, z)$  conditioning on  $v_x$ . We notice that the coordinates  $y, z$  that are hit at each shooting have normal distribution, in particular:

$$\begin{aligned} y &\sim \mathcal{N}(\bar{v}_y t, t^2 \sigma_y^2) \\ z &\sim \mathcal{N}\left(\bar{v}_z t - \frac{g}{2} t^2, t^2 \sigma_z^2\right). \end{aligned}$$

This allows us to write the joint density function of  $y$  and  $z$  as a bivariate normal. And we should rewrite this

function in polar coordinates  $(\rho, \theta)$ :

$$\begin{aligned} f_{\bar{y}, \bar{z}}(\rho, \theta) &= f_{\bar{y}, \bar{z}}(y(\rho, \theta), z(\rho, \theta)) \rho \\ &= \rho \frac{1}{2\pi\sigma_y\sigma_z t^2} \exp \left\{ -\frac{(\rho \cos\theta - \bar{y}_y t)^2}{2t^2\sigma_y^2} - \frac{(\rho \sin\theta - \bar{v}_z t + \frac{g}{2}t^2)^2}{2t^2\sigma_z^2} \right\}. \end{aligned} \quad (2)$$

## 2.1 Score optimization on a single throw

The most convenient strategy during the first part of the match consists of trying to decrease the score as fast as possible until 170 points are reached. Here we enter the final part of the match which we will examine later. So, at first, we want to get a high average score, and we have to find what is the most suitable target region depending on the player variance. We maximize the expected value, varying the target region, in order to find which sector gives the best strategy for the player. Hence, we can use the density function, that we have found in the previous section, to calculate the expected value as a function of the target point region. We use the following integral:

$$\begin{aligned} E_{\bar{y}_y, \bar{v}_z}(s) &= \int_{R_1 \cup \dots \cup R_{83}} S(\rho, \theta) f_{y,z}(\rho, \theta) d\rho d\theta = \\ &= \sum_{i=1}^{83} \int_{R_i} S(\rho, \theta) f_{y,z}(\rho, \theta) d\rho d\theta = \\ &= \sum_{i=1}^{83} S|_{R_i} \int_{R_i} f_{y,z}(\rho, \theta) d\rho d\theta \end{aligned}$$

where  $s$  is the target score, the function  $S$  gives the score correspondent to the hit point  $(\rho, \theta)$  on the dartboard and  $f$  is the density function (2). Notice that the integral splits in a sum of integrals over the 83 possible score regions  $R_i$  and that the function  $S$  is constant when restricted to each of them.

## 3 Construction of the closing process

At 170 points we enter the final part of the match, when it is possible to win the game using at most 3 darts. There are just a few cases (e.g., 159, 169, ...) for which this is not possible and we will see later how to deal with them. As we are interested in studying the closing strategies and the average time duration of the game, we decided to model this part of the match with a stochastic process.

- Let  $E = \{0, 2, 3, 4, \dots, 170\}$  be the state space, containing all the possibile total scores up to 170. Notice that 1 is excluded, because rules cancel scores that lead to 1, state which make it impossible winning with a double score.
- Let  $S = \{1, 2, 3, \dots, 20\} \cup \{2, 4, 6, \dots, 40\} \cup \{3, 6, 9, \dots, 60\} \cup \{25, 50\}$  be the set of all the total scores achievable with a single throw.
- Let  $\{X_n\}_{n \geq 0}$  be the stochastic process that describes the total score achieved after the  $n$ -th throw and such that  $X_0$  is the maximum of the total scores  $s$  reached during the match such that  $s \leq 170$ .

The process  $\{X_n\}_{n \geq 0}$  is a discrete time Markov chain with state space  $E$ . Indeed, the probability of reaching the state  $j \in E$  after the  $(n+1)$ -th throw depends only on the total score at the  $n$ -th one and on the last throw's score; as we can see the Markov condition is verified:

$$\mathbb{P}[X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0] = \mathbb{P}[X_{n+1} = j | X_n = i].$$

Indeed, defined the random variable  $L_i$  of the score achieved at the  $i$ -th throw, we get

- $\mathbb{P}[X_{n+1} = j | X_n = i] = \mathbb{P}[X_n - X_{n+1} = i - j | X_n = i]$   
 $= \mathbb{P}[L_{n+1} = i - j | X_n = i]$   
 $= \mathbb{P}[L_{n+1} = i - j];$
- $\mathbb{P}[X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0] = \frac{\mathbb{P}[X_{n+1} = j, X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0]}{\mathbb{P}[X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0]}$   
 $= \frac{\mathbb{P}[L_{n+1} = i - j, X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0]}{\mathbb{P}[X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0]}$   
 $= \frac{\mathbb{P}[L_{n+1} = i - j] \mathbb{P}[X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0]}{\mathbb{P}[X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0]}$   
 $= \mathbb{P}[L_{n+1} = i - j].$

The state 0 is an absorbing state and marks the end of every match. We now need to study the transition matrix of the process.

### 3.1 Transition matrix

In order to construct the transition matrix of the closing process, we have to notice that the transition probability between states in  $E$  depends on the chosen strategy, or in other words on the target score decided for the next step. We will chose an optimal strategy associated with each state  $i \in E$ , so that is now possible constructing, one row at a time, the transition matrix  $P$ . It has the following structure:

- $p_{0j} = \delta_{0j}$  since 0 is an absorbing state;
- $p_{ij} > 0$  if and only if it exists  $s \in S$  such that  $j = i - s$  and the states are not  $(i, j)$  with  $i$  odd and  $j = 0$ . Transition probability is non zero if the player can move from  $i$  total score to  $j$  with a single throw (double score if we are going towards the absorbing state);
- $p_{ij} = 0$  when  $i < j$  as follows from the previous point. Notice that  $P$  is a lower triangular matrix;
- given  $i \in E$  and the next target score  $s \in S$ , we find  $p_{ij}$  integrating the density function (2) relative to  $s$  on the regions that give  $i - j$  score. Thus, we say when  $i \geq j$  that:

$$p_{ij} = \int_{R_{(i-j)}} f_{\bar{\rho}, \bar{\theta}}(\rho, \theta) d\rho d\theta.$$

Here  $\bar{\rho}, \bar{\theta}$  is the target point correspondent to the optimal strategy and  $R_{(i-j)}$  is the union of all regions that allow to score as wanted. Notice also that when we consider  $p_{i0}$  integration area must not contain all score regions that give  $i$  points not double.

Using the matrix  $P = (p_{ij})_{i,j \in E}$  we are interested in studying absorbing properties in 0.

### 3.2 Choice of strategies

We want, hence, to choose a unique optimal strategy (or better target score) for each total score in the state space. We suppose that the player will find the best strategy taking on account sequences of 3 consecutive throws, as happens during a multi-player match. We decided to make the following hypothesis:

- if a certain state admits a single-throw closing strategy we choose that one as optimal. We decide to always prefer these ones to other strategies with 2 or 3 throws since more throws give more error possibilities. Moreover if the player does not win the match he has 2 more steps to do it before moving to other player's rounds. States that allow a single-throw strategy are  $i \in \{2, 4, 6, \dots, 40, 50\}$ ;
- if we cannot win with at most 3 throws we just maximize the average score as seen in (2.1); e.g.  $i = 169, 159, \dots$ ;
- otherwise we choose as optimal strategy a 2 or 3 throws one that generates the  $i$ -th row in the transition matrix, such that gives the minimum average absorbing time  $k_i^{\{0\}}$  in  $\{0\}$ , starting from  $i$ . We construct the set  $\mathcal{H}^i := \{h \in S : \text{exists strategy from score } i \text{ with } h \text{ as first target score}\}$  and we want to find a solution of the following linear system of equations:

$$\begin{cases} k_l^{\{0\},h} = 1 + \sum_{j \in E} p_{lj}^h k_j^{\{0\},h} \\ k_0^{\{0\},h} = 0 \end{cases} \quad \forall h \in \mathcal{H}^i, l = 0, 1, \dots, i; \quad (3)$$

where  $k_i^{\{0\},h}$  is the average absorbing time in  $\{0\}$  starting from  $i$  and chosen the strategy  $h$ ;  $p_{lj}^h$  is the transition probability between  $l$  and  $j$  supposed  $h$  as target score. When we know the solutions relative to every strategy we can choose the strategy with target score  $\bar{h} \in \mathcal{H}^i$  such that:

$$k_i^{\{0\},\bar{h}} = \min_{h \in \mathcal{H}^i} \{k_i^{\{0\},h}\}.$$

as optimal.

Notice that there exists a solution of the linear systems (3) since the transition matrix is lower triangular; thus we can construct optimal strategies beginning at  $i = 2$ .



---

**Algorithm** Construction of transition matrix.

---

1. Let the first row  $p_{i=0} = \vec{0}$
  2. For  $i = 2, \dots, 170$  :
    - (a) Find, if any exists, the 1-throws strategy and use it to compute the vector of probabilities  $p_i$  equal to the  $i$ -th row of the transition matrix.
    - (b) If 1-thrown strategy does not exist, find the  $n$  2-throws and 3-throws strategies.
    - (c) For  $j = 1, \dots, n$  :
      - i. Compute the vector of probabilities  $p_i^j$  for the  $i$ -th row of transition matrix.
      - ii. Compute the mean absorption time from state  $i$  to 0.
    - (d) Choose the strategy  $\bar{j}$  which  $p_i^{\bar{j}}$  produces the minimum mean absorption time from state  $i$  to 0. Let  $p_i = p_i^{\bar{j}}$ .
    - (e) If  $n = 0$ , compute  $p_i$  choosing *bestdown* as strategy.
- 

**3.2.1 Multi-player hypothesis**

We can also consider a match between two players. In this case, a player that has positive probability to end the match in his round could choose a strategy that allows him to do it, even if it is not the best strategy considering the average absorbing time. If a player has still two darts available, he chooses a good two-throws strategy rather than the best three-throws strategy that increases the possibilities for an other player to play his round and maybe to win the match.

Taking account of this possibility means changing the transition matrix: we triple each state that from now on will be a couple  $(i, j)$ , with  $i$  the total score and  $j$  the remaining darts of the current round. So the states are now  $E = (2, 1)(2, 2)(2, 3)(3, 1) \dots (170, 1)(170, 2)(170, 3)$  and the transition matrix has 508 states (we do not triple the first one since it is an absorbing state). When a player has a single throw left the next state will have second component 3.

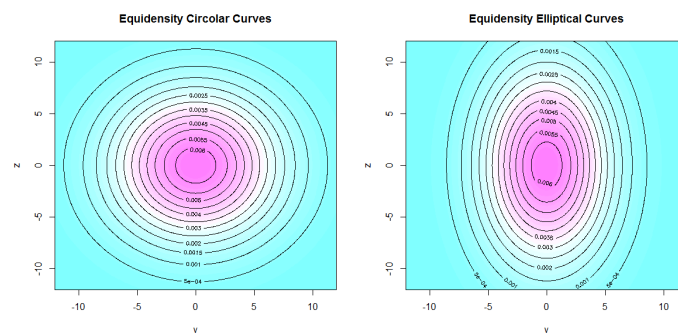
Notice that we have positive transition probability between two states only if remaining darts decrease. Formally if  $\alpha := (i, j)$ ,  $\beta := (l, h)$  are two states the transition probability  $p_{\alpha\beta}$  is non zero if and only if we have same conditions as before on the first component about the score and also  $h = j - 1$  or  $h = 3$  if  $j$  was 1. Thus it is possible to construct iteratively the transition matrix by inserting blocks of three lines such that they get a minimal solution to the linear system for absorbing times. In the place of each equation of the single-player system we have now:

$$\begin{cases} k_{l,1}^{\{0\},h} = 1 + \sum_{j \in E} p_{(l,1)(j,3)}^h k_{j,3}^{\{0\},h} \\ k_{l,2}^{\{0\},h} = 1 + \sum_{j \in E} p_{(l,2)(j,1)}^h k_{j,1}^{\{0\},h} \\ k_{l,3}^{\{0\},h} = 1 + \sum_{j \in E} p_{(l,3)(j,2)}^h k_{j,2}^{\{0\},h} \end{cases}$$

with notation as in the previous case of single player. It can be easily seen that the system has still a solution, as the transition matrix is nearly lower triangular. In this way it is possible to compare all different combinations of three strategies (one for each number of remaining darts) that allow to end a match at a given score. In order to satisfy the initial condition of this paragraph, we imposed that a player in a state of the form  $(i, 2)$  chooses only one-throw or two-throws strategies (if any exists). The choice among the combination of strategies is made by comparing the average of the three absorbing times, because the problem is symmetric on the number of darts in hand. This definition of the problem allows us to extend the modelling of a match to any number of players.

## 4 Simulation results

We supposed in the previous sections the distribution of the single throw to be normal. As it is known the equidensity curves of a normal distribution are circumferences when variances of both components of the aleatory vector coincide. If the components are on a certain ratio we can find other curves such as ellipsis.



Thus we simulated both situations, studying at first the symmetric hypothesis. On the other hand, when considering asymmetric case we supposed variances to be such that  $\sigma_z^2 = 3\sigma_y^2$ , which means a higher error on the vertical component of velocity. Notice that the images above show that the asymmetry is not really pronounced and this is a reason that explains why the two cases give slightly equal results during simulation.

We used Monte Carlo method to simulate the matches, on a basis of 10.000 matches and testing 8 different dispersion radii for both density functions, symmetric and not. So we simulated a total amount of 160.000 matches. We define the *dispersion radius* as the radius that cuts a circular region on the dartboard which the 95 percent of the player's throws hit; so this parameter is a valid precision indicator, that helps studying how results vary at different player's skills. Precisely dispersion radii are  $2\sigma$ , with  $\sigma$  being the common standard deviation in the symmetric case. In the other setting we use:  $\sigma_y = \sqrt{\sqrt{1/3}\sigma^2}$  e  $\sigma_z = \sqrt{\sqrt{3}\sigma^2}$ . We implemented a script that using our model simulates the whole match and constructs the transition matrix as explained. There are some data that are interesting and were studied after simulation:

- *Bestdown*: target score of the first part of the match, that gives the maximum average score at each throw;
- *Pbullseye*: probability of hitting the inner bullseye aiming at it;

- *Peye* probability of hitting the inner or outer bullseye aiming at the inner bullseye.
- *Kmax*, *Kmin*, *Kmean*: maximum, minimum and mean of the mean absorbing times of the transition matrix, respectively.

As noticed before we find similar results in both the different settings and this is due on one hand to approximation errors and on the other to the slightly small asymmetry of the variance. Moreover variances were chosen such that the level curves of the bivariate normal distribution define the same area in both corresponding circular or elliptic shapes, in order to better compare results.

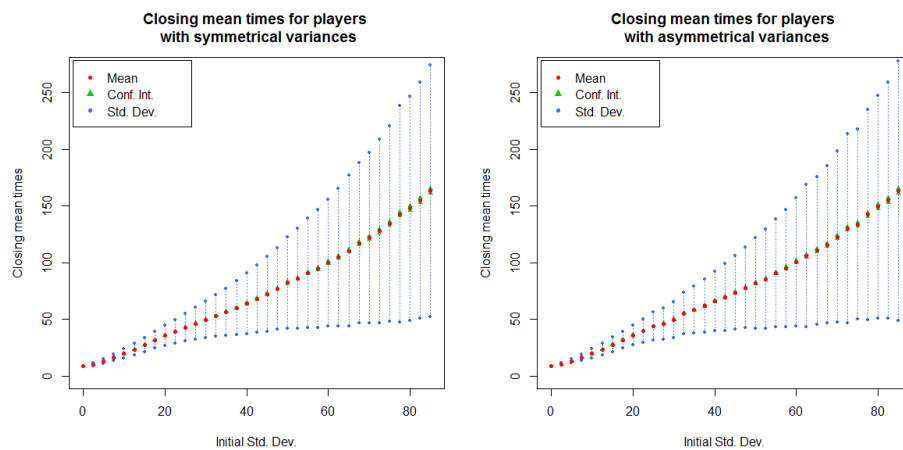


Figure 1: Average match duration

Results are reasonable since it is evident that the higher player's precision (or equivalently the smaller dispersion radius) the less steps are needed to end a match, as expected; we see that also standard deviation associated to the average duration of a match drops. That behaviour is evident in graphs of Figure (1).

Looking at *Bestdown* values (see tables at the end), we can understand how heavily the initial variance of a player influences the strategy, also during the first part of the match. Each player will choose, in our hypothesis, a constant strategy until they get 170 points or less for the first time. In particular, we see that if  $r \leq 31.875$  mm the best average score is obtained aiming at triple 20. Otherwise for player's who have a high dispersion radius the best average score is achieved aiming at the bullseye in order to be quite certain at least of hitting the dartboard. Between these particular cases there are a lot of intermediate data which can be seen in the table (2) at the end of this paper.

An idea of the evolution of the whole match in our simulation is given by graphs of Figure (2) that are respectively referred to players with dispersion radii 25 mm and 50 mm. Here the two different parts of a match are evident; indeed in the first part of the graphs the target score is constant and so the strategy, which is the one that leads as fast as possible to reach 170 points or less. Then the strategy varies depending also on mistakes occurred and this happens in the ending part. Since we have avoided the trivial case of a very precise player there is sometimes a gap between target score and hit region.

It is interesting saying something else on the relation between average match duration and player's dispersion radius. We would like to see what distribution have the number of throws needed to win the match,

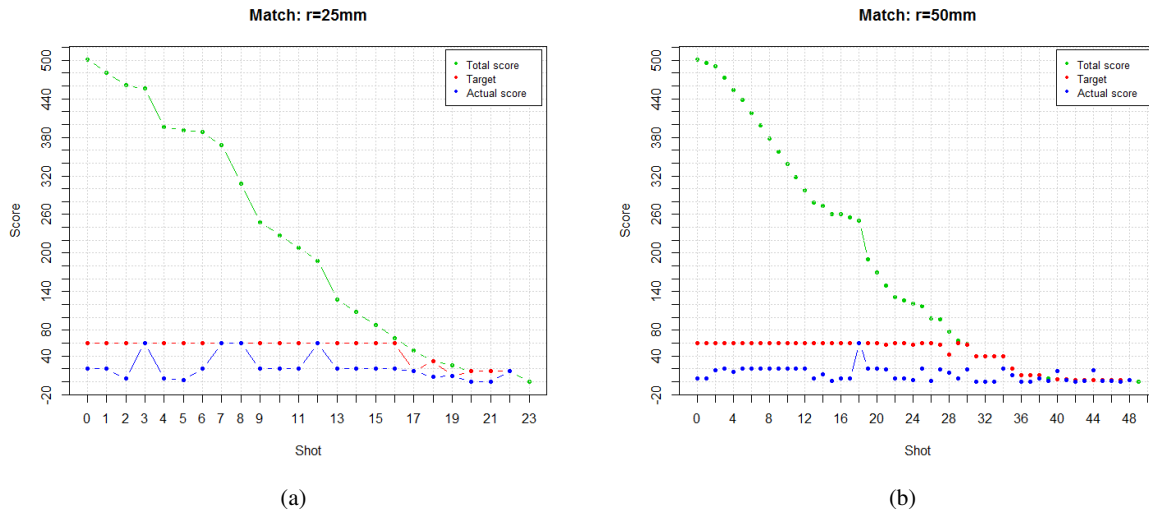


Figure 2: Evolution of the whole match

varying the player’s ability. As we know the higher the dispersion radius the higher number of steps will last a match. We could also expect that the average duration range widens when there are a lot of independent repeated trials. Indeed looking at the graphs in Figure (3), we see that the empiric frequencies converge at a continuous curve when the player is not precise, while they keep discrete for good players. The reason is that the last ones win the match with always nearly the same number of throws.

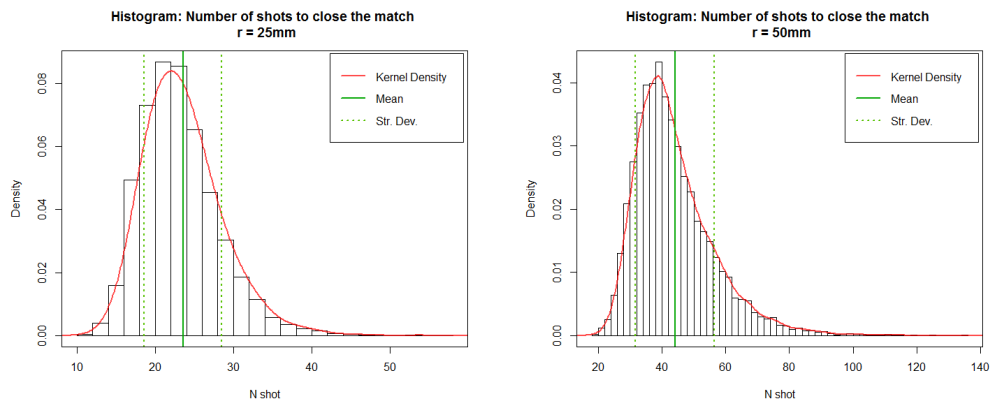


Figure 3: Histograms of number of throws

Finally, we can see a comparison between the different curves that approximately fit the histograms:

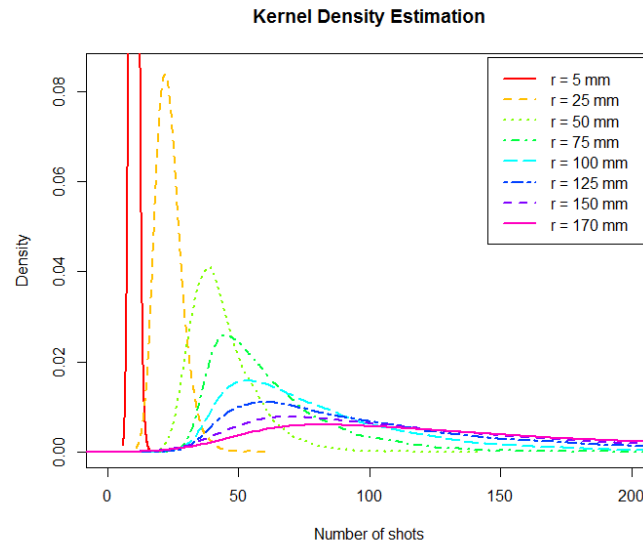


Figure 4: Non parametric kernel density estimation of number of throws

#### 4.1 A simulation with empirical data

The aim of the paper was to prove how the Markovian approach to darts increases the performances choosing the best possible strategy in every round of the match. Thus, in this section, we show an example of how our algorithm works if the probability matrix that describes the distribution on the dartboard for every possible target is known. Our data were collected by the repetition of one hundred throws for every target on the dartboard. Hence, we compare the performances of two virtual players with the same probability matrix, that means the same precision on the single throw, with a Monte Carlo simulation of 10000 matches for both of them. The only difference between the two players is the choice of strategies for every dart throw: the first player (*Smart Player*) takes advantage of the Markovian approach, whereas the second player (*Naive player*) uses a more intuitive and naive strategies in this order:

- if any 1-throw strategies exists, player chooses it;
- if any 2-throws strategies exists, player chooses one of them;
- if any 3-throws strategies exists, player chooses one of them;
- otherwise player targets T20.

The outcome, presented in table 1, shows a remarkable difference in closing time, shorter and less variable for *Smart Player*. This result could make the difference between winning or losing a real match. Thus, we are able to assert that the approach explained in this paper can give an actual competitive advantage to the player, whatever will be his single-throw distribution if it allows to define probabilities on the dartboard's sectors for every target region.

	<b>M</b> throws	<b>95% CI for M</b> throws	<b>sd</b> throws	<b>P</b>	<b>95% CI for P</b>
<b>Smart Player</b>	32.19	(32.03; 32.35)	7.98	0.623	(0.614; 0.632)
<b>Naive Player</b>	36.38	(36.20; 32.56)	9.01		

Table 1: Comparison of results in a simulation with real data.

*M*: average of number of throws to close the match.*P*: probability of winning for *Smart player* versus *Naive player*

## 5 Conclusions

We tested our model and it proved to work and produce reasonable data. We studied empirically how the initial variance influences the choice of strategies during a dart match and average duration of a match. It is important to notice that once we have constructed the transition matrix of a player we can know exactly what his best strategy should be, varying how the match goes on. In our paper we have chosen as reasonable optimization criterion the mean time of absorption, but we have not given a proof that it is the best criterion among the several indexes of the distribution of times absorption. These may be investigated in future.

Dispersion Radius mm	Average Throws	Variance Throws <sup>2</sup>	Std. Dev. Throws	K max Throw	K mean Throws	K min Throws	P bullseye	P eye	BD p	BD t
<b>5</b>	10,25	1,13	1,06	4,11	2,47	1,08	0,9603	1,0000	60	3
<b>10</b>	13,28	3,97	1,99	5,32	3,48	1,74	0,5536	0,9940	60	3
<b>15</b>	16,46	8,48	2,91	6,83	4,55	2,46	0,3012	0,8973	60	3
<b>20</b>	19,81	15,43	3,93	8,44	5,68	3,25	0,1826	0,7220	60	3
<b>25</b>	23,61	25,33	5,03	10,21	6,97	4,14	0,1211	0,5592	60	3
<b>40</b>	35,68	76,47	8,74	16,56	11,95	7,81	0,0492	0,2739	57	3
<b>55</b>	46,26	200,27	14,15	24,03	18,53	13,16	0,0263	0,1557	57	3
<b>70</b>	56,58	430,13	20,74	32,71	26,67	20,21	0,0163	0,0992	21	3
<b>95</b>	77,09	1288,64	35,90	50,73	43,93	35,73	0,0089	0,0552	21	3
<b>120</b>	99,86	3105,55	55,73	72,97	65,75	55,97	0,0056	0,0349	50	2
<b>145</b>	127,86	6533,17	80,83	100,29	92,56	80,91	0,0038	0,0241	50	2
<b>170</b>	162,99	12281,48	110,82	133,01	124,57	110,56	0,0028	0,0176	50	2

Table 2: Results of simulations with symmetric variance, on the axes  $y$  e  $z$ .  $K$  stands for the total number or steps of a match. The last two columns show the target region before 170 points; there are score and multiplicity.

Score	Sym Variance Player			Asym Variance Player		
	1-throw str.	2-throw str.	3-throw str.	1-throw str.	2-throw str.	3-throw str.
48	16	16	16	16	16	16
49	T 14	T 11	T 14	T 14	17	T 14
50	D 25	D 25	D 25	D 25	D 25	D 25
51	19	19	19	19	19	19
52	T 16	T 16	T 16	T 16	T 16	T 16
53	T 17	T 17	T 17	T 17	T 17	T 17
54	T 14	T 14	T 14	T 11	T 14	T 11
55	T 17	T 17	T 17	19	19	19
56	T 16	T 16	T 16	20	20	20
57	T 17	T 17	T 17	T 17	T 17	T 17
58	T 18	T 18	T 18	T 18	T 18	T 18
59	T 19	T 19	T 19	T 19	T 19	T 19
60	T 16	T 16	20	T 8	T 8	T 8
61	T 19	T 19	T 19	T 19	T 19	T 19
62	T 14	T 14	T 14	T 14	T 14	T 14
63	T 14	T 11	D 25	T 14	T 11	T 14
64	T 16	T 16	T 16	T 8	T 8	T 8
65	T 19	T 19	T 19	T 14	T 11	T 14
66	T 16	T 16	T 16	T 16	T 16	T 16
67	T 14	T 11	T 14	T 11	T 11	T 11
68	T 14	T 14	T 9	T 14	T 14	T 14
69	T 7	T 19	T 7	T 11	T 11	T 11
70	T 14	T 14	T 14	T 11	T 14	T 11
71	T 7	T 7	T 7	T 7	T 7	T 7
72	T 16	T 16	T 16	T 8	T 16	T 8
73	T 19	T 19	T 19	T 14	T 11	T 14
74	T 7	T 16	T 7	T 7	T 14	T 7
75	T 7	T 19	T 7	T 8	T 19	T 8
76	T 20	T 20	T 20	T 20	T 20	T 20
77	T 7	T 19	T 7	T 11	T 19	T 11
78	T 19	T 14	T 19	T 11	T 14	T 11

Table 3: Comparison between the optimal strategies of two players with a dispersion radius of 70mm for some scores. We can notice that *asymmetric player* prefers to target the more vertically elongated sectors (e.g. T11 rather than T14)



*A Markovian approach to darts*

*Bortolon Francesco, Castiglione Cristian,  
Parolini Luca, Schiavon Lorenzo*

## References

- [1] <http://www.dartbase.com/Sect1/17.html>.
- [2] R. J. Tibshirani, A. Price, J. Taylor. A Statistician Plays Darts. *Journal of the Royal Statistic Society*, 174: pp.213-226, 2011.
- [3] James R. Norris *Markov Chains*. Cambridge University Press, 1997.
- [4] Sheldon M. Ross *A First Course in Probability*. Pearson Prentice Hall, 1976.
- [5] J. Jacod, P. Protter *Probability Essentials*. Springer, 1999.
- [6] C. P. Robert, G. Casella *Introducing Monte Carlo Methods with R*. Springer, 2010.
- [7] D. Kleppner, R. Kolenkow *An Introduction to Mechanics*. Cambridge University Press, 1973.

# Real time measurement of individual influence in T20 cricket

Paul J. Bracewell<sup>a,b,c</sup>, Ankit K. Patel<sup>a,b</sup>, Jason D. Wells<sup>a</sup>

<sup>a</sup> *DOT loves data*

<sup>b</sup> *Victoria University of Wellington*

<sup>c</sup> *Corresponding author: paul@dotlovesdata.com*

## Abstract

A framework for quantifying the influence of an individual competing in a limited overs game of cricket is outlined. Using ball-by-ball data, a resource model is constructed using an isotonic regression and MCMC Gibbs sampling procedure. As a consequence, the impact of an action relative to the current match state can be evaluated. This comparison of observed outcome to an expected outcome on a ball-by-ball basis enables real time player tracking. The cumulative impact of these batsmen-bowler interactions allows individual influence in a match to be quantified. The developed finalized model is validated through live and static application.

## 1 Introduction

Cricket is a team game based on the balance of two key resources: 1) balls and 2) wickets. Simply, the batting team that utilizes these two resources most effectively will win the match. As an inning progresses, the total number of resources allocated to the batting team decreases. The batting team aims to score as many runs as possible given an allocated number of resources (balls and wickets), while the bowling teams aims to restrict the total number of runs conceded, by taking wickets. The bowling teams overall goal is to deplete the batting team resources as quickly as possible for the least number of runs. This is evident in limited overs matches, where each team bats only once with a maximum number of overs specified.

The first innings batting team is assigned the task of optimizing the total number of runs scored given two resource constraints: 1) *balls* and 2) *wickets*. As wickets are lost batsmen *value* decreases as they go down the batting order after the initial 4-5 batsmen (Duckworth & Lewis, 1998). The team batting second is assigned the task of outscoring the first batting team, given the allocated resources. The first innings reaches completion when all resources have been depleted, while the second innings reaches completion when all resources are depleted *or* the team batting second has achieved the target score. “The optimisation exercise in either team’s task involves choosing some compromise between scoring fast and hence taking higher risks of losing wickets, and playing carefully and hence risking making insufficient runs” (Duckworth & Lewis, 1998, pg. 220).

The core of cricket is intertwined with numerical values that ultimately translate to a match result. However, given its numerical depth there is scarce academic and commercial literature regarding the application of analytical techniques within cricket, relative to other sports. The most notable application of analytics within cricket is the Duckworth Lewis (1998) resource allocation method. Duckworth and Lewis (1998) developed a framework which mathematically allocates resources, balls and wickets, in order to appropriately reset or recalculate target scores during interrupted one-day cricket matches. Generally, this framework is currently implemented by the International Cricket Council (ICC) as the primary method to recalculate the target score during an interrupted limited overs cricket match.

## 2 Literature Review

Although there are a small number of published articles on the application of analytics within cricket, there is increasing analytical literature and the adoption of predictive methodologies at the professional level. It has been noted that “during the past decade a large number of academic papers have been published on cricket performance measures and predictive methods” (Lemmer, 2011, pg. 1).

Critically, there remains an academic and commercial gap surrounding real-time or dynamic predictive rating systems. Proceeding is a review of the most notable academic literature outlining the application of statistical techniques to ball-by-ball cricketing data:

Clarke (1988) applied a dynamic programming model to one-day cricket to: 1) calculate the optimal scoring rate, 2) estimate the total number of runs to be scored in the first innings and 3) estimate the probability of winning in the second innings. These estimates are derived during any stage of an innings. The first innings formulation allowed the development an ‘*optimal scoring model*’ outlining a team’s optimal scoring rate (i.e. runs per over) to obtain a given expected total, for any given number of wickets in hand and balls remaining. The second innings formulation enabled the development of a ‘*probability scoring table*’ outlining the probability of the second innings batting team scoring the target total, for any given number of wickets in hand and balls remaining.

Similarly, Davis, Perera and Swartz (2015) developed a T20 simulator that calculated the probability of a first-innings batting outcomes dependent on batsmen, bowler, and number of overs consumed and total wickets lost. These probabilities were based on an amalgamation of standard classical estimation techniques and a hierarchical empirical Bayes approach, where the probabilities of batting outcomes borrow information from related scenarios (Davis *et al.*, 2015). Simulation suggested that batting teams were not incrementally increasing aggressiveness when falling behind the required run rate.

Swartz, Gill and Muthukumarana (2009) developed a discrete generator simulator for one-day cricket. Applying a Bayesian Latent model, ball-by-ball outcome probabilities were estimated using historical ODI data and were dependent on batsmen, bowler, total wickets lost, total balls bowled and current match score. It was found that the proposed simulator produced reasonably realistic results, with the actual runs and simulated runs revealing an excellent agreement. Moreover comparing wickets taken, the actual results compared favorably with simulated results.

Duckworth & Lewis (2005) developed real time player metrics, using the Duckworth-Lewis methodology, to evaluate player contribution at any given stage of an innings, producing context based measures. The developed metrics were 1) Batsmen average run contribution per unit of resources consumed and 2) Bowlers’ average runs contribution per unit resources consumed. Applying these measures to the 2003 VB series final (Australia vs. England) it was shown that the Duckworth-Lewis based contribution measure were less susceptible to distortions compared to traditional performance metrics.

Brown, Patel and Bracewell (2016) investigated the likelihood of an opening batsman surviving (i.e. not being dismissed) each ball faced over the course of an innings. Using model formulation and selection techniques, Brown et al. (2016) developed a contextually and statistically significant Cox proportional hazard model that was capable of predicting the probability of survival for any opening batsmen (i.e. opener), given certain model conditions. Practically and statistically significant predictors were: 1) cumulative number of runs scored, 2) cumulative number of consecutive dot balls faced and 3) cumulative number of balls faced in which less than two runs in four balls had been scored. The results illustrated that as the magnitude of the three predictors increased for a given opener, the associated survival probabilities for the batsman either remained constant or decrease on a ball-by-ball basis.

### 3 Research Question

The objective of this research was to develop a real-time predictive system that measures the amount of influence a player exerts on a T20 match during any stage of an innings. The development of this real-time system has considerable implications for a variety of stakeholders, such as players, coaches, managers and franchise owners. Addressing this research question players can isolate specific match scenarios in which individual performance drops or increases, indicated through match influence score. Moreover coaches and managers can apply the model results to create effective, player specific training regimes, determine optimal batting line-ups, player-selection decisions and in-game strategies. The goal of this research was to develop a real-time predictive system that focused on measuring a player's T20 influence within the team context.

### 4 Research Methodologies

#### 4.1 Influence Definition

To identify a player's influence during any stage of a match, performance metrics that affect match outcome needed to be established. The type of performance metrics needed to accurately calculate players match influence were identified through expert opinions. Expert referral suggested three key dimensions in order to identify influence: 1) *volume* of contribution, 2) *efficiency* of contribution and 3) contribution under *pressure*.

#### 5.2 Research Methodologies

To effectively measure a player's ball-by-ball influence a number of conventional performance metrics were not applicable, as many are undefined during a match, while others rely on the occurrence of a particular event in order to be defined. Moreover performance metrics that are utilized in cricket are not considered '*advanced*' metrics; primary metrics are very traditional and are unobservable on a ball-by-ball basis. Therefore the main challenge facing the success of this research was the creation and implementation of features which were: 1) observable on a ball-by-ball basis and 2) defined during various stages of an inning. Appendix A lists performance metrics and associated definitions.

### 5 Data

The model development process implemented ball-by-ball observations from the Indian Premier League (2014, 2015, and 2016), Caribbean Premier League (2014/2015, 2015/2016), English NatWest T20 Blast (2015/2016) and Australian Big Bash League (2014, 2015). These are T20 tournaments, where each team can bat a maximum of 20 overs. The commentary logs for matches were programmatically extracted; approximately 95,000 observations. Ball-by-ball data was extracted from ESPNcricinfo (<http://www.espncricinfo.com/>). An automated scripting process was developed to extract and parse the commentary log, and provide a more convenient data structure. The process extracted relevant details on a ball-by-ball basis and stored the data in a tabular form for easy access; appendix B illustrates data structure post extraction. Rain interrupted and abandoned matches were removed from the dataset. Moreover, the models outlined throughout section 6.1-6.4 were built using 50% of the data for training and 50% for testing.

### 6 Feature Creation Methodologies

This section outlines the methodologies implemented to extract features on a ball-by-ball basis. This section contains the following subsection: *Subsection 6.1* develops a T20 ball-by-ball resource table using

the methodology outlined in Bhattacharya, Gill & Swartz (2011). *Subsection 6.2* outlines a novel method to calculate the expected number of runs scored by the batting team. *Subsection 6.3* applies and extends the survival analysis methodology, described in Brown *et al.* (2016), to second innings and non-opening batsmen, generating survival probabilities for all batting positions. *Subsection 6.4* outlines, the method to derive an individual's match influence.

## 6.1 T20 Ball-by-Ball Resource Table

The resources available during any stage of an innings was calculated using a modified Duckworth-Lewis system developed by Bhattacharya, Gill & Swartz (2011), because the Duckworth & Lewis (1998) resource allocation method was originally designed for one-day (50-over) cricket. Bhattacharya, Gill & Swartz (2011) applied a Gibbs sampling scheme relating to isotonic regression to observed scoring rates to produce a non-parametric ball-by-ball resource table. Over-by-over resource table results were interpolated on a ball-by-ball basis. Isotonic regression is a technique to deal with such constraints. The following minimization equation was considered:

$$F = \min_{y_{uw}} \sum_{u=1}^{20} \sum_{w=0}^9 q_{uw} (r_{uw} - y_{uw})^2, \quad (1)$$

where  $r_{uw}$  is the estimated percentage of resources remaining where  $u$  overs are available and  $w$  wickets have been lost:

$$r_{uw} = \frac{\text{mean}[x(u,w)]}{\text{mean}[x(20,0)]}, \quad (2)$$

$x(u, w)$  represents the runs scored from the stage in the first innings where  $u$  are overs available and  $w(u)$  wickets have been taken until end of the first innings. The optimization is with respect to matrix  $y_{uw}$ , where the double summation corresponds to  $u = 1, \dots, 20$  and  $w = 0, \dots, 9$  and  $q_{uw}$  are weights.

The optimization problem is subject corresponds to the following constraints:

1.  $y_{uw} \geq y_{u,w+1}$
2.  $y_{uw} \geq y_{u+1w}$
3.  $y_{10,0} = 100$ ; for  $w = 0, \dots, 9$
4.  $y_{0,w} = 0$
5.  $y_{u,10} = 0$ ; for  $u = 1, \dots, 20$

A minimization with squared error discrepancy corresponds to the method of constrained maximum likelihood estimation where the data  $r_{uw}$  are independently normally distributed with means  $y_{uw}$  and sample variance  $\frac{1}{q_{uw}}$ <sup>1</sup>. The results of the isotonic regression revealed a few limitations: 1. many adjacent entries had the same value, due to the various fitted  $y$ 's occur on the boundaries imposed by the monotocity constraint. 2. The resultant resource table suffered from incompleteness, missing values corresponding to match situations where data are unavailable. Recognizing that the optimisation problem arises from a normal likelihood, a Bayesian model, with  $y$ 's as unknown parameters was adopted. Assigning a flat default prior to the  $y$ 's subject to the monotocity constraints. A posterior density with form:

$$\exp \left\{ -\frac{1}{2} \sum_{u=1}^{20} \sum_{w=0}^9 q_{uw} (r_{uw} - y_{uw})^2 \right\}, \quad (3)$$

and Gibbs sampling can be carried out via sampling from the full conditional distributions:

---

<sup>1</sup>  $\frac{1}{q_{uw}} = \frac{r_{uw}}{n}$ , where  $n$  = sample size

$$[y_{uw} | \cdot] \sim \text{Normal}\left(r_{uw}, \frac{1}{q_{uw}}\right), \quad (4)$$

subject to the local constraints on  $y_{uw}$  in the given iteration of the algorithm. In the spirit of Bayesian statistics prior information is applied to impute missing data by imputing missing  $y$ 's with Duckworth-Lewis table entries. Knowing [ball-by-ball] resource availability allows the calculation of resources based metrics, incorporating a 'time' factor and reveals how players are performing during various stage of an inning and varying resource constraints. Resource availability will be a heavily used metric throughout the paper.

## 6.2 Expected Runs Model

To understand the ball-by-ball contribution an individual player creates, each ball outcome affected the overall [expected] total was measured. For example, calculating the difference between the *expected total<sub>i</sub>* and *expected total<sub>i-1</sub>*, ball-by-ball contribution measurements can be evaluated. For example, if the current ball expected total = 129, and batsmen A hits a boundary four off the following delivery, increasing the expected total to 135, the batsmen runs contributed = 6 (135 -129), while the bowlers runs contributed = -6. Such performance metrics indicate player contribution at a team-level as oppose to an individual level; explained by metrics such as batsmen strike rate, bowlers' and economy rate etc.

To generate contribution metrics, a [ball-by-ball] expected runs model was developed. Applying a gradient boosted machine (GBM) algorithm with a Poisson distribution and 25,000 iterations, two separate models (first and second innings) predicting the expected runs on a ball-by-ball basis were developed:

1. Model 1 utilized first innings data, total wickets, total balls, team strike rate, projected total, team percentage boundaries, team runs, current run rate and resources remaining.
2. Model 2 utilized second innings data, team runs, total wickets, team strike rate, resources available, projected total, team percentage dots, current run rate and team percentage boundaries.

Actual innings total results were benchmarked against predicted total. Table 1 outlines accuracy measurements across the models.

Model	Accuracy Measurements			
	Correlation	Adjusted- $R^2$	RMSE	MAE
Model 1	0.73	0.53	18.6	14.6
Model 2	0.57	0.33	20.1	16.6

**Table 1 Accuracy metrics for expected runs model**

## 6.3 Batsmen Survival Probabilities

Applying and extending the methodology outlined in Brown *et al.* (2016) to second innings and non-opening batsmen (i.e. top, middle, lower and tail), a batsmen's survival probability (i.e. probability of dismissal) for any given ball of an innings can be calculated. Brown *et al.* (2016) established 3 criteria during the model development process: 1) Estimated model coefficients had to make practical sense, 2)

decrease in resources availability leads to a decrease in the likelihood of a batsman surviving the next ball. (As resources decrease, pressure increase causing greater risk to be taken on a ball-by-ball basis leading to lower levels of survival), and 3) Probability of batsman survival decreased on a ball-by-ball basis as resources are monotonically decreasing.

Models that complied with these criteria were kept in the candidate set. Table 2 outlines the metrics that met the three requirements, and were statistically significant, across each of the five batting positions (opener, top, middle, lower and tail), for the first and second innings.

Utilising a Cox proportional hazard modeling technique, the time taken to dismissal for a given batsmen was implemented to represent the response variable. The total number of balls faced by a given batsmen represented the time till failure (i.e. dismissal). The Cox proportional hazard model has the following form:

Innings	Batsman Type				
	Opener	Top	Middle	Lower	Tail
First Innings	Contribution	Strike rate	Strike rate	Strike rate	Dots faced
	Less than 2 in 4 total	Dots faced	Dots faced	Dots faced	%Dots
	%Dots	%Dots	%Dots	%Dots	%Boundaries
	Total contribution	Total contribution	Total contribution	Total contribution	Total contribution
	Absolute Pressure	Contribution	%Boundaries	%Boundaries	Runs
Second Innings	Contribution	Runs	Contribution	Contribution	Strike rate
	Consecdottotal	Contribution	Strike rate	Strike rate	Dots faced
	%Dots	Strike rate	Dots faced	Dots faced	Runs
	Absolute pressure	Dots faced	%Dots	%Dots	%Dots
	Strike rate	%Dots	Total contribution	Total contribution	Total contribution

**Table 2 List of 5 most statistically significant metrics across different batting types, across innings 1 and 2**

$$h(t, \mathbf{X}) = h_0(t, \boldsymbol{\alpha})e^{(\boldsymbol{\beta}'\mathbf{X})}, \quad (5)$$

where  $h_0(t, \boldsymbol{\alpha})$  represents the hazard function at baseline levels of covariates, and varies over time, and  $\boldsymbol{\alpha}$  is a vector of parameters influencing the baseline hazard function. The Cox model has the following survival function:

$$S(t, \mathbf{X}) = S_0(t, \mathbf{X}, \boldsymbol{\beta})e^{(\boldsymbol{\beta}'\mathbf{X})}, \quad (6)$$

where  $S_0(t, \mathbf{X}, \boldsymbol{\beta})$  represents the survival function at baseline levels of covariates. A right censoring methodology was adopted as observations occurring at particular times but finish before the outcome of

interest occurs are referred to as right censored observations (i.e. a batsman may not necessarily be dismissed).

The validity of the Cox model relies on two assumptions: 1) the effect of each covariate is linear in the log hazard function, and 2) The ratio of the hazard function for two individuals with different sets of covariates do not depend on time. Both assumptions were met across all models.

To identify the optimal model for each batsmen type (i.e. open, top, middle, lower and tail) across the two innings the study implemented the *glmulti* R package. The models implemented an exhaustive genetic algorithm to explore the candidate set in conjunction with an *AIC* criterion to dictate model selection. Genetic algorithms are very efficient at exploring highly discrete spaces, and have been used successfully in related optimisation problems (Calcagno & Mazancourt, 2010). No interactions were considered and a model with a minimum set of 3 and maximum set of 5 predictors were required. Final models declared convergence.

A batsmen's ball-by-ball survival probability was cumulatively aggregated to generate an area under the curve (AUC) measure, indicating a batsmen's overall match contribution (i.e. volume of contribution). Brown *et al.* (2016) found that the AUC metric suitably measured a batsmen's in-game contribution.

#### 6.4 Player Influence Model

Given a player's match influence is measured through their ability to effect match outcome (i.e. win or lose), individual influence is dictated by match outcome. Each ball-by-ball observation was allocated a '*first-inning match outcome*' and '*second-inning match outcome*' indicator, indicating whether the first or second inning batting team won or lost. A logistic regression and naïve Bayes technique was applied to identify the predictors that had a significant effect on match outcome, among batsmen and bowlers, respectively. Four different models were developed: 1) 1<sup>st</sup> inning batting logistic regression model, 2) 1<sup>st</sup> inning bowling naïve Bayes model, 3) 2<sup>nd</sup> inning batting logistic regression model, and 4) 2<sup>nd</sup> inning bowling naïve Bayes model. Performance metrics that had statistical and practical significance on match outcome are indicated in appendix A with and \*. Examining the significant predictors it can be observed that the created features have a statistical and practical effect on player influence score, validating the hypothesis that the created features provide sufficient information relating to match outcome.

Given that the dependent variable, match outcome, is a binary variable, the model predicts a probability of winning given a players current performance metric. The influence score is multiplied by 100 to generate a dynamic [real-time] influence score, with higher values indicating better performance. The influence models incorporate the 3 key contribution metrics: 1) volume, 2) efficiency, and 3) pressure.

Given the influence score represents a player's propensity to influence match outcome, the average [match] influence score, for each batsmen and bowler, across both the winning and losing teams, and across innings 1 and 2 were compared. Conducting an ANOVA analysis on the different groups found statistically significant differences among the average influence score across the eight groups (table 3). Transforming the influence scores into binary variables (i.e. influence scores > 50, then influence score = 1, otherwise 0), the overall classification rate = 57% and Gini coefficient = 0.2, indicating limited discrimination.

		Match Result					
		Win			Loss		
Innings		Batting	Bowling	<i>p-value</i>	Batting	Bowling	<i>p-value</i>



<b>First</b>	56.5	47.5	< 0.002	52.6	47.6	< 0.002
<b>Second</b>	58.4	45.6	< 0.002	52.1	42.5	< 0.002

**Table 3 ANOVA Influence scores winning vs. losing teams, across innings**

Consequently, an examination of the time (i.e. over) during an inning in which a player's influence score begins to significantly (i.e. statistically) affect [final] match outcome was carried out. Because a player's match influence is generated from balls (i.e. balls faced or balls bowled), it is assumed that earlier inning scores are not indicative of actual outcome, as a player would exhaust few deliveries to exert significant match influence. Therefore, an ANOVA analysis was conducted on influence scores occurring after the 5<sup>th</sup> over. These results showed an increase in overall classification rate = 60% and Gini coefficient = 0.3 (AUC = 0.65), revealing the model results become increasingly accurate as the number of balls faced (batsmen), and number of balls bowled (bowler), increases. This is an intuitive result.

These results show that the influence scores are indicative of an individual's ability to influence match outcome, and capable of producing [dynamic] predictive player ratings. Moreover, the influence methodology is currently being applied and evaluated in real-time, to the 2017 Indian Premier League (IPL), by DOT loves data<sup>2</sup>. If the application of the models is deemed successful, the models will become a coaching, player recruitment and risk-management product<sup>3</sup>.

## 6.5 Influence Model Static Application

Given that a player's influence score is an indication of current match influence, it was assumed that the average match influence and average season influence provides an indication of a player's overall influence rating. To test this hypothesis, the authors participated in the 2016/2017 BBL fantasy league competition. Prior to competition commencement the average influence scores for all players participating in the 2016/2017 Big Bash competition were calculated<sup>4</sup>. These ratings were utilized to select the optimal team (Patel & Bracewell, 2016) for round 1 of the competition, the ratings were updated at the completion of each round and the suggested changes, as recommended by the optimisation system, were implemented. The competition consisted of 8 rounds and 35 matches. This methodology for team selection was carried out throughout the entire competition; overall the author's account finished in the top 1%. These results validated the hypothesis revealing that the model can be used in a static manner to aggregate player ratings, generate a player's overall value, and applied to recruit players.

## 7 Discussion and Conclusion

Through the model development of a model capable of calculating the ball by ball match influence of a batsman and bowlers, a novel perspective of assessing dynamic player contribution to match outcome was established. Applying the model to ball-by-ball data it was established that the model possessed the capability to [dynamically] evaluate a player's ability to influence match outcome. Moreover, model results could be aggregated to produce static measures, allowing the quantification of a player's match or seasonal value. These match or seasonal influence scores were used to participate in Australia's 2016 BBL fantasy

<sup>2</sup> A New Zealand based data science company

<sup>3</sup> Model results will be presented during the conference

<sup>4</sup> This was accomplished by running the influence model on various domestic and international T20 competitions.

league competition, overall finishing in the top 1%. This rating statistic extends understanding of player performance from conventional metrics to encapsulate the risk and in-game strategies adopted by player and teams. As this is inherently linked the manner in which a player approaches an innings, it is suitable for further research for optimizing team selection, scouting youth talent, and player development. The availability of machine readable access of ball-by-ball data enables deeper understanding and derivation of in-game strategies. This research has highlighted the use of advanced in-game metrics for investigating live player contribution and identified areas of future research.

## 8 Appendix A

Performance Metric	Definitions	Type
Batsmen Contribution	Batsmen Runs/ Team Total	Volume
Batsmen Runs	Total Number of runs a has contributed to the batting side total	Volume
Batsmen Strike Rate	Batsmen Runs/ Balls Faced	Efficiency
Runs Contributed	$expected\ runs_i - expected\ runs_{i-1} > 0$	Volume
Batsmen percentage boundaries	Total Boundaries hit by batsmen / Total Balls faced	Efficiency
Batsmen percentage dots	Total dots faced by batsmen / Total Balls faced	Efficiency
Consecutive dot balls	Cumulative number of consecutive dot balls faced by batsmen	Efficiency
Less than 2 in 4 total	Cumulative number of balls faced with less than 2 runs in 4 balls	Efficiency
Batsmen Total Contribution	Summation of batsmen runs contributed	Volume
Absolute pressure	Summation of pressure by individual batsmen	Pressure
Pressure	$(1/(resources\ available))^{0.4}$	Pressure
Runs saved	$expected\ runs_{i-1} - expected\ runs_i < 0$	Efficiency
Bowler Total contribution	Summation of batsmen runs saved	Volume
Balls bowled	Total deliveries bowled by a bowler	Volume
Runs Conceded	Total number of balls in which at least 1 run was conceded	Volume
Bowler Percentage dots	Total Bowler dots / Total balls bowled by bowler	Efficiency
Bowler Percentage boundaries	Total boundary balls / Total balls bowled by bowler	Efficiency
Economy Rate	Bowler runs conceded / total balls bowled by bowler	Efficiency
Total Wickets	Total batsmen dismissed by a team	Volume
current run rate	Total team runs / total team balls faced	Efficiency
Team percentage boundaries	Total Team boundaries / Total team balls faced	Efficiency
Team Dots	Total balls bowled by team in which no runs were scored	Volume
Team dot percentage	Total Team dots / Total team balls faced	Efficiency
Projected total	$(Current\ Total)/(Resource\ Available)$	Volume/ pressure
Area under the curve (AUC)	A running aggregation of a batsmen's survival probability	Volume

## 9 References

- [1] Bhattacharya, R., Gill, P. S., & Swartz, T. B. (2011). Duckworth–Lewis and twenty20 cricket. *Journal of the Operational Research Society*, 62(11), 1951-1957.
- [2] Brown, P., Patel, A.K., & Bracewell, P.J. (2016, July 12). Real Time Prediction Of Opening Batsmen Dismissal in Limited Overs Cricket. Paper presented at The Proceedings of the 13th Australian Conference on Mathematics and Computers in Sports. (pp. 80-85). Melbourne, Victoria, Australia: ANZIAM Mathsport. ISBN: 978-0-646-95741-8
- [3] Calcagno, V., & de Mazancourt, C. (2010). glmulti: an R package for easy automated model selection with (generalized) linear models. *Journal of statistical software*, 34(12), 1-29.
- [4] Clarke, S. R. (1988). Dynamic programming in one-day cricket-optimal scoring rates. *Journal of the Operational Research Society*, 39(4), 331-337.
- [5] Davis, J., Perera, H., & Swartz, T. B. (2015). A simulator for Twenty20 cricket. *Australian & New Zealand Journal of Statistics*, 57(1), 55-71.
- [6] Duckworth, F. C., & Lewis, A. J. (1998). A fair method for resetting the target in interrupted one-day cricket matches. *Journal of the Operational Research Society*, 49(3), 220-227.
- [7] Duckworth, F. C., & Lewis, A. J. (2005). Comment on Carter M and Guthrie G (2004). Cricket interruptus: fairness and incentive in limited overs cricket matches. *The Journal of the Operational Research Society*, 56(11), 1333-1337.
- [8] Lemmer, H. H. (2011). The single match approach to strike rate adjustments in batting performance measures in cricket. *Journal of sports science & medicine*, 10(4), 630.
- [9] Patel, A.K., & Bracewell, P.J. (2016, July 12). Team Rating Optimisation for T20 Cricket. Paper presented at The Proceedings of the 13th Australian Conference on Mathematics and Computers in Sports. (pp. 80-85). Melbourne, Victoria, Australia: ANZIAM Mathsport. ISBN: 978-0-646-95741-8
- [10] Swartz, T. B., Gill, P. S., & Muthukumarana, S. (2009). Modelling and simulation for one-day cricket. *Canadian Journal of Statistics*, 37(2), 143-160.
- [11] TotalSprtTrek (April, 2017). 25 World's Most Popular Sports (Ranked by 13 factors). Retrieved from <http://www.totalsportek.com/most-popular-sports/>

# Optimising a Batting Order in Limited Overs Cricket using Survival Analysis

Patrick Brown <sup>a,b,c</sup>, Paul J. Bracewell <sup>a,b</sup>, Ankit K. Patel <sup>a,b</sup>

<sup>a</sup>*DOT loves data*

<sup>b</sup>*Victoria University of Wellington*

<sup>c</sup>*Corresponding author: [patrick@dotlovesdata.com](mailto:patrick@dotlovesdata.com)*

*tel.no. +6448941857*

## Abstract

Several within game metrics exist to summarise individual batting performances in cricket. However, these metrics summarise individual performance and do not account for real time nor partnership performance. Previous research has successfully formulated models capable of calculating how likely a partnership is to survive each ball, for different partnerships based on within-game events. Those results are extended to optimise batting order. An expectation of how likely a batting partnership is to survive each ball within an innings can aid the development of more effective partnership strategies to optimise a team's final total. Using Cox proportional hazard models, each New Zealand partnership was assigned a measure of effectiveness. This measure of effectiveness was used to optimally position New Zealand batsmen. New Zealand captain, KS Williamson, is suggested as the optimal batsman to bat in position three regardless of which opener is dismissed. Reviewing New Zealand's loss against Australia on 4<sup>th</sup> December 2016, indicates a suboptimal order was used with JDS Neesham and BJ Watling batting at four and five respectively. Given the circumstances, C Munro and C de Grandhomme were quantified as a more optimal order.

## 1. Introduction

In cricket, the better an individual batsman or batting partnership performs, the more likely the team is to win. Quantifying batting performance is therefore fundamental to help with in-game decisions, to optimise team performance and maximise chances of winning. Several within-game metrics exist to summarise individual batting performances in cricket. However, these metrics summarise individual performance and do not account for partnership performance. An expectation of how likely a batting partnership is to survive each ball within an innings can enable more effective partnership strategies to optimise a team's final total.

Swartz, Gill, Beaudoin and De Silva (2006) used a combination of simulation, Bayesian log-linear modelling and simulated annealing to determine two potentially optimal batting orders in the Indian national cricket team. Based on a comparison with the Indian batting order adopted in the 2003 World Cup final, these batting order suggestions were found to potentially improve One-Day International (ODI) performance by approximately six runs.

Kachoyan and West (2016) applied Kaplan-Meier estimation techniques (Kaplan and Meier (1958)) to show how batting careers can be illustrated using survival functions. The batsman's innings was described as a lifespan with a 'death' referring to a dismissal, as suggested by Ibrahim (2005). Observations during which players were not dismissed were referred to as censored observations. This

methodology was used to construct distinct survival probability curves to illustrate the completed career performance of cricketers SR Waugh and SR Tendulkar.

Brown, Patel and Bracewell (2017) formulated a predictive model capable of calculating the ball-by-ball probability of a batting partnership being dismissed in the first innings of a limited overs cricket game. Cox proportional hazard models were implemented to consider the potential effect of eight batting partnership performance predictor variables on the ball-by-ball probability of a batting partnership facing the next ball without being dismissed. Some of these variables are modifications to those used in previous work involving cricket analytics (e.g Patel, Bracewell and Rooney (2016), Bracewell et al. (2016), Brown, Patel and Bracewell (2016) and Bracewell and Ruggiero (2009)).

Data were split according to the wicket at which the partnership was played. In each subset, a pragmatic model selection methodology was utilised to find an appropriate set of candidate models. Firstly, the estimated Cox model coefficients had to be practical according to whether they were expected to increase or decrease the ball-by-ball likelihood of survival. The predictors were also required to be statistically significant. The ball-by-ball survival probabilities for all partnerships considered were calculated using

$$\log\left(\frac{p}{1-p}\right) = \exp(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n), \quad (1)$$

where  $p$  represented the probability of survival and  $\beta_1, \beta_2, \dots, \beta_n$  represented the weights for each attribute,  $x_1, x_2, \dots, x_n$ , respectively. These probabilities were used to calculate the Area Under the Curve (AUC) for each partnership. The AUC was implemented as a performance measure used to rank the batting partnerships. Based on ODI games played between 26<sup>th</sup> December 2013 and 14<sup>th</sup> February 2016, the model for opening batting partnerships ranked Pakistani's A Ali and S Aslam as the optimal opening batting partnership. This method of calculating batting partnership rankings was also positively correlated with typical measures of success: average runs scored, proportion of team runs scored and winning. South African's, HM Amla and AB de Villiers were ranked as the optimal partnership at wicket two. As at 28<sup>th</sup> February 2016, these batsmen were rated 6<sup>th</sup> equal and 2<sup>nd</sup> in the world, respectively. More importantly, these results highlighted that this pair enable South Africa to maximise South Africa's chances of winning, by setting a total in an optimal manner.

In the work by Brown, Patel and Bracewell (2016), survival analysis methodology was successfully applied to investigate the effect of within-game events on the ball-by-ball likelihood of an opening batsman being dismissed in the first innings of a limited overs cricket game. In the work by Brown, Bracewell and Patel (2017), this framework was extended to non-opening batsmen and batting partnerships. The objective of that work was to develop an original approach capable of optimising batting partnership strategy in the first innings of a limited overs cricket match, in an attempt to increase a team's scoring rate and chances of winning. The objective of this work was to demonstrate the practical application of the framework developed by Brown, Bracewell and Patel (2017), through a case study involving optimising the New Zealand batting order.

## 2. Methods

### 2.1 Data Collection

Ball-by-ball data was extracted from Cricinfo commentary ([www.espncricinfo.com](http://www.espncricinfo.com)) for ODI cricket matches contested between 26<sup>th</sup> December 2013 and 29<sup>th</sup> October 2016. For each ball faced, data consisting of a number of variables were collected. These included the match, innings and player identifiers, over and ball numbers, bowler and batsman-facing metrics and outcomes from that ball. Those outcomes included if there was a dismissal, number of runs scored and number of extras (only

wides and no balls are considered due to the audit trail within the data extract). Matching this transactional information with the scorecard data enabled batting position to be established. Data collection was restricted to within-game events.

## 2.2 Data Manipulation

For ODI data, variables that potentially have an effect on the probability of a batsman, or a batting partnership, being dismissed and could be derived from the estimates of the ball-by-ball data were then calculated. These metrics were identified leveraging expert opinion from current and former international first class players and coaches. At both the batsman and partnership level, the cumulative number of runs scored was included. Other factors included dot ball and consecutive dot ball effects. Another factor was the number of balls faced by the batsman or partnership in which less than two runs in four balls had been scored. Contribution to the team total and boundaries scored were also considered. Calculation and incorporation of these variables followed on from data collection.

The data were split into multiple subsets as shown in Table 1. Each set consisted of data associated with the first innings of games.

Table 1: Categorized Batsman

<b>Batsman class</b>	<b>Batting positions</b>
Openers	1 and 2
Top order	3 and 4
Middle order	5, 6 and 7
Low order	8 and 9
Tail	10 and 11

The data were split into further subsets defined by the wicket that each partnership was played. This resulted in ten further datasets, one associated with each wicket.

## 3. Results

Using Cox models and the optimisation framework developed by Brown, Bracewell and Patel (2017), this research investigated the optimal batting partnership strategy in the New Zealand team at a deeper level. For each partnership at each wicket, the average AUC, average number of runs scored, average proportion of team runs scored and win percentage were combined into an overall measure of effectiveness, used to determine the most effective (optimal) partnerships. Figure 1 shows a decision tree illustrating the optimal partnership strategy, for New Zealand at each wicket, depending on which batsman in the partnership is dismissed.

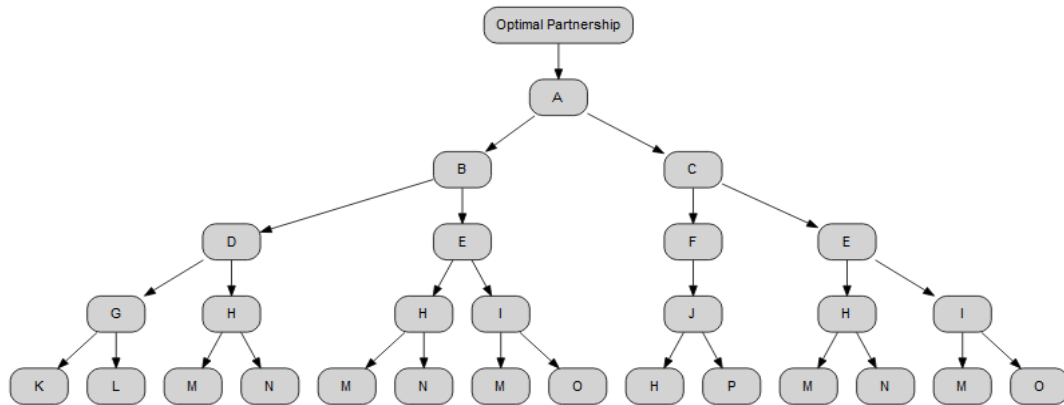


Figure 1: Decision tree illustrating optimal New Zealand batting partnership strategy

Table 2 describes the partnerships, denoted as letters, in the decision tree in Figure 1.

Table 2: Partnership correspondence

Diagram notation	Batting partnership	
	P1	P2
A	MJ Guptill	DG Brownlie
B	MJ Guptill	KS Williamson
C	DG Brownlie	KS Williamson
D	MJ Guptill	LRPL Taylor
E	KS Williamson	LRPL Taylor
F	DG Brownlie	BB McCullum
G	MJ Guptill	CJ Anderson
H	LRPL Taylor	GD Elliott
I	KS Williamson	GD Elliott
J	BB McCullum	LRPL Taylor
K	MJ Guptill	GD Elliott
L	CJ Anderson	HM Nichols
M	GD Elliott	CJ Anderson
N	LRPL Taylor	JDS Neesham
O	KS Williamson	TWM Latham
P	BB McCullum	CJ Anderson

Figure 1 shows that MJ Guptill and DG Brownlie are the optimal opening partnership. MJ Guptill and DG Brownlie had a 100% win percentage when batting together as openers. In the event that either of these opening batsmen are dismissed, KS Williamson is suggested as the optimal batsman to bat in position three. DG Brownlie and KS Williamson had a 100% win percentage when batting together at second wicket. MJ Guptill and KS Williamson had a 70% win percentage when batting together at second wicket. Together with the fact that KS Williamson is the current New Zealand captain and was ranked as number five in the ICC ODI player rankings on 8<sup>th</sup> February 2016, these results suggest that the approach developed by Brown, Bracewell and Patel (2017) to optimise batting partnership strategy is valid.

In addition, KS Williamson was acknowledged as “the most important player to his team in the world” with the biggest contribution to success in test cricket, spanning the last three years (retrieved from <http://www.foxsports.com.au/cricket/by-the-numbers-why-new-zealand-superstar-kane-williamson-is-cricket-s-most-valuable-batsman/news-story/241fb534736c1d475d3c28fe053d9754>). KS Williamson is the third top century scorer of any batsman over the last three years, while he is only one of two players to score over 30% of his team's hundreds. Further, KS Williamson has scored 3011 runs in the last three years, considerably higher than the next best New Zealand batsman, TWM Latham, with 2031. This further highlights the validity of the optimisation procedure.

Based on the optimal partnership strategy illustrated in Figure 1, an optimal top six New Zealand batting line up is suggested in Table 3.

Table 3: Optimal New Zealand top six batsmen

Batting position	Batsman
1	MJ Guptill
2	DG Brownlie
3	KS Williamson
4	LRPL Taylor
5	GD Elliott
6	JDS Neesham

### 3.1 Optimisation Case Study

On 4<sup>th</sup> December 2016, New Zealand played Australia in a ODI and lost by 68 runs. Former Black Caps all-rounder and Auckland A cricket coach, AR Adams, criticised the New Zealand coaching staff for their batting order changes in that game (retrieved from [http://www.nzherald.co.nz/sport/news/article.cfm?c\\_id=4&objectid=11760500](http://www.nzherald.co.nz/sport/news/article.cfm?c_id=4&objectid=11760500)). AR Adams questioned the choice of JDS Neesham as number four. C Munro batted as number six and C de Grandhomme batted as number eight. Given how successfully Aucklanders, C Munro and C de Grandhomme, batted in domestic cricket in 2016, AR Adams suggested that C Munro and C de Grandhomme should have been played in positions four and five, followed by BJ Watling and JDS Neesham.

The objective of this case study was to determine the optimal New Zealand batting order for the game against Australia to assess whether it aligns with AR Adams's suggestions, and demonstrate the practical application of the work developed by Brown, Bracewell and Patel (2017).

The final models developed by Brown, Bracewell and Patel (2017) were fitted to data from the first innings of limited overs cricket games. In the ODI between Australia and New Zealand, New Zealand



batted in the second innings. To account for this, the models were applied to data from the most recent ODI game prior to 4<sup>th</sup> December 2016, in which New Zealand batted in the first innings. New Zealand's opponents in this game were India, with the game contested on 26<sup>th</sup> October 2016. The intent was to use the performance of batsmen in the ODI against India as an indication of how these batsmen would have performed in the ODI against Australia. As such, the optimal New Zealand batting order against India could be used as an indicator to suggest the order that would have likely optimised the scoring rates and chances of winning against Australia.

Each batsman was assigned a measure of effectiveness based on the average AUC, total number of runs scored, proportion of team runs scored and strike rate.

### 3.2 Bootstrapping

Bootstrapping is a technique used to re-sample data without replacement and allows estimation of the sampling distribution of a statistic (Mooney, Duval and Duval, 1993). As the analysis in this case study was based on one game, bootstrapping was used to generate 1000 bootstrapped samples of batsman effectiveness.

To determine the optimal batting order, the process was repeated with a different batsman removed for each iteration. Table 4 illustrates the New Zealand batting order used in the ODI game against India.

Table 4: New Zealand batting order against India 26<sup>th</sup> October

Batting position	Batsman
1	MJ Guptill
2	TWM Latham
3	KS Williamson
4	LRPL Taylor
5	JDS Neesham
6	BJ Watling
7	AP Devcich
8	MJ Santner
9	TG Southee

Each bootstrapped sample consisted of the effectiveness of each batsman from Table 4, with one batsman removed. In this particular game, New Zealand only used nine batsmen. As such, each bootstrapped sample contained eight observations. The bootstrapping procedure was repeated to give 1000 samples. The mean effectiveness for the team was calculated for each sample. The mean of those means was used as a rating of effectiveness for the team with the batsman excluded from analysis. The rating was then used to determine where that batsman was positioned in the optimal batting order. The optimal batting order is the order that would have maximised the team's final score and chances of winning. The smaller the rating associated with each batsman, the less effective the team would have been without the batsman and the higher the batsman was optimally positioned.

### 3.3 Optimal New Zealand Order Against India 26<sup>th</sup> October 2016

Table 5 illustrates New Zealand batsmen based on their position in the optimal order, compared with where they were actually positioned in the ODI against India.

Table 5: Suggested New Zealand order compared with actual New Zealand order against India 26<sup>th</sup> October

Batting position	Optimal batsman	Actual batsman
1	MJ Guptill	MJ Guptill
2	TWM Latham	TWM Latham
3	KS Williamson	KS Williamson
4	LRPL Taylor	LRPL Taylor
5	<b>BJ Watling</b>	<b>JDS Neesham</b>
6	<b>AP Devcich</b>	<b>BJ Watling</b>
7	<b>JDS Neesham</b>	<b>AP Devcich</b>
8	MJ Santner	MJ Santner
9	TG Southee	TG Southee

Table 5 illustrates that optimally, JDS Neesham should have been positioned as number seven, behind both BJ Watling and AP Devcich.

### 3.4 Optimal New Zealand Order Against Australia 4<sup>th</sup> December 2016

The optimal order suggested in Table 5 was used to optimally position New Zealand batsmen in the primary ODI of interest against Australia.

Given that C Munro and C de Grandhomme were not involved in the ODI between India and New Zealand or any previous ODI contested in 2016, a rating of effectiveness from the optimisation procedure could not be derived for these batsmen.

As such, a different approach was taken to determine the optimal position for C Munro and C de Grandhomme, relative to BJ Watling and JDS Neesham. The Plunket Shield is New Zealand's domestic first-class cricket competition. The 2016-2017 season is the most recent competition in which C Munro, C de Grandhomme and JDS Neesham had all batted. As such, in order to compare these batsmen, this research investigated their domestic performances in the Plunket Shield. However, the Plunket Shield does not categorise as limited overs cricket. Consequently, the final models and optimisation procedure could not be applied to games from this competition. The optimisation procedure used to rate the New Zealand batsmen against India accounted for AUC, total runs scored, proportion of team runs scored and strike rate. Given this, batting averages in the 2016-2017 Plunket Shield competition were used to compare the effectiveness of C Munro and C de Grandhomme with that of BJ Watling and JDS Neesham. The intent was to compare the most recent performance of these batsmen prior to the ODI between Australia and New Zealand, within the same competition for the same period. This is likely to have been indicative of where to position these batsmen, relative to each other, in the optimal New Zealand order against Australia. Despite being from different competitions and formats, these performances are an adequate proxy of batting performance, primarily due to the timeliness of observations.

During the period of the 2016-2017 Plunket Shield season prior to New Zealand's ODI on 4<sup>th</sup> December, C Munro and C de Grandhomme were scoring at a considerably higher rate compared with JDS Neesham. C Munro was averaging 84.50, while C de Grandhomme was averaging 54.00 (retrieved from <http://www.foxsports.com.au/cricket/what-were-they-thinking-andre-adams-left-dazed-and-confused-by-new-zealands-tactics/news-story/d9e119fe95695102d0b6c9b38102c187>). JDS

Neesham averaged 8.00 in the same competition for the same period (retrieved from [http://www.stats.espncricinfo.com/plunket-shield-2016-17/engine/records/batting/highest\\_career\\_batting\\_average.html?id=11507;team=2621;type=tournament](http://www.stats.espncricinfo.com/plunket-shield-2016-17/engine/records/batting/highest_career_batting_average.html?id=11507;team=2621;type=tournament)). BJ Watling had not played any domestic cricket during the 2016-2017 season prior to 4<sup>th</sup> December 2016. However, in ODIs, BJ Watling was averaging 26.09 compared with 21.75 for JDS Neesham. As such, comparative ratings of effectiveness for C Munro and C de Grandhomme would likely have been higher than the actual ratings calculated for BJ Watling and JDS Neesham, based on the optimisation procedure. This supports AR Adams's suggestion to play C Munro and C de Grandhomme as number four and five, respectively, with BJ Watling and JDS Neesham as number six and seven, respectively.

Similarly, lower order batsmen, MJ Santner, MJ Henry, LH Ferguson and TA Boult did not bat in the ODI between India and New Zealand on 26<sup>th</sup> October. With the exception of LH Ferguson, the other eight batsmen from Table 5 batted in the previous game between India and New Zealand on 23<sup>rd</sup> October 2016. As such, the final models were applied to this game and a measure of effectiveness was assigned to each batsman. The bootstrapping procedure, was applied to rate each batsman. The rating was then used to position each batsman in the optimal order. LH Ferguson completed his ODI debut against Australia. Given his lack of ODI experience, LH Ferguson is positioned as number eleven in the optimal order.

Table 6 illustrates the New Zealand order that would have likely optimised the scoring rates and chances of winning against Australia. This is compared with the actual batting order.

Table 6: Optimal New Zealand order compared with actual New Zealand order against Australia 4<sup>th</sup> December

Batting position	Optimal batsman	Actual batsman
1	MJ Guptill	MJ Guptill
2	TWM Latham	TWM Latham
3	KS Williamson	KS Williamson
4	<b>C Munro</b>	<b>JDS Neesham</b>
5	<b>C de Grandhomme</b>	<b>BJ Watling</b>
6	BJ Watling	C Munro
7	JDS Neesham	MJ Santner
8	MJ Henry	C de Grandhomme
9	MJ Santner	MJ Henry
10	TA Boult	LH Ferguson
11	LH Ferguson	TA Boult

Based on the findings in Table 6, New Zealand appear to have utilised a suboptimal order with JDS Neesham and BJ Watling batting at four and five respectively. Given the circumstances, C Munro and C de Grandhomme were quantified as a more optimal order.

In the ODI between New Zealand and Australia, BJ Watling struggled, only scoring 6 from 13 before being dismissed. JDS Neesham scored 34 from 62. C Munro recorded the highest score, 49 from 59, of the three batsmen. The optimal batting order is consistent with this, suggesting C Munro should have batted before BJ Watling and JDS Neesham, as number four. This would have improved New Zealand's chances of winning, based on previously observed individual batting strategies.

## 4. Discussion and Conclusion

Reviewing New Zealand's loss against Australia on 4<sup>th</sup> December 2016, indicates a suboptimal order was used, with JDS Neesham and BJ Watling batting at four and five respectively. Given the circumstances, C Munro and C de Grandhomme were quantified as a more optimal order. This supported the batting order suggestions made by former Black caps all-rounder and Auckland A coach, AR Adams. This demonstrates a practical application of the framework developed by Brown, Bracewell and Patel (2017).

To ensure complete case analysis, performance from other competitions was used. This demonstrated the wider applicability and usefulness of the methodology developed by Brown, Bracewell and Patel (2017) and for scouting and selection purposes.

Given the increased interest in short forms of the game, particularly T20, extending this research to T20 matches is clearly of interest. Applying the final models to T20 data from the 2016 Indian Premier League (IPL) may provide a suitable technique to validate the final models. It is expected that the survival probabilities for ODI batting partnerships would be higher than those calculated for batting partnerships contesting IPL games.

The developed models were fitted to data from the first innings of a selection of ODI games. The emphasis on first innings was due to the lack of previous research into what a match winning total should be. Another area of future work could involve extending the framework developed by Brown, Bracewell and Patel (2017) and the techniques applied in this research, to investigate the performance of batsmen and partnerships participating in the second innings of limited overs cricket games.

## References

- [1] Bracewell, P.J., Coomes, M., Nash, J., Meyer, D., and Rooney, S.J. (2016). Rating the attacking performance of a non-wicket taking bowler in limited overs cricket. Paper presented at The Proceedings of the 13<sup>th</sup> Australian Conference on Mathematics and Computers in Sport. (74-79). Melbourne, Victoria, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.
- [2] Bracewell, P.J. and Ruggiero, K. (2009). A parametric control chart for monitoring Individual batting performances in cricket. *Journal of Quantitative Analysis in Sports*, 5(3): 1-19.
- [3] Brown, P., Patel, A.K. and Bracewell, P.J. (2016). Real time prediction of opening batsman dismissal in limited overs cricket. Paper presented at The Proceedings of the 13<sup>th</sup> Australian Conference on Mathematics and Computers in Sport. (80-85). Melbourne, Victoria, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.
- [4] Brown, P., Bracewell, P.J. and Patel, A.K. (2017). Optimising batting partnership strategy in the first innings of a limited overs cricket match. Manuscript submitted for publication.
- [5] By the numbers: Why New Zealand superstar Kane Williamson is cricket's most valuable batsman. <http://www.foxsports.com.au/cricket/by-the-numbers-why-new-zealand-superstar-kane-williamson-is-cricket-most-valuable-batsman/news-story/241fb534736c1d475d3c28fe053d9754>. Accessed: 2017-01-19.
- [6] Cricinfo (2016) Retrieved from <http://www.espncricinfo.com/>.
- [7] Ibrahim, J.G. (2005) *Applied Survival Analysis*. The 21st Annual Summer Workshop of the North-eastern Illinois Chapter of the American Statistical Association. Available from <http://www.amstat.org/chapters/northeasternillinois/pastevents/summer05.htm>
- [8] Kachoyam, B. and West, M. (2016). Cricket as life and death. Paper presented at The Proceedings of the 13<sup>th</sup> Australian Conference on Mathematics and Computers in Sport. (85-90). Melbourne, Victoria, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.
- [9] Kaplan E.L. and Meier P. (1958) Non-parametric estimation from Incomplete Observations. *American Statistical Association Journal*, 53(28): 457-481.

- [10] Mooney, C. Z., Duval, R. D. and Duval, R. Bootstrapping: A nonparametric approach to statistical inference. No. 94-95. Sage, 1993.
- [11] Patel, A.K., Bracewell, P.J. and Rooney, S.J. (2016). Team rating optimisation for T20 Cricket. Paper presented at The Proceedings of the 13<sup>th</sup> Australian Conference on Mathematics and Computers in Sport. (91-96). Melbourne, Victoria, Australia: ANZIAM MathSport. ISBN: 978-0-646-95741-8.
- [12] Plunket shield, 2016/17 - Otago / records / highest averages [http://www.stats.espncricinfo.com/plunket-shield-2016-17/engine/records/batting/highest\\_career\\_batting\\_average.html?id=11507;team=2621;type=tournament](http://www.stats.espncricinfo.com/plunket-shield-2016-17/engine/records/batting/highest_career_batting_average.html?id=11507;team=2621;type=tournament). Accessed: 2017-04-11.
- [13] Swartz, T.B., Gill, P.S., Beaudoin, D., and De Silva, B.M. (2006) Optimal batting orders in one-day cricket. *Computers & operations research* 33, 7, 1939–1950.
- [14] What were they thinking? - Andre Adams on batting line up. [http://www.nzherald.co.nz/sport/news/article.cfm?c\\_id=4&objectid=11760500](http://www.nzherald.co.nz/sport/news/article.cfm?c_id=4&objectid=11760500). Accessed: 2017-04-11.
- [15] What were they thinking? Andre Adams left dazed and confused by New Zealand's tactics. <http://www.foxsports.com.au/cricket/what-were-they-thinking-andre-adams-left-dazed-and-confused-by-new-zealands-tactics/news-story/d9e119fe95695102d0b6c9b38102c187>. Accessed:2017-04-11.

# Optimization of Harmony in Team Formation Problem for Sports Clubs: A real life volleyball team application

Gerçek Budak\*, İmdat Kara\*\*, Yusuf Tansel İç\*\*\*, Refail Kasımbeyli\*\*\*\*

\* Adana Science and Technology University, Department of Industrial Engineering: [gbudak@adanabtu.edu.tr](mailto:gbudak@adanabtu.edu.tr)

\*\* Başkent University, Department of Industrial Engineering: [ikara@baskent.edu.tr](mailto:ikara@baskent.edu.tr)

\*\*\* Başkent University, Department of Industrial Engineering: [ytansel@baskent.edu.tr](mailto:ytansel@baskent.edu.tr)

\*\*\*\* Anadolu University, Department of Industrial Engineering: [rkasimbeyli@anadolu.edu.tr](mailto:rkasimbeyli@anadolu.edu.tr)

## Abstract

Sports team coaches' main concern is forming the best team to win the upcoming match. Even if a team squad is comprised of limited number of players, the combination of them makes out a complicated problem with huge number of possible line-ups. Academic researches on this subject increase in the last decade since this decision became important financially and solvable as the parameters are reachable. There are many aspects that define the best team such as team harmony, player performance, team strategy, opponent suitability etc. This research proposes a new mathematical model which aims to form the best team with respect to team harmony.

## 1 Introduction

Team formation problem for sports clubs is an assignment problem that the team coach tries to form the best team by assigning the players in the squad to the identified positions to win the match (Boon and Sierksma, 2003). The identification of the positions differs from one sport branch to other and is made according to the responsibilities and/or locations of the players in the game field (Budak et al., 2017; Ben Abdelkrim et al., 2010). For the decision makers, the most important question to be answered is that "what is the best team".

The best team can be formed with respect to different perspectives which are total performance, team and player harmony, suitability of coach's strategy, opponent team suitability and other dimensions. These perspectives are also affected by uncontrollable factors such as home/away court advantages and disadvantages (Anderson et al., 2012), psychological and physical status of players (Nippert and Smith, 2008), audience affect (Lavier and Sukthankar, 2014) etc.

Team formation problem and player selection has been handled by many academic researchers in last decades since the industrialized sport sector extremely grew financially and international and national player databases are established on player performances. For example, Boon and Sierksma (2003) proposed an assignment based mathematical model which maximizes the assigned players' total performance. Tavana et al. (2013) assembled a two-phase fuzzy inference system that ranks and selects the alternative players in the first phase and then locates them in-field positions. Dadelo et al. (2014) used Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) to find out the best players by using the data of players' physical measurements. Ahmet et al. (2013) developed a multi-objective genetic algorithm for skills to find out the best team before the season stage on cricket sport. Chen et al. (2014) applied Analytical Hierarchy Process (AHP) and TOPSIS for selecting the best pitcher for the team. There are other researches that support the combination selection and player selection in sports clubs (for further information see Winter et al. (2006), Caro (2012), Villa and Lozano (2016) and Lorains et al. (2013)).

Considering the previous researches and techniques used, this paper develops a new mathematical model for team formation problem of sports clubs. The proposed model considers player preferences that constructs team and player harmony dimension of the team formation problem. This research initializes this consideration in the decision making process. To achieve this, team harmony is initially quantified by questionnaires to understand player preferences. The player preferences became programmable as it became quantifiable.

## 2 Searching the Best Team: Team and Player Harmony

The decision makers of the team formation problem should also concern the formed team's harmony. Team harmony and harmony between players have not been considered by previous researches on this problem. This component of the problem affects the players' performance and accordingly to match result (Stevens and Bloom, 2003; Saavedra, 2013).

On the other hand, team harmony issue are studied by several researches such as Yukelson (1997), Murray and Chao (2005), Hinsz et al. (1997), Hutchins (1991) and Levine et al. (1993). Yukelson (1997) output psychological aspects which are common purpose, communication skills, trust and friendship define the team harmony which. In addition to this, during the season the responsibility and role sharing in the field must be made clearly and effectively (Yukelson, 1997). Murray and Chao (2005) defined team harmony as whole of the pairwise relationships with each of the players. A team with excellent relationships with each other skill and emotion wisely creates a harmonic team (Murray and Chao, 2005). Hinsz et al. (1997), Hutchins (1991) and Levine et al. (1993) proposed that the team coordination, responsibility and skill sharing during the game show how the team harmony is.

To recap the topic, team coach's assigned players' satisfaction level based on the role sharing and the preferences on each position summation gives the team harmony level. As these factors are related with the preferences of players, these data should be obtained by questionnaires to the players.

## 3 A New Formulation for Team Formation Problem of Sports Clubs

In this section, a new mathematical model for team formation before the match is introduced. This model considers team and player harmony dimensions of the problem. This proposed mathematical model aims to form the best team by maximizing the total of the satisfaction level of the players from the formed team. Coach's strategy is also going to be provided by the constraints of player-position capabilities and skill level thresholds.

There are three sets in this problem which are player set, position set and skill set. Player set is comprised of the players in the squad which are decided before the season stage and can be shown as  $P = \{1, 2, 3 \dots n\}$ .

Position set is constituted depending on the sport and positions differentiate with each other on responsibilities and location in the field. This set can be found by the handbooks of each team sport branch from the international federations and can be shown as  $R = \{1, 2, 3 \dots m\}$ .

The last set is skills set of the sport and consists of the special movements and activities that typical to the sport branch. These movements and activities are determined according to the necessities of accomplishing the positions' responsibilities. Skill set can be shown as  $S = \{1, 2, 3 \dots k\}$ .

The main parameters of proposed model are importance levels of positions, player-position capabilities, and coach's thresholds on each positions skill, player gladness among other player and team and players' performance forecasts on each skill for the upcoming match.

Importance levels of position are found by objectively as it depends on many qualitative aspects so that using the experts' opinions is going to avoid subjectivity (Eckenrode, 1965). AHP is one of the best techniques to obtain these weights since the properties of AHP are suitable (Saaty, 1990). There are also other methodologies that could help the decision maker to obtain these parameters such as SMART, point allocation, DELPHI method etc. Importance level of  $j^{\text{th}}$  position on the match result is represented as  $PR_j$ . They are valued between 0 and 1 and their total for all  $j$  is equal to one.

The parameters that the coach is going to compose according to his/her preferences are player-position capabilities and skill level thresholds. Player-position capabilities parameter is for each player and position and identified by the coach that whether a player can play in that particular position. This parameter is symbolized as  $PP_{ij}$ . It is valued 1 if the  $i^{\text{th}}$  player is able to play on  $j^{\text{th}}$  position and it is valued 0 for the other case. Thresholds for each position's skill level are identified by the coach according to his/her strategic mentality. This parameter is represented with  $HT_{jy}$  which shows the minimum expectation of the team coach on  $j^{\text{th}}$  position's  $y^{\text{th}}$  skill and valued between 0 and 100.

The parameters that the players are going to identify are the gladness level from the other players and gladness from the formed team. Gladness level of a player shows how much a player is glad from the other player and this parameter is valued between 0 and 1.  $i^{\text{th}}$  player's gladness level for  $p^{\text{th}}$  player is shown as  $PH_{ip}$ . Second parameter that players define is team harmony level according to their wish list on other positions. Each player forms a team according to their preferences and this parameter is shown as  $TH_{ipj}$ . If  $i^{\text{th}}$  player assigned  $p^{\text{th}}$  player to the  $j^{\text{th}}$  position the  $TH_{ipj}$  is equal to 1 and other cases  $TH_{ipj}$  is valued 0.

Players' performance forecasts on each skill for the upcoming match are symbolized as  $P_{iy}$  which represents the  $i^{\text{th}}$  player's  $y^{\text{th}}$  skill forecast and they must be valued between 0-100 as percentage ratios.

The decision variable of team formation problem before the match stage is represented with  $x_{ij}$  binary variable that shows whether  $i^{\text{th}}$  player is assigned to position  $j$  or not. If  $i^{\text{th}}$  player assigned to  $j^{\text{th}}$  position,  $x_{ij}$  is equal to 1 and other case it equals to 0. Another decision variable in this proposed model is  $y_i$  which is a dependent binary variable that shows if  $i^{\text{th}}$  player is assigned to a position or not.  $q_{ij}$  is a dependent decision variable that is defined as the gladness level of player  $i$  if  $i^{\text{th}}$  player is assigned to  $j^{\text{th}}$  position. Total team harmony of the players related to positions is defined as a dependent decision variable and symbolized with  $TM$ .

The following model is proposed for the team formation problem before the match stage by aiming to maximize the total harmony among the players:

Objective function;

$$\text{Max } z = \sum_i^m \sum_j^n q_{ij} PR_j + TM \quad (1)$$

Subject to;

$$\sum_{j=1}^n x_{ij} \leq 1, \quad i = 1, \dots, m \quad (2)$$

$$\sum_{i=1}^m x_{ij} = 1, \quad j = 1, \dots, n \quad (3)$$

$$\sum_i P_{iy} x_{ij} \geq HT_{jy}, j = 1, \dots, n, y = 1, \dots, k \quad (4)$$

$$x_{ij} \leq PP_{ij}, i = 1, \dots, m, j = 1, \dots, n \quad (5)$$



$$\sum_{j=1}^n x_{ij} = y_i, \quad i = 1, \dots, m \quad (6)$$

$$\sum_{p=1}^m PH_{ip} y_p x_{ij} = q_{ij}, \quad i = 1, \dots, m, \quad j = 1, \dots, n, \quad (7)$$

$$\sum_{i=1}^m \sum_{p=1}^m \sum_{j=1}^n TH_{ipj} x_{pj} = TM \quad (8)$$

$$y_i, x_{ij} \in \{0, 1\}, i = 1, \dots, m, j = 1, \dots, n \quad (9)$$

Objective function of the proposed model (1) is maximizing the total team harmony. Constraint (2) guarantees that a player could be assigned to one position as a player cannot play in more than one position. There must not be an unassigned position so that constraint (3) establishes each position has an assignment. An assigned player's skill level cannot be less than the thresholds that team coach identified for the assigned position and constraint (4) is added to the model. Constraint (5) provides that the players could only be assigned to the positions that they are capable to play. As defined,  $y_i$  is a dependent decision variable which is obtained with constraint (6). Constraint (8) obtains the total gladness of the players from the formed team position-wisely. Constraint (9) defines the binary decision variables.

Constraint (7) obtains each assigned player's gladness level from the formed team; however, it is structured as non-linear. To use linear programming literature and knowledge this constraint must be linearized for this reason McCormick Envelopment is a useful tool for linearization (McCormick, 1976). By using this technique, constraint (7) is going to be excluded from the model and the following constraints (7a, 7b, 7c and 7d) shown below is going to be included to the mathematical model:

$$q_{ij} \geq 0, \quad i = 1, \dots, m, \quad j = 1, \dots, n \quad (7a)$$

$$q_{ij} \geq \sum_p^m PH_{ip} x_{ij} + \sum_p^m PH_{ip} y_p - \sum_p^m PH_{ip}, \quad i = 1, \dots, m, j = 1, \dots, n \quad (7b)$$

$$q_{ij} \leq \sum_p^m PH_{ip} x_{ij}, \quad i = 1, \dots, m, \quad j = 1, \dots, n \quad (7c)$$

$$q_{ij} \leq \sum_p^m PH_{ip} y_p, \quad i = 1, \dots, m, \quad j = 1, \dots, n \quad (7d)$$

The total number of decision variables in this mathematical model is  $(2mn + n(m^2 - m) + 2m + 1)$  and total number of constraints is  $(6mn + nk + 2m + n + 3)$ . Therefore, assuming  $m \geq n$  and  $m \geq k$  the model greatness is  $O(m^3)$ .

## 4 Conclusion

Team formation problem for the sports clubs before the match stage is crucial problem for the team coaches. This paper achieves the quantification of the team harmony concept by considering the psychological aspect of team players. Team harmony is defined as the players' preference in field and

positions. By assuming this definition, player's performances are dependent on the team harmony expression is regarded.

This paper initializes a novel approach to solve the team formation problems with a harmony based mathematical model which also integrates the coach's strategy, mentality and expectations to the decision process. This model is a useful tool for the team coaches to observe the team with the best harmony which is over the thresholds determined by the team coach.

## References

- Abdelkrim, N. B., Chaouachi, A., Chamari, K., Chtara, M., & Castagna, C. (2010). *Positional role and competitive-level differences in elite-level men's basketball players*. The Journal of Strength & Conditioning Research, 24(5), 1346-1355.
- Ahmed, F., Deb, K., & Jindal, A. (2013). *Multi-objective optimization and decision making approaches to cricket team selection*. Applied Soft Computing, 13(1), 402-414.
- Anderson, M., Wolfson, S., Neave, N., & Moss, M. (2012). *Perspectives on the home advantage: A comparison of football players, fans and referees*. Psychology of Sport and Exercise, 13(3), 311-316.
- Boon, B. H., & Sierksma, G. (2003). *Team formation: Matching quality supply and quality demand*. European Journal of Operational Research, 148(2), 277-292.
- Budak, G., KARA, İ., & İÇ, Y. T., (2017). *Weighting the Positions and Skills of Volleyball Sport by Using AHP: A real life application*. IOSR Journal of Sports and Physical Education, 4(1), 23-29.
- Caro, C. A. (2012). *College football success: The relationship between recruiting and winning*. International Journal of Sports Science & Coaching, 7(1), 139-152.
- Chen, C. C., Lee, Y. T., & Tsai, C. M. (2014). *Professional baseball team starting pitcher selection using AHP and TOPSIS methods*. International Journal of Performance Analysis in Sport, 14(2), 545-563.
- Dadelo, S., Turskis, Z., Zavadskas, E. K., & Dadeliene, R. (2014). *Multi-criteria assessment and ranking system of sport team formation based on objective-measured values of criteria set*. Expert Systems with Applications, 41(14), 6106-6113.
- Eckenrode, R. T. (1965). *Weighting multiple criteria*. Management science, 12(3), 180-192.
- Hinsz, V. B., Tindale, R. S., & Vollrath, D. A. (1997). *The emerging conceptualization of groups as information processors*. Psychological bulletin, 121(1), 43.
- Hutchins, E. (1991). *The social organization of distributed cognition*. L. B. Resnick, J. M. Levine, and S. D. Teasley (Eds.), Perspectives on socially shared cognition, pp. 283-307. Washington, DC: APA.
- Laviers, K. R. & Sukthankar, G. (2014), *Chapter 13 - Using Opponent Modeling to Adapt Team Play in American Football. Plan, Activity, and Intent Recognition*, Boston, pp. 313-341.
- Levine, J. M., Resnick, L. B., & Higgins, E. T. (1993). *Social foundations of cognition*. Annual review of psychology, 44(1), 585-612.
- Lorains, M., Ball, K., & MacMahon, C. (2013). *Performance analysis for decision making in team sports*. International Journal of Performance Analysis in Sport, 13(1), 110-119.
- McCormick G.P. (1976). *Computability of global solutions to factorable nonconvex programs: part I - convex underestimating problems*. Mathematical Programming 10, 147-175, North-Holland Publishing Company.
- Murray, J. Y., & Chao, M. C. (2005). *A cross-team framework of international knowledge acquisition on new product development capabilities and new product market performance*. Journal of International Marketing, 13(3), 54-78.
- Nippert, A. H., & Smith, A. M. (2008). *Psychologic stress related to injury and impact on sport performance*. Physical medicine and rehabilitation clinics of North America, 19(2), 399-418.
- Saaty, T. L. (1990). *Decision making for leaders: the analytic hierarchy process for decisions in a complex world*. RWS publications.
- Saavedra, L. K. (2013). *Effective team building: The role of coaches*. Strategies, 26(4), 3-6.

- Stevens, D., & Bloom, G. (2003). *The effect of team building on cohesion*. AVANTE-ONTARIO-, 9(2), 43-54.
- Tavana, M., Azizi, F., Azizi, F., & Behzadian, M. (2013). *A fuzzy inference system with application to player selection and team formation in multi-player sports*. Sport Management Review, 16(1), 97-110.
- Villa, G., & Lozano, S. (2016). *Assessing the scoring efficiency of a football match*. European Journal of Operational Research, 255(2), 559-569.
- Winter, E. M., Jones, A. M., Davison, R. R., Bromley, P. D., & Mercer, T. H. (Eds.). (2006). *Sport and Exercise Physiology Testing Guidelines: Volume I—Sport Testing: The British Association of Sport and Exercise Sciences Guide*. Routledge.
- Yukelson, D. (1997). *Principles of effective team building interventions in sport: A direct services approach at Penn State University*. Journal of Applied Sport Psychology, 9(1), 73-96.

# Formula 1 lap time modeling using generalized additive models

C. Casella\* and P. Vidoni\*\*

\*University of Udine

email: casella.claudio@spes.uniud.it

\*\* Department of Economics and Statistics - University of Udine, via Tomadini 30/A, I-33100 Udine, Italy

email: paolo.vidoni@uniud.it

## Abstract

In this paper we aim at describing the Formula 1 team and driver performances, during a particular race, using a suitable generalized additive model for the representation of lap time evolution. Although the modeling of Formula 1 races is an extremely complicated task, our model gives a rather flexible specification of the lap times as a function of some relevant numeric and categorical predictor variables. We fit the model using the freely available data regarding the Formula 1 season 2015 and we find that, for a grand prix without unpredictable events such as safety car, virtual safety car or race suspension, the model provides an accurate description of the race development. Moreover, it can be fruitfully considered for specifying alternative race strategies, with particular regard to the pit stop choices.

## 1 Introduction

The modeling of Formula 1 races is an extremely complicated task, since car and driver performances depend on a number of mutually interacting explicative variables and on the occurrence of unpredictable events such as safety car or virtual safety car, driver's error, breakdown or car crash. In this paper we define a fairly simple semiparametric regression model with the aim of describing team and driver performances by considering the driver's lap time as response variable. More precisely, the logarithmic transformation of the lap time is modeled as a function of each lap of the race, taking into account also tyre degradation, type and age of the tyres, fuel consumption, team and driver, traffic and interactions between drivers, pit stop and first lap after pit stop, occurrence of the safety car. In order to provide a flexible specification of the dependence of the response on the numeric and the categorical predictor variables, we consider a suitable generalized additive mixed model, where the linear predictor involves also smooth functions of some covariates. A random effect is introduced in order to account for the random variability observed from driver to driver and to describe the general effect of the remaining explanatory variables on the response.

We fit the model using the freely available data regarding the Formula 1 season 2015 and we find that, for a grand prix without unpredictable events such as safety car, virtual safety car or race suspension, the model provides an accurate description of the race development. Moreover, the estimated model may be fruitfully considered also for evaluating and simulating alternative race strategies, with particular regard to the pit stop choices. A suitable extension of the model is defined in order to describe the more challenging situation of a

race with safety car deployment, since it is well-known that the safety car could be an extremely important factor for deciding pit stop strategies.

The paper is organized as follows. Section 2 briefly reviews the main features of the generalized additive (mixed) models. Indeed, a model is specified by considering numeric and categorical explanatory variables, potentially useful for describing the log transformed lap times without the safety car occurrence. In Section 3 we fit the models by considering Formula 1 data related to some races of the season 2015 and we discuss the relevance of the estimated model for both inference and prediction purposes. An further model is specified for considering also the lap times observed in the safety car regime. Finally, we conclude in Section 4 with a short discussion, highlighting some further research lines.

## 2 A generalized additive model for lap time description

Generalized additive models (GAMs) were originally defined by Hastie and Tibshirani (1986, 1990) as semi-parametric regression models where the relationship between the dependent and the explanatory variables may follow smooth nonlinear patterns. They can be viewed as suitable extensions of generalized linear models where the linear predictor involve also a sum of smooth functions of the covariates. In general the model structure can be described as

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_j x_j + s_1(x_{j+1}) + s_2(x_{j+2}) + s_3(x_{j+3}, x_{j+4}) + \dots$$

where  $Y$  is the response variable, following an exponential family distribution,  $x_1, x_2, \dots$  are the explanatory variables,  $\beta_0, \beta_1, \dots$  are unknown scalar parameter and  $s_r, r = 1, 2, \dots$  are smooth functions of the covariates. Indeed,  $g$  is the link function which describes the relationship between the expected response  $\mu = E(Y)$  and the explanatory variables (see also Wood, 2006, for a general introduction to GAMs).

Since Formula 1 data, such as telemetry information, are usually not publicly available, we specify and fit our model by considering only freely available data on lap by lap drivers performance and car status with regard to some Formula 1 races. We consider also information on relevant events such as pit stop occurrence and safety car deployment. The response variable is the logarithmic transformation of the lap time (Time) in seconds, namely the time taken by the driver to cover the current lap, while the explanatory variables are:

- Lap, discrete-valued variable indicating the lap of the race, measured by considering the fraction of laps already covered for a given race;
- Team, categorical variable indicating the constructor of the car;
- Driver, categorical variable indicating the pilot of the car;
- Compound, categorical variable indicating the compound of the tyres used in the current lap (supersoft, soft, medium, hard);
- TyreLaps, discrete-valued variable giving the portion of race completed by the current set of tyres;
- TyreState, categorical variable indicating the tyre condition when they have been installed (new or used);
- Pit, a dummy variable indicating whether the pit stop was carried out in the current lap;
- FollowingPit, a dummy variable indicating whether the pit stop was carried out during the previous lap;
- SC, a dummy variable indicating whether the safety car is on the track in the current lap;

- Distance, continuous variable indicating the gap in seconds from the car that precedes at the end of the current lap;
- Circuit, categorical variable indicating where the grand prix takes place.

With regard to the observed values for the response variable Time, we take into account only lap times under a given threshold, since high values are usually related to unpredictable events, such as driver's error, car breakdown or car crash, and they can be interpreted as outliers which could generate a distortion in the model fitting procedure. Moreover, we do not consider as well the time of the first lap of the race since it can be viewed as a further outlier, strongly related to the qualifying position, and not particularly relevant for the specification of a good pit stop strategy.

Concerning the covariates, we underline that the explanatory variable Lap is measured by evaluating the fraction of laps already covered by a particular driver in a specific grand prix. In this way, we can consider data from different races. A further useful explanatory variable could be the quantity of fuel on board, but unfortunately this information is not available and it can not be inferred using variables such as traffic, driving strategy of the pilot, engine and car features. We account indirectly for the modification of the quantity of fuel on board by considering the covariate Lap, since the percentage of fuel on board decreases as the race lap number increases. Moreover, the covariate Distance is considered since the traffic could be a problem for drivers. We have empirically noticed that this effect is detectable only for distances in seconds less than a suitable value and then we decide to use this value also for greater distances, assuming that this is substantially equivalent to a clear track state.

We aim at specifying a model for describing a race without safety car deployment and, for this reason, in the model fitting procedure, we omit all data related to laps completed under safety car regime. Indeed, since the influence of the covariate Driver on the response variable is characterized by a random variability observed from driver to driver, we do not consider this effect as a fixed systematic effect, as for the other covariates. Then, for this purpose, we define a Generalized Additive Mixed Model (GAMM) with a random intercept term describing the driver effect. More precisely, the model structure is defined as

$$g(\mu_b) = b_k + \beta_1 \text{Team} + \bar{\beta}_2 \text{Pit} * \text{Circuit} + \bar{\beta}_3 \text{FollowingPit} * \text{Circuit} + s_1(\text{Lap}) + s_2(\text{TyreLaps}; \text{Compound}, \text{TyreState}) + s_3(\text{Distance}), \quad (1)$$

where  $b_k$  is the random effect of the  $k$ -th driver,  $k = 1, \dots, K$ , modelled as a normal distributed random variable with mean 0 and variance  $\sigma_D^2$ . The random intercepts are independent and identically distributed and the conditional distribution of the response  $Y$  given  $b_k$ ,  $k = 1, \dots, K$ , is defined within the exponential family class, with  $\mu_b = E(Y|b_k, k = 1, \dots, K)$ . Moreover, with the symbol  $*$  we highlight that we consider both the main effects and the interaction effects of the covariates taken into account, with  $\bar{\beta}_j$  the vector of the associated regression parameters.

The effects of the non-categorical covariates Lap, TyreLaps and Distance are estimated as smooth curves, while the other terms in the model formula are defined by considering linear functions of the explanatory variables, admitting also interaction effects. In principle, all the covariates can interact and influence each other, but some of them are more strongly related. For example, in order to evaluate the degradation of a set of tyres, it is indispensable to know not only the number of laps completed by the current set of tyres during the race, but also the compound and whether the tyres are new or already used. For this reason, different smooth functions of the same covariate TyreLaps are defined for each level of the factors Compound and TyreState. Similarly, if we are interested in the time lost during the pit stop or during the lap after

the pit stop, it is fundamental to consider the information on the particular circuit where the race is running. Moreover, the introduction of the variable indicating the lap after the pit stop is due to the fact that a pit stop modifies significantly the times registered during both the lap of the pit stop and the one that follows, because the pit lane is always crossed by the finish line.

At the end of the following section, we attempt to specify a further model, describing also the lap times under the safety car regime. The aim is to evaluate how this condition can modify the time lost during a pit stop; although the model requires a substantial improvement, this aspect could be very important from the race strategy point of view.

### 3 An application to Formula 1 data season 2015

#### 3.1 Model fitting

We estimate the proposed models using the statistical software R (R Core Team, 2017) and in particular the add-on package `mgcv` (see Wood, 2006). The data are obtained from the Ergast Developer API ([www.ergast.com/mrd](http://www.ergast.com/mrd)), which is an experimental web service providing motor racing data, and from the Pirelli web page ([www.pirelli.com](http://www.pirelli.com)).

At first we fit and analyse model (1) using data of five races of the Formula 1 season 2015; namely, Australia, Malaysia, Italy, Russia and United Arab Emirates. Ignoring all the laps completed under the safety car regime, we consider 4053 laps and we assume that the response variable, namely the logarithmic transformation of the lap time, follows a Gaussian distribution with the canonical link function. The specification of the GAMM model in the final form (1) has required a considerable effort. The parametric and the smooth regression terms, which are finally considered, assure a satisfactory description of the response, avoiding an excessive overfitting. The model is estimated following a penalized likelihood approach, where the smoothing parameters are selected using the Generalized Cross Validation (GCV) criterion, and we find that the significance of the individual regression parameters of the parametric terms and of the smooth terms is substantial. With regard to the smooth terms, we consider thin plate regression splines and the estimated degrees of freedom of  $s_1(\text{Lap})$ ,  $s_2(\text{TyreLaps}; \text{Compound}, \text{TyreState})$  and  $s_3(\text{Distance})$  are 8.13, 24.92 and 3.26, respectively. Thus, the total number of the effective degrees of freedom used in model (1) is 60.31, indicating that a relatively complex regression function has been estimated. Simpler model structures may achieve a more parsimonious fit, but they usually provide an unsatisfactory description of the response variable. Notice that we have six different smooth functions for the covariate `TyreLaps`, according to the six combinations of the observed values `new` and `used` of the factor `TyreState` and `supersoft`, `soft` and `medium` of the factor `Compound` (in the five races taken into account the `hard` compound tyres have never been used). Finally, with regard to the random intercept term, corresponding to the driver effect, we find that the common variance parameter  $\sigma_D^2$  is significantly different from zero.

Model checking is performed by considering the model residuals. In particular, the Q-Q plot of the standardized residuals shows a significant skewness and an heavy right tail in the distribution, and this is confirmed by the associated histogram (see Figure 1). The right tail represents the laps that are sensibly slower than expected, due to lapped cars or small driver errors; the slowest laps, caused by crashes and breakdowns, have already been deleted from the dataset imposing an upper threshold, as mentioned in Section 2. In order to mitigate these difficulties, we may consider as response variable an alternative Box-Cox

type transformation of the response Time, instead of  $\log(\text{Time})$ .

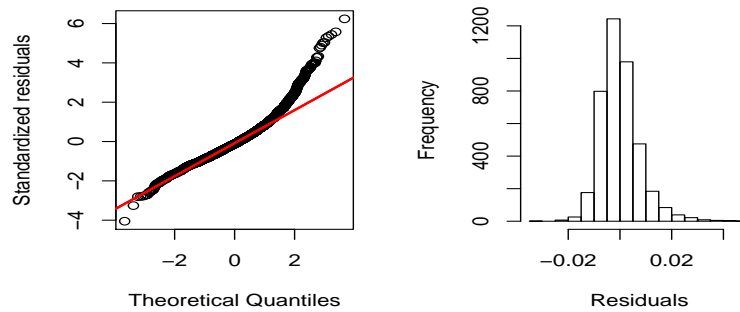


Figure 1: The Q-Q plot and the histogram of the standardized residuals relative to model (1).

Furthermore, in Figure 2 we show also the estimated residuals plotted against the fitted values and we observe that their variance is almost constant, as it should be. Moreover, the existence of clusters is a consequence of the fact that in the dataset we consider data from five different races and then the time for completing a single lap is not necessarily the same for all of them. Indeed, the small cluster on the right hand side is relative to some laps in which a pit stop is carried out and the predicted time turns out to be extremely high.

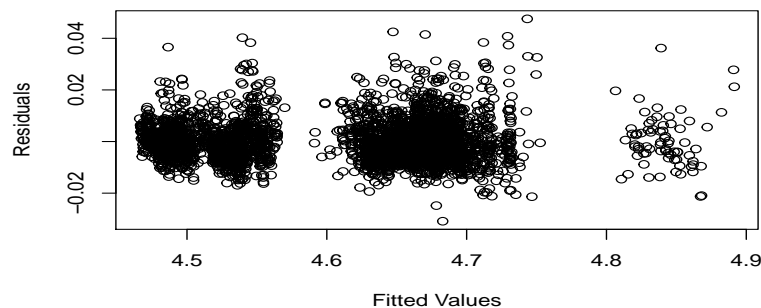


Figure 2: Estimated residuals versus fitted values relative to model (1).

It can be instructive to represent the partial effect of all the estimated smooth terms considered in the fitted the model. The six panels of Figure 3 show the smooth estimated effect of the covariate TyreLaps, with the 95% confidence limits, for the six different values observed for the pairs of factors Compound and TyreState. The range on the  $x$ -axis roughly corresponds to the observed values for the covariate, in order



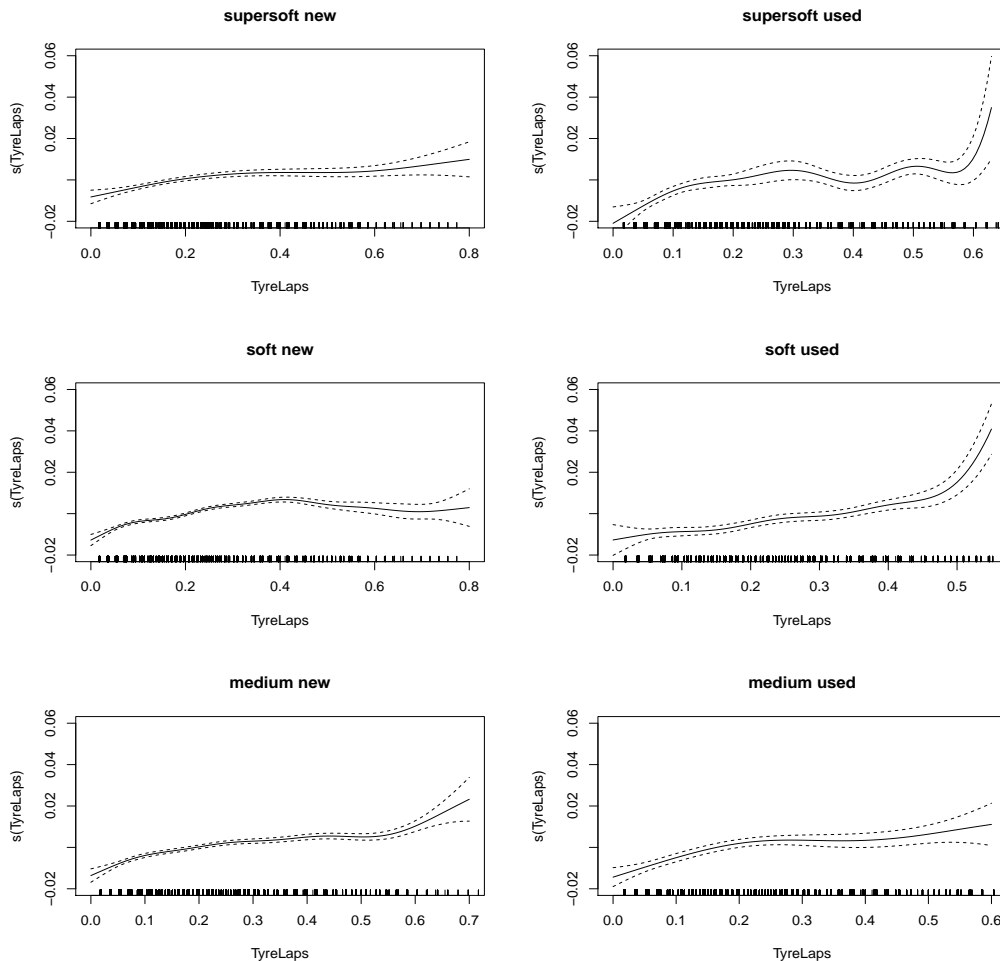


Figure 3: Estimated smooth functions of the covariate `TyreLaps` for the fitted model (1), without the observed values of the partial residuals; the 95% confidence limits are represented as dashed lines.

to avoid unreliable extrapolations. We observe that the estimated function tends to increase as `TyreLaps` increases, for each type of compound and tyre state. This provides a clear description of the effect of tyre degradation on the lap time. In particular, for a set of `supersoft used` tyres or `soft used` tyres, a sudden worsening of performance occurs whenever `TyreLaps` reaches half of the race. For the other types of tyres this effect does not usually appear, since the Formula 1 teams may guess more easily the critical degradation point and they usually change the tyres before reaching it. With regard to the `supersoft` compound, the use for more than 50% of the grand prix can be seen as quite unexpected. However, this circumstance has been observed only in the Russian grand prix, where the circuit has a peculiar type of asphalt, which causes a reduced tyre degradation effect. Moreover, also the fact that the smooth functions may sometimes present

a slight decrease, as `TyreLaps` increases, can be explained by recalling that we consider the overall effect of `TyreLaps` and we do not account for the particular effect that a specific circuit may have on tyre duration.

Two further plots are given in Figure 4 and they describe the estimated smooth effect due to the covariates `Lap` and `Distance`. The number of covered laps gives an indirect information about the amount of fuel on board, which conditions the weight of the car. Since function  $s(\text{Lap})$  decreases as `Lap` increases, we obviously conclude that a lighter car is usually faster. However, we notice that at the end of the race the values tend to increase. Perhaps, this can be explained as a confounding phenomenon with the tyre deterioration effect, which can be substantial in the final part of the race. Furthermore, it may simply describe that the drivers push less in the last laps, especially when it is not possible to improve the final rank position. Finally, regarding the effect of the covariate `Distance`, it is possible to observe that being too close to the driver that precedes leads to a lap time increase. This can be due to aerodynamics reasons or to the fact that the car ahead is slower than the one behind and the overtake is not easy. In addition, we mention the fact that battling with another car always causes a waste of time. This traffic effect decreases as the distance increases, reaching its minimum at 3.7 seconds, as already explained in Section 2.

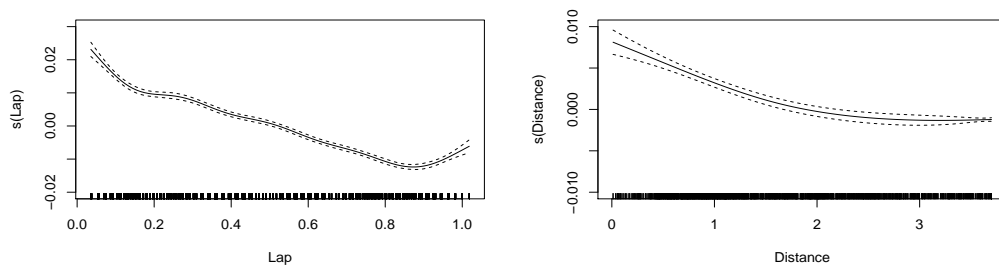


Figure 4: Estimated smooth functions of the covariates `Lap` (left) and `Distance` (right) for the fitted model (1), without the observed values of the partial residuals; the 95% confidence limits are represented as dashed lines.

### 3.2 Prediction

By considering the fitted model presented above, it is possible to make predictions on the random intercept terms, which describe the specific effect of each driver on the lap time evolution. In particular, we consider as point predictor for  $b_k$ ,  $k = 1, \dots, K$ , the estimated conditional expectation  $\hat{b}_k = \hat{E}(b_k|Y)$ . Using these predicted values, we may conclude, for example, that the difference between the expected  $\log(\text{Time})$  values for the Ferrari's drivers, Raikkonen and Vettel, is  $1.8 \cdot 10^{-3}$ . This means that, given the same observed values for the covariates, the lap times recorded by Raikkonen are approximately 1.002 times higher than those recorded by Vettel; that is, Raikkonen is expected to lose about 0.1 seconds per minute with respect to Vettel. In Figure 5 are represented both the times recorded by Vettel (in black) and Raikkonen (in purple) in the first 24 laps of the 2015 Italian grand prix and the corresponding predicted times in red and blue, respectively. Concerning the difference between the red and the blue lines, we emphasize that the two predicted patterns

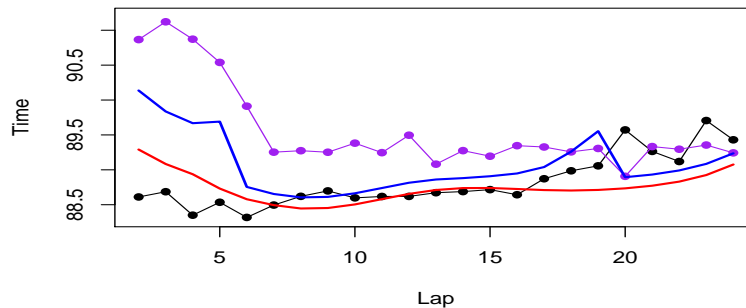


Figure 5: Times recorded by Vettel (black) and Raikkonen (purple) in the first 24 laps of the 2015 Italian grand prix and the predicted times for Vettel (red) and Raikkonen (blue).

present the same distance whenever the corresponding covariates have the same values, and then the difference is only due to the driver effect. The peaks observed for Raikkonen correspond to situations with a presence of traffic, so that the observed values for the covariate *Distance* differ significantly.

The GAMM model (1) could be usefully considered also for making predictions on the lap times, by assuming different values for the covariates. This can be crucial for evaluating and simulating alternative race strategies, with particular regard to the pit stop choices. Let us consider again the 2015 Italian grand prix, where almost all the drivers carried out only one pit stop. Then, we implicitly assume that this is the common decision for the teams, but we want to find which is the best lap for making the pit stop, taking also into account the 2015 Formula 1 rules. In that specific grand prix, the available compounds were the *soft* and the *medium* and each driver had to use both of them; moreover, by regulation, the first ten drivers on the grid had to start the race with the tyres used in the Q2 qualifying session, namely the *soft* compound. This is the case, for example, of the Williams' driver Massa, who started the race with *soft* used tyres and pitted at the 19th lap installing *medium* new tyres. In order to make predictions for the lap times of Massa, we fit model (1) deleting all the Williams' lap time data relative to 2015 Italian grand prix. We adopt the same basic principle underlying cross-validation procedures, in order avoid overfitting problems that may occur when we make predictions on data values already used for estimating the predictive model.

The total time spent by Massa to complete the race was 4728.323 seconds and the predicted total time given by the fitted model, imposing the pit stop at the 19th lap, turns out to be 4734.983 seconds. By simulating 52 different races, changing the lap in which the pit stop is carried out, we find out that the best choice is to stop at the 23rd lap, thus completing the whole race in 4732.81 seconds: more than two seconds can be gained, in the predicted total time, if the pit stop is postponed from the 19th to the 23rd lap. The estimated standard deviation for the predicted log time per lap is lower than 0.002 seconds. However, adopting this best strategy wouldn't change the race final outcome, since Massa arrived 3rd with a gap of 22 seconds from the 2nd. In Figure 6 we represent the real lap times recorded by Massa (in black), the lap times prediction with a pit stop at the 19th lap (in red) and the lap times prediction with a pit stop at the 23rd lap (in blue).

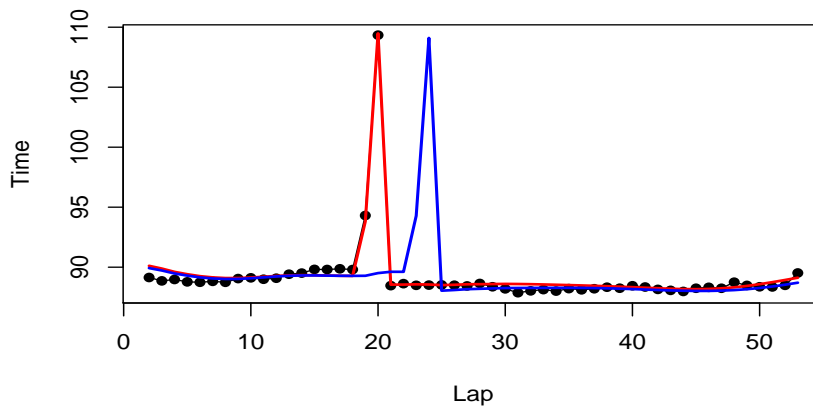


Figure 6: Lap times recorded by Massa in the 2015 Italian grand prix (black) and lap times prediction with a pit stop at the 19th lap (red) and at the 23rd lap (blue).

### 3.3 Extension to the safety car regime

The GAMM model (1) may be extended for describing also the lap times under the safety car regime. As emphasized before, this aspect is very important from the race strategy point of view. The model structure is necessarily more complex and it involves at least two additional model terms: the first one describing the combined effect of SC and Pit and the second one describing the combined effect of SC and FollowingPit. The model is estimated using also the lap times completed under the safety car regime, except the first one. Since the moment in which the safety car enters the track is completely random, the first lap in the safety car regime could be almost as fast as a normal lap, whenever it enters in the latest turns, or strongly slower, whenever it enters in the first turns. Then we consider 4211 lap times.

It is well-known that the teams usually carry out the pit stop when the safety car enters the track and this fact makes challenging the evaluation of the actual effect of the safety car presence on the lap time. Moreover, we usually observe an increase in the number of unexpected lap time data, indicating that the race is in a completely different regime. This is confirmed by the histogram of the standardized residuals and, in particular, by the graph with the estimated residuals plotted against the fitted values, where we observe that the variability of the lap times completed with the safety car (corresponding to the cluster in the right hand side) is significantly higher than that of the normal ones (see Figure (7)). As a consequence of this unsatisfactory model fitting, the lap times prediction specified under safety car deployment are usually characterized by a substantial error.

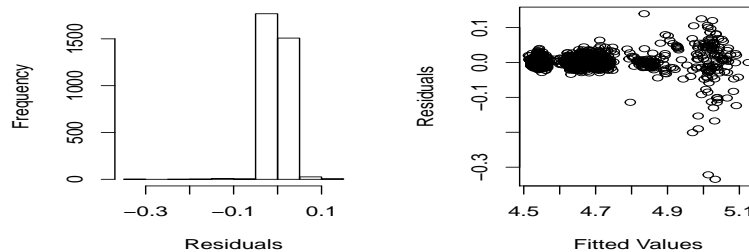


Figure 7: Histogram of the standardized residuals (left) and plot of the estimated residuals versus the fitted values (right) relative to the model describing also the lap times under the safety car regime.

## 4 Conclusions

Although the GAMM model (1) provides a satisfactory description of the lap time pattern of a Formula 1 grand prix, and it may be effective for prediction purposes, it can be further improved by considering the following potential developments. In particular, in order to deal with the asymmetry and the kurtosis observed in the distribution of the standardized residuals, we may consider a flexible extension of the GAM models, where also the scale and the shape of the distribution depend on suitable predictor variables (see, for example, Rigby and Stasinopoulos, 2005, and Kneib, 2013). Moreover, the goodness of the model may be increased by considering additional covariates, even if a potential problem with a more complex model can be related to the large number of main and interaction fixed effects. This may lead to overfitting, thus reducing the predictive ability of the estimated model. In order to mitigate this problem, a suitable covariate selection procedure has to be considered. One possibility is to use a boosting algorithm such as that one defined by Tutz and Binder (2006). Finally, with regard to the challenging situation with the safety car deployment, we empathize that an alternative model has to be defined, since the presence of the safety car determines a different race regime, which is not adequately described by the models presented in this paper.

## References

- [1] Hastie, T. and Tibshirani, R. (1986) *Generalized additive models*. Statistical Science **1**, 297-318.
- [2] Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. Chapman and Hall.
- [3] Kneib, T. (2013) *Beyond mean regression*. Statistical Modelling **13**, 275-303.
- [4] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>.
- [5] Rigby, R.A. and Stasinopoulos D.M. (2005) *Generalized additive models for location, scale and shape*. Journal of the Royal Statistical Society Series C (Applied Statistics) **54**, 507-554.
- [6] Tutz G. and Binder H. (2006) *Generalized additive modeling with implicit variable selection by likelihood-based boosting*. Biometrics **62**, 961-971.
- [7] Wood, S.N. (2006) *Generalized Additive Models: an introduction with R*. Chapman & Hall/CRC.

# A comparison of the Performance of the generalized Google PageRank model (GeM) to other Ranking Systems on the NCAA Basketball Tournament

P. Coletti\* and A. Pilkington\*\*

\*ACE Program, University of Notre Dame, Notre Dame, IN 46556: Paul.E.Coletti.2@nd.edu

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

## Abstract

In this paper we compare the performance of a generalized version of Google's PageRank(GeM) ranking system on predicting outcomes in the NCAA Basketball tournament over the years from 2010 to 2016 to other popular ranking systems, namely Colley, Massey, and Keener. We use both raw and weighted versions of these rankings. We use the score obtained by a ranking method for the ESPN Tournament Challenge in each year as a measure of its effectiveness. The results show that although the GeM algorithm outdoes the other systems for 2015, its scores on the challenges are not significantly different from those obtained by algorithms based on the Massey and Colley methods. On the other hand, algorithms based on Keener's method show significantly lower scores than those based on other methods. There is also significant variation in results from year to year, with some years showing significantly lower or higher scores across methods than other years.

## 1 Introduction

The challenge of predicting the outcomes of all of the games in the NCAA Division I Basketball Tournament each March is a popular but difficult challenge. After many attempts to use ranking systems from linear algebra to predict the winners, we noticed that there is much variation in the performance of the different ranking systems in any given year. There is also much variation in the performance of a given system from year to year; frequently a system that excels in one year does badly in the next or vice versa. This sparked our curiosity and led us to examine the variation in the performance of four main ranking systems over a period of seven years.

Each March thousands of pools are created on ESPN before the tournament begins. Each pool member fills out their predictions and then watches the story unfold over the subsequent weeks, during what is commonly known as March Madness. Since there are sixty four teams in this knockout tournament (after the round involving the eight play-in teams, four of which make it into the tournament), it is clearly impossible to fill out a bracket for each possible scenario. It is also highly unlikely that even the most ardent basketball fans will be familiar with the strengths and weaknesses of all 64 teams, making it necessary to employ an algorithm in order to reasonably attempt a prediction of the tournament's outcomes.

In the ESPN Tournament Challenge, points are awarded to a bracket if it has predicted the correct winner for a game. There are six rounds in the tournament. Ten points are awarded for each correct prediction of

a game winner in round 1. The number of points awarded for a correct prediction doubles as the rounds progress, ending with an award of 320 points for correctly predicting the winner of the final game. Clearly, maximizing the number of points gained for such a bracket depends on many factors. An upset in the first round that takes out a strong team can easily take a large toll on the points gained from a quarter of a bracket, and the elimination of strong teams in the final three rounds can render one's bracket completely void for those rounds. Because one must predict the winners of all games before the tournament begins, a method of prediction that is 70% accurate on predicting the outcomes of the games that are actually played can have a greatly reduced level of accuracy overall.

In this paper, we compare the performance of some popular computer ranking models in the ESPN Tournament Challenge, using the points gained in the challenge as a measure of the effectiveness of each system. Each computer ranking model creates a rating for each NCAA Division I team, and each such rating can be used to rank the teams. We assign number 1 to the team with the highest rating, number 2 to the team with the second highest rating, etc. The ranking systems considered are based on algorithms developed by W. Colley (Colley (2002)), J.P. Keener (Keener (1993)) and K. Massey (Massey (1997)), and on an adaptation of Google's PageRank algorithm (Brin and Page (1998)) developed by A. Govan, C. Meyer and R. Albright (Govan et al. (2008)) to rank sports teams. We compared the performance of the basic ranking methods along with the performance of the ranking methods with various weighting systems. The methods were implemented using R from the CRAN website.

In Section 2 we describe our working example and where to find the relevant game data. In Section 3 we present a brief introduction to the ranking systems and demonstrate their implementation on the working example. In Section 4 we present and analyze the results of the study for the NCAA tournament.

*Remark 1.* Hyperlinks appear in blue throughout.

## 2 Our Working Example

For our working example, we have chosen the games played in the Big 12 conference in 2015 prior to the conference tournament. The teams, along with their win loss record and total point differential for these conference games, are shown in Table 1. We also assign an index to each team, which are shown in alphabetical order in the table. The assigned indices are shown in the left hand column of Table 1.

Index	Team	W-L	PD
1	Baylor	11-7	91
2	Iowa State	12-6	85
3	Kansas	13-5	123
4	Kansas State	8-10	-78
5	Oklahoma	12-6	106
6	Oklahoma State	8-10	-3
7	TCU	4-14	-92
8	Texas	8-10	16
9	Texas Tech	3-15	-243
10	West Virginia	11-7	-5

Table 1: Win-loss record, point differential and assigned index of Big 12 conference teams for games played prior to 2015's conference tournament.

Our data for this study was obtained from the website <http://masseyratings.com/>, maintained by K. Massey. Under Information/Data, one can currently obtain archived records of games played in many sports leagues, dating back to the late 1990's. You can find a list of the games played in our working example under [College basketball 2015/NCAA/NCAAI/Big 12/All/Intra](#).

### 3 The Ranking Systems

#### 3.1 Colley's Method

Colley's method (Colley (2002), Langville and Meyer (2012)) of ranking is based only on the ratings for the teams derived from Laplace's rule of succession:

$$r_i = \frac{1 + w_i}{2 + t_i},$$

where  $w_i$  is the number of wins for team  $i$  and  $t_i$  is the number of games played by team  $i$ . Colley makes the observation that the Laplace ratings derived from this rule for a team's set of opponents,  $O_i$ , should average to roughly  $1/2$ . Thus  $t_i/2 \approx \sum_{j \in O_i} r_j$ . Colley then cleverly manipulates wins to get the following approximation for  $w_i$ :

$$\begin{aligned} w_i &= \frac{w_i - l_i}{2} + \frac{w_i + l_i}{2} \\ &= \frac{w_i - l_i}{2} + \frac{t_i}{2} \\ &= \frac{w_i - l_i}{2} + \sum_{j=1}^{t_i} \frac{1}{2} \\ &\approx \frac{w_i - l_i}{2} + \sum_{j \in O_i} r_j, \end{aligned}$$



where  $l_i$  denotes the number of losses for team  $i$ . He then substitutes this approximation for  $w_i$  into the right hand side of the ratings from Laplace's rule to derive the following system of linear equations:

$$\left\{ (2 + t_i)r_i - \sum_{j \neq i} n_{ij}r_j = 1 + \frac{w_i - l_i}{2} \right\}_{1 \leq i \leq n}$$

where  $n_{ij}$  denotes the number of times team  $i$  has played team  $j$ . There is one linear equation in this system for each team where the teams are assigned an index in the set  $\{1, \dots, n\}$ . Colley's ratings for the teams are the solutions to the above system of linear equations.

For our running example, the resulting system of equations  $Cr = B$  (to be solved for Colley's ratings) is shown below:

$$\begin{bmatrix} 20 & -2 & -2 & -2 & -2 & -2 & -2 & -2 & -2 & -2 \\ -2 & 20 & -2 & -2 & -2 & -2 & -2 & -2 & -2 & -2 \\ -2 & -2 & 20 & -2 & -2 & -2 & -2 & -2 & -2 & -2 \\ -2 & -2 & -2 & 20 & -2 & -2 & -2 & -2 & -2 & -2 \\ -2 & -2 & -2 & -2 & 20 & -2 & -2 & -2 & -2 & -2 \\ -2 & -2 & -2 & -2 & -2 & 20 & -2 & -2 & -2 & -2 \\ -2 & -2 & -2 & -2 & -2 & -2 & 20 & -2 & -2 & -2 \\ -2 & -2 & -2 & -2 & -2 & -2 & -2 & 20 & -2 & -2 \\ -2 & -2 & -2 & -2 & -2 & -2 & -2 & -2 & 20 & -2 \\ -2 & -2 & -2 & -2 & -2 & -2 & -2 & -2 & -2 & 20 \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \\ r_6 \\ r_7 \\ r_8 \\ r_9 \\ r_{10} \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ 5 \\ 0 \\ 4 \\ 0 \\ -4 \\ 0 \\ -5 \\ 3 \end{bmatrix}.$$

One can solve these equations in  $\mathbb{R}^1$  to get the ratings for the teams, which are rounded off and shown with the corresponding rankings in Table 2.

Team	Rank	Rating
Kansas	1	0.68
Iowa State	2	0.64
Oklahoma	3	0.64
Baylor	4	0.59
West Virginia	5	0.59
Kansas State	6	0.45
Oklahoma State	7	0.45
Texas	8	0.45
TCU	9	0.27
Texas Tech	10	0.23

Table 2: Rankings for the Big 12 teams, 2015, using Colley's method.

<sup>1</sup>One can solve the system  $Cr=B$  in  $\mathbb{R}$  with the command `solve(C,B)`

*Remark 2.* One sees that when we have the full results of a round robin tournament, the rankings derived from Colley’s system will coincide with those derived from the statistic wins minus losses. We let  $\mathbf{e}$  denote a column vector of 1’s with dimension equal to  $n$ , where  $n$  is the number of teams in the tournament. We then see that  $B-\mathbf{e}$  and  $\mathbf{e}$  are both eigenvectors of  $C$ , with eigenvalues  $2+n$  and  $2$  respectively. This follows easily from the fact that the sum of wins minus losses for all teams must be 0. Thus we see that the the solution to the system of equations is  $\frac{1}{2+n}(B-\mathbf{e}) + \frac{1}{2}\mathbf{e}$ , which gives the same ranking as wins minus losses. Excepting an analysis of a single or double round robin tournament’s results, Colley’s rankings can and most often do give a different hierarchy than wins minus losses.

### 3.2 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. It should not be a surprise that, with this approach the probability of getting a system of equations that is inconsistent is very high. However, Massey takes the least squares solution to this system to derive a system of equations with infinitely many solutions. He then replaces the last equation 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 is the system of equations

$$\left\{ \begin{array}{l} t_i r_i - \sum_{j \neq i} n_{ij} r_j = P_i, \\ \sum_j r_j = 0 \end{array} \right\}_{1 \leq i \leq n-1, 1 \leq j \leq n}$$

for the  $n$  teams in the tournament. For this system  $t_i$  and  $n_{ij}$  have the same meaning as in the previous section,  $P_i$  is the total point differential for team  $i$ , and its solution gives us Massey’s ratings.

When applied to our working example, we get the following matrix equation  $\mathbf{M}\mathbf{r} = \mathbf{P}$ :

$$\begin{bmatrix} 18 & -2 & -2 & -2 & -2 & -2 & -2 & -2 & -2 & -2 \\ -2 & 18 & -2 & -2 & -2 & -2 & -2 & -2 & -2 & -2 \\ -2 & -2 & 18 & -2 & -2 & -2 & -2 & -2 & -2 & -2 \\ -2 & -2 & -2 & 18 & -2 & -2 & -2 & -2 & -2 & -2 \\ -2 & -2 & -2 & -2 & 18 & -2 & -2 & -2 & -2 & -2 \\ -2 & -2 & -2 & -2 & -2 & 18 & -2 & -2 & -2 & -2 \\ -2 & -2 & -2 & -2 & -2 & -2 & 18 & -2 & -2 & -2 \\ -2 & -2 & -2 & -2 & -2 & -2 & -2 & 18 & -2 & -2 \\ -2 & -2 & -2 & -2 & -2 & -2 & -2 & -2 & 18 & -2 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \\ r_5 \\ r_6 \\ r_7 \\ r_8 \\ r_9 \\ r_{10} \end{bmatrix} = \begin{bmatrix} 91 \\ 85 \\ 123 \\ -78 \\ 106 \\ -3 \\ -92 \\ 16 \\ -243 \\ 0 \end{bmatrix}$$

We can solve for Massey’s ratings in R. The resulting ratings, with corresponding rankings, are given in Table 3.

Team	Rank	Rating
Kansas	1	6.15
Oklahoma	2	5.30
Baylor	3	4.55
Iowa State	4	4.30
Texas	5	0.80
Oklahoma State	6	-0.15
West Virginia	7	-0.30
Kansas State	8	-3.90
TCU	9	-4.60
Texas Tech	10	-12.15

Table 3: Rankings for the Big 12 teams, 2015, using Massey's method.

*Remark 3.* Massey's method gives the same ranking as the point differential when applied to the full results of a round robin tournament. Since the point differentials for all of the teams must add to 0, it is not difficult to see that a constant multiple of the vector of point differentials is a solution to the above system  $\mathbf{M}\mathbf{r} = \mathbf{P}$ . As with Colley's method, if we do not have the full results of a round robin tournament, the rankings obtained will most likely deviate from those obtained via the point differential.

### 3.3 Keener's Method

Keener's method is based on the existence of a unique eigenvector (up to multiplication by a constant) associated to the dominant eigenvalue for a non-negative, irreducible matrix  $K$  as guaranteed by the Perron-Frobenius Theorem (Meyer (2005), Section 8.3). Here non-negative implies that  $K_{ij} \geq 0$  for all  $i$  and  $j$ , and irreducible implies that for each pair of indices  $(i, j)$ , there is a sequence of entries in  $K$  such that  $K_{i,j_1}K_{j_1,j_2}\dots K_{j_k,j} \neq 0$ . The matrix  $K$  is created using any non-negative game statistic, such as points scored. We let  $s_{ij}$  be the value of the statistic for team  $i$ , for the game played between team  $i$  and team  $j$ . Keener then applies a smoothing function,  $h(x)$ , to avoid having the ratings influenced by outliers and manipulation. The matrix  $K$  is given by:

$$K_{i,j} = \begin{cases} h\left(\frac{s_{ij}+1}{s_{ij}+s_{ji}+2}\right) & \text{if team } i \text{ has played team } j \\ 0 & \text{otherwise} \end{cases},$$

where  $h(x) = \frac{1}{2} + \frac{\text{sgn}(x-1/2)\sqrt{|2x-1|}}{2}$ . If team  $i$  has played team  $j$  more than once, then  $s_{ij}$  is the sum of the values of the relevant statistics for team  $i$  over all games played between the teams. The Keener ratings are given by the eigenvector associated to the dominant eigenvalue of  $K$ .

When we apply Keener's method, using the statistic "points scored", to our working example, we get the following matrix:

$$K = \begin{pmatrix} 0 & 0.59 & 0.40 & 0.66 & 0.53 & 0.36 & 0.66 & 0.64 & 0.59 & 0.66 \\ 0.41 & 0 & 0.42 & 0.57 & 0.44 & 0.58 & 0.65 & 0.59 & 0.67 & 0.64 \\ 0.60 & 0.58 & 0 & 0.56 & 0.56 & 0.57 & 0.60 & 0.63 & 0.73 & 0.57 \\ 0.34 & 0.43 & 0.44 & 0 & 0.58 & 0.43 & 0.40 & 0.36 & 0.39 & 0.40 \\ 0.47 & 0.56 & 0.44 & 0.42 & 0 & 0.65 & 0.64 & 0.65 & 0.71 & 0.46 \\ 0.64 & 0.42 & 0.43 & 0.57 & 0.35 & 0 & 0.45 & 0.61 & 0.64 & 0.37 \\ 0.34 & 0.35 & 0.40 & 0.60 & 0.36 & 0.55 & 0 & 0.29 & 0.70 & 0.40 \\ 0.36 & 0.41 & 0.37 & 0.64 & 0.35 & 0.39 & 0.71 & 0 & 0.66 & 0.64 \\ 0.41 & 0.33 & 0.27 & 0.61 & 0.29 & 0.36 & 0.30 & 0.34 & 0 & 0.34 \\ 0.34 & 0.36 & 0.43 & 0.60 & 0.54 & 0.63 & 0.60 & 0.36 & 0.66 & 0 \end{pmatrix}$$

We can find the eigenvalues and eigenvectors of this matrix in  $\mathbb{R}^2$ . Then picking out the largest eigenvalue and the corresponding eigenvector, we get the ratings and corresponding rankings of the teams shown in Table 4.

Team	Rank	Rating
Kansas	1	0.3735
Baylor	2	0.3518
Oklahoma	3	0.3455
Iowa_St	4	0.3428
Oklahoma_St	5	0.3136
Texas	6	0.3132
West_Virginia	7	0.3111
TCU	8	0.2776
Kansas_St	9	0.2740
Texas_Tech	10	0.2333

Table 4: Rankings for the Big 12 teams, 2015, using Keener's method.

*Remark 4.* The irreducibility of the matrix  $K$  depends on the number of games played in the league and the interconnectivity of the associated graph. If the matrix is not irreducible, one can always add a small perturbation matrix to it (as we do in the GeM model below) to force irreducibility.

### 3.4 Generalized Page Rank (GeM)

The original PageRank algorithm (Brin and Page (1998)) 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 it has a unique eigenvector with eigenvalue 1 and norm 1. This is a special case of the Perron Frobenius theorem. The method normalizes the adjacency matrix of the directed graph of web page links to get a hyperlink matrix  $H$ . The method then adjusts  $H$  by adding a rank 1 matrix to deal with dangling nodes

<sup>2</sup>To create a table of eigenvalues and eigenvectors for a matrix  $K$  using R, we use the command `eigen(K)`

(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) can be computed as an eigenvalue of the transpose matrix 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 Govan et al. (2008). These researchers represent a sport season 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:

- 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  $\alpha$  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  $\alpha$  give a larger perturbation of the GeM matrix.

For our running example with  $\alpha = 0.85$  , the matrix  $H$  looks like:

$$H = \begin{pmatrix} 0 & 0 & 11/45 & 2/45 & 2/9 & 4/9 & 0 & 2/45 & 0 & 0 \\ 1/4 & 0 & 13/40 & 1/40 & 11/40 & 0 & 0 & 0 & 1/8 & 0 \\ 0 & 1/4 & 0 & 7/20 & 1/10 & 1/4 & 0 & 0 & 0 & 1/20 \\ 27/116 & 3/58 & 11/116 & 0 & 0 & 7/58 & 7/58 & 17/116 & 17/116 & 5/58 \\ 11/52 & 7/52 & 7/52 & 3/26 & 0 & 0 & 0 & 0 & 0 & 21/52 \\ 0 & 7/87 & 10/87 & 10/87 & 25/87 & 0 & 5/29 & 0 & 1/87 & 19/87 \\ 27/158 & 15/79 & 6/79 & 5/158 & 19/158 & 6/79 & 0 & 41/158 & 0 & 6/79 \\ 23/95 & 11/95 & 18/95 & 0 & 23/95 & 13/95 & 0 & 0 & 0 & 7/95 \\ 4/133 & 37/266 & 27/133 & 1/38 & 7/38 & 10/133 & 37/266 & 12/133 & 015/133 & \\ 2/7 & 22/105 & 1/15 & 019/105 & 0 & 0 & 9/35 & 0 & 0 & \end{pmatrix}$$

No adjustment for unbeaten teams is necessary, and the matrix  $G$  is the above matrix with  $(1 - 0.85)/10 = 3/200$  added to each entry:

$$G = \begin{bmatrix} 3/200 & 3/200 & 401/1800 & 19/360 & 367/1800 & 707/1800 & 3/200 & 19/360 & 3/200 & 3/200 \\ 91/400 & 3/200 & 233/800 & 29/800 & 199/800 & 3/200 & 3/200 & 3/200 & 97/800 & 3/200 \\ 3/200 & 91/400 & 3/200 & 5/16 & 1/10 & 91/400 & 3/200 & 3/200 & 3/200 & 23/400 \\ 2469/11600 & 171/2900 & 1109/11600 & 3/200 & 3/200 & 341/2900 & 341/2900 & 1619/11600 & 1619/11600 & 64/725 \\ 1013/5200 & 406/3137 & 406/3137 & 147/1300 & 3/200 & 3/200 & 3/200 & 3/200 & 3/200 & 977/2727 \\ 3/200 & 1451/17400 & 1961/17400 & 1961/17400 & 4511/17400 & 3/200 & 937/5800 & 3/200 & 431/17400 & 3491/17400 \\ 633/3950 & 2787/15800 & 1257/15800 & 331/7900 & 463/3950 & 1257/15800 & 3/200 & 1861/7900 & 3/200 & 1257/15800 \\ 839/3800 & 431/3800 & 669/3800 & 3/200 & 839/3800 & 499/3800 & 3/200 & 3/200 & 3/200 & 59/760 \\ 1079/26600 & 443/3325 & 4989/26600 & 71/1900 & 163/950 & 2099/26600 & 443/3325 & 2439/26600 & 3/200 & 2949/26600 \\ 361/1400 & 811/4200 & 43/600 & 3/200 & 709/4200 & 3/200 & 3/200 & 327/1400 & 3/200 & 3/200 \end{bmatrix}$$

The ratings and rankings produced are shown in Table 5.

Team	Rank	Rating
Oklahoma	1	0.148
Kansas	2	0.136
Baylor	3	0.132
Oklahoma_St	4	0.117
Iowa_St	5	0.1114
West_Virginia	6	0.1110
Kansas_St	7	0.091
Texas	8	0.069
TCU	9	0.046
Texas_Tech	10	0.039

Table 5: Rankings for the Big 12 teams, 2015, using the GeM method.

## 4 Performance of Ranking systems on NCAA tournament

We look at the performance of the above four ranking systems in the yearly ESPN Tournament Challenge from the 2009-2010 season to the 2015-2016 season. As mentioned in the introduction, points are awarded to a bracket if you have predicted the correct winner for a game, with 320 points awarded in each round. There are six rounds in the (single-elimination) tournament. Ten points are awarded for each correct prediction of a game winner in round one and the number of points awarded for a correct prediction doubles as the rounds progress. Thus the tournament challenge ends with an award of 320 points for correctly predicting the winner of the final game. The total number of points possible is 1,920.

We applied the ranking systems described above to the games played in each season prior to the tournament. We also applied weighted versions of the ranking systems. The weights, chosen based on experimentation, are described below. In developing these weights, we note that early season games are important in that they are the only source of interaction between the conferences. On the other hand, the early season is a time of experimentation for many teams, so their play may not be representative of their best performances. We

chose to weight the early season games highly with a uniform weight. We then weighted the in-conference games later in the season more highly than earlier ones since a team's performance in the tournament is likely to be similar to that at the end of its own conference games. Since the home team has a home team advantage, away wins/home losses were weighted higher than home/neutral wins. In summary we used the following rules to weight games in our weighted rankings:

- We assigned a weight of 10 to out-of-conference, early season games.
- For conference season games, we increased the weight from 1 to 10 in 10 steps over the course of the season, having carved the season into 10 time intervals of equal length.
- Away wins were weighted as twice the weight previously assigned.

The ratings vectors were created prior to the start of the tournament and the corresponding rankings used to decide the winner of each game before the tournament began.

## 4.1 Results

The results in Table 6 below show the total score of a bracket by year and the ranking method used. The results of a number of statistical tests run on the data are discussed in this section. In Figures 1 and 2, we show summary statistics and boxplots for all methods.

Year	Massey	Weighted Massey	Colley	Weighted Colley	Keener	Weighted Keener	PageRank	Weighted PageRank
2010	810	790	660	860	400	400	770	670
2011	480	490	520	480	300	300	550	540
2012	1340	1270	850	870	540	550	600	620
2013	850	850	600	1050	360	360	590	570
2014	570	590	680	690	560	560	580	620
2015	1060	1100	840	800	750	740	1300	1230
2016	750	680	860	610	730	750	730	650

Table 6: Bracket scores by year and method (2016 marks end of 2015/2016 season).

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
WeightedColley	7	765.71	188.579	480	1050	610.00	800.00	870.00
WeightedMassey	7	824.29	278.080	490	1270	590.00	790.00	1100.00
WeightedGooglePR	7	701.43	241.622	540	1240	570.00	620.00	670.00
WeightedKeener	7	522.86	178.952	300	750	360.00	550.00	740.00
Colley	7	715.71	135.629	520	860	600.00	680.00	850.00
Massey	7	837.14	291.531	480	1340	570.00	810.00	1060.00
GooglePR	7	731.43	263.908	550	1300	580.00	600.00	770.00
Keener	7	520.00	176.730	300	750	360.00	540.00	730.00

Figure 1: Summary statistics for ranking methods by method.

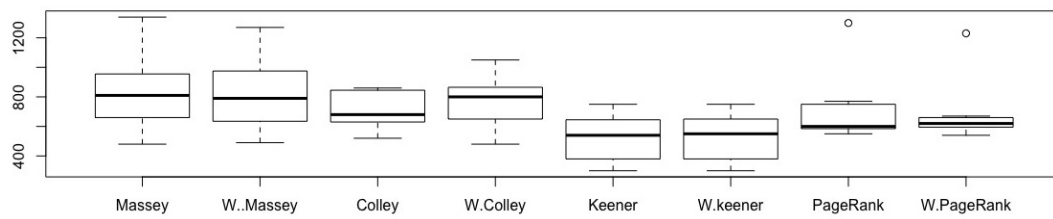


Figure 2: Boxplots for ranking methods (using yearly data).

The heat map in Figure 3 below uses Pearson’s method to show pairwise correlations between ranking methods, colored and ordered according to the strength of the correlation. This shows some surprising results. Three out of the four ranking systems are strongly correlated with their weighted version, whereas Colley’s method has a weak correlation with its weighted version. In fact, the weighted Colley method is very weakly correlated with everything, except both the Massey and weighted Massey methods; it even has a very weak negative correlation with both versions of Keener’s method. On the other hand, Colley’s method without the weighting is strongly correlated with both versions of Keener’s method. Less notable but strong correlations also exist between the PageRank methods and the Keener methods, and between Colley’s method and both versions of Massey’s method.

We also looked at pairwise differences in performance of methods, using a T-test for paired data (paired by year). The p-values resulting from these tests are shown in Table 7, with an asterisk next to those p-values that fall below an  $\alpha = 0.05$  level of significance. We see that both the Keener and weighted Keener methods have significantly lower scores than the regular and weighted versions of Massey’s method and the GeM(PageRank) method.

A Friedman test also showed significant differences in scores depending on which type of method was used, resulting in values of  $\chi^2(7) = 19.274$  and  $p = 0.007$ . In addition, follow up Wilcoxon signed-rank tests were carried out. We give a summary of the results of these Wilcoxon tests in Table 8. The lower triangle of the table gives the Wilcoxon p-value of the difference between the results of the row method and those of the column method. The upper triangle gives the number of years (out of 7) in which the row method



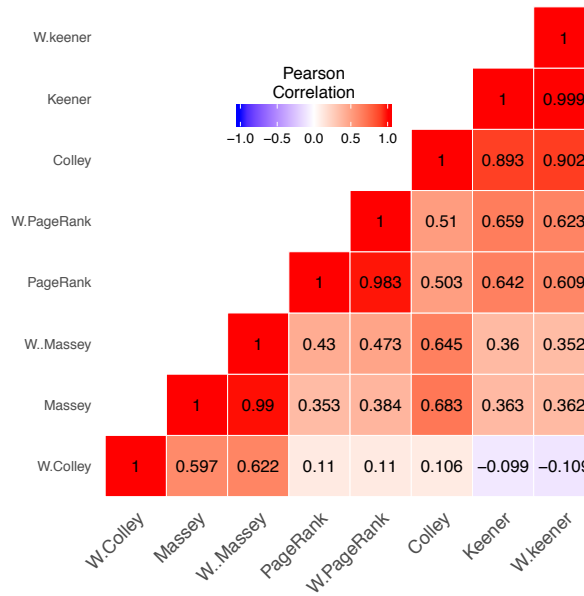


Figure 3: Heat map showing pairwise correlations (Pearson) between ranking methods.

gave a score larger than the column method. Although the effect from the Wilcoxon results is cancelled by a Bonferroni adjustment (used in testing the hypothesis that all means are the same) because of the number of comparisons, we see again that there is a significant difference between the Keener and weighted Keener methods in a number of pairwise comparisons to other methods. As with the T-tests for paired data, we see no significant differences between the methods based on Colley’s method, Massey’s method, and the PageRank method.

A Comparison of Computer Rankings for NCAA Basketball

Coletti, Pilkington

	Massey	W.Massey	Colley	W.Colley	Keener	W.Keener	PageRank
W.Massey	0.45995	-	-	-	-	-	-
Colley	0.19825	0.23360	-	-	-	-	-
W.Colley	0.45063	0.50439	0.57016	-	-	-	-
Keener	0.02438*	0.02475*	0.00077*	0.05328	-	-	-
W.Keener	0.02552*	0.02646*	0.00077*	0.05730	0.45705	-	-
PageRank	0.41106	0.42859	0.86141	0.77764	0.03271*	0.03920*	-
W.PageRank	0.26784	0.26451	0.84605	0.56668	0.03831*	0.04753*	0.16806

Table 7: p-values from paired t-tests comparing methods two at a time.  
A \* indicates a p-value below the 0.05 level of significance.

	W.Colley	W.Massey	W.PageRank	W.Keener	Colley	Massey	PageRank	Keener
W.Colley	-	3	4	6	4	4(1)	4	6
W.Massey	0.672	-	4	6	4	4(1)	4	6
W.PageRank	0.398	0.446	-	6	3	3	2	6
W.Keener	0.063	0.043*	0.063	-	0	1(1)	1	6(4)
Colley	0.866	0.237	0.499	0.018*	-	3	4	7
Massey	0.463	0.527	0.310	0.028*	0.176	-	4	7
PageRank	0.735	0.735	0.204	0.034*	0.866	0.499	-	7(1)
Keener	0.043*	0.043*	0.051	0.414	0.018*	0.018*	0.028*	-

Table 8: Results of signed rank Wilcoxon test applied to methods in pairs. p-values appear in the lower triangle. Number of years for which (row method score - col. method score) ≥ 0 appears in upper triangle. Number of ties (if relevant) in brackets.

As mentioned in the introduction, part of the motivation for studying these results was because we noticed large fluctuations in the performance of ranking methods based on linear algebra from year to year. In Figure 4 we show summary statistics for the scores by year.

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
Y2010	8	670.00	179.682	400	860	465.00	720.00	805.00
Y2011	8	457.50	100.676	300	550	345.00	485.00	535.00
Y2012	8	830.00	319.643	540	1340	562.50	735.00	1170.00
Y2013	8	653.75	245.120	360	1050	412.50	595.00	850.00
Y2014	8	606.25	52.355	560	690	562.50	585.00	665.00
Y2015	8	978.75	224.654	740	1300	762.50	950.00	1205.00
Y2016	8	720.00	75.782	610	860	657.50	730.00	750.00

Figure 4: Summary statistics for ranking methods by year.

In Table 9 below, we show the p-values from T-tests for paired data comparing the results of the set of

methods between two years. We see that the scores for 2011 are significantly different from those in other years when compared using the paired T-tests. Furthermore the scores for 2015 are significantly different from other years, with the exception of the scores for 2012.

We also applied a Friedman test to the data in Table 6, which showed significant differences in the scores depending on which year the methods were applied. This test resulted in values of  $\chi^2(6) = 27.04$  and  $p = 0.000$ . We show the results (without adjustment) of follow up Wilcoxon tests in Table 10, with the number of methods for which the row year had a score greater than or equal to that of the column year in the upper triangle (and the number of draws in brackets if relevant). In the lower triangle, we recorded the p-value for the Wilcoxon test comparing both years. As we saw with the paired T-tests, the year 2011 has significantly different scores than all of the other years, and the scores in the year 2015 are significantly different than all other years except 2012. One interesting feature of the table is that, with the exception of 2013 and 2014, each year has significantly different scores on the ranking methods than adjacent years.

	2010	2011	2012	2013	2014	2015
2011	0.0010*	-	-	-	-	-
2012	0.1058	0.0098*	-	-	-	-
2013	0.6947	0.0314*	0.0565	-	-	-
2014	0.3030	0.0019*	0.0885	0.5650	-	-
2015	0.0031*	$4.1e - 05^*$	0.2779	0.0182*	0.0031*	-
2016	0.5366	0.0008*	0.3768	0.5358	0.0121*	0.0229*

Table 9: p-values from paired t-tests comparing years two at a time with no adjustment. A \* indicates a p-value below the 0.05 level of significance.

	2010	2011	2012	2013	2014	2015	2016
2010	-	8	2	5	5	1	5
2011	0.012*	-	0	0	0	0	0
2012	0.123	0.012*	-	7	6(1)	4	3
2013	0.622	0.012*	0.042*	-	4	1	3
2014	0.161	0.012*	0.075	0.575	-	0	1
2015	0.017*	0.012*	0.401	0.042*	0.012*	-	6
2016	0.779	0.012*	0.674	0.575	0.025*	0.042*	-

Table 10: Results of signed rank Wilcoxon test applied to years in pairs. p-values appear in the lower triangle. Number of methods for which (row method score - col. method score)  $\geq 0$  appears in upper triangle. Number of ties (if relevant) in brackets.

## 5 Conclusion

Our results indicate that our weighting of the ranking systems did not produce a significant change in performance. Although our research showed that this weighting system was one of the better ones among those we studied, further research is necessary in order to find an optimal weighting system for each method. In retrospect, it might be best to either ignore early season games, when teams are experimenting with strategies, or give them less weight than games played near the start of the in-conference games.

We see that, when compared pairwise, the ranking methods based on Keener's method have significantly smaller scores than those based on the other methods. By contrast, although each of the other methods had good years and bad years, there was no significant difference between the other methods when compared pairwise. It is conceivable that a different combination of game statistics used in Keener's method would yield better results. The generalized PageRank model allows for much greater flexibility in the choice and use of game statistics, a combination of which may yield better results than the version we used. It is also possible that different values of the scaling parameter  $\alpha$  in the generalized PageRank model would yield better results.

We saw that, in a pairwise comparison, the results for 2011 were significantly lower than those for other years and with the exception of 2012, the results for 2015 were significantly higher than those for other years. The Wilcoxon signed rank test also revealed an interesting pattern of significant differences in results between consecutive years, with the exception of 2013 and 2014. The fluctuations in performance of the ranking systems may be reflective of the cyclic nature of college basketball teams' prowess, due to the graduation of senior players, the loss of players leaving early to join the NBA, and the influx of freshmen who have no experience in the tournament. Whatever the reason, this data suggests that it might be wise to supplement linear models with other methods such as regression analysis of team/game statistics when making predictions for the tournament.

## References

- [1] Colley, W. N. (2002) *Colley's Bias Free College Football Ranking Method*.
- [2] Brin, Sergey and Page, Lawrence (1998) *The Anatomy of a Large Scale Hypertextual Web Search Engine*. Computer Networks and ISDN Systems, 33: 107-17.
- [3] Govan, Anjela Y., Meyer, Carl D. and Albright, Russell (2008) *Generalizing Google's PageRank to Rank National Football League Teams*. SAS Global Forum 2008, Paper 151.
- [4] Keener, James P. (1993) *The Perron-Frobenius Theorem and The Ranking of Football Teams*. SIAM Review, Vol. 35, No. 1 (Mar. 1993), pp 80-93.
- [5] Langville, Amy M. and Meyer, Carl D. (2006) *Google's Page Rank and Beyond: The Science of Search Engine Rankings*. Princeton University Press.
- [6] Langville, Amy M. and Meyer, Carl D. (2012) *Who's # 1?: The Science of Rating and Ranking*. Princeton University Press.
- [7] Massey, Kenneth (1997) *Statistical Models Applied to The Rating of Sports Teams*. Bluefield College.
- [8] Meyer, Carl D. (2005) *Matrix Analysis and Applied Linear Algebra*. SIAM, Philadelphia.
- [9] *The Comprehensive R Archive Network*. <https://cran.r-project.org/>

# Bayesian hierarchical models for predicting individual performance in football (soccer)

L. Egidi\* and J. S. Gabry\*\*

\*Dipartimento di Scienze Statistiche, Università degli Studi di Padova, email address: egidi@stat.unipd.it

\*\* Department of Statistics, Columbia University, New York, email address: jgabry@gmail.com

## Abstract

The task of predicting the performance of football (soccer) players is gaining increasing attention in the sports and statistical communities. We discuss the merits and flaws of a variety of hierarchical Bayesian models for detecting factors relevant to player performance in the presence of noisy data, and we compare the models on their predictive accuracy on hold-out data. We apply our analyses to the 2015–2016 season in the top Italian league, Serie A, and use the player ratings provided by a popular Italian fantasy football game as a motivating example. Our central goals are to explore what can be accomplished with a simple freely available dataset and to focus on a small number of interesting modeling and prediction questions that arise. We validate our models through graphical posterior predictive checks and we provide out-of-sample predictions for the second half of the season, using the first one as training set.

## 1 Introduction

In most of the published statistical research on football — Baio and Blangiardo (2010), Dixon and Coles (1997), Karlis and Ntzoufras (2009) — the authors primarily focus on modeling some aspect of the global result of a match between opposing teams (e.g., goal differential), or on predicting the order of the league table at the end of a season, and rarely on the performance of individual players over the course of a season. One reason for not focusing on predictions at the individual player level is that the performance of individual football players is noisy and hard to predict. The dimensions of the pitch combined with the number of players, the difficulty of controlling the ball without the use of hands, and many other factors all contribute to the predictive challenge. In fact, as far as we can tell from reviewing the current literature, there have been no published attempts to use a hierarchical Bayesian framework to address the challenges of modeling this kind of data.

Nevertheless, we suspect that even in football —in fantasy football at least (Bonomo et al., 2014)— a prediction task for individual performance could be well posed. In this paper we present and critique several Bayesian hierarchical models (Gelman et al., 2013, Gelman and Hill, 2006) designed to predict the results of an Italian fantasy football game with players nested within position and team. All models are estimated via Markov chain Monte Carlo using RStan, the R (R Core Team, 2016) interface to the Stan C++ library (Stan Development Team, 2016a).

The outcome of interest is the fantasy rating of each player in Italy’s top league, Serie A, for each match of the 2015–2016 season. In some sense, we are using these data with a dual purpose: we would like to provide estimates and predictions both for the fantasy game and for the sport itself. That is, we use the fantasy

ratings as both an outcome of interest and also as a (crude) proxy for the quality of a player's performance. Although we take Fantacalcio, an Italian fantasy football product, as our example, the process of developing these models and comparing them on predictive performance does not depend on the idiosyncrasies of this particular fantasy system and is applicable more broadly.

Our central goals are to explore what can be accomplished with a simple freely available dataset (comprising only a few variables) and to focus on a small number of interesting modeling and prediction questions that arise. For this reason we also gloss over many issues that we believe should be of interest in subsequent research, for instance variable selection, additional temporal correlation structures, and the possibility of constructing more informative prior distributions.

The rest of the paper is structured as follows. In Section 2 we briefly introduce the Italian fantasy football game Fantacalcio. We then describe our dataset and present the models we fit in Section 3, where a mixture model (Section 3.3) is explained in detail and the other models derived as consequence. Preliminary results are presented in Section 4, along with a variety of posterior predictive checks as well as out-of-sample prediction tasks. Section 5 concludes.

## 2 Overview of the game

Fantasy sports games typically involve roster selection and match-by-match challenges against other participants with the results determined by the collective performance of the players on the fantasy rosters. In Italy, fantasy football was popularized by the brand Fantacalcio edited by Riccardo Albini in the 1990s (see <http://www.fantacalcio.it> for further details) and in the rest of the paper we use the original denomination for referring at the Italian game.

At the beginning of the season, the virtual managers are allocated a limited amount of virtual money with which to buy the players that will comprise their roster. Each player in the Italian Serie A league has an associated price determined by various factors including past performance and forecasts for the upcoming season. After every match in Serie A, the prominent Italian sports periodicals assign each player a rating, a so-called *raw score*, on a scale from one to ten. In practice there is not much variability in these scores; they typically range from four to eight, with the majority between five and seven. These raw scores are very general and largely subjective performance ratings that do not account for significant individual events (goals, assists, yellow and red cards, etc.) in a consistent way.

As a means of systematically including specific in-game events in the ratings, Fantacalcio provides the so-called *point scoring* system. Points are added or deducted from a player's initial raw score for specific positive or negative events during the match. The point scores are more variable than the raw scores, especially across positions (e.g., when comparing defending and attacking players). Goalkeepers suffer the most from the point scoring system, as they are deducted a point for every goal conceded. On the other extreme, forwards (attacking players) typically receive the highest point scores because every goal scored is worth three points.

For player  $i$  in match  $t$  the total rating  $y_{it}$  is

$$y_{it} = R_{it} + P_{it}, \quad (1)$$

where  $R$  is the raw score and  $P$  is the point score. Table 1 lists the game features that contribute to a player's point score  $P_{it}$  for a given match.

Event	Points
Goal	+3
Assist	+1
Penality saved*	+3
Yellow card	-0.5
Red Card	-1
Goal conceded*	-1
Own Goal	-2
Missed penalty	-3

Table 1: Bonus/Malus points in Fantacalcio. The symbol \* denotes an event only applicable to goalkeepers.

Importantly, there are two general ways we observe an outcome of  $y_{it} = 0$ . First, player  $i$ 's rating for match  $t$  will be zero if the player does not play in the match — because of injury, disqualification, coach's decision, or some other reason — or he does not participate in the match for long enough for their impact to be judged by those tasked with assigning the subjective raw score ( $R_{it} = 0$ ). We will refer to this first type of zero as a *missing* observation because the player did not enter the match. Second, due to the nature of the Fantacalcio scoring system, a player can also receive a score of zero even if he does play in the match. For example, a goalkeeper who receives a raw score of four and concedes four goals will have a score of zero for the match. We will refer to this second type of zero — quite uncommon — as an *observed* zero.

One of the main aims of this paper is the attempt to model the missing values which naturally arise over the season.

## 3 Data and models

### 3.1 Data

All data for this paper are from the 2015–2016 season of the Italian Serie A and were collected from the Italian publication La Gazzetta dello Sport (<http://www.gazzetta.it>). We decided to select those players which participated in at least a third of matches during the *andata* (the first half of the season); this results in a dataset containing ratings for 237 players (18 goalkeepers, 90 defenders, 78 midfielders, and 51 forwards). For illustration purposes of the data at hand, Figure 1 displays the average ratings for the players of our dataset plotted against the initial standardized prices for each player, discussed in Section 2. For a wider overview on the data we used, see <http://www.gazzetta.it/calcio/fantanews/statistiche/serie-a-2015-16/>.

There are  $N = 237$  players and  $T = 38$  matches in the dataset. When fitting our models we use only the  $T_1 = 19$  matches from the first half of the 2015–2016 Serie A season. The remaining matches are used later for predictive checks. The players are grouped into  $J = 4$  positions (forward, midfielder, defender, goalkeeper) and  $K = 5$  team clusters. The five clusters (not listed here) were determined using the official Serie A rankings at the midpoint of the season. The purpose of the team clustering is both to use a grouping structure that has some practical meaning in this context and also to reduce the computational burden somewhat by

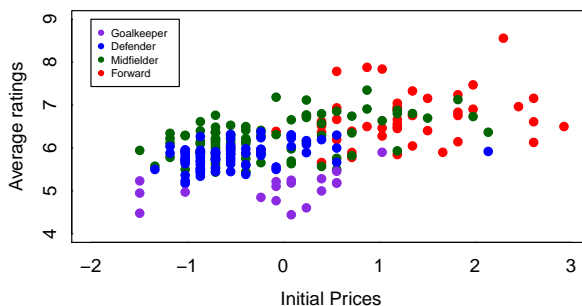


Figure 1: Average ratings plotted against the initial standardized prices for each of the 237 players of the dataset, taking into account the four different positions.

including cluster-specific parameters rather than team-specific parameters.

### 3.2 General framework and notation

The notation we use for data and parameters is similar to the convention adopted by Gelman and Hill (2006) for multilevel models. For match  $t \in \{1, \dots, T\}$ , let  $y_{ijkt}$  denote the value of the total rating for player  $i \in \{1, \dots, N\}$ , with position (role on the team)  $j \in \{1, \dots, J\}$ , on a team in team-cluster  $k \in \{1, \dots, K\}$ . To ease the notational burden, throughout the rest of the paper the subscripts  $j$  and  $k$  will often be implicit and we will use  $y_{it}$  in place of  $y_{ijkt}$ . We denote by  $\mathbf{Z}$  the  $N \times T$  binary matrix in which each element  $z_{it}$  is 1 if player  $i$ 's team plays match  $t$  at its home stadium and 0 otherwise. And let  $q_i$  denote the initial standardized price for player  $i$ . These values are assigned by experts and journalists at the beginning of the season based on their personal judgement and then updated throughout the season to reflect each player's performance.

Let  $\alpha_i$  denote the individual intercept for each player, with  $i = 1, \dots, N$ . We denote with  $\gamma_{k[i]}$  the team-cluster intercept and with  $\beta_{k[i],t}$  the team-cluster of the opponent in match  $t$ , with  $k = 1, \dots, K$ . In our simplified framework we set the number of team-clusters  $K = 5$ .  $\rho_{j[i]}$  is the position intercept, with  $j = 1, \dots, J$  and  $J = 4$ . The standardized prices are multiplied by a coefficient  $\delta_{j[i]}$ , which also varies over the  $J$  positions. Because we are interested in detecting trends in player ratings, we also incorporate the average rating up to the game  $t - 1$ ,  $s_{i,t-1}$ , multiplied by a factor  $\lambda_{j[i]}$  estimated from the data. For the mixture model in Section 3.3, the same average rating  $s_{i,t-1}$  is also multiplied by a coefficient  $\zeta_{j[i]}$  in order to model the probability of participating in the match  $t$ .

For illustration purpose, here we present in detail the mixture model (hereafter, MIX), and we gloss over the technical details for the other two models we fit, which may be conceptually derived from the first one: the hierarchical autoregressive model (HAR), whose estimates are carried out by replacing all the missing values (see Section 2) with some zeros; and the hierarchical autoregressive missing model (HAr-mis), which actually treats the unobserved ratings as modeled parameters — we wrote a simple Stan program implementing the joint model for the observed and missing observations —. It is worth noticing that the MIX and the HAr-mis model are actual attempts for modeling the missingness in our dataset.



### 3.3 Mixture model (MIX)

Even if we found that some players have a tendency to be ejected from matches due to red cards, for instance, or tend to suffer injuries at a high rate, it would still be very challenging to arrive at sufficiently informative probability distributions for these events. Even with detailed player histories over many seasons, it would be hard to predict the number of missing matches in the current season. Nevertheless, we can try to incorporate the *missingness* behavior intrinsic to the game into our models. Assuming that it is very rare for a player to play in every match during a season, we can try to model the overall propensity for missingness. A general way of doing this entails introducing a latent variable, which we denote  $V_{it}$  and define as

$$V_{it} = \begin{cases} 1, & \text{if player } i \text{ participates in match } t, \\ 0, & \text{otherwise.} \end{cases}$$

If for each player  $i$  we let  $\pi_{it} = Pr(V_{it} = 1)$ , then we can specify a mixture of a Gaussian distribution and a point mass at 0 (Gottardo and Raftery, 2008)

$$p(y_{it} | \eta_{it}, \sigma_y^2) = \pi_{it} \text{Normal}(y_{it} | \eta_{it}, \sigma_y^2) + (1 - \pi_{it}) \delta_0, \quad (2)$$

where  $\delta_0$  is the Dirac mass at zero,  $\sigma_y^2$  is the variance of the error in predicting the outcome and  $\eta_{it}$  is the linear predictor

$$\eta_{it} = \alpha_i + \beta_{k[i],t} + \gamma_{k[i]} + \rho_{j[i]} + \delta_{j[i]} q_i + \theta z_{it} + \lambda_{j[i]} s_{i,t-1}. \quad (3)$$

The probability  $\pi_{it}$  is modeled using a logit regression,

$$\pi_{it} = \text{logit}^{-1}(p_0 + \zeta_{j[i]} s_{i,t-1}), \quad (4)$$

which takes into account predictors that are likely to correlate with player participation.  $s_{i,t-1}$  is the average rating for player  $i$  up to match  $t - 1$  and  $p_0$  is the intercept for the logit model.

For the new parameters introduced in (4) we use the weakly informative priors

$$(p_0, \zeta) \stackrel{iid}{\sim} \text{Normal}(0, 5^2).$$

The models for the group-level and individual parameters are

$$\alpha_i \sim \text{Normal}(\mu_\alpha, \sigma_\alpha^2), \quad i = 1, \dots, N \quad (5)$$

$$\gamma_k \sim \text{Normal}(0, \sigma_\gamma^2), \quad k = 1, \dots, K \quad (6)$$

$$\beta_k \sim \text{Normal}(0, \sigma_\beta^2), \quad k = 1, \dots, K \quad (7)$$

$$\rho_j \sim \text{Normal}(\mu_\rho, \sigma_\rho^2), \quad j = 1, \dots, J \quad (8)$$

with weakly informative prior distributions for the remaining parameters and hyperparameters.

In this formulation, the parameters  $\mu_\alpha$  and  $\mu_\rho$  are the prior means of the individual intercepts and of the position-specific intercepts.

The HAR and the HAR-mis models — which differ only concerning how they use and code the missing values — may be easily defined through the distribution  $\text{Normal}(y_{it} | \eta_{it}, \sigma_y^2)$ , with the same  $\eta_{it}$  as in (3).

## 4 Preliminary results, posterior predictive checks and predictions

### 4.1 Results

We fit the models via Markov chain Monte Carlo using RStan, the R interface to the Stan C++ library (Stan Development Team, 2016a), and monitored convergence as recommended in Stan Development Team (2016b). Figure 2 shows the parameter estimates for the HAR, the HAR-mis and the MIX model. At a first glance, the magnitude and the sign of the parameters for the MIX model and the HAR-mis are quite close. According to all the models, the beta's, gamma's and delta's coefficients are almost all shrunk towards their grand mean 0, with a low variability.

As it is evident, the largest source of variation for the three models is represented by the position. For what concerns the lambda's, the estimates obtained through the HAR model are greater than those obtained under the HAR-mis and the MIX model. We recall that, for every  $t$ , these coefficients are multiplied by the lagged average rating  $s_{i,t-1}$ ; then, we strongly believe that the greater HAR values are mainly due to coding the missing values as zeros, instead of modeling as parameters, as for the HAR-mis model. All the models recognize a slight advantage due to playing at home ( $\theta > 0$ ).

### 4.2 Posterior predictive checks

Now that we have estimated all of the models, we turn our attention to evaluating the fit of the models to the observed data. We use the 19 match days comprising the first half of the Serie A season — the *andata* — as training data, and for every player we make in-sample predictions for those 19 matches.

Figure 3 shows an example of a graphical posterior predictive check focusing on the *cumulative* ratings for each player over the matches in the training data. For illustration purposes, here we only show the results for one team, Napoli: the dashed black lines represent the observed values, while the red, green, blue lines represent predictions from the HAR, MIX and HAR-mis models, respectively. HAR and MIX models make predictions quite close to the observed values for many of the players. In correspondence of players with a non-trivial amount of missing (here zero) values, these models result to be preferable to the HAR-mis (see the plots for El Kaddouri, for instance).

We are also interested in the calibration of the model. In Figure 4 we display the median predictions and 50% posterior predictive intervals under the MIX for our selected team Napoli, overlaying the observed data points. In a well-calibrated model we expect half of the observed values to lie outside the corresponding 50% intervals. By this measure the MIX model has decent but not excellent calibration, since for most of the players — especially for the goalkeeper and the defenders— the 50% intervals cover more than 50% of the observed (blue) points. Conversely, for the volatile superstar Higuaín (an outlier even among forwards) a few points fall inside the intervals.

### 4.3 Out of sample predictions

As usual in a Bayesian framework, the prediction for a new dataset may be directly performed via the posterior predictive distribution for our unknown set of observable values. Following the same notation of Gelman et al. (2013), let us denote with  $\tilde{y}$  a generic unknown observable. Its distribution is then conditional on the observed  $y$ ,

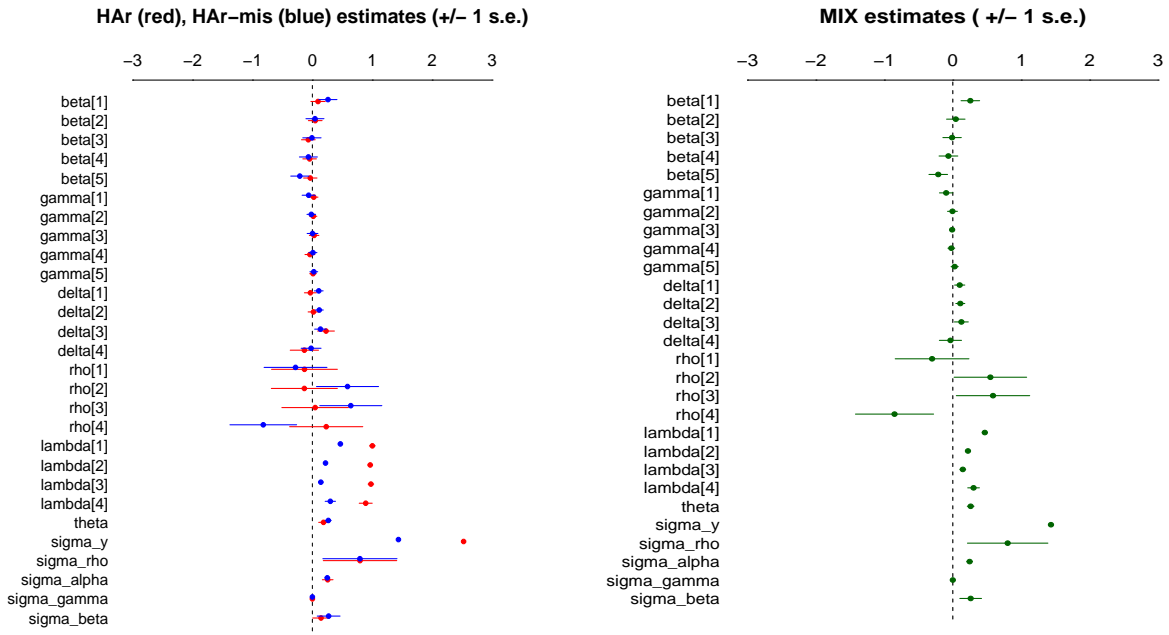


Figure 2: Posterior summary statistics for the HAR, HAR-mis and MIX model.  $\beta_k$ ,  $k = 1, \dots, 5$  are the coefficients for the clusters opponent team (5=good, 4 = quite good, 3= medium, 2=low, 1=very low);  $\gamma_k$ ,  $k = 1, \dots, 5$  are the coefficients for the clusters own team, same classification as before;  $\delta_j$ ,  $j = 1, \dots, J$  are the coefficients for the initial prices of the players;  $\lambda_j$ ,  $j = 1, \dots, J$  are the coefficients of the lagged observed average rating;  $\rho_j$ ,  $j = 1, \dots, J$  are the positions parameters (1 = Forward, 2=Midfield, 3=Defender, 4=Goalkeeper);  $\theta$  is the coefficient for the home/away predictor;  $\sigma_y$  is the individual standard deviation;  $\sigma_\alpha$  is the standard deviation for the individual intercepts  $\alpha_i$ ,  $i = 1, \dots, N$ ;  $\sigma_\rho$  is the position's parameters standard deviation;  $\sigma_\gamma$  is the clusters own teams standard deviation;  $\sigma_\beta$  is the clusters opponent teams standard deviation. The further set of parameters for the MIX model, represented by  $\zeta_j$ ,  $j = 1, \dots, J$  and  $p_0$ , is not shown here.

$$p(\tilde{y}|y) = \int_{\Theta} p(\tilde{y}, \theta|y) d\theta = \int_{\Theta} p(\theta|y) p(\tilde{y}|\theta) d\theta$$

where the conditional independence of  $y$  and  $\tilde{y}$  given  $\theta$  is assumed. We fit the models over the  $T = 19$  matches in the first half of the season and then generate predictions for the  $T^* = 19$  matches in the second half of the season.

Based on average predicted ratings for the held-out data from the second half of the 2015–2016 Serie A season, Figure 5 displays the best teams of eleven players that can be assembled from the available players according to each of the models. Also shown is the best team assembled using the observed ratings from the same set of matches. As is evident at a first glance, the predictions obtained through the HAR model are quite inefficient: this model tends to overestimate the players' rating, which are quite far from the observed ratings of the second part of the season. The team created based on the predictions from the HAR-mis and

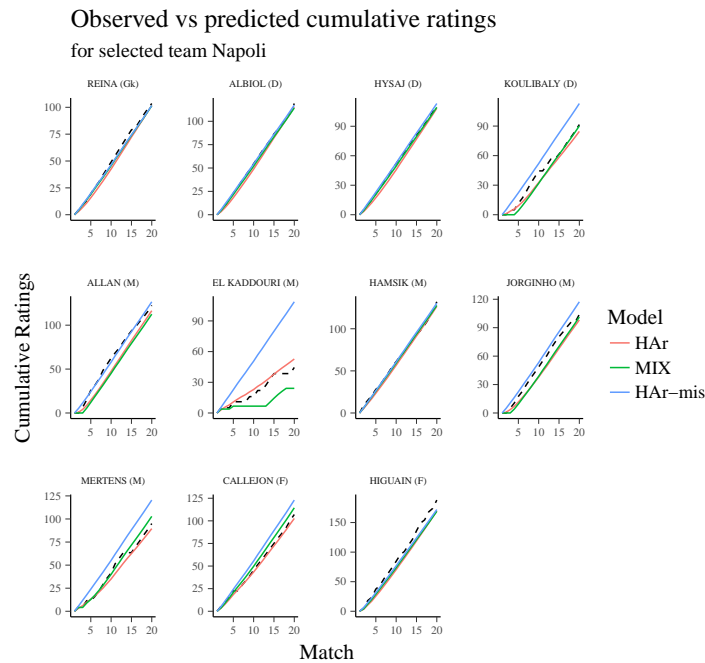


Figure 3: Posterior predictive validation of the HAR model against MIX and HAR-mis models for selected team Napoli, throughout the first half of the 2015–2016 Serie A season. The dashed black line represents the observed cumulative ratings, while the red, green, and blue lines show the medians of the predictions from the HAR, MIX and HAR-mis models, respectively.

the MIX model include four of the eleven players (Acerbi, Pogba, Hamsik, Higuaín) from the team based on the actual ratings. Dybala, who is the third best forward according to these models, is also rated highly (fifth best forward) according the observed ratings. And Rudiger, the second best defender according to the models, is also rated highly (eighth best defender).

Informally, the teams selected by the MIX and the HAR-mis models appear to be quite competitive: from this section, it is evident that modeling the missingness allows to obtain better predictions.

## 5 Discussion

The recent successes of so-called football (soccer) analytics are due in large part to the increasing number of available metrics for analyzing and describing the game. According to our current knowledge, the only attempt to using these and many other metrics for measuring player performance is the OPTA index. Compared to attempts like the OPTA index, our ratings may seem like very crude approximations to player performance—and they are—since they gloss over many games events. But the formulation of an index based on as many variables as possible has not been the aim of this paper. The attractiveness of our general approach is that it is based on a coherent statistical framework: we have an outcome variable  $y$  (the player rating) that is actually available, probability models relating the outcome to predictors, the ability to add

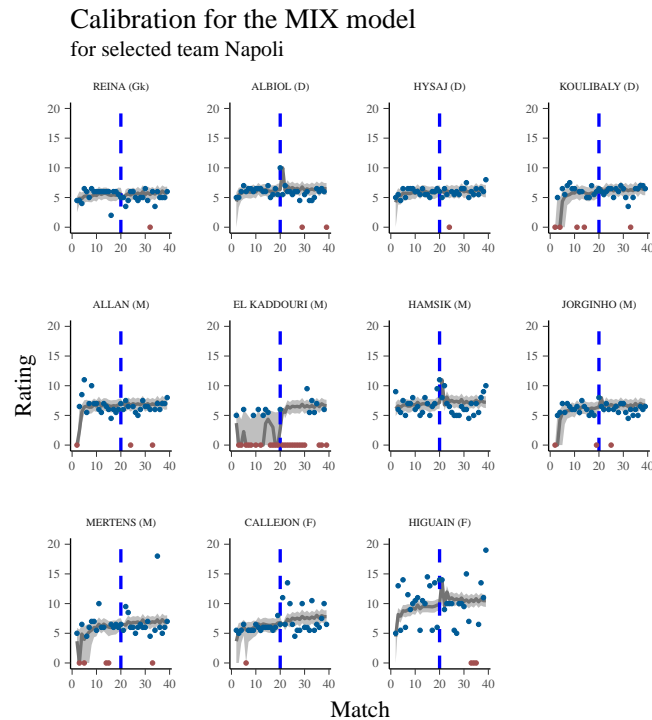


Figure 4: Calibration check for the MIX model for selected team Napoli . Blue points are observed values, red points are the zeros (missing values). The light gray ribbons represent 50% posterior predictive intervals and the dark gray lines are the median predictions. The dashed vertical blue line delimits the in-sample predictions from the out-of sample predictions.

prior information into an analysis in a principled way, and the ability to propagate our uncertainty into the predictions by drawing from the posterior predictive distribution.

We proposed some hierarchical models for predicting player ratings, taking care of the missingness as a part of the models. As expected, we preliminarily found that a player's position is, in most cases, an important factor for predicting performance (as measured by the Fantacalcio ratings). However, it is somewhat counterintuitive that the inferences from these models suggest that the quality of a player's team and the opposing team and the initial price of the players do not account for much of the variation in player ratings. It is also notable that the association between the current and lagged performance ratings—expressed by the average lagged rating—is slightly different from zero after accounting for the other inputs into the models. Future research should consider whether other functional forms for describing associations over time are more appropriate, to what extent the inclusion of other variables in the models could improve the predictive performance, and if more informative priors can be developed at the position and team levels of the models. Another future issue should concern the choice of the training and the test set: for simplicity, in this paper we considered only the first part of the season as training set and the second one as test set; however, we strongly believe that our models may be used in a dynamic way, using data at match day  $t$  for predicting the players' performances at match day  $t + 1$ .

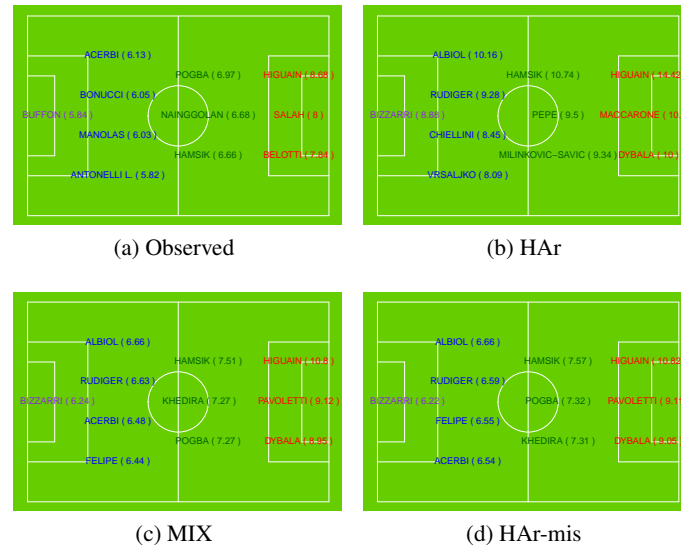


Figure 5: Best teams according to out-of-sample prediction of average player ratings for the HAr, MIX and HAr-mis model compared to the observed best team for the second part of the season. The averaged ratings are computed for those players who played at least 15 matches in the second half of the season.

## References

- Baio, G. and Blangiardo, M. (2010), ‘Bayesian hierarchical model for the prediction of football results’, *Journal of Applied Statistics* **37**(2), 253–264.
- Bonomo, F., Durán, G. and Marengo, J. (2014), ‘Mathematical programming as a tool for virtual soccer coaches: a case study of a fantasy sport game’, *International Transactions in Operational Research* **21**(3), 399–414.
- Dixon, M. J. and Coles, S. G. (1997), ‘Modelling association football scores and inefficiencies in the football betting market’, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **46**(2), 265–280.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B. and andDonald B. Rubin, A. V. (2013), *Bayesian Data Analysis*, third edn, Chapman & Hall/CRC.
- Gelman, A. and Hill, J. (2006), *Data analysis using regression and multilevel/hierarchical models*, Cambridge University Press.
- Gottardo, R. and Raftery, A. E. (2008), ‘Markov chain monte carlo with mixtures of mutually singular distributions’, *Journal of Computational and Graphical Statistics* **17**(4), 949–975.
- Karlis, D. and Ntzoufras, I. (2009), ‘Bayesian modelling of football outcomes: using the skellam’s distribution for the goal difference’, *IMA Journal of Management Mathematics* **20**(2), 133–145.
- R Core Team (2016), ‘R: A language and environment for statistical computing’.  
**URL:** <https://www.R-project.org/>
- Stan Development Team (2016a), ‘The Stan C++ library, version 2.14.0’.  
**URL:** <http://mc-stan.org>
- Stan Development Team (2016b), *Stan Modeling Language User’s Guide and Reference Manual, Version 2.14.0*. <http://mc-stan.org/>.

# How long does a tennis game last?

M. Ferrante\*, G. Fonseca\*\* and S. Pontarollo\*

\*Dip. di Matematica "Tullio Levi-Civita", Università di Padova, Via Trieste 63, 35121-Padova, Italy  
email: ferrante@math.unipd.it, spontaro@math.unipd.it

\*\* Dip. di Sc. Econom. e Stat., Univ. di Udine, via Tomadini, 30/A, 33100-Udine, Italy  
email: giovanni.fonseca@uniud.it

## Abstract

In this paper we present a generalisation of previously considered Markovian models for Tennis that overcome the assumption that the points played are i.i.d and includes the time into the model. Firstly we postulate that in any game there are two different situations: the first 6 points and the, possible, additional points after the first deuce, with different winning probabilities. Then we assume that the duration of any point is distributed with an exponential random time. We are able to compute the law of the (random) duration of a game in this more general setting.

## 1 Introduction

Markovian framework is particularly suitable to describe the evolution of a tennis match. The usual assumption is that the probability that a player wins one point is independent of the previous points and constant during the match. Under these hypotheses the score of a game, set and match can be described by a set of nested homogeneous Markov chains. Hence, theoretical results concerning winning probabilities and mean duration of a game, set and match can be easily obtained. A complete account on this approach can be found in Klaassen and Magnus (2014).

Anyway, some authors criticise the assumption that the point winning probability is constant along the match and independent of the previous points played, see e.g. Klaassen and Magnus (2001). In particular, it is pointed out that playing decisive points, i.e. points after a deuce score, modifies players attitude and this reflects heavily on the probability to win these points.

In Carrari *et al.* (2017), we propose a modification of the model at the game's level. Indeed, we assume that during any game there are two different situations: the first points and the, possible, additional points played after the (30,30) score that in our model coincide with the "Deuce". Under this hypothesis, following the approach used in Ferrante and Fonseca (2014), we computed the winning probabilities and the expected number of points played in a game.

In the present work we include in the model the time needed to play a single point. The aim of such a modification is to obtain the computation of the expected length of a match in terms of actual time and not just as number of points. Indeed there is a concern about the length of tennis matches and several modification to the game rules are nowadays proposed in order to fix the length of a match or at least to avoid too long matches. In Section 2. we present the model, in Section 3. we compute the game winning probabilities and in Section 4. we obtain the expected length of a game.

How long does a tennis game last?

M. Ferrante, G. Fonseca, S. Pontarollo

## 2 The continuous time model

In this paper we model the tennis game as a continuous-time Markov chain (see Norris (1998) for a complete account on this topic). We define the state space  $S$  of the chain, which collects all the possible scores in the game, and the generator matrix  $Q$  on  $S$ . In order to determine the matrix  $Q$ , we define independently the transition matrix of the associated Jump chain, which is a discrete-time Markov chain, and the exponential holding times.

The transition matrix of the associated Jump chain follows the model defined in Carrari *et al.* (2017) for a discrete-time Markov chain of the tennis. The classical assumptions previously considered in the literature (see e.g Newton and Keller (2005)) were that the probability to win any point by the player on service was independent of the previous points and constant during the game. In Carrari *et al.* (2017) we assume that  $p$ , the probability to win a point, does not remain the same during the game. As empirical data on the matches confirm, the estimated winning probability of the first 6 points of a game is different from that of the additional played points from the “Deuce” on. For this reason, we consider a second parameter  $\bar{p}$ , that describe this part of the game and, to avoid trivial cases, we assume that both  $p$  and  $\bar{p}$  belong to  $(0, 1)$ .

Regarding the holding times, it is not easy to find in the literature data sets to test different scenarios that better describe the true course of a tennis game (see Morante and Brotherhood (2007) for some statistics on the duration). For this reason in this paper we assume that all the holding times have the same distribution, i.e. share the same rate  $\lambda$  of their Exponential Laws.

Let us now define precisely our model: the state space is the set  $S = \{1, 2, \dots, 17\}$  which describes the score of a game as defined in Table 1.

Score	(0,0)	(15,0)	(0,15)	(30,0)	(15,15)	(0,30)	(40,0)	(30,15)	(15,30)
State	1	2	3	4	5	6	7	8	9
Score	(0,40)	(40,15)	(15,40)	<i>Deuce</i>	<i>Adv<sub>A</sub></i>	<i>Adv<sub>B</sub></i>	<i>Win<sub>A</sub></i>	<i>Win<sub>B</sub></i>	
State	10	11	12	13	14	15	16	17	

Table 1: Scores and corresponding states used in equations

Note that in the present model the scores  $(30, 30)$  and *Deuce* are represented by the single state 13, since they share the same mathematical properties, as it happens to the pairs  $(40, 30)$ –*Adv<sub>A</sub>* and  $(30, 40)$ –*Adv<sub>B</sub>*. The graph representing the transition probabilities of the Jump process is presented in Fig. 1, where  $q = 1 - p$  and  $\bar{q} = 1 - \bar{p}$ .

By our construction, we define the transition rates in the generator matrix  $Q$  by  $\lambda_1 = \lambda p$  and  $\lambda_2 = \lambda q$  (see Norris (1998)) and in Fig. 2 we present the graph of the continuous time Markov chain describing a tennis game. Note that the expected length of a point is equal to  $1/\lambda$  and it does not depend on who is the winner of the point.

From the graph it is immediate to write down the generator matrix  $Q = (q_{ij})_{i,j \in S}$  and to prove that the states



How long does a tennis game last?

M. Ferrante, G. Fonseca, S. Pontarollo

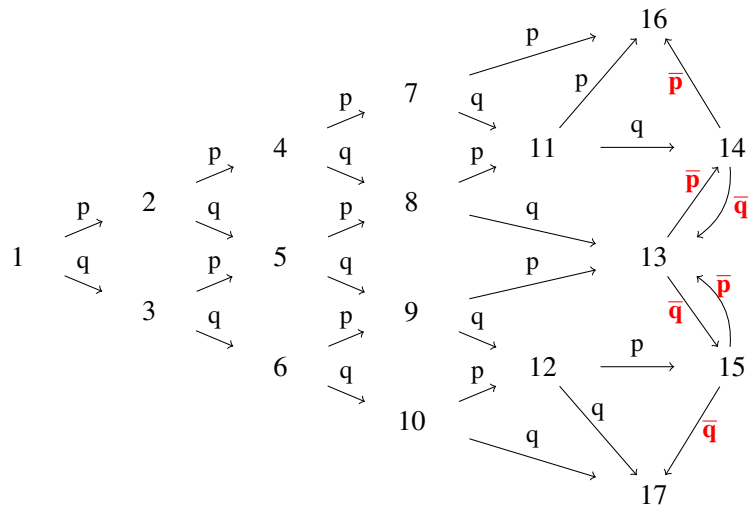


Figure 1: Graph of the Jump Markov chain

16 and 17 are absorbing, while all the other states are transient.

$$Q = \begin{bmatrix} -\lambda & \lambda_1 & \lambda_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\lambda & 0 & \lambda_1 & \lambda_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\lambda & 0 & \lambda_1 & \lambda_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\lambda & 0 & 0 & \lambda_1 & \lambda_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\lambda & 0 & 0 & \lambda_1 & \lambda_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\lambda & 0 & 0 & \lambda_1 & \lambda_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\lambda & 0 & 0 & \lambda_2 & 0 & 0 & 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\lambda & 0 & \lambda_1 & 0 & \lambda_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\lambda & 0 & \lambda_1 & 0 & 0 & 0 & 0 & \lambda_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\lambda & 0 & \lambda_2 & 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\lambda & 0 & 0 & \lambda_1 & 0 & \lambda_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\bar{\lambda} & \bar{\lambda}_1 & \bar{\lambda}_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \bar{\lambda}_2 & -\bar{\lambda} & 0 & \bar{\lambda}_1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \bar{\lambda}_1 & 0 & -\bar{\lambda} & 0 & \bar{\lambda}_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

In order to compute the winning probabilities, we need to determine the absorption probabilities in states 16 and 17 for the transition matrix of the Jump process, while to investigate the distribution of the expected length of a game, we need to evaluate the exponential matrix of  $Q$ , which is in general not very simple.

How long does a tennis game last?

M. Ferrante, G. Fonseca, S. Pontarollo

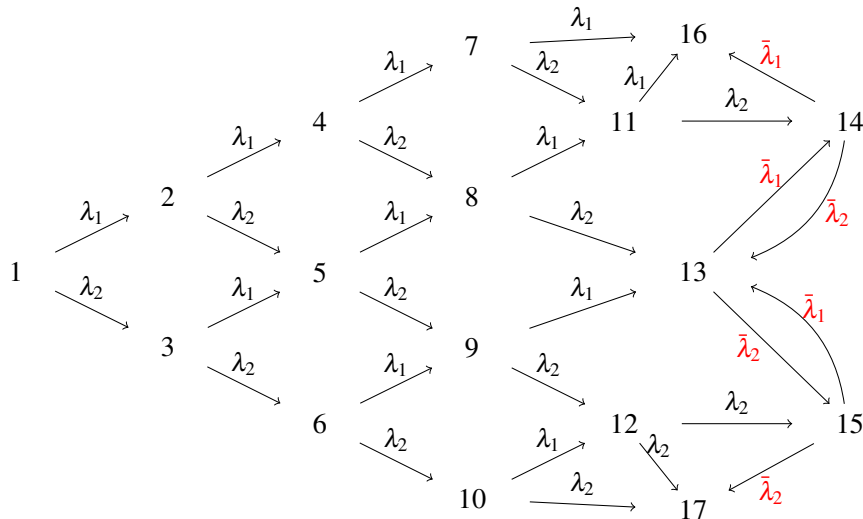


Figure 2: Graph of the continuous time Markov chain describing a tennis game, with its transition rates.

### 3 Winning probabilities

In this section we recall some of the result proved in Carrari *et al.* (2017). The transition matrix of the Jump process is

$$P = \begin{bmatrix} 0 & p & q & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & p & q & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & p & q & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & p & q & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p & q & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q & 0 & 0 & 0 & 0 & p & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p & 0 & q & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q & p & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p & 0 & 0 & 0 & 0 & q \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & q & 0 & p & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & p & 0 & q \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \bar{p} & \bar{q} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \bar{q} & 0 & 0 & \bar{p} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \bar{p} & 0 & 0 & 0 & \bar{q} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The winning probability of the game for the player A on service, denoted by  $h_1$ , coincides with the absorption probability in the state 16 of the previous Markov chain starting from 1, which can be obtained (see e.g.

How long does a tennis game last?

M. Ferrante, G. Fonseca, S. Pontarollo

Norris (1998)) as the minimal, non negative solution of the linear system

$$h_i = \sum_{j \in S} p_{ij} h_j \quad \text{for } 1 \leq i \leq 15, h_{16} = 1, h_{17} = 0.$$

The solution can be easily calculated and we obtain that

$$h_1 = p^2 \left[ 5p^2 - 4p^3 + 4(p-1)^2 p \bar{p} - \frac{2(p-1)^2 \bar{p}^2 (p(4\bar{p}-2) - 2\bar{p}-3)}{2\bar{p}^2 - 2\bar{p} + 1} \right].$$

Denoting by  $G(p, \bar{p}) = h_1$ , by  $A$  and  $B$  the two players, and by  $P_{XY}^G$  the probability that the player  $Y$  wins a game when  $X$  serves, thanks to the symmetry of the model we obtain that:

$$\begin{aligned} P_{aA}^G &= G(p_A, \bar{p}_A) \\ P_{aB}^G &= G(1 - p_A, 1 - \bar{p}_A) \\ P_{bB}^G &= G(p_B, \bar{p}_B) \\ P_{bA}^G &= G(1 - p_B, 1 - \bar{p}_B) \end{aligned} \tag{1}$$

Note that, since  $P_{xX}^G + P_{xY}^G = 1$ ,  $G(1 - p_X, 1 - \bar{p}_X) = 1 - G(p_X, \bar{p}_X)$  and that for  $p_X = \bar{p}_X$ , the previous probabilities coincides with those well known in the literature (see e.g. Newton and Keller (2005)). In Table 2 we report the values of  $G$  for increasing  $p$  and  $\bar{p}$ .

	$\bar{p}$									
$p$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.1	0.001	0.004	0.011	0.021	0.034	0.049	0.062	0.071	0.078	0.081
0.2	0.011	0.022	0.043	0.076	0.119	0.165	0.204	0.233	0.252	0.263
0.3	0.040	0.061	0.099	0.158	0.234	0.312	0.378	0.425	0.455	0.472
0.4	0.102	0.132	0.185	0.264	0.363	0.464	0.549	0.607	0.643	0.663
0.5	0.206	0.242	0.302	0.391	0.500	0.609	0.697	0.758	0.794	0.812
0.6	0.357	0.392	0.451	0.535	0.636	0.736	0.815	0.868	0.898	0.913
0.7	0.545	0.575	0.622	0.688	0.766	0.842	0.901	0.939	0.960	0.969
0.8	0.748	0.767	0.795	0.835	0.881	0.924	0.957	0.978	0.989	0.993
0.9	0.922	0.929	0.938	0.951	0.965	0.979	0.989	0.995	0.998	0.999
1.0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 2: Winning probabilities of a game

## 4 Duration of a game

In this section we evaluate the duration of a game and of a game given that the player on serve wins the game. By the Markovian structure of the present model, the problem we face is equivalent to evaluate the distribution of the absorption time to one of the states 16 and 17. These distributions are called in the literature Phase-type distributions (see e.g. Neuts (1981)) and we have explicit formulas for their densities and moments. The only drawback is that the computation of the density passes through the evaluation of the matrix exponential of a  $16 \times 16$  matrix, which is usually not feasible. On the contrary, for the moments we only need to be able to evaluate the inverse of the same matrix and its powers.

How long does a tennis game last?

M. Ferrante, G. Fonseca, S. Pontarollo

### 4.1 The unconditioned case

Due to the difficulties described above, in this section we consider only the case where  $\bar{p} = p$ . Starting from  $Q$ , the generator matrix defined above, we can compute the distribution of the duration time of a game. Indeed, let  $T$  be equal to the matrix  $Q$  where the last two rows and two columns have been erased, that is

$$T = \begin{bmatrix} -\lambda & \lambda_1 & \lambda_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\lambda & 0 & \lambda_1 & \lambda_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\lambda & 0 & \lambda_1 & \lambda_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\lambda & 0 & 0 & \lambda_1 & \lambda_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\lambda & 0 & 0 & \lambda_1 & \lambda_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -\lambda & 0 & 0 & \lambda_1 & \lambda_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\lambda & 0 & 0 & 0 & \lambda_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\lambda & 0 & 0 & \lambda_1 & 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\lambda & 0 & 0 & \lambda_2 & \lambda_1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\lambda & 0 & \lambda_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\lambda & 0 & 0 & \lambda_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\lambda & 0 & 0 & \lambda_1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\lambda & \lambda_1 & \lambda_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_2 & -\lambda & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_1 & 0 & -\lambda \end{bmatrix}$$

and let  $\mathbf{T}^0$  be a vector of length 15 where each entries is equal to the jumping time needed for each state (excluding  $Win_A$  and  $Win_B$ ) to reach one of the two absorbing states, that is

$$\mathbf{T}^0 = (0, 0, 0, 0, 0, 0, \lambda_1, 0, 0, \lambda_2, \lambda_1, \lambda_2, 0, \lambda_1, \lambda_2)^\top .$$

Then, if  $\alpha = (1, 0, \dots, 0)$ , the density function of the duration time, denoted by  $f_{p,\lambda}$ , can be computed as

$$f_{p,\lambda} = \alpha e^{Tt} \mathbf{T}^0$$

(see Neuts (1981) for the simple proof). Therefore, if we set

$$A_{p,\lambda}(t) = -3\lambda^2 p^4 t^2 (\lambda t - 4) + 6\lambda^2 p^3 t^2 (\lambda t - 4) - p^2 (4\lambda^3 t^3 - 21\lambda^2 t^2 + 30) + p (\lambda^3 t^3 - 9\lambda^2 t^2 + 30) - 15 + \lambda^2 t^2$$

and

$$B_{p,\lambda}(t) = 15\sqrt{2} (2p^2 - 2p + 1) \left( e^{2\lambda t \sqrt{2(1-p)p}} - 1 \right) ,$$

the time distribution is given by

$$f_{p,\lambda}(t) = \frac{\lambda}{24\sqrt{(1-p)p}} e^{-\lambda t (1 + \sqrt{2(1-p)p})} \left( B_{p,\lambda}(t) + \left( 4\lambda t \sqrt{(1-p)p} e^{\lambda t \sqrt{2(1-p)p}} \right) A_{p,\lambda}(t) \right) .$$

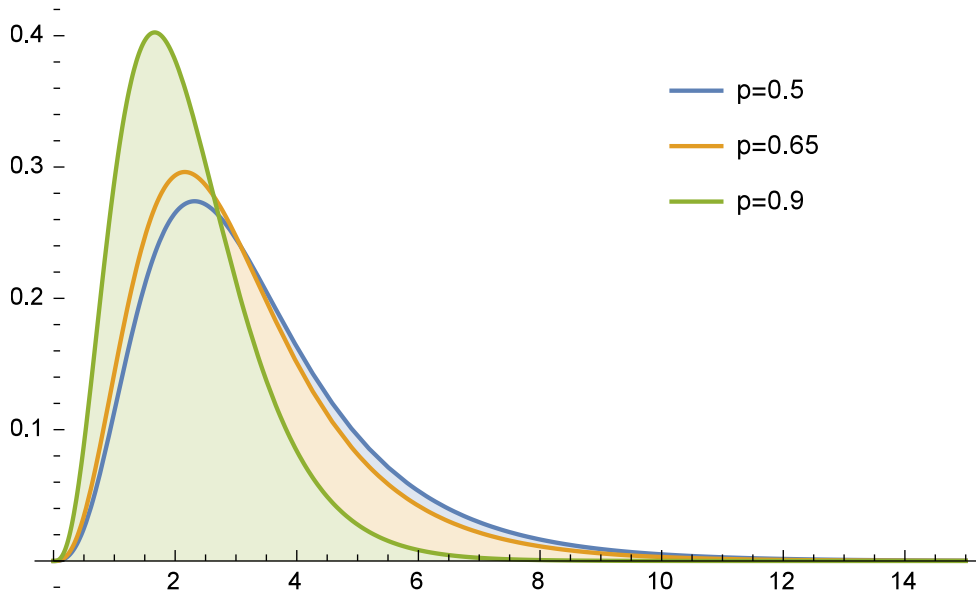
In Figure 3. we plot the density  $f_{p,\lambda}$  in the cases  $p = 0.5, 0.65$  and  $0.9$ , while  $\lambda = 2$ .

We are also able to evaluate the moments of these random times. Indeed, the expected time  $\mu$  needed to win a game can be computed as  $\alpha(-T)^{-1}\mathbf{e}$ , where  $\mathbf{e} = (1, \dots, 1)^T$ . Therefore, the average time is given by

$$\mu = \frac{4(-6p^6 + 18p^5 - 18p^4 + 6p^3 + p^2 - p + 1)}{\lambda(2p^2 - 2p + 1)} .$$

How long does a tennis game last?

M. Ferrante, G. Fonseca, S. Pontarollo


 Figure 3: Graph of the distribution of a game duration time for three different values of  $p$ , and for  $\lambda = 2$ .

Finally the moment of order two, computed as  $\mu^2 = 2\alpha(-T)^{-2}\mathbf{e}$ , allows us to obtain the variance

$$\sigma^2 = \frac{4}{\lambda^2(2p^2 - 2p + 1)^2} (-144p^{12} + 864p^{11} - 2160p^{10} + 2880p^9 - 2232p^8 + 1152p^7 - 618p^6 + 414p^5 - 197p^4 + 40p^3 + 3p^2 - 2p + 1) .$$

## 4.2 The conditioned case

Let us now compute the expected duration of a game given that the player on serve wins the game. In this case it is easy to prove that the Jump chain of the conditioned chain is the matrix  $P'$  on the state space  $\{1, \dots, 16\}$  given by:

$$p'_{ij} = p_{ij} \frac{h_j}{h_i} \quad \text{with } i, j \in \{1, \dots, 16\},$$

where the  $h_i$  are the absorption probabilities in 16. Moreover, the conditional generator matrix is the matrix  $Q'$  obtained from  $P'$  and the original exponential holding times of parameter  $\lambda$ . The mean and variance computed above are now

$$\mu = \frac{4(20p^5 - 84p^4 + 148p^3 - 143p^2 + 79p - 21)}{\lambda(2p^2 - 2p + 1)(8p^3 - 28p^2 + 34p - 15)},$$

$$\sigma^2 = \frac{4}{\lambda^2(2p^2 - 2p + 1)^2(-8p^3 + 28p^2 - 34p + 15)^2} (320p^{10} - 2880p^9 + 11616p^8 - 27744p^7 + 43608p^6 - 47460p^5 + 36746p^4 - 20540p^3 + 8287p^2 - 2288p + 336) .$$

How long does a tennis game last?

M. Ferrante, G. Fonseca, S. Pontarollo

In Table 3 we show the  $Mean \pm StandardDeviation$  of the actual time of a game when the point winning probabilities for the serving player are  $p$  and  $\bar{p} = p$  (the rate  $\lambda$  is set arbitrarily equal to 2 for ease of exposition). For different values of  $p$  (recall that we have set  $\bar{p} = p$ ) the first row shows the ( $mean \pm StandardDeviation$ ) time distribution for finishing a game when the player with probability of winning a point equal to  $p$  is serving. The second row shows ( $mean \pm StandardDeviation$ ) when the distribution is conditioned to the winning of the player that is serving. Note that, due to symmetry of the problem, when  $p < 0.5$  and the model is conditioned as above, the results represent both the average time for the player (characterized by  $p < 0.5$ ) to win a game on his turn of serving, and the average time to win a game for the player with same  $p$  when the other player is serving. The same empirical quantities are derived in Morante and Brotherhood (2007). This is one of the few quantitative studies about the duration of points and games in Tennis, although they are more interested in relating playing time with performance indicators. They compute the average duration using a sample of Grand Slam matches for both male and female professional players.

<b>p = 0.9</b>	<b>p = 0.8</b>	<b>p = 0.7</b>	<b>p = 0.6</b>	<b>p = 0.5</b>	<b>p = 0.4</b>	<b>p = 0.3</b>	<b>p = 0.2</b>	<b>p = 0.1</b>
2.23±1.13	2.54±1.34	2.92±1.60	3.24±1.82	3.37±1.90	3.24±1.82	2.92±1.60	2.54±1.34	2.23±1.13
2.23±1.13	2.53±1.33	2.87±1.58	3.18±1.80	3.37±1.90	3.40±1.86	3.30±1.71	3.13±1.52	2.96±1.35

Table 3: ( $mean \pm StandardDeviation$ ) time for finishing a game (first row) and for winning a game while serving.

## 5 Conclusions

In this paper we present a model for Tennis including non i.i.d. point winning probabilities and the time of play. Indeed, we allow winning point probabilities to change depending on the score of the game. Moreover, we are interested in describing the time of play since there is some concerns about the excessive length of matches, especially in male Grand Slam competitions. In particular, in the present work, we obtain the distribution of the actual time of a game.

## References

- [1] Carrari, A., Ferrante M. and Fonseca G. (2017) *A new Markovian model for tennis matches*. Electronic Journal of Applied Statistical Analysis, to appear.
- [2] Ferrante, M. and Fonseca G. (2014) *On the winning probabilities and mean durations of volleyball*. Journal of Quantitative Analysis in Sports **10**, 91-98.
- [3] Klaassen, F. and Magnus, J.R. (2001) *Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model*. Journal of the American Statistical Association **96**, 500-509.
- [4] Klaassen, F. and Magnus, J.R. (2014) *Analyzing Wimbledon*. Oxford University Press, Oxford.
- [5] Morante, S. and Brotherhood, J. (2007) *Match Characteristics of Professional Singles Tennis*. [http://www.cptennis.com.au/pdf/CooperParkTennisPDF\\_MatchCharacteristics.pdf](http://www.cptennis.com.au/pdf/CooperParkTennisPDF_MatchCharacteristics.pdf).
- [6] Neuts, M.F. (1981) *Matrix-Geometric Solutions in Stochastic Models: An algorithmic approach*. Johns Hopkins University Press, Baltimore.

How long does a tennis game last?

M. Ferrante, G. Fonseca, S. Pontarollo

- [7] Newton, P.K. and Keller, J.B. (2005) *Probability of winning at tennis (I). Theory and data*. Studies in Applied Mathematics **114**, 241-269.
- [8] Norris, J.R. (1998) *Markov chains*. Cambridge University Press, Cambridge.

# Optimal Shot Selection Strategies for the NBA

M. Fichman\* and J. O'Brien\*\*

\*Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA 15213-3815 (e-mail: mf4f@andrew.cmu.edu).

\*\* Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA 15213-3815 (e-mail: jo0x@andrew.cmu.edu).

## Abstract

In this paper we conduct an equilibrium analysis of the 2015/16 NBA season culminating in the dramatic final series upset produced by the Cleveland Cavaliers (CLE) over the Golden State Warriors (GSW). The constant mixed equilibrium strategy for shot selection is constructed for each pair of NBA teams. This strategy jointly optimizes offense and defense defined over a mutually exclusive set of eleven shot location zones. This strategy prescribes what proportion of attempted shots is taken from each of the eleven shot zones. Aggregate results predict a higher proportion of 3-point (0.379) than actual 2015/16 season and playoff proportions (0.286 and 0.309) respectively, suggesting that NBA 3-point averages are still increasing. At the individual team level, the results highlight the importance of a team's defensive strengths because variation in predicted strategies is contingent upon the opposing team. In the final playoff series GSW started close to their predicted optimal strategy and then almost monotonically shifted further away whereas CLE drifted closer to their predicted optimal strategy. Final outcomes were consistent with these equilibrium predictions.

## 1 Introduction

Recently a high school team grabbed headlines in the Wall Street Journal:

“The NBA's most efficient offenses seek out layups and threes. A high school in Minnesota takes the idea to the extreme.” (Cohen, (2017))

Although the trends in high school basketball have significant implications for the future of basketball, similar strategic trends have been observed in NBA basketball since the introduction of the 3-point shot in 1979. The 2015-16 NBA season was notable for the fact that the Golden State Warriors (GSW) produced a record breaking 89% regular season win percentage only to lose a dramatic 7-game final series against the Cleveland Cavaliers (CLE). Understanding the underlying strategies may have profound implications for the future of basketball coaching. The objective of this paper is to conduct an equilibrium analysis of the 2015-16 season by constructing the optimal offensive/defensive constant mixed strategy defined over a set of eleven mutually exclusive shot location zones. A strategy is defined as the set of relative proportions of

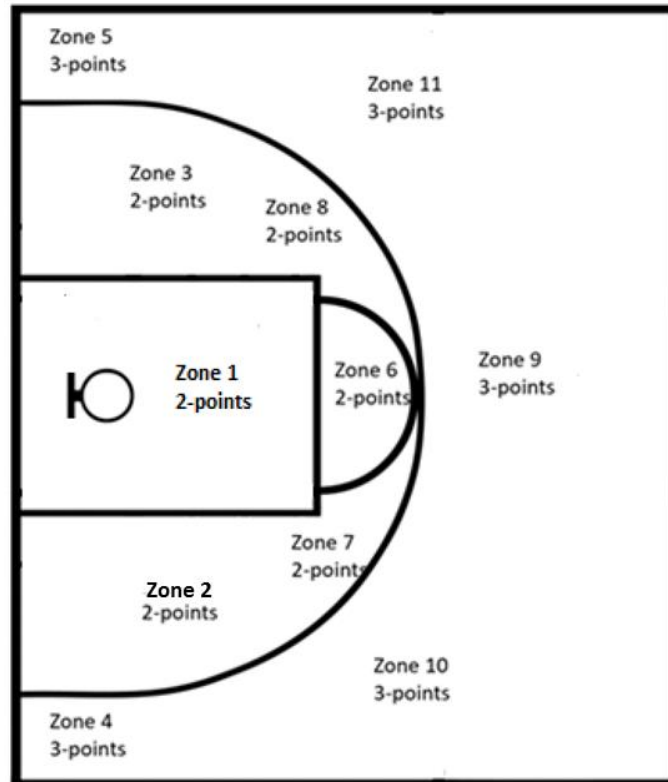


shots taken in a game from each shot location zone. For example, dropping mid-range shots and only taking layups and 3-pointers is an example of a constant mixed strategy where the mid-range zones' relative proportions are zero. The results of this analysis are then used to predict post season results and analyze the final 7-game series.

This paper is organized as follows. In section 2 we develop the theory behind constructing the efficient constant mixed strategies. In this paper we solve for the optimal constant mixed equilibrium strategy for each pair of teams assuming each team wants to win their matches. In section 3 our 2015/16 season database is introduced and used to conduct the equilibrium analysis. In section 4 the results from our analysis are summarized and applied to predicting post season playoff game outcomes. Finally, in section 5 the individual team results are used to interpret the results from the championship series between CLE and GSW followed by a short discussion and conclusion.

## **2 Efficient Constant Mixed Basketball Strategies**

The game of basketball has two simple objectives for any team. Each team wants to score points (offense) and prevent their opponent from scoring points (defense). The team that scores the most points wins the game. Strategically, the court can be viewed as a set of shot locations, which vary in terms of their risk and expected points from shots taken by location. A coach faces a pair of problems; designing and implementing a mixed strategy of shots taken from different parts of the court and defending against the opposing team's mixed strategy of shots. In this paper we solve this problem for the set of shot locations identified in figure 1.



**Figure 1: Court shot locations**

There are two equilibrium mixed strategies associated with each pair of teams with the property that points generated from Team A's offensive equilibrium strategy equals the points given up by Team B's defensive equilibrium and vice versa. To solve this problem, we build upon the work of Fichman and O'Brien (2016), to construct a Nash Equilibrium for every pair of team's offense versus defense by finding the mixed strategy that maximizes the difference between the natural log of the offensive and defensive Sharpe Ratios. Using the natural logarithm has two major advantages. First, the problem when viewed from an expected utility perspective is a member of the class of iso-elastic utility functions, Norstad (2011), which permits solving for the constant mixed strategy. Second, this problem is equivalent to maximizing the growth rate of point production per unit of risk taken by the offensive team net of the opposing team defense's attempt's to minimize point production. Our results generate a predicted outcome for the game from the difference between the pair of equilibria; a positive difference is a predicted win, negative a loss and zero a draw.

## 2.1 Identifying the Equilibrium Mixed Strategy

The maximization problem is an extension of modern portfolio theory, Elton and Gruber (1995) adapted to basketball as follows:

$$\frac{\text{Maximize}}{\text{w.r.t. } \omega} \ln\left(\frac{\mu_{\text{Offense}}}{\sigma_{\text{Offense}}}\right) - \ln\left(\frac{\mu_{\text{Defense}}}{\sigma_{\text{Defense}}}\right) \quad (1)$$

$$\text{Subject to: } \sum_j \omega_j = 1$$

$$\omega_j \geq 0$$

Where  $\omega$  = a vector of eleven shot location weights (figure 1),  $\bar{p}$  the probability of success from each shot location, and  $s^T$  the possible points from each location. Expected points from the mixed strategy is defined as:  $\mu = s^T \omega^T \bar{p}_o$  and the variance covariance matrix for different shot locations is:  $\sigma = \sqrt{\omega^T \Sigma_o \omega}$ ,  $\Sigma$ . Similarly, the maximization problem for defense follows by changing the subscript “o” for offense to “d” for defense.

We solve for the constant mixed strategy that maximizes the growth rate in generating points per unit of risk from the offense net of defense. This problem is solved by setting the problem up as a standard Lagrange problem with the inequality constraints handled as Kuhn-Tucker constraints. First order conditions for the regular Lagrange problem ignoring the Kuhn-Tucker constraints imply that the Lagrange multiplier is zero which follows by multiplying equation 2 by  $\omega^T$  which implies that  $\lambda = 0$ .

$$\left(\frac{1}{\omega^T s^T \bar{p}_o} s^T \bar{p}_o - \frac{1}{\omega^T \Sigma_o \omega} \Sigma_o \omega\right) - \left(\frac{1}{\omega^T s^T \bar{p}_d} s^T \bar{p}_d - \frac{1}{\omega^T \Sigma_d \omega} \Sigma_d \omega\right) = \lambda 1. \quad (2)$$

Second, the optimal solution must also satisfy the complementary slackness condition which implies for each shot zone either the multiplier is either non negative or zero if the shot zone has slack (e.g., Bryson and Ho (1975)). This implies there are three possible Nash Equilibria solutions: interior mixed strategy solution where shot zone weights are all greater than zero (e.g., shots are taken from every shot zone), mixed strategy solution with some shot zone weights equaling zero (e.g., drop mid-range shots and only take layups and 3-pointers) and a pure strategy solution with a single shot zone (e.g., 100% 3-points shots from a single location). Each equilibrium strategy prescribes the proportion of shots taken from each zone relative to the total number of shots in a game. Two equilibria are associated with each pair of teams with the following interpretation. Team A’s offensive equilibrium is Team B’s defensive equilibrium and vice versa Team B’s offensive equilibrium is Team A’s defensive equilibrium when playing each other. Finally, in section 4 we

solve this problem numerically for every pair of NBA teams and all solutions are mixed strategies with a combination of zero and interior weights.

## 2.2 Example: Applying the Theory to Conference Extremes 2015/16

During the regular 2015 season the Cleveland Cavaliers and the Golden State Warriors clinched their respective Eastern and Western conferences. The Warriors had a record breaking win percentage of 89.5% whilst the Cavaliers had a very respectable 69.5% win percentage. At the other end of their conference tables the Lakers could only manage 20.9% and the 76ers 12.2% regular season win percentages. It is instructive to compare equilibrium predictions for these extreme performances. In table 1 below, the equilibrium value for the pair CLE offense and PHI defense is 0.376, which is greater than 0.209, the equilibrium for CLE defense and PHI offense. This implies that CLE is predicted to defeat PHI. For the case of GSW and LAL this difference is much greater 0.709 versus 0.311. In both cases the conference leaders are predicted to beat the weakest same conference team, but for the case of GSW versus LAK the game is more likely hinge upon GSW's offensive performance because of the larger difference between offensive and defensive equilibria than is the case for CLE versus PHI.

	Offensive Equilibrium	Defensive Equilibrium
Eastern Conference		
Cleveland Cavaliers	0.376	0.209
Philadelphia 76ers	0.209	0.376
Western Conference		
Golden State Warriors	0.709	0.311
Los Angeles Lakers	0.311	0.709

**Table 1: Equilibrium Analysis of the Top versus Bottom Teams in each Conference**

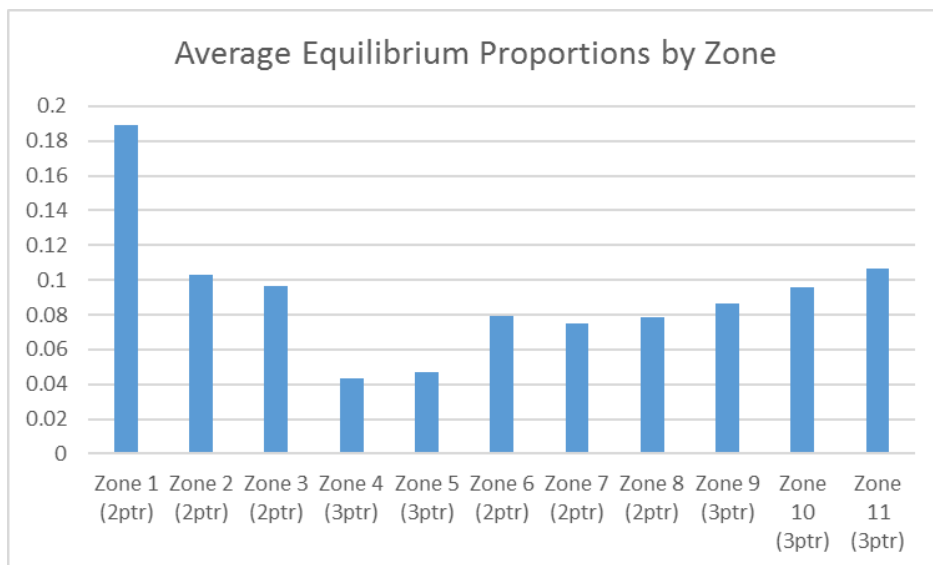
## 3 Data

The data used is play-by-play log data from nbastuffer.com. This database provides plays tagged by game clock time, 24-second clock time elapsed, play descriptions, shot types, and shot location for every regular season and playoff game for 2015/16. We preprocessed this data by shot location to calculate the statistical distributions by shot location (see figure 1 for shot locations) for every team separately for the regular season and the playoffs. This provided the input data for optimization problem (1). This optimization problem

was analyzed and solved for every pair of teams (30x29) for the regular season. This provided the base set of results used to analyze and predict post season performance.

## 4 Equilibrium Results

The results reveal that the aggregate predicted proportions of 2- and 3-point shots for the NBA as a whole are 0.621 and 0.379. The actual proportions for 2015/16 regular season were 0.714 and 0.286. This suggests that the NBA is still headed higher with future 3-point shot proportions and in the playoffs teams did shoot higher proportions of 3-point shots respectively at 0.691 (2-points) and 0.309 (3-point shots). Figure 2, below provides a summary of the average equilibrium weights by the shot zones defined in figure 1.



**Figure 2: Equilibrium Proportions by Shot Zones**

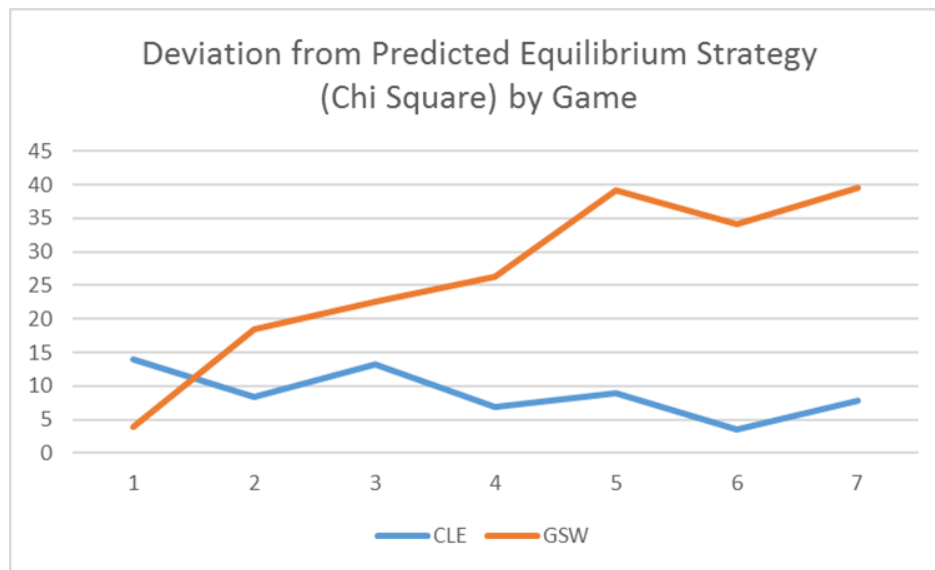
Using regular season data, a pair of regressions were run to predict winning points spread per regular season game from the eleven shot zones. Regression 1 (significant at  $p < 0.0000$ ) predicted points spread (winner minus loser) from the *average points given up* per zone by the winning team's defense, against the losing team's offense. Regression 2 (significant at  $p < 0.015$ ) changed the predictor variable to *average points made* per zone by the winning team's offense, against the losing team's defense. Regression 1 results indicate that the losing team's offense takes more shots in the 2-point zones, especially the 2-point left center box, zone 3 (positive coefficient,  $p < 0.001$ ), cutting off the higher payoff from the 3-point left center box, zone 5. The other significant zone was driving to the posts, zone 1 (positive coefficient,  $p < 0.10$ ). Results from regression 2 suggest that the winning team takes less shots from the 2-point circle box, zone 6 (negative

coefficient,  $p < 0.01$ ), and instead shoots more 3-point shots from the right hand top center, zone 11 (positive coefficient,  $p < 0.05$ ). Within season there is a highly significant relationship ( $p < 0.0001$ ) between the predicted winning equilibrium difference and the realized winning margin in points. The question is, does this relationship forecast out of sample post season results? To answer this question, we test the null hypothesis of no relationship between realized points spread (winning score minus the losing score) and the winning team's equilibrium score minus the losing team's equilibrium score. To test this hypothesis, we tested for the significance of the product moment correlation coefficient between these two difference scores in a one-tail test against the alternative hypothesis that predicts a positive relationship. The results reveal that the correlation coefficient is  $r = 0.20$  ( $p = 0.033$ ). Next we focus on our major objective of interpreting the final series in the playoffs between the Warriors and the Cavaliers using results from our equilibrium analysis.

## 5 2016 Final Series: GSW versus CLE

As is well known in the final series CLE upset the favorite, GSW. GSW were predicted to win and this was reinforced by the equilibrium analysis. The difference between GSW's offensive equilibrium (0.362) and CLE's offensive equilibrium (0.277) favors GSW for the win. In addition, both CLE and GSW when playing each other's defense lowers their average payoff per unit of risk from their respective averages against the NBA as a whole (e.g., 0.471 to 0.277 for CLE and 0.496 to 0.362 for GSW). This is expected because both teams are clearly above average but a curious feature is revealed. To support the 0.362 prediction, the equilibrium analysis implies *a significant shift in the 2- and 3-point strategy for GSW that is much larger than what is predicted for CLE*. The average predicted proportion of 3-point shots for GSW taken over the NBA as a whole is 0.743 but when playing against CLE this needs to drop to 0.212 from our mixed strategy analysis. In other words, the ability of CLE to defend against the 3-point shot is very strong when playing against GSW and this impacts GSW's optimal strategy. That is, GSW when playing CLE *should shift towards a much higher proportion of two point shots*. For CLE, the equilibrium analysis calls for 0.547 2-point shots versus their average across NBA teams of 0.474 2-point shots. If our analysis is correct, then GSW moving in the direction of our equilibrium strategy should increase their chances of winning while not changing or increasing the number of 3 point shots attempted should reduce their chance of winning. So the above raises the interesting question regarding to what degree adjustments actually took place in this series?

We conduct a Chi Square analysis of each team's actual versus predicted 2- and 3-point strategy employed in the final playoff series. The Chi square test is calculated with the Yates correction for continuity from the difference between the observed frequency of 2- and 3-point shots minus the frequency of 2- and 3-point shots that are predicted from equilibrium offensive mixed strategy for each team. In Game 1 GSW's observed deviation from the equilibrium strategy is not significant at the 5% level (Chi square = 3.43) whereas CLE's deviation is significant (13.184,  $p < 0.01$ ). The results for this game and the remaining games are provided in figure 3 below.



**Figure 3: Deviation from Predicted Equilibrium Strategy**

However, it is interesting to observe that for the case of GSW the Warriors moved almost monotonically away from the predicted equilibrium as the series unfolded. As noted earlier GSW *commenced the series relatively close to their predicted equilibrium* but then subsequently moved further and further away from this outcome. On the other hand, Cleveland gradually moved closer to the predicted equilibrium strategy. If we compute a product moment correlation coefficient between each team's deviation from the predicted strategy (as measured from the Chi square statistic) and the points spread (CLE – GSW),  $r = 0.50$  for GSW and  $r = 0.02$  for CLE. That is, GSW's deviation is positively correlated with CLE's success (0.50).

## 6 Conclusion

The objective of this paper was to conduct an equilibrium analysis of the 2015/16 season that was marked by the record breaking season by the Golden State Warriors and the major upset in the final series by the

Cleveland Cavaliers. As a result, this season has some potentially important strategic implications that warrant closer attention in an attempt to understand this season. The analysis was conducted by decomposing the basketball court into a set of eleven shot location areas and compiling the risk and average payoff statistics associated with each location for every team in the NBA. The constant mixed shot location equilibrium strategy was estimated for every pair of teams in the NBA and then the properties from this equilibrium analysis were estimated.

This analysis generated some interesting insights into the season. First basketball is embracing, as predicted, the increasing use of the 3-point shot. The current results of actual versus predicted for 2- and 3-point shots suggest that the percentage of 3-point shots will continue to trend even higher in the NBA to the current predicted equilibrium levels of 37.9% compared to the 2015/16 season average of 28.6% and the post season average of 30.9%. Second, the court location analysis by game realization yielded some equally interesting insights into the importance of defense. A significant difference in performance was observed for teams whose defense could force the offensive team into taking shots in the 2-point locations away from zones 4 and 5 (figure 1). Similarly, for the successful offense they were able to take their shots from the 3-point zones 9, 10 and 11 (figure 1) as opposed to being forced into the 2-point zones. In other words, at the margin a successful defense is shifting shot rate from 3-point to 2-point zones in a few key areas of the court.

Analyzing the final series also yielded some important strategic insights. The takeaways from the Wall Street Journal article referred to in the introduction, which is what Daryl Morey is doing with the Houston Rockets, is to drop mid-range shots and only take layups and 3-pointers. The shot zone analysis at the margin lent support to this by associating it with teams that win. A second idea in the article was that the optimal offensive strategy is constant:

“Pine City takes 59% of its shots from behind the 3-point line because it makes sense statistically ...”

The results from our equilibrium analysis do not support this type of assertion. To the contrary the results suggest that a team’s offensive strategy is very much dependent upon the team they are playing. For example, GSW have the highest predicted relative frequency of 3-point shots compared to any other team in the NBA, a predicted average of 74.3%. This high predicted percentage is consistent with their within season performance because they could generally come out shooting threes and make the other team try to match their point production. But if GSW is playing CLE this predicted average dropped to 21.1% for 3-point shots. In the final series GSW started near the predicted equilibrium when playing against CLE with



68.5% 2-point shots but then almost monotonically GSW moved in the direction *away from the equilibrium prediction* towards their predicted average percentages. On the other hand, over the series CLE drifted closer to their predicted equilibrium. These results support the conclusion that these different strategic choices had a major impact on the final series outcome and thus reinforce the idea that the relative proportion of 2- and 3-point shots taken in a game is strongly influenced by the opposing team's defense.

## References

- [1] Bryson, A. E. and Y.-C. Ho (1975). *Applied Optimal Control: Optimization, Estimation, and Control*. Hemisphere Publishing Corporation.
- [2] Cohen, B. (2017). The Basketball Team That Never Takes a Bad Shot. *Wall Street Journal*, [online], Available at <https://www.wsj.com> [Accessed 28 April 2017].
- [3] Elton, E. J. and Gruber. M. (1995) *Modern Portfolio Theory and Investment Analysis*, 5th Ed. New York: Wiley.
- [4] Fichman, M., and O'Brien, J.R. (2016) Three Point Shooting and Efficient Mixed Strategies: A Portfolio Management Approach Working Paper Tepper School of Business, Carnegie Mellon University.
- [5] Norstad, J., (2011) *An Introduction to Utility Theory* Working Paper Northwestern University.

# A Statistical Investigation of Factors Influencing the Results of One-Day Internationals in Cricket

Chris Frankland\* and Gordon Hunter\*\*

Department of Mathematics, Kingston University, Kingston upon Thames, KT1 2EE, U.K.

\* Chrissie.Frankland2@outlook.com

\*\* G.Hunter@kingston.ac.uk

## Abstract

The effect of playing “home” or “away” and many other factors, such as batting first or second, winning or losing the toss, have been hypothesised as influencing the outcome of major cricket matches. Anecdotally, it has often been noted that Subcontinental sides (India, Pakistan, Sri Lanka and Bangladesh) tend to perform much better on the Subcontinent than away from it, whilst England do better in Australia during cooler, damper Australian Summers than during hotter, drier ones. In this paper, focusing on results of men’s One Day International (ODI) matches involving England, we investigate the extent to which a number of factors – including playing home or away (or the continent of the venue), batting or fielding first, winning or losing the toss, the weather conditions during the game, the condition of the pitch, and the strength of each team’s top batting and bowling resources – influence the outcome of matches. By employing a variety of Statistical techniques, we find that the continent of the venue does appear to be a major factor affecting the result, but winning the toss does not. We then use the factors identified as significant in an attempt to build a Binary Logistic Regression Model that will estimate the probability of England winning at various stages of a game. Finally, we use this model to predict the results of some England ODI games not used in training the model.

## 1 Introduction

For many years, people have speculated over the optimal strategy to win a cricket match, given a particular team make-up and similarly for the opposition. Does winning the toss make a significant difference and if so, should one opt to bat or bowl first? Do the pitch and/or weather conditions play an influential role? However, until relatively recently, very few quantitative analyses had been carried out to try to answer questions.

Joshi (2009) studied the effect of prevailing weather conditions on England’s performance in Ashes test series in Australia, concluding that England teams performed much better during “La Niña” (cooler, damper) Australian Summers than in “El Niño” (hotter, drier) ones, an observation which attracted considerable attention in the press after England’s 2010-11 Ashes series victory in Australia – their first there since 1986-87 (Alleyne 2011). It would appear that England players performed much better in conditions closer to those with which they were more familiar when playing in England. Conversely, touring teams from tropical countries tend not to perform well when playing in England during a cool damp May. These observations prompted us to ask whether weather conditions and/or the nature of the pitches might be more influential on the outcome than the actual quality of the players in the teams.

With this in mind, we decided to perform a quantitative statistical investigation, using data from real matches, to study factors which seem to affect England's performances in One Day International (ODI) matches. The decision was made to focus on ODIs since many of them have been played since the first in 1971, but (unlike Test matches) very few (less than 1%) end without a decisive result. Furthermore, the shorter Twenty-Twenty (T20) format can be dominated by good performances by just one or two players and are widely considered to be very unpredictable. Although detailed match by match data can be obtained from internet sites such as [www.cricinfo.com](http://www.cricinfo.com), a set of readily formatted data was kindly provided by Professor Steven Stern of Queensland University of Technology. We investigated whether playing home or away, and the continent on which the match was played, winning or losing the toss, and batting first or second made a difference to the probability of England winning the match. Preliminary investigation indicated that the continent of the match venue did appear to be a significant factor, with England tending to perform particularly poorly on the Indian subcontinent (in India, Pakistan, Sri Lanka or Bangladesh) compared to elsewhere in the World. We followed this up by creating logistic regression models for the probability of an England win based on various factors, applying these models to matches not used in training, and testing whether using them with various gambling strategies to place "virtual bets", studying whether each of these could lead to a net profit when used over a series of matches.

The remainder of the paper is structured as follows. A review of relevant related work is presented in section 2, followed by details of the data we used, including the variables we are taking into account. The results of our exploratory data analysis including descriptive statistics, are given in section 4, and a description of how we produced our logistic regression models, and their success in predicting match strategies, are given in section 5. Finally, in section 6 we discuss our findings, draw our conclusions and suggest possible future work.

## 2 Related Previous Work

Modelling of One-Day cricket scores has extensively been carried out by Duckworth (2001), Duckworth & Lewis (2012), Schall & Weatherall (2013), and latterly Stern (2016). The primary purposes of these studies was to develop and improve a fair method of deciding which side should win a match in the case it should be interrupted by rain or a similar disruption, or to set fair modified targets should it be possible to continue with a (shortened) match following such an interruption, based on the state of the game at the enforced break point. The methods they developed were not designed to predict or explain the outcomes of matches based on factors known before the start of the match.

The role of home advantage has been discussed in the general context of sports matches by Stefani (2008), and specifically in relation to English one day cricket by Morley & Thomas (2005). The factor of travel fatigue noted by Stefani is probably not a major factor in ODI cricket, since teams tend to travel to the match venue a day or two before each match, whilst the factor of larger attendance by home team fans was deemed to be rather unimportant by Morley & Thomas. Nevertheless, the fact that, when playing away abroad, teams may find themselves playing in somewhat unfamiliar conditions, with very different weather and pitches of different nature to those which they are used to experiencing when playing at home may lead to the visiting team being at a disadvantage. For example, on average, runs are scored at 5.24 per over in Lahore, Pakistan, but only at 4.84 per over at the Oval, London. Similarly, many pitches in England and New Zealand are quite grassy (i.e. have a layer of fresh grass on the top, and are colloquially known as "green tops") except during very long dry periods, and those, particularly in overcast conditions, tend to favour fast or fast-medium seam and

swing bowling. This contrasts with the very dry dusty wickets particularly common on the subcontinent, which tend to favour batsmen early in the match, but spin bowlers later in the game. These may also prove to be factors affecting whether the home team has an advantage over the visitors.

In most formats of cricket, the team winning a toss of a coin immediately before the start of the match have the choice of whether to bowl or bat first. This might be considered to be an important factor influencing the final result, since damp or overcast conditions early in a game might favour the side bowling first, whilst having to chase a known total might assist the team batting second. However, poorer light or a pitch becoming dusty and/or cracked later in the game would again tend to help the side bowling at the end. The issue of what benefit winning the toss gives has been discussed by de Silva & Swartz (1997) and by Ishan Mukherjee (2014), but the latter study suggested that winning the toss was of little if any value in terms of giving a team a better chance of winning the game.

Attanayake & Hunter (2015) used data-driven Monte Carlo simulations, based on actual batting and bowling statistics of cricketers playing in the international teams at that time, to model team scores in international Twenty-Twenty (T20) cricket. Swartz et al (2009) have applied conditional Bayesian models to investigate the ball by ball progression of one day cricket matches based on the current state of the game. However, it is believed that the previously mentioned paper by Joshi (2009) is the one attempt to include weather conditions into a model for match or series outcomes. In that paper, the author took an “El Niño” index (essentially measuring how hot and dry a given Australian Summer was) as the only independent variable, and had the test series margin of victory (matches won less matches lost) for England as the dependent variable. Thus, he did not attempt to predict the results on individual matches, not take any other factors, such as the perceived qualities of the teams, “spin-friendliness” of the pitches, nor state of play (e.g. runs scored or wickets lost by one particular side) at any point in a match.

With these issues in mind, in this paper we try to create and test a model which can predict the results of a One Day International cricket match, in which England are one team, based on factors known prior to the match or measurable during the early to middle stages of the match. The methodology should be easily generalisable to model other teams and/or other formats of cricket for which sufficient data on previous matches is available.

### **3 Data Used in this Study, and Hypotheses to be Investigated**

We have used data on 120 ODI matches involving England as one team from the period 1<sup>st</sup> January 2005 to 31 December 2016. Of these, 40 were the most recent “home” games for England (played in England and Wales, from 2012), 40 were from England’s last 40 matches played on the Indian subcontinent (since 2005, not including games against Pakistan played in the UAE) and the remainder the most recent “away” matches for England played elsewhere in the World. Only matches against the “major” (i.e. the nine other Test-playing) nations were included, and World Cup, ICC Champions Trophy and other matches at “neutral” venues were excluded. Equal numbers of matches of each of the three categories were used in order to facilitate a set of Analysis of Variance (ANOVA) calculations. Although all the data were available on [www.cricinfo.com](http://www.cricinfo.com), we were kindly provided with a set of pre-formatted data by Professor Steven Stern. Only matches which reached a result were considered – matches abandoned due to bad weather were excluded. The data for each match noted the team totals, wickets lost and overs bowled for each innings, which team batted first, the venue of the game, and which team won the toss.

For each match considered, a margin of victory (MOV) was computed, treating an England win as a positive score and one won by their opponents as negative. If the team batting first won the game, the MOV was calculated as :

$$MOV = \frac{\textit{Winning team runs} - \textit{Losing team runs}}{\textit{Aggregate runs scored in game}}$$

whereas for a match won by the team batting second, the MOV was taken as :

$$MOV = \frac{\textit{Overs remaining in winning team's innings}}{\textit{Total overs allocation of winning team}}$$

unless the match was won on the last scheduled ball of the match, in which case the first formula was used. Whilst this latter formula does not take account of the number of wickets left intact at the end of either team's innings, it does emphasize that using the available overs appropriately is more important than retaining a large number of wickets, especially towards the end of an ODI innings. Trying to incorporate wickets left in hand would considerably complicate the calculations, although this is carried out by the Duckworth-Lewis and Duckworth-Lewis-Stern approaches (Stern, 2016). Using an ANOVA approach in SPSS, we test the following hypotheses, inspired by previous work by other authors discussed in sections 1 and 2 above, recalling that the MOV would be negative for matches which England lose :

Hypothesis 1 : "Winning the toss will significantly increase the Margin of Victory (MOV)."

Hypothesis 2 : "Choosing to bat first will not significantly affect the Margin of Victory (MOV)."

Hypothesis 3a : "England perform significantly better (i.e. tend to have higher MOVs) when playing at home, compared to playing anywhere else."

Hypothesis 3b : "England perform significantly worse (i.e. tend to have lower MOVs) when playing on the Indian subcontinent."

The match venues were grouped into three categories : "Home" (i.e. played in England or Wales), "Away – Subcontinent" (i.e. played against India, Pakistan, Sri Lanka or Bangladesh, where the opposition were at home), and "Away – Rest of the World" (i.e. played against Australia, South Africa, West Indies, New Zealand or Zimbabwe, with England's opponents being at home).

## 4 Descriptive Statistics, Exploratory Data Analysis and ANOVA

Summary information on the distribution of England's MOVs relative to winning or losing the toss, batting or bowling first, and the category of the match venue can be seen in figures 1, 2 and 3 below. Further details of the analyses can be found in Frankland (2017).

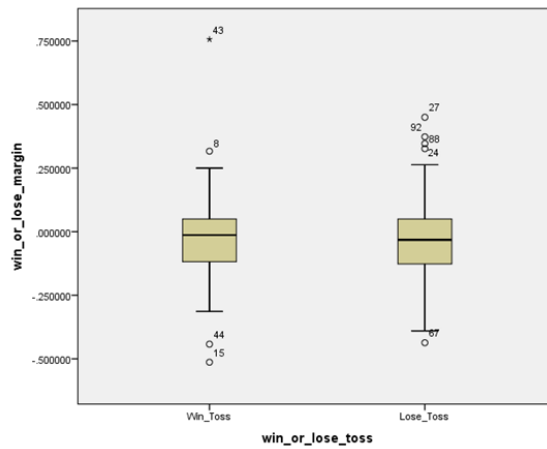


Figure 1. Distribution of England’s Margins of Victory (here recorded as “win\_or\_lose\_margin), with a positive score indicating an England win, relative to whether they won or lost the toss of the coin before the match. England won the toss in very close to 50% of games, as expected : 59 out of 120.

It can be observed that, in line with the findings of the previous study on ODIs by de Silva & Swartz (1997), winning the toss (Figure 1) does not appear to have much effect on England’s margin of victory (or defeat), nor does choosing to bowl or bat first (Figure 2).

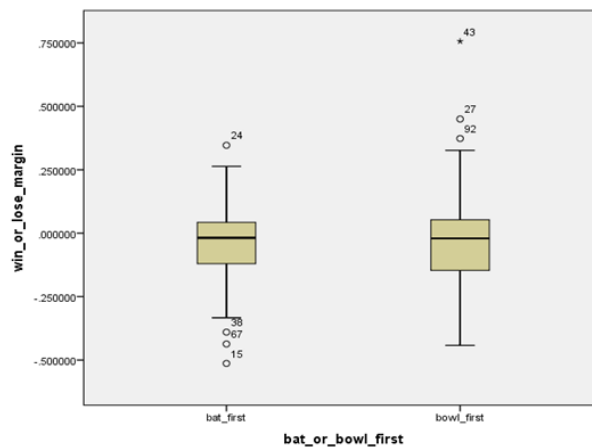


Figure 2. Distribution of England’s Margins of Victory (here recorded as “win\_or\_lose\_margin), with a positive score indicating an England win, relative to whether they batted or bowled first. The interquartile range of the MOV was a little higher when England bowled first and the distribution of outliers is rather different between the two situations, but the median values are almost identical.

In their last 40 home matches, England have won the toss 19 times, but of those occasions have only chosen to bat first 6 times. At home, England captains may prefer to chase a total. In contrast, when playing on the subcontinent, England have lost the toss 18 times in the last 40 games, but of those they have been require to bowl first 13 times. Subcontinental captains may believe that England are vulnerable when chasing a total, particularly against high quality spin bowling on deteriorating dry dusty wickets.

The trends shown in the distributions of Margins of Victory with respect to the venues of the matches are rather different (see Figure 3). Although the distribution of England's margins of victory is somewhat wider when playing at home than elsewhere in the World, the median values are similar except for matches on the subcontinent, where the median is negative, indicating that England tend to lose matches there. This is also born out by a direct comparison of the proportions of matches England win at home and away against the various subcontinental teams (see Figure 4).

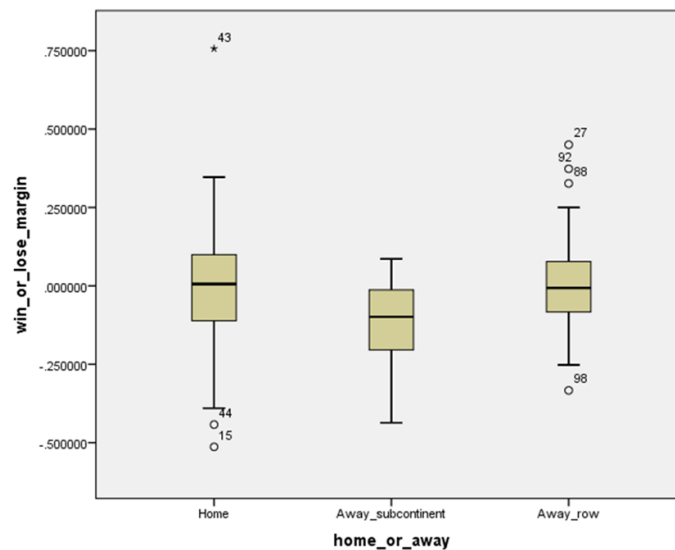


Figure 3 : The distributions of England's Margins of Victory (here recorded as "win\_or\_lose\_margin), with a positive score indicating an England win, relative to the venue of the match. It appears they tend to perform rather worse on the Indian subcontinent than elsewhere in the World, home or away.

The observations made by visual inspection of the box plots in Figures 1, 2 and 3 are confirmed by the ANOVA analysis. Treating the MOV as the dependent variable, and proposing the result of the toss, home or away, batting or bowling first, and pairwise interactions between these as the independent variables, the overall "corrected model" proved to be significant ( $p = 0.032$ ) at the 5% level, but the intercept did not ( $p = 0.191$ ), indicating that there was no statistically significant "offset" from zero to England's Margin of Victory, if all the independent variables take their default values. The only factor which did have a statistically significant impact on the MOV was whether the match was played at home or away ( $p = 0.014$ ). This is broadly in agreement with the findings of de Silva & Swartz (1997). A Tukey HSD multiple comparisons post-hoc test was carried out on the influence of this variable which, as noted previously, could take three possible values. It was found that the mean MOV was significantly different for matches played on the subcontinent relative to either matches played at home or elsewhere in the World. However, no significant difference was found between the MOVs for home matches and those played elsewhere in the World (i.e. neither in England or Wales, nor on the subcontinent) - see Table 1 below.

	Home	Subcontinent	Rest of World
Home	-	0.1183±0.0403 ( $p = 0.011$ )	-0.0052±0.0403 ( $p = 0.991$ )
Subcontinent	-0.1183±0.0403 ( $p = 0.011$ )	-	-0.1235±0.0403 ( $p = 0.008$ )
Rest of World	0.0052±0.0403 ( $p = 0.991$ )	0.1235±0.0403 ( $p = 0.008$ )	-

Table 1 : Results of a Tukey HSD multiple comparisons post-hoc test on how the match venue affects England's margin of victory. The values in each cell indicate the mean difference (row – column) ± standard error, with the statistical significance value in parentheses. It can be seen that significant differences (at 2% level or better) occur between the categories for games played on the subcontinent and those at home or in the rest of the World, but not between home games and those played in the rest of the World (excluding the subcontinent).

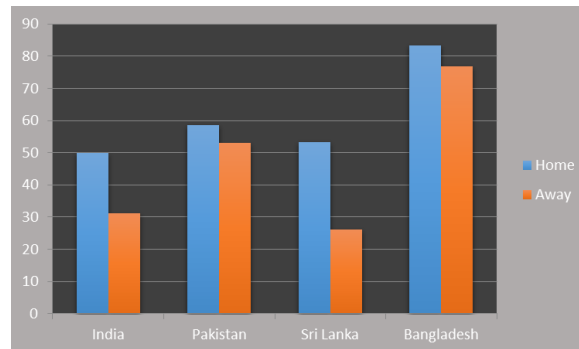


Figure 4. Proportion (as percentages) of ODI matches won by England at home or away against each of the subcontinental sides. It can be observed that in all cases, England perform better at home against each of these teams, with the difference between home and away results being particularly marked for matches against India or Sri Lanka.

These results confirm the anecdotal wisdom that England tend to underperform when playing on the Indian subcontinent, and further evidence is shown by comparing how they perform at home and away against each subcontinental side (see figure 4). Although England tend to win most of their matches against Bangladesh, their record playing in Bangladesh is inferior to playing Bangladesh in England. This underperformance on the subcontinent may be due to a number of factors, such as nature of the pitches, quality of spin bowling, hot or humid weather, to be investigated in the next section.

## 5 Logistic Regression Model for England Winning or Losing

Since it was found that only one categorical variable, namely the venue, significantly affected England's Margin of Victory, it was decided that it was not appropriate to create a predictive regression model for this quantity. However, as noted previously, a very low proportion (less than 1%) of ODIs which run to completion end in a tie, so effectively the result of a completed ODI match is a binary outcome. For the purposes of our study which focuses on England's results, we will treat the outcome as either a win or a loss for England. We can therefore create a logistic regression model, to calculate the log odds of England winning a given match as a linear function of various factors relating to that match, ideally factors which would be known ideally in advance of the match, or at least during the early stages of the match. Further details of the models and calculations are given in Frankland (2017).



If the probability of the team of interest (in this case, England) winning a given match is  $p$ , then we can model the log odds of that team winning, based on the values of various “predictor” variables  $x_1, x_2, \dots, x_N$  as follows :

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_Nx_N$$

where each of the  $\beta_i$  is a parameter associated with the corresponding predictor  $x_i$  and  $\beta_0$  is a constant. The values of the  $\{\beta_i\}$  can be estimated from evidence in given data by maximum likelihood estimation. In our case, we carried out this procedure using the SAS system. Our proposed predictor variables were chosen based on the results of our ANOVA analysis described in section 4 above. For simplicity, it was decided to focus only on games between England and subcontinental sides, namely India, Pakistan, Sri Lanka and Bangladesh. The first proposed predictor variable was therefore taken to represent the venue, where 0 meant the game took place on the subcontinent and 1 meant it was held in England or Wales. Another proposed factor, due to subcontinental sides tending to have a large number of high quality spin bowlers, was the “spin friendliness” of pitches at that venue, defined as a fraction between 0 and 1, computed by

$$S_F = \frac{\text{Number of wickets taken by spin bowlers}}{(\text{Total number of wickets} - \text{Run outs})}$$

at that venue over the three previous matches at that venue. Since England cricketers would not be so used to playing in hot or humid conditions, it was anticipated that England would perform less well when it was hot and humid than when it was cooler and drier. (Many matches in India are played between October and February, during which time Northern India at least does experience something of a “Winter” in relative terms, when the weather tends to be more like an English Spring or Autumn than tropical conditions.) Data on mean temperature and humidity at each venue for each match date was obtained from the website Wunderground.com. Initially, these were tested as separate independent variables, but it became clear that England performed worse when both temperature and humidity were high, so it was decided to combine them into a single predictor by multiplying the temperature in °C by the relative humidity (expressed as a percentage). In order to take account of the quality or strength of the actual teams playing in that particular game, noting that this could vary considerably from match to match due to injuries or players being unavailable for various other reasons, the ICC ranking points for the top three bowlers in each team were summed to give a bowling strength score, and similarly for the ICC ranking points for the top three batsmen in each team. These were computed in this manner, since the ICC team rankings are based on a long term weighted moving average for each international team, and do not take account of the precise set of individual players participating in each game. Although the ICC player rankings are also based on long-term weighted moving averages, these will make some allowance for relatively radical changes to a team’s composition, such as occurred when several England players refused to tour Bangladesh during the Autumn of 2016. The difference in total ICC ranking points of the top three bowlers between the sides, and similarly for the top three batsmen in each side, were calculated for use as “relative bowling strength” and “relative batting strength” respectively. Only the top three players were included for each team, since the ICC published rankings, updated each month, only list 100 bowlers and 100

batsmen from across the cricketing World at any one time, so not all players are included. The final predictor included is one not known prior to the start of the match, namely the relative positions of the two teams after 30 overs of their innings. Over recent years, ODIs have been consistently limited to a maximum of 50 overs per side, and it is relatively rare for a side to be bowled out, or for a team batting second to achieve their target, in less than 30 overs. The relative state of play 30 overs into each innings, calculated as

$$\left( \frac{\text{EngRuns after } N \text{ overs}}{\text{EngWickets after } N \text{ overs} + 1} \right) - \left( \frac{\text{OppRuns after } N \text{ overs}}{\text{OppWickets after } N \text{ overs} + 1} \right)$$

where *EngRuns* is the total number of runs scored by England, *EngWickets* is the number of wickets lost by England, both up to that point (*N* overs of the innings completed), etc., and here *N* = 30, should provide a useful predictor of the outcome of the match. Furthermore, with many bookmakers and betting exchanges now allowing in-play gambling, making a prediction after 30 overs of the second innings could still be useful to gamblers seeking to place a good bet.

The optimal parameters for the model was computed using a backwards stepwise selection procedure in SAS, where initially all proposed predictor variables were included, then least significant variable removed at each stage until all remaining variables were statistically significant at the 10% level. The final overall model was found to be highly statistically significant ( $p < 0.0001$ ), but only the intercept and the influence of four predictor variables were found to be statistically significant. The final logistic regression model was found to be

$$\ln\left(\frac{p}{1-p}\right) = 1.47523 - 0.83202x_1 + 0.00326x_2 + 0.00099843x_3 - 0.00043326x_4$$

where  $x_1$  is the proportion of wickets that have fallen to spin bowlers at that venue ( $p = 0.0202$ ),  $x_2$  the relative difference in scores between England and the opposition after 30 overs of each innings ( $p = 0.0040$ ),  $x_3$  the difference in the total ICC ranking scores of the top three bowlers in each side between England and the opposition prior to the match ( $p = 0.0128$ ), and  $x_4$  the product of the mean temperature (in °C) and the relative humidity (as a percentage) during the match ( $p = 0.0006$ ).

The model was then tested on data not used in the calculation of its parameters by predicting the probabilities of England winning each of the ODIs they played in India in early 2013. (The series in India in early 2017 was not used, since bookmakers' odds for that series were not publically available at the time of writing.) One match from the 2013 series (that held on 19<sup>th</sup> January 2013) was excluded since it was won within the first 30 overs of the second innings (and hence  $x_2$  would not be defined), but our regression model successfully predicted the results of all the four other games in that series. Backing the bookmaker's favorites (using odds from Bet365.com given at the 30<sup>th</sup> over of the second innings of each match) would have predicted the correct winner in 3 out of 4 cases.

## 6 Conclusions and Future Work

Our investigations relating to factors affecting England's performance in ODI cricket matches has indicated that neither winning the toss nor batting or bowling first has any significant effect on match outcomes. However, the venue of a match – in particular, whether England is playing on the Indian subcontinent – does affect England's chances of winning. These findings are consistent with those of de Silva & Swartz (1997). Building a logistic regression model enables us to predict the results of matches better (on the matches used here for testing) than following the bookmakers' favorite each

time. Our final model did not explicitly include the continent of the venue, but the “spin friendliness” of the pitch and the (temperature times humidity) variable would both, at least in part, act as proxies for that. The final model also did not include the “relative team batting strength” variable, although it did include the corresponding relative bowling strength. This is possibly because only the ICC points scores of the top three ranked players of either type in each team were included in these. Typically, each specialist bowler will bowl 10 overs in a 50 over ODI innings, so the top three bowlers will deliver 30 out of 50 overs, or 60% of the ball bowled. In contrast, all eleven players may be required to bat, and the top three rated batsmen are less likely to score 60% or more of the team’s total, which may explain that observation.

Work on applying our model to betting strategies (Norton et al, 2015), and investigating whether it could be used to generate a profit (which use of the bookmakers’ favorite each time is unlikely to do) is currently in progress. Future work could include further refining the model, incorporating additional factors, or generalising it to apply to matches involving any two specified times, to different formats of cricket, or even to other sports involving two players or teams in direct competition.

## Acknowledgements

We would like to thank to Professor Steven Stern for providing the ODI dataset, and to Dr. Maurice Beck and Dr. Tim Paulden for providing valuable insights regarding the sports betting industry.

## References

- [1] Alleyne, R. (2011) *The Ashes : Was Weather Key to England’s Historic Cricket Victory over Australia ?* The Daily Telegraph, U.K., 7<sup>th</sup> January 2011.
- [2] Attanayake, D. & Hunter, G. (2015) *Probabilistic Modelling of Twenty-Twenty (T20) Cricket : An Investigation into Various Metrics of Player Performance and their Effects on the Resulting Match & Player Scores*, Proceedings of the 5th International Conference on Mathematics in Sport, Loughborough, U.K.
- [3] de Silva, B. & Swartz, T. (1997) *Winning the Coin Toss and the Home Team Advantage in One-Day International Cricket Matches*, New Zealand Statistician, 32, 16-22
- [4] Duckworth, F. (2001) *A Role for Statistics in International Cricket*, Teaching Statistics, 23 (2), pp. 38-44. DOI: 10.1111/1467-9639.00048.
- [5] Duckworth, F. & Lewis, A. (2012) *The D/L Method: Answers to frequently asked questions* (updated September 2012). Available at: <http://www.espncriinfo.com/ci/content/page/581925.html>
- [6] Frankland, C. (2017) *A Statistical Investigation of Factors Influencing the Results of One-Day Internationals in Cricket*, Final Year BSc Project Dissertation, Kingston University, U.K.
- [7] Joshi (2009) *Could El Niño Southern Oscillation affect the results of the Ashes series in Australia ?* Weather, 64 (7) 178 – 180
- [8] Morley, B. & Thomas, D. (2015) *An Investigation of Home Advantage and Other Factors Affecting Outcomes in English One-Day Cricket Matches*, Journal of Sports Sciences, 23 (3), 261-268, DOI: 10.1080/02640410410001730133.
- [9] Mukherjee, Ishan (2014) *Time for cricket to do away with toss-tradition*. Available at: <https://www.sportskeeda.com/amp/cricket/toss-alternative-method-cricket>
- [11] Norton, H., Gray, S. & Faff, R. (2015) *Yes, one-day international cricket ‘in-play’ trading strategies can be profitable !*, Journal of Banking & Finance, 61, S164-S176
- [12] Schall, R. & Weatherall, D. (2013) *Accuracy and fairness of rain rules for interrupted one-day cricket matches*, Journal of Applied Statistics, 40 (11), 2462-2479.
- [13] Stefani, R. (2008) *Measurement and Interpretation of Home Advantage*, In *Statistical Thinking in Sports* (Albert, J., Koning, R.H., Eds.) Boca Raton, FL, USA: CRC Press, pp. 81-93.
- [14] Stern, S.E. (2016) *The Duckworth-Lewis-Stern method: Extending the Duckworth-Lewis methodology to deal with modern scoring rates*, Journal of the Operational Research Society, 67 (12), 1469-1480, DOI: 10.1057/jors.2016.30.
- [15] Swartz, T, Gill, P. and Muthukumarana, S. (2009) *Modelling and Simulation for One-Day Cricket*, Canadian Journal of Statistics, 37 (2), 143 -160

# On Reducing Sequence Effects in Competitions

Y. Gerchak\* and E. Khmel'nitsky\*\*

\*Department of Industrial Engineering, Tel-Aviv University, Tel-Aviv 69978, Israel, yigal@post.tau.ac.il

\*\*Department of Industrial Engineering, Tel-Aviv University, Tel-Aviv 69978, Israel, xmel@tau.ac.il

## Abstract

Some important sequencing decisions in sports are made by a coin toss. While that constitutes ex-ante fairness, it can generate ex-post unfairness, which is undesirable. We propose a scheme where the teams bid by degree of difficulty of their attempt for place in the sequence. In American football that would be relevant to which team starts an overtime period and from where. In soccer, it would be relevant to order of penalty shots in tied games, when a winner has to be determined, and the degree of difficulty is the distance from which shot is attempted. We propose and analyze an auction for order of penalty shots in soccer where the bids are distances. We address both discrete and continuous ability distributions. In our scheme, the higher bidder attempts to score first from a distance which is a weighted average of the bids, and the lower bidder attempts either from 11m ("Rule 1"), or from a distance that is also a weighted average of the bids ("Rule 2").

## 1 Introduction

Many sports include a phase where teams or athletes take turns performing a task and the winner is the one who has the most successes. Examples include penalty shots where a soccer game results in a tie and a winner has to be declared, playoff series in various sports where a "turn" is a home game, as well as the manner of playing an overtime period in a tied American Football game. The sequence in which the attempts are played in many of these tie-breakers has a real or perceived effect on the outcome.

A common way to determine the order or sequence in such situations is by coin-toss. However, while that constitutes ex-ante fairness, it does not generate ex-post fairness, and it is considered rather undesirable to allow a coin-toss to significantly affect the outcome of a match (Brams and Sanderson 2013). The question thus is: Is there a better way to do that?

In tied soccer games that require a clear winner, teams typically take turns shooting a total of 5 penalty shots each from a distance of 11m: ABABABABAB. Similar tie breakers are now used in the NHL in every tied game. Many fans and sports reporters believe team A to have a "psychological advantage" in this shoot-out and there is some empirical research supporting that (Palacio-Huerta et al., 2010). Other empirical research, however, refutes those findings (Kocher et al., 2012).

An interesting idea is bidding for the place in the sequence by the difficulty of the task. For example, in NFL overtimes if team A bids "75 yard line" (i.e., to start from its own 25 yard line) and team B bids "70 yard line", team A would start on offence from, say,  $(75 + 70)/2 = 72.5$  (or a different weighted average) yards away from team B's end zone. A kickoff spot may be adjusted accordingly (Granot and Gerchak, 2014). That is an auction with positive externality, since the team with losing bid is still influenced (positively) by the magnitude of the winning bid. In soccer, assuming for simplicity that each team is allowed only one attempt (kick), the bid could be for the distance of its kicking point from the

goal. If A bids 12.5m and B 11.5m then A would kick first, from a distance of some average of 12.5m and 11.5m, while B would kick second from either 11m ("Rule 1") or from an average of 11.5m and 12.5m ("Rule 2").

We analyze the resulting games for both rules, for discrete and continuous ability distributions. We then compare the bids resulting from each rule.

## 2 Bidding for Kicking First in Soccer

Suppose the teams are evenly matched, risk neutral and characterized by their abilities to score on a penalty kick. Such abilities are denoted by  $x$  for A and  $y$  for B,  $x, y \in [0, \infty)$ . Each team knows its own ability and forms a probability distribution function,  $F(\cdot)$ , over the other team's ability. Assume that if both attempts are from 11m, the following success probabilities for a team with ability  $y$  are known to be:

$p(y)$  – the probability that the first attempt is successful;

$q(y)$  – the probability that the second attempt is successful if the first attempt by the other team was successful;

$r(y)$  – the probability that the second attempt is successful if the first attempt by the other team failed.

One would expect that  $p(y) > r(y) > q(y) \forall y$ . (1)

It could also be the case that  $r(y) > p(y)$ . If so,  $r(y)$  should be closer to  $p(y)$  than  $q(y)$ , so the team kicking first will have an advantage. The probabilities decrease with the kick distance. If a team with ability  $y$  attempts from a distance  $(11+z)m$ , then we assume that the probability of success is  $p(y)e^{-\theta z}$ , where  $\theta > 0$  is given. The other probabilities,  $q(y)$  and  $r(y)$  change similarly.

Suppose the teams follow a symmetric strategy that determines the bid a team submits given its ability,  $\beta(y)$ ,  $\beta: [0, \infty) \rightarrow [0, \infty)$ . The "infinite" distance could be, for example, the length of the field. We seek a monotone increasing strategy  $\beta(y)$ , which maximizes the expected probability that A wins (or its expected utility if we normalize the utility of win to 1 and of a loss or a tie to 0). Note that we focus on an outright win – a tie is given no value.

We denote the bids of teams A and B by  $a$  and  $b$ , respectively, and assume that if  $a > b$ , then team A attempts first from a distance of  $\alpha a + (1 - \alpha)b$ ,  $1/2 \leq \alpha \leq 1$ , where  $\alpha$  is fixed by the organizers and known to the teams. The two alternative rules discussed below determine the sequence of the kicks and the kicks' distances. As far as the team with a lower bid is concerned, Rule 1 has it attempting a shot from 11m (i.e., zero extra distance), while Rule 2 requires it to attempt its shot from a distance of  $\alpha b + (1 - \alpha)a$ , with the same  $\alpha$  as used for the team attempting first. If  $a < b$  the roles of the teams are reversed. Note that the assumption that  $\alpha \geq 1/2$  guarantees that the higher bidder kicks from farther than the other.

### 2.1 Rule 1

The objective of team A is to maximize its expected probability of winning,

$$\begin{aligned} \max_{\beta(\cdot)} J = & \int_0^x p(x) e^{-\theta(\alpha\beta(x) + (1-\alpha)\beta(y))} (1 - q(y)) f(y) dy \\ & + \int_x^\infty (1 - p(y) e^{-\theta(\alpha\beta(y) + (1-\alpha)\beta(x))}) r(x) f(y) dy. \end{aligned} \quad (2)$$

### 2.1.1 Two values distribution of abilities

Suppose that

$$X, Y = \begin{cases} c & \text{with probability } \varepsilon \\ d & \text{with probability } 1 - \varepsilon \end{cases} \quad 0 \leq \varepsilon \leq 1, \quad (3)$$

and without loss of generality  $c > d$ . We first note that if condition (1) holds, then the theory that the team attempting first has an advantage is correct. This is proven in the next lemma.

**Lemma 1.** *If (1) holds then the expected probability of success of a team attempting first is greater than the expected probability of a team attempting second.*

*Proof.* The lemma states that

$$\varepsilon p(c) + (1 - \varepsilon)p(d) > \varepsilon[q(c)(\varepsilon p(c) + (1 - \varepsilon)p(d)) + r(c)(1 - \varepsilon p(c) - (1 - \varepsilon)p(d))] \\ + (1 - \varepsilon)[q(d)(\varepsilon p(c) + (1 - \varepsilon)p(d)) + r(d)(1 - \varepsilon p(c) - (1 - \varepsilon)p(d))].$$

Equivalently,

$$[\varepsilon(r(c) - q(c)) + (1 - \varepsilon)(r(d) - q(d))][\varepsilon p(c) + (1 - \varepsilon)p(d)] > \varepsilon(r(c) - p(c)) \\ + (1 - \varepsilon)(r(d) - p(d)).$$

If (1) holds, then the left-hand side of the latter inequality is positive, while the right-hand side is negative. This proves the lemma.  $\square$

Suppose that in case of ties, which are quite likely here, the winner is selected randomly. Suppose first that  $x = c$ , i.e.,  $\beta(c) = a$ , then,

$$J = \frac{\varepsilon}{2} p(c) e^{-\theta a} (1 - q(c)) + \frac{\varepsilon}{2} r(c) (1 - p(c) e^{-\theta a}) \\ + (1 - \varepsilon) p(c) e^{-\theta(aa + (1 - \alpha)\beta(d))} (1 - q(d)). \quad (4)$$

Since  $J$  decreases in  $\beta(d)$  for each  $a$ , then the value of  $\beta(d)$  that maximizes  $J$  is  $\beta(d) = 0$ . This satisfies the assumption made with respect to the monotonicity of  $\beta(\cdot)$ , here  $\beta(d) \leq \beta(c)$ . After substituting  $\beta(d) = 0$  in (4), the maximization of  $J$

$$J = \frac{\varepsilon}{2} p(c) e^{-\theta a} (1 - q(c)) + \frac{\varepsilon}{2} r(c) (1 - p(c) e^{-\theta a}) \\ + (1 - \varepsilon) p(c) e^{-\theta a a} (1 - q(d))$$

w.r.t. parameter  $a$ , is carried out by solving the equation  $dJ/da = 0$ . The result is,

$$a = \begin{cases} \frac{1}{\theta(1 - \alpha)} \ln \frac{\varepsilon(q(c) + r(c) - 1)}{2\alpha(1 - \varepsilon)(1 - q(d))}, & \text{if } q(c) + r(c) > 1 + \frac{2\alpha(1 - \varepsilon)(1 - q(d))}{\varepsilon} \\ 0, & \text{otherwise} \end{cases}. \quad (5)$$

Suppose now that  $x = d$ , i.e.,  $\beta(d) = a$ , then,

$$J = \frac{1 - \varepsilon}{2} p(d) e^{-\theta a} (1 - q(d)) + \frac{1 - \varepsilon}{2} r(d) (1 - p(d) e^{-\theta a}) \\ + \varepsilon (1 - p(c) e^{-\theta(\alpha\beta(c) + (1 - \alpha)a)}) r(d).$$

Since  $J$  increases in  $\beta(c)$  for each  $a$ , then the value of  $\beta(c)$  that maximizes  $J$  is  $\beta(c) = \infty$ . Again, this satisfies the monotonicity assumption, as  $\beta(d) \leq \beta(c)$ . Substituting, the bid  $a$  that maximizes the objective  $J$ ,

$$J = \frac{1-\varepsilon}{2} p(d) e^{-\theta a} (1 - q(d) - r(d)) + \frac{1+\varepsilon}{2} r(d),$$

is

$$a = \begin{cases} \infty, & \text{if } q(d) + r(d) > 1 \\ 0, & \text{otherwise} \end{cases} \quad (6)$$

This result is independent of  $\alpha$  and of the high ability parameters. The bid  $a$  obtained in (5) and (6) for the two-value ability distribution leads to the following strategy:

- the team whose ability is at least as high as of the other team (case  $x = c$ ), submits either zero bid (11m kick), or the bid determined in the first line of (5).
- the team whose ability is at most as high as of the other team (case  $x = d$ ) submits either zero bid, or an “infinity” bid. The team which submits an “infinity” bid has a positive probability of success in the scenario where the ability of the other team turns out to be also  $d$ , i.e.,  $y = d$ , as it then submits the “infinity” bid as well, and, if it is selected to kick first, it fails for sure.

### 2.1.2 Asymmetric beliefs of the teams

Suppose that

$$X = \begin{cases} c & \text{with probability } \varepsilon_1 \\ d & \text{with probability } 1 - \varepsilon_1 \end{cases} \quad Y = \begin{cases} c & \text{with probability } \varepsilon_2 \\ d & \text{with probability } 1 - \varepsilon_2 \end{cases}$$

and  $c > d$ . In this case, the conclusions are the same as in Section 2.1.1, with

$$a_1 = \begin{cases} \frac{1}{\theta(1-\alpha)} \ln \frac{\varepsilon_2(q(c) + r(c) - 1)}{2\alpha(1-\varepsilon_2)(1-q(d))}, & \text{if } q(c) + r(c) > 1 + \frac{2\alpha(1-\varepsilon_2)(1-q(d))}{\varepsilon_2} \\ 0, & \text{otherwise.} \end{cases}$$

$a_2$  is similar with  $\varepsilon_1$  replacing  $\varepsilon_2$ .

### 2.1.3 Symmetric beliefs with a three values distribution

Let

$$X, Y = \begin{cases} c & \text{with probability } \varepsilon \\ d & \text{with probability } \delta \\ k & \text{with probability } 1 - \varepsilon - \delta \end{cases} \quad c > d > k.$$

The optimal strategy of the teams, calculated in the same way as in Section 2.1.1., is as follows.

If  $x = c$ , then

$$a = \frac{1}{(1-\alpha)\theta} \ln \left[ \frac{\varepsilon(q(c) + r(c) - 1)}{2\alpha((1-\varepsilon)(1-q(k)) - \delta(q(d) - q(k)))} \right]$$

if the argument of the ln-function is greater than 1; otherwise,  $a = 0$ .

If  $x = d$ , then

$$a = \frac{1}{(1-\alpha)\theta} \ln \left[ \frac{\delta(q(d) + r(d) - 1)}{2\alpha(1-\varepsilon-\delta)(1-q(k))} \right]$$

if the argument of the ln-function is greater than 1; otherwise,  $a = 0$ .

If  $x = k$ , then

$$a = \begin{cases} \infty, & \text{if } q(k) + r(k) > 1 \\ 0, & \text{otherwise} \end{cases}$$

### 2.1.4 Continuous exponential distribution of abilities

Suppose that the beliefs of the teams with regard to the abilities of the other team are distributed exponentially,  $X, Y \sim Exp(\lambda)$  and the probabilities  $p(y)$ ,  $q(y)$  and  $r(y)$  grow with the team ability in the form,

$$p(y) = 1 - e^{-py}, \quad q(y) = 1 - e^{-qy}, \quad r(y) = 1 - e^{-ry}, \quad p > r > q. \quad (7)$$

This section limits the solution of problem (2) to the linear  $\beta(\cdot)$ ,

$$\beta(y) = ty. \quad (8)$$

which is simple to implement. In such a case,

$$J = \lambda e^{-\lambda x} \left( \frac{e^{-(q+t\theta)x}(1-e^{-px})(e^{(\lambda+q+t(1-\alpha)\theta)x}-1)}}{\lambda+q+t(1-\alpha)\theta} - \frac{e^{-t\theta x}(1-e^{-rx})}{\lambda+t\alpha\theta} + \frac{1-e^{-rx}}{\lambda} + \frac{e^{-(p+t\theta)x}(1-e^{-rx})}{\lambda+p+t\alpha\theta} \right). \quad (9)$$

In the numerical experiment below, we use  $\theta = 0.1$ , which implies that the probability of success drops down exponentially with a factor close to 2 for each 7m increment of the shot distance. This follows, since  $e^{-0.7} \approx 1/2$ . We use  $p = 0.1$ ,  $r = 0.09$ ,  $q = 0.08$  and  $x = 17$ , which imply that the probabilities of success from 11m are  $p(x) \approx 0.82$ ,  $r(x) \approx 0.78$ , and  $q(x) \approx 0.74$ . Also, take  $\alpha = 0.5$ . The results of this case, as well as the results of a similar case considered in the Rule 2 section, are presented in Figures 1 and 2. We observe that team's bid increases with the expected ability of the other team for the two rules. However, the dependence of a team's bid on its own ability differs whether Rule 1 or Rule 2 is applied: the bid decreases for Rule 1 and increases for Rule 2.

### 2.1.5 Uniform distribution of abilities

Let

$$X, Y \sim U[0,1].$$

We seek a solution of problem (2) of the linear form (8). The objective is

$$J = \int_0^x p(x)e^{-\theta(\alpha x+(1-\alpha)ty)}(1-q(y))dy + \int_x^1 (1-p(y)e^{-\theta(\alpha y+(1-\alpha)tx)})r(x)dy. \quad (10)$$

We numerically maximize (10) for  $\theta = 0.1$ ,  $\alpha = 0.5$ ,  $p = 1.8$ ,  $r = 1.6$ ,  $q = 1.4$ . The optimal bid and the maximum probability of success as a function of  $x$  are presented in Figure 3.

We observe that for a low ability the bid increases up to some level, then decreases and drops down to zero for a high ability. The team with a high ability takes a conservative approach.

## 2.2 Rule 2

As defined above, this rule requires both teams to shoot from a distance, which is a linear combination of the two bids. The expected probability of A winning is now,

$$\max_{\beta(\cdot)} J = \int_0^x p(x)e^{-\theta(\alpha\beta(x)+(1-\alpha)\beta(y))}(1-q(y)e^{-\theta(\alpha\beta(y)+(1-\alpha)\beta(x))})f(y)dy + \int_x^\infty (1-p(y)e^{-\theta(\alpha\beta(y)+(1-\alpha)\beta(x))})r(x)e^{-\theta(\alpha\beta(x)+(1-\alpha)\beta(y))}f(y)dy. \quad (11)$$



### 2.2.1 Two values distribution of abilities

Suppose that

$$X, Y = \begin{cases} c & \text{with probability } \varepsilon \\ d & \text{with probability } 1 - \varepsilon \end{cases}$$

Suppose that  $x = c$ , i.e.,  $\beta(c) = a$ . Then,

$$J = \frac{\varepsilon}{2} p(c) e^{-\theta a} (1 - q(c) e^{-\theta a}) + \frac{\varepsilon}{2} r(c) e^{-\theta a} (1 - p(c) e^{-\theta a}) + (1 - \varepsilon) p(c) e^{-\theta(aa+(1-\alpha)b)} (1 - q(d) e^{-\theta(ab+(1-\alpha)a)}). \quad (12)$$

By solving  $dJ/db = 0$  w.r.t.  $b$ , we get that  $b$  depends linearly on  $a$ ,

$$b^*(a) = \frac{1}{\alpha\theta} \ln \frac{q(d)}{1-\alpha} - a \frac{1-\alpha}{\alpha}.$$

Substituting in (12), we obtain

$$J = \frac{\varepsilon}{2} p(c) e^{-\theta a} (1 - q(c) e^{-\theta a}) + \frac{\varepsilon}{2} r(c) e^{-\theta a} (1 - p(c) e^{-\theta a}) + (1 - \varepsilon) \alpha p(c) \left( \frac{1-\alpha}{q(d)} \right)^{\frac{1-\alpha}{\alpha}} e^{-\theta a \frac{2\alpha-1}{\alpha}}.$$

After changing variable,  $z = e^{-\theta a}$ ,

$$J = c_1 z - c_2 z^2 + c_3 z^{\frac{2\alpha-1}{\alpha}}, \quad (13)$$

where  $c_1 = \frac{\varepsilon}{2} (p(c) + r(c))$ ,  $c_2 = \frac{\varepsilon}{2} p(c) (q(c) + r(c))$  and  $c_3 = (1 - \varepsilon) \alpha p(c) \left( \frac{1-\alpha}{q(d)} \right)^{\frac{1-\alpha}{\alpha}}$ .

In the particular case when  $\alpha = 1/2$ ,  $J$  is maximized at  $z = c_1/2c_2$ . That is,

$$a = \begin{cases} \frac{1}{\theta} \ln \frac{2p(c)(q(c)+r(c))}{p(c)+r(c)}, & \text{if } q(c) + r(c) > \frac{1}{2} \left( 1 + \frac{r(c)}{p(c)} \right) \\ 0, & \text{otherwise} \end{cases}. \quad (14)$$

In the particular case when  $\alpha = 1$ ,  $J$  is maximized at  $z = (c_1 + (1 - \varepsilon)p(c))/2c_2$ . That is,

$$a = \begin{cases} \frac{1}{\theta} \ln \frac{2\varepsilon p(c)(q(c)+r(c))}{2p(c)-\varepsilon(p(c)-r(c))}, & \text{if } q(c) + r(c) > \frac{1}{\varepsilon} - \frac{1}{2} \left( 1 - \frac{r(c)}{p(c)} \right) \\ 0, & \text{otherwise} \end{cases}. \quad (15)$$

Note that the condition  $b^*(a) \leq a$  holds under the assumptions made on  $p(\cdot)$ ,  $q(\cdot)$  and  $r(\cdot)$ .

Suppose now that  $x = d$ , i.e.,  $\beta(d) = a$ . Then,

$$J = \frac{1-\varepsilon}{2} p(d) e^{-\theta a} (1 - q(d) e^{-\theta a}) + \frac{1-\varepsilon}{2} r(d) e^{-\theta a} (1 - p(d) e^{-\theta a}) + \varepsilon (1 - p(c) e^{-\theta(ab+(1-\alpha)a)}) q(d) e^{-\theta(aa+(1-\alpha)b)}. \quad (16)$$

By solving  $dJ/db = 0$ , we get that  $b$  again depends linearly on  $a$ ,

$$b^*(a) = \frac{1}{\alpha\theta} \ln \frac{p(c)}{1-\alpha} - a \frac{1-\alpha}{\alpha}.$$

Substituting in (16), we obtain

$$J = \frac{1-\varepsilon}{2} p(d) e^{-\theta a} (1 - q(d) e^{-\theta a}) + \frac{1-\varepsilon}{2} r(d) e^{-\theta a} (1 - p(d) e^{-\theta a})$$

$$+\varepsilon\alpha q(d) \left(\frac{1-\alpha}{p(c)}\right)^{\frac{1-\alpha}{\alpha}} e^{-\theta a^{\frac{2\alpha-1}{\alpha}}}.$$

After changing variable,  $z \equiv e^{-\theta a}$ ,

$$J = c_1 z - c_2 z^2 + c_3 z^{\frac{2\alpha-1}{\alpha}}, \tag{17}$$

where  $c_1 = \frac{1-\varepsilon}{2}(p(d) + r(d))$ ,  $c_2 = \frac{1-\varepsilon}{2}p(d)(q(d) + r(d))$  and  $c_3 = \varepsilon\alpha r(d) \left(\frac{1-\alpha}{p(c)}\right)^{\frac{1-\alpha}{\alpha}}$ .

In the particular case when  $\alpha = 1/2$ ,  $J$  is maximized at  $z = c_1/2c_2$ . That is,

$$a = \begin{cases} \frac{1}{\theta} \ln \frac{2p(d)(q(d)+r(d))}{p(d)+r(d)}, & \text{if } q(d) + r(d) > \frac{1}{2} \left(1 + \frac{r(d)}{p(d)}\right) \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

In the particular case when  $\alpha = 1$ ,  $J$  is maximized at  $z = (c_1 + \varepsilon r(d))/2c_2$ . That is,

$$a = \begin{cases} \frac{1}{\theta} \ln \frac{2(1-\varepsilon)p(d)(q(d)+r(d))}{2\varepsilon r(d)+(1-\varepsilon)(p(d)+r(d))}, & \text{if } q(d) + r(d) > \frac{1}{2} \left(1 + \frac{r(d)}{p(d)} \frac{1+\varepsilon}{1-\varepsilon}\right) \\ 0, & \text{otherwise} \end{cases}. \tag{19}$$

Summarizing, the two-value ability distribution leads to the following strategy:

- the team whose ability is at least as high as of the other team (case  $x = c$ ), submits the bid that maximizes (13);
- the team whose ability is at most as high as of the other team (case  $x = d$ ), submits the bid that maximizes (17);
- in the particular case  $\alpha = 1/2$ , the team submits a bid that depends neither on  $\varepsilon$ , nor on the ability of the other team (see (14) and (18)), a dominant strategy.
- in the particular case  $\alpha = 1$ , a team submits a bid that depends on  $\varepsilon$ , but does not depend on the ability of the other team (see (15) and (19)), again a dominant strategy.
- Table 1 compares the bids of the two rules.

### 2.2.2 Exponential distribution of abilities

By continuing the case considered in Section 2.1.4, we assume that the beliefs of the teams with regard to the abilities of the other team are distributed exponentially,  $X, Y \sim \text{Exp}(\lambda)$  and seek a solution of problem (11) in the linear form (8). The objective is

$$J = \lambda e^{-t\theta x} \left( (1 - e^{-rx}) \left( -\frac{e^{-(\lambda+t\theta)x}}{\lambda+t\theta} + \frac{e^{-(\lambda+p+t\theta)x}}{\lambda+p+t\theta} + \frac{e^{-\lambda x}}{\lambda+t(1-\alpha)\theta} \right) + (1 - e^{-px}) \left( \frac{1 - e^{-(\lambda+q+t\theta)x}}{\lambda+q+t\theta} + \frac{e^{t(1-\alpha)\theta x} - e^{-\lambda x}}{\lambda+t(1-\alpha)\theta} - \frac{1 - e^{-(\lambda+t\theta)x}}{\lambda+t\theta} \right) \right) \tag{20}$$

We maximize (20) numerically for the parameters presented in Section 2.1.4, and plot the results in Figures 1 and 2.

**2.2.3 Uniform distribution of abilities**

In a uniform case,

$$X, Y \sim U[0,1]$$

the objective is

$$J = \int_0^x p(x)e^{-\theta(atx+(1-\alpha)ty)}(1 - q(y)e^{-\theta(aty+(1-\alpha)tx)})dy + \int_x^1 (1 - p(y)e^{-\theta(aty+(1-\alpha)tx)})r(x)e^{-\theta(atx+(1-\alpha)ty)}dy. \tag{21}$$

We numerically maximize (21) for  $\theta = 0.1, \alpha = 0.5, p = 1.8, r = 1.6, q = 1.4$  The optimal bid and the maximum probability of success as a function of  $x$  are presented in Figure 3.

Table 1. The bid under Rule 1 ( $a_1^*$ ) vs. the bid under Rule 2 ( $a_2^*$ ).

	$\alpha = 1/2$	$\alpha = 1$
$x = c$	$a_1^* > a_2^*$ if $q(c) + r(c) > 1 + \frac{T + \sqrt{T^2 + 4TS}}{2S}$ , where $S = \left(\frac{\varepsilon}{(1-\varepsilon)(1-q(d))}\right)^2$ and $T = \frac{2p(c)}{p(c)+r(c)}$ $a_1^* < a_2^*$ , if $\frac{1}{2}\left(1 + \frac{r(c)}{p(c)}\right) < q(c) + r(c) < 1 + \frac{T + \sqrt{T^2 + 4TS}}{2S}$ $a_1^* = a_2^* = 0$ , if $0 < q(c) + r(c) < \frac{1}{2}\left(1 + \frac{r(c)}{p(c)}\right)$	$a_1^* > a_2^*$ if $q(c) + r(c) > 1 + \frac{\varepsilon}{2(1-\varepsilon)(1-q(d))}$ , $a_1^* < a_2^*$ , if $\frac{1}{2}\left(1 + \frac{r(c)}{p(c)}\right) < q(c) + r(c) < 1 + \frac{\varepsilon}{2(1-\varepsilon)(1-q(d))}$ $a_1^* = a_2^* = 0$ , if $0 < q(c) + r(c) < \frac{1}{2}\left(1 + \frac{r(c)}{p(c)}\right)$
$x = d$	$a_1^* > a_2^*$ , if $q(d) + r(d) > 1$ $a_1^* < a_2^*$ , if $\frac{1}{2}\left(1 + \frac{r(d)}{p(d)}\right) < q(d) + r(d) < 1$ $a_1^* = a_2^* = 0$ , if $0 < q(d) + r(d) < \frac{1}{2}\left(1 + \frac{r(d)}{p(d)}\right)$	$a_1^* > a_2^*$ , if $q(d) + r(d) > 1$ $a_1^* < a_2^*$ , if $\min\left\{1, \frac{1}{2}\left(1 + \frac{r(d)}{p(d)}\frac{1+\varepsilon}{1-\varepsilon}\right)\right\} < q(d) + r(d) < 1$ $a_1^* = a_2^* = 0$ , otherwise

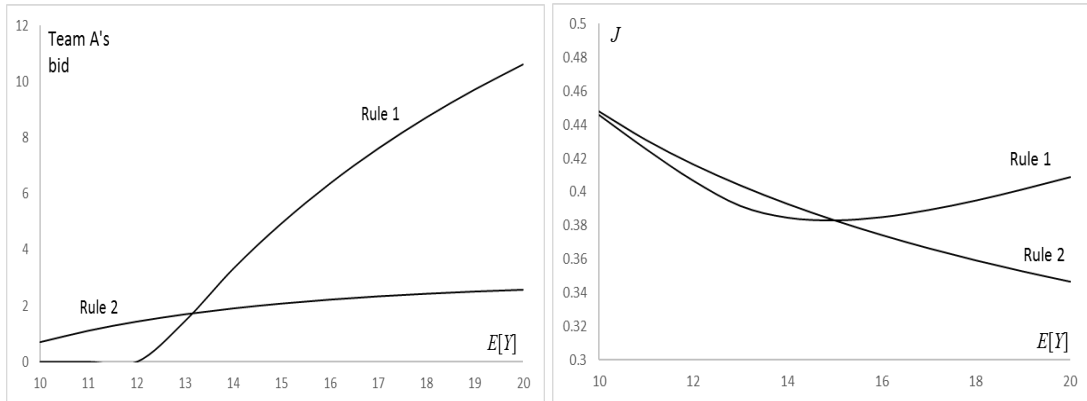


Figure 1. Team A's bid (left side) and the maximum probability of success (right side) of a team with ability  $x = 17$ , as a function of the expected ability of the other team for the two rules and exponential distribution of abilities.

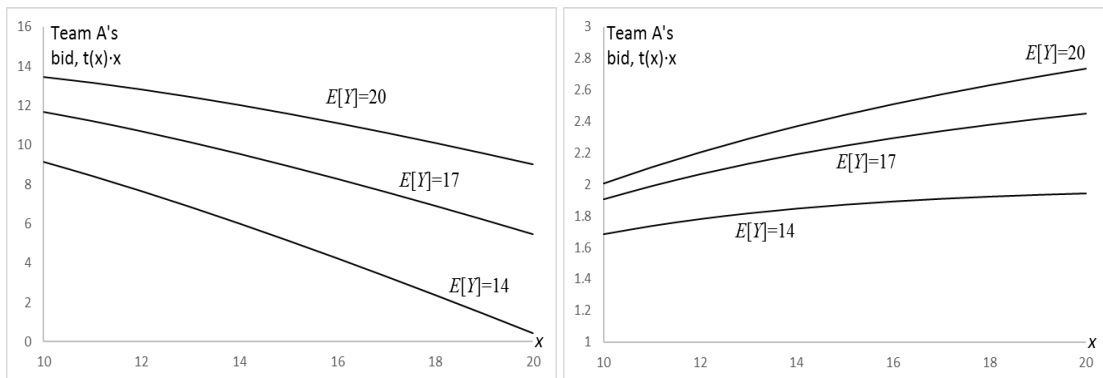


Figure 2. Team A's bid, as a function of its own ability (left side for Rule 1, right side for Rule 2) and exponential distribution of abilities.

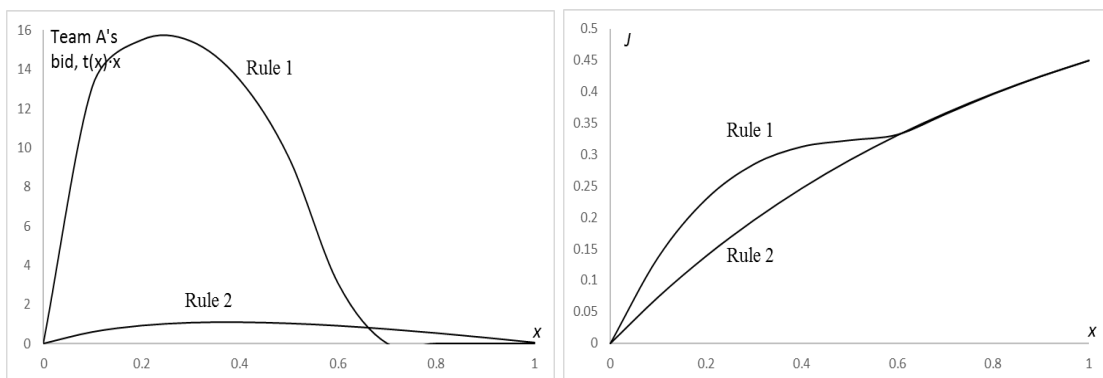


Figure 3. Team A's bid (left side) and the probability of success (right side), as a function of its own ability for the two rules and uniform distribution of abilities.

### 3 Concluding remarks

The auction with positive externalities we model turns out not to be simple to analyze, and we could do so only for special cases, and occasionally only numerically. Nevertheless, various mechanisms are used in practice without knowledge of the optimal policy or equilibrium. We believe our scheme is, in principle, implementable. Presumably, teams who have a player (players) who is good in free kicks from, say, 16-25m will bid higher than ones who have no player with such talent. If Rule 2 is used, that is likely to require the team, which lost the bid, to kick from a distance from which they are not very good, introducing another positive externality.

If one would try to apply the bidding approach to overtimes of American or Canadian football games (Granot and Gerchak 2014), the bids would be for the kickoff spots of the opposing team. If team A bids that team B will kick off from, say, 30 yards (from its own goal line) and team B bids that A will kick off from, say, 25 yards, the overtime period starts with B kicking off from 30 yards, and team A thus being the first on offence. From then on the game continues as usual. That is, if team A fails to score or is intercepted, team B takes over; if team A scores a touchdown it wins the game; if it scores a field goal, then, by current NFL rules, the other team is given one possession to match or better the field goal. Other football leagues' overtime rules award each team at least one possession.

### References

- [1] Brams, S. J. and Sanderson, Z. N. (2013) *Why you shouldn't use a coin toss for overtime*. + Plus Magazine.
- [2] Granot, D. and Gerchak, Y. (2014) *An auction with positive externality and possible application to overtime rules in football, soccer and chess*. Operations Research Letters **42**, 12-15.
- [3] Kocher, M. G., M. V. Lenz and Sutter, M. (2012) *Psychological pressure in competitive environments: New evidence from randomized natural experiments*. Management Science **58**, 1585-1591.
- [4] Palacio-Huerta, I. and Apesteguia, J. (2010) *Psychological pressure in competitive environments: evidence from randomized natural experiment*. American Economic Review **100**, 2548-2564.

# On the Dependency of Soccer Scores - A Sparse Bivariate Poisson Model for the UEFA European Football Championship 2016

A. Groll\* and T. Kneib\*\* and A. Mayr† and G. Schaubberger‡

\*Department of Statistics, Georg-August-University Goettingen: agroll@uni-goettingen.de

\*\*Department of Statistics, Georg-August-University Goettingen: tkneib@uni-goettingen.de

†IMBE, Friedrich-Alexander-University Erlangen-Nuernberg: andreas.mayr@fau.de

‡Department of Statistics, Ludwig-Maximilians-University Munich: gunther@stat.uni-muenchen.de

## Abstract

When analyzing and modeling the results of soccer matches, one important aspect is to account for the correct dependence of the scores of two competing teams. Several studies have found that, marginally, these scores are moderately negatively correlated. Even though many approaches that analyze the results of soccer matches are based on two (conditionally) independent pairwise Poisson distributions, a certain amount of (mostly negative) dependence between the scores of the competing teams can simply be induced by the inclusion of covariate information of both teams in a suitably structured linear predictor. One objective of this article is to analyze if this type of modeling is appropriate or if additional explicit modeling of the dependence structure for the joint score of a soccer match needs to be taken into account. Therefore, a specific bivariate Poisson model for the two numbers of goals scored by national teams competing in UEFA European football championship matches is fitted to all matches from the three previous European championships, including covariate information of both competing teams. A boosting approach is then used to select the relevant covariates. Based on the estimates, the tournament is simulated 1,000,000 times to obtain winning probabilities for all participating national teams.

## 1 Introduction

Many approaches that analyze and predict the results of soccer matches are based on two (conditionally) independent Poisson distributions. Both numbers of goals scored in single soccer matches are modeled separately, assuming that each score follows its own Poisson distribution, see, e.g., Lee (1997) or Dyte and Clarke (2000). For example, Dyte and Clarke (2000) predict the distribution of scores in international soccer matches, treating each team's goals as conditionally independent Poisson variables depending on two influence variables, the FIFA ranking of each team and the match venue. Poisson regression is used to estimate parameters for the model and based on these parameters the matches played during the 1998 FIFA World Cup were simulated.

However, it is well-known that, marginally, the scores of two competing teams in a soccer match are correlated. One of the first works investigating the topic of dependency between scores of competing soccer

teams is the fundamental article of Dixon and Coles (1997). There it has been shown that the joint distribution of the scores of both teams cannot be well represented by the product of two independent marginal Poisson distributions of the home and away teams. They suggest to use an additional term to adjust for certain under- and overrepresented match results. Along these lines, Rue and Salvesen (2000) propose a similarly adjusted Poisson model with some additional modifications. After all, it needs to be noted that the findings in Dixon and Coles (1997) are based on the marginal distributions and, hence, only hold for models where the predictors of both scores are uncorrelated. However, the model proposed by Dixon and Coles (1997) includes team-specific attack and defense ability parameters and then uses independent Poisson distributions for the numbers of goals scored, conditioned on these ability parameters. Therefore, the linear predictor for the number of goals of a specific team depends both on parameters of the team itself and its competitor. Groll et al. (2015) have already pointed out that when fitting exactly the same model to FIFA World Cup data the estimates of the attack and defense abilities of the teams are negatively correlated. Therefore, although (conditionally) independent Poisson distributions are used for the scores in one match, the linear predictors and, accordingly, the predicted outcomes are (negatively) correlated.

These findings already indicate that up to a certain amount the dependence between the scores of two competing teams can simply be displayed by the inclusion of the covariate information of both teams. For example, Groll and Abedieh (2013) use a pairwise (conditionally independent) Poisson model for the number of goals scored by national teams in the single matches of the UEFA European football championship (EURO), but incorporate several potential influence variables of *both* competing teams in the linear predictors of the single Poisson distributions together with additional team-specific random effects. Furthermore, in order to additionally account for the matched-pair design, they include a second match-specific random intercept, following Carlin et al. (2005), which is assumed to be independent from the team-specific random intercept. However, it turns out that this additional random intercept is very small ( $< 1 \cdot 10^{-5}$ ) and, hence, can be ignored. This provides further evidence that if highly informative covariates of both competing teams are included into the linear predictor structure of both independent Poisson distributions, this might already appropriately model the dependence structure of the match scores.

These results are further confirmed in Groll et al. (2015). Following Groll and Abedieh (2013), an  $L_1$ -regularized independent Poisson model is used on FIFA World Cup data. There, the linear predictors of the single independent Poisson components include, in addition to team-specific attack and defense abilities, the differences of several covariates of both competing teams. In an extensive goodness-of-fit analysis it is investigated if the obtained dependence structure between the linear predictors of the two scores of a match represents the actual correlations in an appropriate manner. For this purpose the correlations between the real outcomes and the model predictions are compared and it turned out that the correlations within the linear predictors for both teams competing in a match fully accounted for the correlation between the scores of those teams and that there was no need for further adjustment.

In contrast to positive correlation between two Poisson variables, which can easily be induced by using a suitable bivariate distribution (for example, the bivariate Poisson distribution proposed below), negative correlation cannot be induced that simple without any information on the two competing teams. In fact, if no covariate information about the competing teams is available, a simple model would be  $\lambda_1 = \lambda_2 = \exp(\beta_0)$ , where  $\lambda_1$  and  $\lambda_2$  represent the expectations of two independent Poisson random variables, i.e.  $E[y_k] = \lambda_k = \exp(\beta_0), k = 1, 2$ , with  $y_1$  and  $y_2$  denoting the goals scored by team 1 and team 2, respectively. Hence, the estimate  $\hat{\beta}_0$  will be chosen such that  $\exp(\hat{\beta}_0)$  reflects the overall average amount of goals scored by all teams.

In this model, the scores of competing teams are actually marginally independent and, hence, the model is not really suitable for modeling soccer scores.

One way to induce negative correlation is to decide which of the team's  $\lambda_k$  is increased, while at the same time decreasing the other team's  $\lambda_l, k \neq l, k, l \in \{1, 2\}$ . This can be done by using covariate information on the competing teams. Similar to Groll et al. (2015), let now the model be extended such that both linear predictors contain the differences of several informative covariates of both competing teams, i.e.

$$\lambda_1 = \exp(\beta_0 + (\mathbf{x}_1 - \mathbf{x}_2)^T \boldsymbol{\beta}), \quad \lambda_2 = \exp(\beta_0 + (\mathbf{x}_2 - \mathbf{x}_1)^T \boldsymbol{\beta}), \quad (1)$$

with  $\mathbf{x}_1$  and  $\mathbf{x}_2$  denoting the covariates of team 1 and team 2. Then it is obvious that, potentially, almost any magnitude of negative correlation could be achieved for the corresponding numbers of goals  $y_1$  and  $y_2$ , depending on the size of the estimated regression coefficients  $\boldsymbol{\beta}$ ; namely, if for suitably chosen signs the regression coefficients would tend to plus or minus infinity, one of the linear predictors of the competing teams would tend to plus, and the other one to minus infinity. Consequently, either  $\lambda_1$  or  $\lambda_2$  would tend to zero, while the other parameter would tend to plus infinity and, hence, one team would score zero goals, while its opponent would score infinitely many goals; this yields (accounting for the ties on all teams with zero goals) very extreme and negative correlations. To sum up, if large values for the regression parameters  $\hat{\boldsymbol{\beta}}$  are estimated, large negative magnitudes of the (marginal) correlation for the scores are obtained, even though, conditioned on the covariate information, independent Poisson distributions are assumed.

However, note that in (1) there is no possibility for introducing any type of positive correlation; for example, if two very similar (in the sense of their covariates) teams are competing, there might be some increased probability for the teams to score a similar number of goals or for the match to end in a draw. To allow for such positive correlations, a more flexible model is needed. So one major objective of this article is to analyze if the type of (conditionally) independent modeling in (1) is appropriate or if additional explicit modeling of the dependence structure for the joint score of a soccer match needs to be taken into account.

One possibility to explicitly model (positive) dependence within the Poisson framework is the bivariate Poisson distribution. One of the first works dealing with this distribution in the context of soccer data is Maher (1982). An extensive study for the use of the bivariate Poisson distribution for the modeling of soccer data is found in Karlis and Ntzoufras (2003). There, the three parameters  $\lambda_1$ ,  $\lambda_2$  and  $\lambda_3$  of the bivariate Poisson distribution are modeled by linear predictors depending on team-specific attack and defense abilities as well as team-specific home effect parameters. In particular, it is illustrated how also the third parameter  $\lambda_3$ , which represents the covariance between the two scores, can be explicitly structured in terms of covariate effects (here: simply team-specific home effects). We adopt this approach in the present work and extend the linear predictors of the three parameters  $\lambda_1$ ,  $\lambda_2$  and  $\lambda_3$  of the bivariate Poisson distribution to include several covariate effects. We set up a specific bivariate Poisson model for the two numbers of goals scored by national teams competing in EURO tournaments including covariate information of both competing teams.

In addition to the bivariate Poisson, also alternative approaches to handle correlation of soccer matches have been proposed in the literature. For example, McHale and Scarf (2006, 2011) model the dependence by using bivariate discrete distributions and by specifying a suitable family of dependence copulas.

A second objective of this work is to provide predictions of the EURO 2016. Therefore, the proposed specific bivariate Poisson model is fitted to all matches from the three previous EUROS 2004 - 2012, including covariate information of both competing teams. A suitable boosting approach is then used to select a small set of relevant covariates. Based on the obtained estimates, the EURO 2016 tournament is simulated 1,000,000 times to obtain winning probabilities for all participating national teams.



The rest of the article is structured as follows. In Section 2 we introduce the bivariate Poisson model for soccer data. The boosting methodology for fitting the bivariate Poisson model for the number of goals is introduced in Section 2.4. Next, we present a list of several possible influence variables in Section 3.1 that will be considered in our regression analysis. Based on the boosting approach a selection of these covariates is determined yielding a sparse model, which is then used in Section 4 for the prediction of the EURO 2016.

## 2 A Bivariate Poisson Model for Soccer Data

In the present section, we set up a specific bivariate Poisson model for the two numbers of goals scored by national teams competing in EURO tournaments including covariate information of both competing teams.

### 2.1 The Bivariate Poisson Distribution

In the following, we consider random variables  $X_k, k = 1, 2, 3$ , which follow independent Poisson distributions with parameters  $\lambda_k > 0$ . Then the random variables  $Y_1 = X_1 + X_3$  and  $Y_2 = X_2 + X_3$  follow a joint bivariate Poisson distribution, with a joint probability function

$$\begin{aligned} P_{Y_1, Y_2}(y_1, y_2) &= P(Y_1 = y_1, Y_2 = y_2) \\ &= \exp(-(\lambda_1 + \lambda_2 + \lambda_3)) \frac{\lambda_1^{y_1}}{y_1!} \frac{\lambda_2^{y_2}}{y_2!} \sum_{k=0}^{\min(y_1, y_2)} \binom{y_1}{k} \binom{y_2}{k} k! \left( \frac{\lambda_3}{\lambda_1 \lambda_2} \right)^k. \end{aligned} \quad (2)$$

Note that in the following sections  $Y_1$  and  $Y_2$  will represent the numbers of goals scored by two soccer teams competing in a soccer match. The bivariate Poisson distribution allows for dependence between the two random variables  $Y_1$  and  $Y_2$ . Marginally each random variable follows a univariate Poisson distribution with  $E[Y_1] = \lambda_1 + \lambda_3$  and  $E[Y_2] = \lambda_2 + \lambda_3$ . Moreover, the dependence of  $Y_1$  and  $Y_2$  is expressed by  $\text{cov}(Y_1, Y_2) = \lambda_3$ . If  $\lambda_3 = 0$  holds, the two variables are independent and the bivariate Poisson distribution reduces to the product of two independent Poisson distributions. The notation and usage of the bivariate Poisson distribution for modeling soccer data has been described in detail in Karlis and Ntzoufras (2003).

### 2.2 Incorporation of Covariate Information

In general, each of the three parameters  $\lambda_k, k = 1, 2, 3$ , in the joint probability function (2) of the bivariate Poisson distribution can be modeled in terms of covariates by specifying a suitable response function, similar to classical generalized linear models (GLMs). Hence, one could use, for example,  $\lambda_k = \exp(\boldsymbol{\eta}_k)$ , with linear predictor  $\boldsymbol{\eta}_k = \beta_{0k} + \mathbf{x}_k^T \boldsymbol{\beta}_k$  and response function  $h(\cdot) = \exp(\cdot)$  in order to guarantee non-negative Poisson parameters  $\lambda_k$ . The vectors  $\mathbf{x}_k = (x_{1k}, \dots, x_{pk})^T$  collect all covariate information of predictor  $k$ .

### 2.3 Re-parametrization of the Bivariate Poisson Distribution

In the context of soccer data a natural way to model the three parameters  $\lambda_k, k = 1, 2, 3$ , would be to include the covariate information of the competing teams 1 and 2 in  $\lambda_1$  and  $\lambda_2$ , respectively, and some extra information reflecting the match conditions of the corresponding match in  $\lambda_3$ . However, it is often reasonable to assume that the covariate effects  $\boldsymbol{\beta}_k, k = 1, 2$ , should be the same for both competing teams. In particular,

this is the case for international soccer matches at e.g. FIFA World Cups or European championships, where it is usually of no relevance, which team is the first and which the second mentioned, as no home advantage (except for the host) is to be expected. Then, one obtains the model representation

$$\lambda_1 = \exp(\beta_0 + \mathbf{x}_1^T \boldsymbol{\beta}), \quad \lambda_2 = \exp(\beta_0 + \mathbf{x}_2^T \boldsymbol{\beta}), \quad (3)$$

with  $\mathbf{x}_1$  and  $\mathbf{x}_2$  denoting the covariates of team 1 and team 2. In contrast, the covariance parameter  $\lambda_3$  could generally depend on different covariates and effects, i.e.

$$\lambda_3 = \exp(\alpha_0 + \mathbf{z}^T \boldsymbol{\alpha}), \quad (4)$$

where  $\mathbf{z}$  could contain parts of the covariates  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , or their differences or completely new covariates. If instead in the linear predictors in (3) the differences of the teams' covariates are used, one obtains  $\lambda_1 = \exp(\beta_0 + (\mathbf{x}_1 - \mathbf{x}_2)^T \boldsymbol{\beta})$ ,  $\lambda_2 = \exp(\beta_0 + (\mathbf{x}_2 - \mathbf{x}_1)^T \boldsymbol{\beta})$ , or, with  $\tilde{\mathbf{x}} = \mathbf{x}_1 - \mathbf{x}_2$ , the simpler model

$$\lambda_k = \exp\left(\beta_0 + (-1)^{(k-1)} \tilde{\mathbf{x}}^T \boldsymbol{\beta}\right), \quad k = 1, 2.$$

This allows to re-parametrize the bivariate Poisson probability function from (2) in the following way:

$$\begin{aligned} P_{Y_1, Y_2}(y_1, y_2) &= P(Y_1 = y_1, Y_2 = y_2) \\ &= \exp(-(\gamma_1(\gamma_2 + \gamma_2^{-1}) + \lambda_3)) \frac{(\gamma_1 \gamma_2)^{y_1}}{y_1!} \frac{\left(\frac{\gamma_1}{\gamma_2}\right)^{y_2}}{y_2!} \sum_{k=0}^{\min(y_1, y_2)} \binom{y_1}{k} \binom{y_2}{k} k! \left(\frac{\lambda_3}{\gamma_1^2}\right)^k, \end{aligned} \quad (5)$$

with  $\lambda_1 = \gamma_1 \gamma_2$ ,  $\lambda_2 = \frac{\gamma_1}{\gamma_2}$ . The new parameters  $\gamma_1, \gamma_2$  are then given as functions of the linear predictors  $\gamma_1 = \exp(\beta_0)$ ,  $\gamma_2 = \exp(\tilde{\mathbf{x}}^T \boldsymbol{\beta})$ , with  $\tilde{\mathbf{x}} = \mathbf{x}_1 - \mathbf{x}_2$  denoting the difference of both teams' covariates.

As before, we set  $\lambda_3 = \exp(\alpha_0 + \mathbf{z}^T \boldsymbol{\alpha})$ . In the current analysis, we base the linear predictor of  $\lambda_3$  in general on the same covariate differences. However, as we generally don't want to prefer any specific direction for the effects of these differences, we use their absolute value and set  $\lambda_3 = \exp(\alpha_0 + |\tilde{\mathbf{x}}|^T \boldsymbol{\alpha})$ , where  $|\tilde{\mathbf{x}}| = (|x_{11} - x_{21}|, \dots, |x_{1p} - x_{2p}|)^T$ . Hence, for two competing teams with similar covariate values, an additional potential positive correlation could be induced by a positive value for  $\alpha_0$ , which at the same time could be reduced for teams with rather different covariate realizations by fitting negative covariate effects  $\boldsymbol{\alpha}$ .

At this point it is also worth mentioning that if the variable selection procedure introduced in the next subsection would produce estimates  $\hat{\alpha}_1 = \hat{\alpha}_2 = \dots = 0$ , i.e. exclude all covariate effects of the covariance parameter  $\lambda_3$  from the model, and at the same time estimate a large negative value for  $\hat{\alpha}_0$ , one would actually obtain model (1) from the introduction, with two (conditionally) independent Poisson distributed scores.

Note that the re-parametrization presented in this section is necessary in order to guarantee that the boosting algorithm introduced in the next section, which iteratively cycles through the different predictors of the three distribution parameters, produces equal covariate effects  $\hat{\boldsymbol{\beta}}$  for both competing teams.

## 2.4 Estimation

We apply a statistical boosting algorithm to estimate the linear predictors for  $\gamma_1$ ,  $\gamma_2$  and  $\lambda_3$ . The concept of boosting emerged from the field of machine learning (Freund and Schapire, 1996) and was later adapted to estimate predictors for statistical models (Friedman et al., 2000; Friedman, 2001). Main advantages of

statistical boosting algorithms are their flexibility for high-dimensional data and their ability to incorporate variable selection in the fitting process (Mayr et al., 2014a). Furthermore, due to the modular nature of the algorithm, they are relatively easy to extend to new regression settings (Mayr et al., 2014b). The aim of the algorithm is to find estimates for the predictors

$$\exp(\widehat{\eta}_{\gamma_1}) = \exp(\widehat{\beta}_0) = \widehat{\gamma}_1, \quad (6)$$

$$\exp(\widehat{\eta}_{\gamma_2}) = \exp(\widehat{\mathbf{x}}^T \widehat{\boldsymbol{\beta}}) = \widehat{\gamma}_2, \quad (7)$$

$$\exp(\widehat{\eta}_{\lambda_3}) = \exp(\widehat{\alpha}_0 + |\widehat{\mathbf{x}}|^T \widehat{\boldsymbol{\alpha}}) = \widehat{\lambda}_3 \quad (8)$$

that optimize the multivariate likelihood of  $L(Y_1, Y_2, \gamma_1, \gamma_2, \lambda_3) := P_{Y_1, Y_2}(y_1, y_2)$  with  $P_{Y_1, Y_2}(y_1, y_2)$  from Equation (5), leading to the optimization problem

$$(\widehat{\eta}_{\gamma_1}, \widehat{\eta}_{\gamma_2}, \widehat{\eta}_{\lambda_3}) = \underset{(\widehat{\eta}_{\gamma_1}, \widehat{\eta}_{\gamma_2}, \widehat{\eta}_{\lambda_3})}{\operatorname{argmax}} \mathbb{E} [L(Y_1, Y_2, \exp(\widehat{\eta}_{\gamma_1}), \exp(\widehat{\eta}_{\gamma_2}), \exp(\widehat{\eta}_{\lambda_3}))].$$

The algorithm cycles through the different predictors and carries out one boosting iteration for each. In every boosting iteration, only one component of the corresponding predictor is selected to be updated, leading to automated variable selection for the covariates. For more on boosting for multiple dimensions see Schmid et al. (2010) and Mayr et al. (2012). Let the data now be given by  $(y_{1i}, y_{2i}, \widehat{\mathbf{x}}_i^T), i = 1, \dots, n$ . Then, the following cyclic boosting algorithm is applied:

### (1) Initialize

Initialize the additive predictors with starting values, e.g.  $\widehat{\eta}_{\gamma_1}^{[0]} := \log(\bar{y}_1); \widehat{\eta}_{\gamma_2}^{[0]} := 0; \widehat{\eta}_{\lambda_3}^{[0]} := \log(0.0001)$ . Set iteration counter to  $m := 1$ .

### (2) Boosting for $\gamma_1$

Increase iteration counter:  $m := m + 1$

If  $m > m_{\text{stop}\gamma_1}$  set  $\widehat{\eta}_{\gamma_1}^{[m]} := \widehat{\eta}_{\gamma_1}^{[m-1]}$  and skip step (2).

Compute  $\mathbf{u}^{[m]} = \left( \frac{\partial}{\partial \eta_{\gamma_1}} L(y_{1i}, y_{2i}, \exp(\widehat{\eta}_{\gamma_1}^{[m-1]}), \exp(\widehat{\eta}_{\gamma_2}^{[m-1]}), \exp(\widehat{\eta}_{\lambda_3}^{[m-1]})) \right)_{i=1, \dots, n}$

Estimate  $\widehat{\beta}_0^{[m]}$  for  $\mathbf{u}^{[m]}$  by  $\widehat{\beta}_0^{[m]} = \bar{u}^{[m]}$ .

Update  $\widehat{\eta}_{\gamma_1}^{[m]}$  with  $\widehat{\beta}_0^{[m]} := \widehat{\beta}_0^{[m-1]} + \nu \cdot \widehat{\beta}_0^{[m]}$ , where  $\nu$  is a small step length (e.g.,  $\nu = 0.1$ )

### (3) Boosting for $\gamma_2$

If  $m > m_{\text{stop}\gamma_2}$  set  $\widehat{\eta}_{\gamma_2}^{[m]} := \widehat{\eta}_{\gamma_2}^{[m-1]}$  and skip step (3).

Compute  $\mathbf{u}^{[m]} = \left( \frac{\partial}{\partial \eta_{\gamma_2}} L(y_{1i}, y_{2i}, \exp(\widehat{\eta}_{\gamma_1}^{[m]}), \exp(\widehat{\eta}_{\gamma_2}^{[m-1]}), \exp(\widehat{\eta}_{\lambda_3}^{[m-1]})) \right)_{i=1, \dots, n}$

Fit all components of  $\widehat{\mathbf{x}}$  separately to  $\mathbf{u}^{[m]}$ , leading to  $\widehat{\beta}_1^{[m]}, \dots, \widehat{\beta}_p^{[m]}$ .

Select component  $j^*$  that best fits  $\mathbf{u}^{[m]}$  with

$$j^* = \operatorname{argmin}_{1 \leq j \leq p} \sum_{i=1}^n (u_i^{[m]} - \widehat{\beta}_j^{[m]} x_j)^2$$

Update  $\hat{\eta}_{\gamma_2}^{[m]}$  with  $\hat{\beta}_{j^*}^{[m]} = \hat{\beta}_{j^*}^{[m-1]} + \nu \cdot \hat{\beta}_{j^*}^{[m]}$ , keeping all other components fixed.

(4) **Boosting for  $\lambda_3$**

If  $m > m_{\text{stop}\lambda_3}$  set  $\hat{\eta}_{\lambda_3}^{[m]} := \hat{\eta}_{\lambda_3}^{[m-1]}$  and skip step (4).

Compute  $\mathbf{u}^{[m]} = \left( \frac{\partial}{\partial \eta_{\lambda_3}} L(y_{1i}, y_{2i}, \exp(\hat{\eta}_{\gamma_1}^{[m]}), \exp(\hat{\eta}_{\gamma_2}^{[m]}), \exp(\hat{\eta}_{\lambda_3}^{[m-1]})) \right)_{i=1, \dots, n}$

Fit all components of  $|\tilde{\mathbf{x}}|$  separately to  $\mathbf{u}^{[m]}$ , leading to  $\hat{\alpha}_0^{[m]}, \dots, \hat{\alpha}_p^{[m]}$ .

Select component  $j^*$  that best fits  $\mathbf{u}^{[m]}$  with

$$j^* = \operatorname{argmin}_{0 \leq j \leq p} \sum_{i=1}^n (u_i^{[m]} - \hat{\alpha}_j^{[m]} z_j)^2.$$

Update  $\hat{\eta}_{\lambda_3}^{[m]}$  with  $\hat{\alpha}_{j^*}^{[m]} = \hat{\alpha}_{j^*}^{[m-1]} + \nu \cdot \hat{\alpha}_{j^*}^{[m]}$ , keeping all other components fixed.

**Iterate** steps (2) to (4) until  $m \geq \max(m_{\text{stop}\gamma_1}, m_{\text{stop}\gamma_2}, m_{\text{stop}\lambda_3})$

Note that the presented algorithm reflects the structure for our re-parametrization of the bivariate Poisson distribution, but could also be easily adapted to estimate  $\hat{\eta}_{\lambda_k}$  corresponding to the original parameters  $\lambda_k, k = 1, 2, 3$ . Furthermore, we focused on linear predictors in our approach, however, the algorithm's structure stays the same if non-linear base-learners are applied to estimate additive predictors.

The main tuning parameters of the algorithm are the stopping iterations for the different predictors. They display the typical trade-off between small models with small variance and larger models with higher risk of overfitting. The best combination of stopping iterations  $(m_{\text{stop}\gamma_1}, m_{\text{stop}\gamma_2}, m_{\text{stop}\lambda_3})$  is typically chosen via cross-validation or resampling procedures or by optimizing the underlying likelihood on separate test data. The specification of the step length  $\nu$  is of minor importance as long as it is chosen small enough, it mainly affects the convergence speed (Schmid and Hothorn, 2008). The algorithm is implemented with the R add-on package `gamboostLSS` (Mayr et al., 2012; Hofner et al., 2016).

### 3 Application

In the following, the proposed model is applied to data from the previous EUROS 2004-2012 and is then used to predict the UEFA European championship 2016 in France.

#### 3.1 Data

In this section a description of the covariates is given that are used (in the form of differences) in the bivariate Poisson regression model introduced in the previous sections. As most of these variables have already been used in Groll and Abedieh (2013) a more detailed description is found there. Several of the variables contain information about the recent performance and sportive success of national teams, as it is reasonable to assume that the current form of a national team at the start of an European championship has an influence on the team's success in the tournament, and thus on the goals the team will score. Besides these sportive variables, also most recent results of economic factors, such as a country's GDP and population size, are taken into account. Furthermore, variables are incorporated that describe the structure of a team's squad.

### Economic Factors:

- *GDP<sup>1</sup> per capita*. The GDP per capita represents the economic power and welfare of a nation. Hence, countries with great prosperity might tend to focus more on sports training and promotion programs than poorer countries. The GDP per capita (in US Dollar) is publicly available on the website of The World Bank (see <http://data.worldbank.org/indicator/NY.GDP.PCAP.CD>).
- *Population<sup>2</sup>*. In general, larger countries have a deeper pool of talented soccer players from which a national coach can recruit the national team squad. Hence, the population size might have an influence on the playing ability of the corresponding national team. However, as this potential effect might not hold in a linear relationship for arbitrarily large numbers of populations and instead might diminish (compare Bernhard and Busse, 2004), the logarithm of the quantity is used.

### Sportive Factors:

- *Home advantage*. There exist several studies that have analyzed the existence of a home advantage in soccer (see, for example, Pollard and Pollard, 2005, Pollard, 2008 and Brown et al., 2002 for FIFA World Cups or Clarke and Norman, 1995, for the English Premier league). Hence, there might also exist a home effect in European championships. For this reason a dummy variable is used, indicating if a national team belongs to the organizing countries.
- *ODDSET odds*. The analyses in Groll and Abedieh (2013) and Groll and Abedieh (2014) indicate that bookmakers' odds play an important role in the modeling of international soccer tournaments such as the EURO as they contain a lot of information with respect to the success of soccer teams. They include the bookmakers' expertise and cover big parts of the team specific information and market appreciation with respect to which teams are amongst the tournament's favorites. For the EUROs from 2004 to 2012 the 16 odds of all possible tournament winners before the start of the corresponding tournament have been obtained from the German state betting agency ODDSET.
- *Market value*. The market value recently has gained increasing attention and importance in the context of predicting the success of soccer teams (see, for example, Gerhards and Wagner 2008, 2010; Gerhards et al. 2012, 2014). Estimates of the teams' average market values can be found on the webpage <http://www.transfermarkt.de><sup>3</sup>. For each national team participating in a EURO these market value estimates (in Euro) have been collected right before the start of the tournament.
- *FIFA ranking*. The FIFA ranking provides a ranking system for all national teams measuring the performance of the teams over the last four years. The exact formula for the calculation of the underlying FIFA points and all rankings since implementation of the FIFA ranking system can be found at the official FIFA website (<http://de.fifa.com/worldranking/index.html>). Since the calculation formula of the FIFA points changed after the World Cup 2006, the rankings according to FIFA points

<sup>1</sup>The GDP per capita is the gross domestic product divided by midyear population. The GDP is the sum of gross values added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products.

<sup>2</sup>In order to collect data for all participating countries at the EURO 2004, 2008 and 2012, different sources had to be used. Amongst the most useful ones are <http://www.wko.at>, <http://www.statista.com/> and <http://epp.eurostat.ec.europa.eu>. For some years the populations of Russia and Ukraine had to be searched individually.

<sup>3</sup>Unfortunately, the archive of the webpage was established not until 4th October 2004, so the average market values of the national teams that we used for the EURO 2004 can only be seen as a rough approximation, as market values certainly changed after the EURO 2004.

are used instead of the points<sup>4</sup>.

- *UEFA points*. The associations' club coefficients rankings are based on the results of each association's clubs in the five previous UEFA CL and Europa League (previously UEFA Cup) seasons. The exact formula for the calculation of the underlying UEFA points and all rankings since implementation of the UEFA ranking system can be found at the official UEFA website, compare <http://www.uefa.com/memberassociations/uefarankings/country/index.html>. The rankings determine the number of places allocated to an association (country) in the forthcoming UEFA club competitions. Thus, the UEFA points represent the strength and success of a national league in comparison to other European national leagues. Besides, the more teams of a national league participate in the UEFA CL and the UEFA Europa League, the more experience the players from that national league are able to earn on an international level. As usually a relationship between the level of a national league and the level of the national team of that country is supposed, the UEFA points could also affect the performance of the corresponding national team.

#### Factors describing the team's structure:

- *(Second) maximum number of teammates*<sup>5</sup>. If many players from one club play together in a national team, this could lead to an improved performance of the team as the teammates know each other better. Therefore, both the maximum and the second maximum number of teammates from the same club are counted and included as covariates.
- *Average age*. To include possible differences between rather old and rather young teams, the average age of all 23 players is collected from the website <http://www.transfermarkt.de>.
- *Number of Champions League (Europa League) players*<sup>5</sup>. The European club leagues are assessed to be the best leagues in the world. Therefore, the competitions between the best European teams, namely the UEFA CL and Europa League, can be seen as the most prestigious and valuable competitions on club level. As a measurement of the success of the players on club level, the number of players in the semi finals (taking place only weeks before the respective EURO) of these competitions are counted.
- *Number of players abroad*<sup>5</sup>. The national teams strongly differ in the numbers of players playing in the league of the respective country and players from leagues of other countries. For each team, the number of players playing in clubs abroad (in the season previous to the respective EURO) is counted.

#### Factors describing the team's coach.

Also covariates of the coach of the national team may have an influence on the performance of the team. Therefore, the *age* of the coach is observed together with a dummy variable<sup>6</sup>, indicating if the coach has the same *nationality* as his team or not.

<sup>4</sup>The FIFA ranking was introduced in August 1993.

<sup>5</sup>Note that this variable is not available by any soccer data provider and thus had to be collected "by hand".

<sup>6</sup>These two variables are available on several soccer data providers, see, for example, <http://www.kicker.de/>.

## 4 Bivariate Poisson Regression on the EUROS 2004 - 2012:

We now applied the boosting approach introduced in Section 2.4 with linear predictors as specified in Equations (6)-(8). The vectors of covariate differences  $\tilde{\mathbf{x}}$  and of absolute covariate differences  $|\tilde{\mathbf{x}}|$  in (7) and (8), respectively, incorporate all 15 potential influence variables from Section 3.1. An extract of the design matrix, which corresponds to the covariate differences, is presented in Table 1. For each of the linear predictors (6)-(8) the optimal number of boosting steps needed to be determined. To do so three-dimensional 10-fold cross validation has been used.

	Team 1	Team 2	Goals 1	Goals 2	Year	odds	market value	...
1	Portugal	Greece	1	2	2004	-39.0	7.85	...
2	Spain	Russia	1	0	2004	-33.5	7.67	...
3	Greece	Spain	1	1	2004	38.5	-7.58	...
4	Russia	Portugal	0	2	2004	34.0	-7.94	...
5	Spain	Portugal	0	1	2004	0.5	-0.27	...
6	Russia	Greece	2	1	2004	-5.0	-0.09	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 1: Extract of the design matrix which corresponds to the differences of the covariates.

The intercept  $\hat{\beta}_0$ , corresponding to the linear predictor of  $\hat{\gamma}_1$ , was updated several times by the boosting algorithm. Also the linear predictor of  $\hat{\gamma}_2$  was updated several times and from the set of potential influence variables only the covariates *ODDSET odds*, *market value* and *UEFA points* were chosen.

In contrast, the linear predictor of  $\hat{\lambda}_3$  was not updated. No covariates were chosen and the intercept  $\hat{\alpha}_0$  was set to a large negative value, leading to  $\hat{\lambda}_3 \approx 0$ . This reflects an important result from the modeling perspective. It shows that no additional covariance needs to be considered, if the linear predictors of the two Poisson parameters  $\lambda_1$  and  $\lambda_2$ , in our re-parametrization reflected by  $\gamma_2$ , already contain informative covariate information from both teams and, hence, already induce a certain amount of (negative) correlation. Instead, on the EURO 2004-2012 data two independent Poisson distributions can be used for the two numbers of goals of the matches, if the linear predictor of both Poisson parameters each contains covariate information (here in the form of differences) of both competing teams. Altogether, we obtained a quite simple model with the following estimates (corresponding to scaled covariate information):

- $\gamma_1 = \exp(\beta_0)$  with  $\hat{\beta}_0 = 0.176$   
 $\implies \hat{\gamma}_1 = \exp(\hat{\beta}_0) = 1.192$ ; the parameter reflects the average number of goals, if two teams with equal covariates play against each other
- $\gamma_2 = \exp((\mathbf{x}_1 - \mathbf{x}_2)^T \boldsymbol{\beta}) = \exp(\tilde{\mathbf{x}}^T \boldsymbol{\beta})$  with  $(\hat{\beta}_{odds}, \hat{\beta}_{marketvalue}, \hat{\beta}_{UEFApoints}) = (-0.120, 0.143, 0.029)$
- $\lambda_3 = \exp(\alpha_0 + |\mathbf{x}_1 - \mathbf{x}_2|^T \boldsymbol{\alpha}) = \exp(\alpha_0 + |\tilde{\mathbf{x}}|^T \boldsymbol{\alpha})$  with  $\alpha_0 = -9.21$ ,  $\boldsymbol{\alpha} = 0$   
 $\implies \lambda_3 \approx 0$ ; no (additional) covariance between scores of both teams

Hence, regarding the findings with respect to  $\lambda_3$ , we actually obtain model (1) from the introduction with two (conditionally) independent Poisson distributions with parameters  $\lambda_1 = \gamma_1 \gamma_2$  and  $\lambda_2 = \gamma_1 / \gamma_2$ . This is a very important result from the general modeling perspective with regard to soccer data: at least for the present case of European championship data, no additional modeling of the covariance seems necessary

and for suitably designed linear predictors, which are based on informative covariates, two (conditionally) independent Poisson distributions are adequate.

In the remainder of this article we will use this simple model with (conditionally) independent Poisson distributions for the prediction of the EURO 2016. Based on this final model, in the following different simulation studies were applied in order to obtain probabilities for the EURO 2016 winner. Finally, the prediction power of this simple model is investigated.

#### 4.1 Probabilities for the UEFA European Championship 2016 Winner

For each match of the EURO 2016, the final model from the previous section is used to calculate the two distributions of the scores of the two competing teams. For this purpose, for the two competing teams in a match the covariate differences of the three selected covariates *ODDSET odds*, *market value* and *UEFA points* have to be calculated in order to be able to compute an estimate of the linear predictor of the parameter  $\hat{\gamma}_2$ . Then, the match result can be drawn randomly from these predicted distributions, i.e.  $G_1 \sim \text{Poisson}(\hat{\lambda}_1)$ ,  $G_2 \sim \text{Poisson}(\hat{\lambda}_2)$ , with estimates  $\hat{\lambda}_1 = \hat{\gamma}_1 \hat{\gamma}_2$  and  $\hat{\lambda}_2 = \hat{\gamma}_1 / \hat{\gamma}_2$ . Note here that being able to draw exact match outcomes for each match constitutes an advantage in comparison to several alternative prediction approaches, as this allows to precisely follow the official UEFA rules when determining the final group standings<sup>7</sup>. If a match in the knockout stage ended in a draw, we simulated another 30 minutes of extra time using scoring rates equal to 1/3 of the 90 minutes rates, i.e. using Poisson parameters  $\hat{\lambda}_1/3$  and  $\hat{\lambda}_2/3$ . If the match then still ended in a draw, the winner was calculated simply by coin flip, reflecting a penalty shoot out.

The whole tournament was simulated 1,000,000 times. Based on these simulations, for each of the 24 participating teams probabilities to reach the next stage and, finally, to win the tournament are obtained. These are summarized in Table 2 together with the winning probabilities based on the ODDSET odds for comparison. In contrast to most other prediction approaches for the UEFA European championship favoring France (see, for example, Zeileis et al., 2016; Goldman-Sachs Economics Research, 2016), we get a neck-and-neck race between Spain and Germany, finally with better chances for Spain. The major reason for this is that in the simulations with a high probability both Spain and Germany finish their groups on the first place and then face each other in the final. In a direct duel, the model concedes Spain a thin advantage with a winning probability of 51.1% against 48.9%. The favorites Spain and Germany are followed by the teams of France, England, Belgium and Portugal. This also shows how unlikely in advance of the tournament the triumph of the Portuguese team was assessed. While the bookmaker ODDSET expected a probability of

<sup>7</sup>The final group standings are determined by the number of points. If two or more teams are equal on points after completion of the group matches, specific tie-breaking criteria are applied: if two or more teams are equal on points, the first tie-breaking criteria are matches between teams in question (1. obtained points; 2. goal difference; 3. higher number of goals scored). 4. if teams still have an equal ranking, criteria 1 to 3 are applied once again, exclusively to the matches between the teams in question to determine their final rankings. If teams still have an equal ranking, all matches in the group are considered (5. goal difference; 6. higher number of goals scored). 7. if only two teams have the same number of points, and they were tied according to criteria 1-6 after having met in the last round of the group stage, their ranking is determined by a direct penalty shoot-out (this criterion would not be used if three or more teams had the same number of points.). 7. fair play conduct (yellow card: 1 point, red card: 3 points); 8. Position in the UEFA national team coefficient ranking system.

Note that due to the augmentation from 16 to 24 teams, also the four best third-placed teams qualified for the newly introduced round-of-sixteen. For the determination of the four best third-placed teams also specific criteria are used: 1. obtained points; 2. goal difference; 3. higher number of goals scored; 4. fair play conduct; 5. Position in the UEFA national team coefficient ranking system. Note that depending on which third-placed teams qualify from groups, several different options for the round-of-sixteen have to be taken into account.



only 4.5%, after all our model assigned an increased probability of 5.5% to this event.

Note that, based on the 1,000,000 simulations, one can also determine the most probable tournament outcome. However, for the sake of brevity this is skipped here.

























		Round of 16	Quarter Finals	Semi Finals	Final	European Champion	Oddset
Spain		95.4	72.9	52.3	35.1	21.8	13.9
Germany		99.3	79.5	51.3	34.4	21.0	16.9
France		97.5	71.9	48.2	25.8	13.8	18.9
England		95.2	69.4	43.4	23.9	12.9	9.2
Belgium		93.9	58.7	32.8	18.7	9.5	7.3
Portugal		92.5	52.3	27.4	12.6	5.5	4.5
Italy		87.7	47.6	23.8	11.4	4.8	5.3
Croatia		73.2	35.3	16.8	7.3	2.7	3.2
Poland		86.0	42.2	15.6	5.5	1.6	2
Austria		79.1	34.0	13.4	4.4	1.3	2.7
Switzerland		77.9	35.8	13.3	4.3	1.2	1.6
Turkey		56.1	21.2	8.3	2.8	0.8	1.6
Wales		65.6	27.4	9.6	2.8	0.8	1.6
Russia		62.3	25.1	8.6	2.5	0.6	1.3
Ukraine		71.0	25.8	7.7	2.0	0.4	1
Iceland		61.7	20.0	6.2	1.5	0.3	1.6
Czech Rep.		42.5	13.6	4.6	1.3	0.3	1.6
Slovakia		44.5	13.6	3.6	0.8	0.2	1
Sweden		42.9	11.2	3.3	0.8	0.1	1
Ireland		41.7	10.6	3.1	0.7	0.1	1
Romania		45.6	12.3	2.8	0.5	0.1	0.8
Albania		41.3	10.4	2.2	0.4	0.1	0.8
Hungary		37.1	8.1	1.8	0.3	0.0	0.8
Nor. Ireland		10.0	1.0	0.1	0.0	0.0	0.4

Table 2: Estimated probabilities (in %) for reaching different stages in the EURO 2016 for all 24 teams based on 1,000,000 simulation runs of the tournament and winning probabilities based on ODDSET's odds.

## 4.2 Prediction power

In the following, we try to assess the performance of our model with respect to prediction. From the online bookmaker *Tipico* (<https://www.tipico.de/de/online-sportwetten/>), we collected the “three-way” odds<sup>8</sup> for all 51 matches of the EURO 2016. By taking the three quantities  $\tilde{p}_r = 1/\text{odds}_r, r \in \{1, 2, 3\}$  and by normalizing with  $c := \sum_{r=1}^3 \tilde{p}_r$  in order to adjust for the bookmaker's margins, the odds can be

<sup>8</sup>Three-way odds consider only the tendency of a match with the possible results *victory of team 1*, *draw* or *defeat of team 1* and are usually fixed some days before the corresponding match takes place.

directly transformed into probabilities using  $\hat{p}_r = \tilde{p}_r/c^9$ . On the other hand, let  $G_1$  and  $G_2$  denote the random variables representing the number of goals scored by two competing teams in a match. Then, we can compute the same probabilities by approximating  $\hat{p}_1 = P(G_1 > G_2)$ ,  $\hat{p}_2 = P(G_1 = G_2)$  and  $\hat{p}_3 = P(G_1 < G_2)$  for each of the 51 matches of the EURO 2016 using the corresponding Poisson distributions  $G_1 \sim \text{Poisson}(\hat{\lambda}_1)$ ,  $G_2 \sim \text{Poisson}(\hat{\lambda}_2)$ , where the estimates  $\hat{\lambda}_1$  and  $\hat{\lambda}_2$  are obtained by our regression model. Based on these predicted probabilities, the average probability of a correct prediction of a EURO 2016 match can be obtained. For the true match outcomes  $\omega_m \in \{1, 2, 3\}$ ,  $m = 1, \dots, 51$ , it is given by  $\bar{p}_{\text{three-way}} := \frac{1}{51} \sum_{m=1}^{51} \hat{p}_{1m}^{\delta_{1\omega_m}} \hat{p}_{2m}^{\delta_{2\omega_m}} \hat{p}_{3m}^{\delta_{3\omega_m}}$ , with  $\delta_{rm}$  denoting Kronecker's delta. The quantity  $\bar{p}_{\text{three-way}}$  serves as a useful performance measure for a comparison of the predictive power of the model and the bookmaker's odds. We obtain the quite favorable result that the predictive power of our model ( $\bar{p}_{\text{three-way}} = 42.22\%$ ) outperforms the bookmaker's odds ( $\bar{p}_{\text{three-way}} = 39.23\%$ ), especially if one has in mind that the bookmaker's odds are usually released just some days before the corresponding match takes place and, hence, are able to include the latest performance trends of both competing teams.

If one puts one's trust into the model and its predicted probabilities, the following betting strategy can be applied: for every match one would bet on the three-way match outcome with the highest expected return, which can be calculated as the product of the model's predicted probability and the corresponding three-way odd offered by the bookmakers. We applied this strategy to our model's results, yielding a return of 30.28%, when for all 51 matches equal-sized bets are placed. This is also a very satisfying result.

## 5 Concluding remarks

As several recent studies have shown that, marginally, the scores of two competing soccer teams are (moderately) negatively correlated, an important aspect for an adequate analysis of soccer data is to account for the correct dependence of scores. First, in the paper it is shown that, if suitably structured linear predictors are chosen, a certain amount of (negative) dependence between the scores can be induced, even though, conditioned on the covariate information, independent pairwise Poisson distributions are assumed. Hence, a major objective of this article was to analyze if this type of modeling is appropriate or if an additional explicit modeling of the dependence structure for the joint score of a soccer match needs to be taken into account.

For this purpose, a flexible bivariate Poisson model for the number of goals scored by soccer teams facing each other in international tournament matches is set up. As an application, the UEFA European championships 2004-2012 serve as the data basis for an analysis of the influence of several covariates on the success of national teams in terms of the number of goals they score in single matches. Procedures for variable selection based on boosting methods, implemented in the R-package `gamboostLSS`, are used.

The boosting method selected only three covariates for the two Poisson parameters  $\lambda_1$  and  $\lambda_2$ , namely the *ODDSET odds*, the *market value* and the *UEFA points*, while for the covariance parameter  $\lambda_3$  no covariates were selected and the parameter was in fact estimated to be zero. This reflects an important general result for the modeling of soccer data: on the EURO 2004-2012 data no additional (positive) covariance needs to be considered. Hence, instead of the bivariate Poisson distribution two (conditionally) independent Poisson distributions can be used, if the two corresponding Poisson parameters  $\lambda_1, \lambda_2$  already contain informative covariates from both teams, and, this way already induce a certain amount of (negative) correlation.

<sup>9</sup>The transformed probabilities only serve as an approximation, based on the assumption that the bookmaker's margins follow a discrete uniform distribution on the three possible match tendencies.

The obtained sparse model was then used for simulation of the EURO 2016. According to these simulations, Spain, Germany and France turned out to be the top favorites for winning the title, with an advantage for Spain. An analysis of the predictive power of the model yielded very satisfactory results.

A major part of the statistical novelty of the presented work lies in the combination of boosting methods with a bivariate Poisson model. While the bivariate Poisson model enables explicit modeling of the covariance structure between match scores, the boosting method allows to include many covariates simultaneously and performs automatic variable selection.

## Acknowledgement

We are grateful to Falk Barth and Johann Summerer from the ODDSET-Team for providing us all necessary odds data and to Sven Grothues from the Transfermarkt.de-Team for the pleasant collaboration.

## References

- [1] Bernard, A.B. and Busse, M.R. (2004), Who wins the olympic games: Economic development and medall totals, *The Review of Economics and Statistics* **86**(1), 413–417.
- [2] Brown, T.D., Raalte, J. L.V., Brewer, B.W., Winter, C.R., Cornelius, A.E. and Andersen, M.B. (2002), World cup soccer home advantage, *Journal of Sport Behavior* **25**, 134–144.
- [3] Carlin, J.B., Gurrin, L.C., Sterne, J. A.C., Morley, R. and Dwyer, T. (2005), Regression models for twin studies: a critical review, *International Journal of Epidemiology* **B57**, 1089–1099.
- [4] Clarke, S.R. and Norman, J.M. (1995), Home ground advantage of individual clubs in English soccer, *The Statistician* **44**, 509–521.
- [5] Dixon, M.J. and Coles, S.G. (1997), Modelling association football scores and inefficiencies in the football betting market, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **46**(2), 265–280.
- [6] Dyte, D. and Clarke, S.R. (2000), A ratings based Poisson model for World Cup soccer simulation' *Journal of the Operational Research Society* **51** (8), 993–998.
- [7] Freund, Y. and Schapire, R. (1996), Experiments with a new boosting algorithm, in *Proceedings of the Thirteenth International Conference on Machine Learning Theory*, San Francisco: Morgan Kaufmann Publishers Inc., San Francisco, CA, pp. 148–156.
- [8] Friedman, J.H. (2001), Greedy function approximation: A gradient boosting machine, *The Annals of Statistics* **29**, 1189–1232.
- [9] Friedman, J.H., Hastie, T. and Tibshirani, R. (2000), Additive logistic regression: A statistical view of boosting (with discussion), *The Annals of Statistics* **28**, 337–407.
- [10] Gerhards, J., Mutz, M. and Wagner, G.G. (2012), Keiner kommt an Spanien vorbei - außer dem Zufall, *DIW-Wochenbericht* **24**, 14–20.
- [11] Gerhards, J., Mutz, M. and Wagner, G.G. (2014), Predictable winners. market value, inequality, diversity, and routine as predictors of success in european soccer leagues, *Zeitschrift für Soziologie* **43**(3).
- [12] Gerhards, J. and Wagner, G.G. (2008), Market value versus accident - who becomes European soccer champion?, *DIW-Wochenbericht* **24**, 236–328.
- [13] Gerhards, J. and Wagner, G.G. (2010), Money and a little bit of chance: Spain was odds-on favourite of the football worldcup, *DIW-Wochenbericht* **29**, 12–15.
- [14] Goldman-Sachs Economics Research (2016), The econometrician's take on euro 2016. <http://www.goldmansachs.com/our-thinking/macro-economic-insights/euro-cup-2016/>.

- [15] Groll, A. and Abedieh, J. (2013), Spain retains its title and sets a new record - generalized linear mixed models on European football championships, *Journal of Quantitative Analysis in Sports* **9**(1), 51–66.
- [16] Groll, A. and Abedieh, J. (2014), *A study on European football championships in the glmm framework with an emphasis on UEFA champions league experience*, In New perspectives on stochastic modeling and data analysis (Bozeman, Girardin and Skiadas, Eds.) ISAST, pp. 313–321.
- [17] Groll, A., Schauburger, G. and Tutz, G. (2015), Prediction of major international soccer tournaments based on team-specific regularized Poisson regression: an application to the FIFA World Cup 2014, *Journal of Quantitative Analysis in Sports* **11**(2), 97–115.
- [18] Hofner, B., Mayr, A. and Schmid, M. (2016), gamboostLSS: An R package for model building and variable selection in the GAMLSS framework, *Journal of Statistical Software* . accepted.
- [19] Karlis, D. and Ntzoufras, I. (2003), Analysis of sports data by using bivariate poisson models, *The Statistician* **52**, 381–393.
- [20] Lee, A.J. (1997), Modeling scores in the premier league: is manchester united really the best?, *Chance* **10**, 15–19.
- [21] Maher, M.J. (1982), Modelling association football scores, *Statistica Neerlandica* **36**, 109–118.
- [22] Mayr, A., Binder, H., Gefeller, O. and Schmid, M. (2014a), The evolution of boosting algorithms - from machine learning to statistical modelling, *Methods of Information in Medicine* **53**(6), 419–427.
- [23] Mayr, A., Binder, H., Gefeller, O. and Schmid, M. (2014b), Extending statistical boosting - an overview of recent methodological developments, *Methods of Information in Medicine* **53**(6), 428–435.
- [24] Mayr, A., Fenske, N., Hofner, B., Kneib, T. and Schmid, M. (2012), Generalized additive models for location, scale and shape for high-dimensional data – a flexible approach based on boosting, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **61**(3), 403–427.
- [25] McHale, I.G. and Scarf, P.A. (2006), Forecasting international soccer match results using bivariate discrete distributions, Technical Report 322, Working paper, Salford Business School.
- [26] McHale, I.G. and Scarf, P.A. (2011), Modelling the dependence of goals scored by opposing teams in international soccer matches, *Statistical Modelling* **41**(3), 219–236.
- [27] Pollard, R. (2008), Home advantage in football: A current review of an unsolved puzzle, *The Open Sports Sciences Journal* **1**, 12–14.
- [28] Pollard, R. and Pollard, G. (2005), Home advantage in soccer: A review of its existence and causes, *International Journal of Soccer and Science Journal* **3**(1), 25–33.
- [29] Rue, H. and Salvesen, O. (2000), Prediction and retrospective analysis of soccer matches in a league, *Journal of the Royal Statistical Society: Series D (The Statistician)* **49**(3), 399–418.
- [30] Schmid, M. and Hothorn, T. (2008), Boosting additive models using component-wise P-splines, *Computational Statistics & Data Analysis* **53**, 298–311.
- [31] Schmid, M., Potapov, S., Pfahlberg, A. and Hothorn, T. (2010), Estimation and regularization techniques for regression models with multidimensional prediction functions, *Statistics and Computing* **20**, 139–150.
- [32] Zeileis, A., Leitner, C. and Hornik, K. (2016), Predictive Bookmaker Consensus Model for the UEFA Euro 2016, Working Papers 2016-15, Faculty of Economics and Statistics, University of Innsbruck.  
<http://EconPapers.repec.org/RePEc:inn:wpaper:2016-15>

# An analysis of characteristics of soccer teams using a Markov process model considering the location of the ball on the pitch

N. Hirotsu\*, K. Inoue\*\* and M. Yoshimura\*\*\*

\*Juntendo University, 1-1 Hiragagakuendai, Inzai, Chiba, Japan, nhirotsu@juntendo.ac.jp

\*\*Juntendo University, 1-1 Hiragagakuendai, Inzai, Chiba, Japan, sh1214014@juntendo.ac.jp

\*\*\*Juntendo University, 1-1 Hiragagakuendai, Inzai, Chiba, Japan, msyoshi@juntendo.ac.jp

## Abstract

In this paper, we propose statistical models of a soccer game that will be useful to provide insights of characteristics of teams, based on Markov process models considering the location of the ball on the pitch. Using these models we analyze their characteristics related to such factors as home advantage, offensive and defensive strength, in terms of goals and possession according to the location of the ball. We divide the pitch into up to 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. Using this method, we analyze the characteristics of teams based on annual data from the J-League Division 1 in 2015 and illustrate their characteristics in ways that allow us to understand their offensive and defensive strength, according to the location of the ball on the pitch. As this study is still in progress, we here show a part of result we have obtained to date.

## 1 Introduction

Evaluation of soccer teams' strength has been well studied by using mathematical or statistical models. Maher (1982) analyzed three consecutive seasons of four English football league divisions from the 1971-72 season and estimated the offensive strength for scoring goals and defensive strength against conceding goals for each team in his Poisson regression model by means of the maximum likelihood method. Lee (1997) also analyzed the 1995-96 season of the English Premier League in a similar manner. Dixon and Coles (1997) evaluated the strengths of teams in order to make profits in the betting 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 by incorporating the rate of scoring goals, given the current time and score during the game, using English league and cup football data from 1993 to 1996. Hirotsu and Wright (2002, 2003a, 2003b) proposed a Markov process model of a soccer game, together with a log-linear model representing not only the transition rates of scoring and conceding goals, but also the rates of gaining and losing possession, using explanatory variables of home advantage, offensive strength and defensive strength.

In this paper, we extend the model of Hirotsu and Wright by considering the location of the ball on the pitch, and analyze teams' characteristics related to such factors as home advantage, offensive and defensive strength, in terms of goals and possession according to the location of the ball. We use annual data from the J-League Division 1 in 2015 and show various characteristics of the

teams involved. As this study is still in progress, we here show a part of result we have obtained to date.

## 2 Method

### 2.1 Markov models of a soccer game

A soccer game can be seen as progressing through a set of stochastic transitions occurring due to a change of possession of the ball or the scoring of a goal. Hirotzu and Wright (2002, 2003a, 2003b) assumed a Markov property in these transitions and proposed a Markov process model, which seems appropriate to a soccer game as an approximation. Under this assumption, the transition time until a goal is scored or possession changes follows an exponential distribution and the number of these transitions in any given time interval follows a Poisson distribution. We here introduce the location of the ball on the pitch to the model of Hirotzu and Wright as a states.

According to the level of separation of the pitch, we here present three models, as follows.

- (1) 4-state model: This model consists of the following 4 states (Hirotzu and Wright, 2003a):
  - State  $H_G$ : Home team scores a goal;
  - State  $H_P$ : Home team is in possession of the ball;
  - State  $A_P$ : Away team is in possession of the ball;
  - State  $A_G$ : Away team scores a goal.
  
- (2) 8-state model: State  $H_G$  and  $A_G$  are same as the 4-state model. State  $H_P$  and  $A_P$  of the 4-state model are divided into the following 3 states, respectively:
  - States  $H_P^F$ ,  $H_P^M$ ,  $H_P^D$ : Home team is in possession of the ball and the ball is located in the “forward”, “midfield” and “defense” area from the standpoint of the home team, respectively;
  - States  $A_P^F$ ,  $A_P^M$ ,  $A_P^D$ : Away team is in possession of the ball and the ball is located in the “forward”, “midfield” and “defense” area from the standpoint of the home team, respectively.
  
- (3) 20-state model: State  $H_G$  and  $A_G$  are same as the 4-state model. State  $H_P^F$ ,  $H_P^M$ ,  $H_P^D$ ,  $A_P^F$ ,  $A_P^M$  and  $A_P^D$  of the 8-state model are further divided into the following 3 states, respectively:
  - State  $H_P^I$ : Home team is in possession of the ball and the ball is located in the “I” area ( $I=1, \dots, 9$ );
  - State  $A_P^I$ : Away team is in possession of the ball and the ball is located in the “I” area ( $I=1, \dots, 9$ ).

The “I” area ( $I=1, \dots, 9$ ) on the pitch is defined as shown in Figure 1.

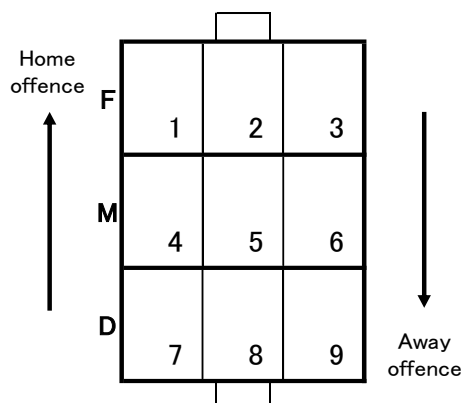
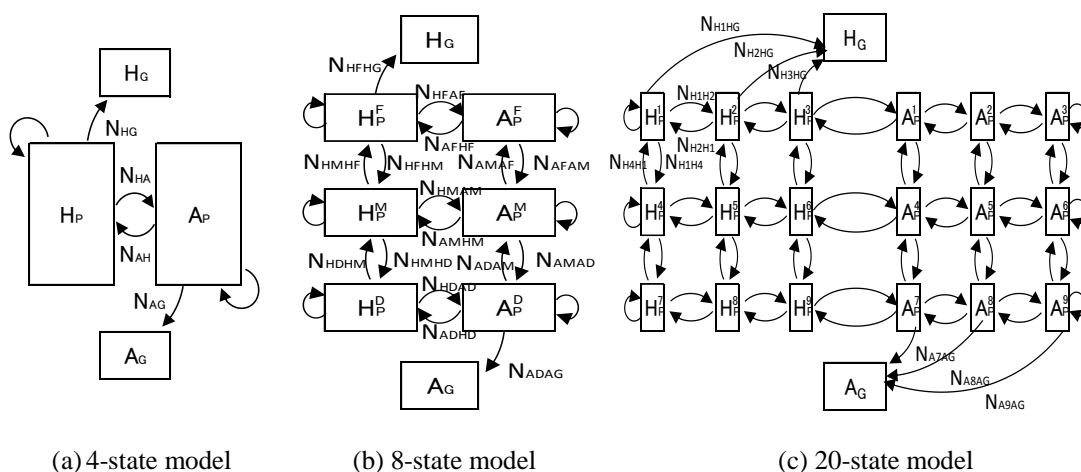


Figure 1. The areas on the pitch.



(a) 4-state model

(b) 8-state model

(c) 20-state model

Figure 2. Markov models of a soccer game.

Figure 2 shows a graphical representation of these models. We note that we omit some arrows which represent transition between states in the figure 2 (b) and 2 (c), for avoiding to be messy. In Figure 2(a),  $N_{HG}$  represents the total number of goals scored by home team in a game.  $N_{HA}$  represents the total number of changes of possession from home team to away team. Here, under the assumption of the Markov property,  $N_{HG}$  and  $N_{HA}$  follow Poisson distributions whose means are proportional to  $T_H$ , the total possession time of the home team.  $N_{AG}$ ,  $N_{AH}$  and  $T_A$  are also defined in the same manner for away team. In Figure 2(b),  $N_{HFAF}$  represents the total number of changes of possession from home team to away team in the “forward” area in the game. Here, under the assumption of the Markov property,  $N_{HFAF}$  follows Poisson distributions whose means are proportional to  $T_{HF}$ , the total possession time of home team in the “forward” area. Other total numbers of changes of possession such as  $N_{AFFH}$  are also defined in a similar manner. In Figure 2(c),  $N_{H1A2}$  represents the total number of changes of possession from home team to away team in the transition from the “1” area to the “2” area in the game. Here, under the assumption of the Markov property,  $N_{H1A2}$  follows Poisson distributions whose means are proportional to  $T_{H1}$ , the total possession time of home team in the “1” area. Other total numbers of changes of possession such as  $N_{A1H1}$  are also defined in a similar manner.

In measurement of possession time, we extracted consecutive possession time from the data provided by Data Stadium Inc. We do not include the time as possession of the ball when the following events occurred; Ball-out, Foul, Penalty, Offside, Substitution and Goal. By excluding the time of these events, we obtained a reasonable distribution of consecutive possession time. We show a histogram of a game of J1 league in the 4-state model, as an example, in Figure 3. Even though this is a simple example, we can see that the Markov assumption seems to be hold. In the case including the above events, the distribution becomes multimodal.

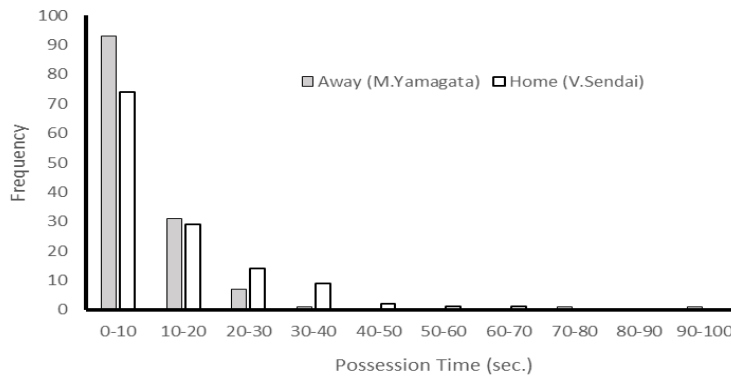


Figure 3. Histogram of possession time in a game (V.Sendai vs.M.Yamagata)

## 2.2 Estimation of factors in the models

We use a generalized linear model to estimate the factors which represent the characteristics of teams. We estimate the following factors based on each of the three models:

(1) 4-state model:

- $\lambda$  : Intercept for scoring goals;
- $\lambda_h$  : Home-team advantage for scoring goals;
- $\lambda_{\text{off}}(\mathbf{X})$ : Offensive strength of team X for scoring goals;
- $\lambda_{\text{def}}(\mathbf{X})$ : Defensive strength of team X against conceding goals;

- $\mu$  : Intercept for gaining possession;
- $\mu_h$ : Home-team advantage for gaining possession;
- $\mu_{\text{off}}(\mathbf{X})$ : Offensive strength of team X for gaining possession;
- $\mu_{\text{def}}(\mathbf{X})$ : Defensive strength of team X against losing possession;

(2) 8-state model: According to Figure 2(b), we define the following factors:

- $\lambda^F, \lambda^M, \lambda^D$ : Intercept for scoring goals from the “F”, “M” and “D” area, respectively;
- $\lambda_{\text{h}}^F, \lambda_{\text{h}}^M, \lambda_{\text{h}}^D$ : Home-team advantage for scoring goals from the “F”, “M” and “D” area, respectively;
- $\lambda_{\text{off}}^F(\mathbf{X}), \lambda_{\text{off}}^M(\mathbf{X}), \lambda_{\text{off}}^D(\mathbf{X})$ : Offensive strength of team X for scoring goals from the “F”, “M” and “D” area, respectively;
- $\lambda_{\text{def}}^F(\mathbf{X}), \lambda_{\text{def}}^M(\mathbf{X}), \lambda_{\text{def}}^D(\mathbf{X})$ : Defensive strength of team X against conceding goals from the “F”, “M” and “D” area, respectively;



- $\mu^{AFHF}$ : Intercept for gaining possession in the “F” area;
- $\mu^{AFHF}_h$ : Home-team advantage for gaining possession in the “F” area;
- $\mu^{AFHF}_{off}(X)$ : Offensive strength of team X for gaining possession in the “F” area;
- $\mu^{AFHF}_{def}(X)$ , Defensive strength of team X against losing possession in the “F” area;
  
- $\mu^{AFHM}$ : Intercept for gaining possession in the transition from the “F” area to the “M” area;
- $\mu^{AFHM}_h$ : Home-team advantage for gaining possession in the transition from the “F” area to the “M” area;
- $\mu^{AFHM}_{off}(X)$ : Offensive strength of team X for gaining possession in the transition from the “F” area to the “M” area;
- $\mu^{AFHM}_{def}(X)$ , Defensive strength of team X against losing possession in the transition from the “F” area to the “M” area;
  
- $\mu^{HFHM}$ : Intercept for keeping possession in the transition from the “F” area to the “M” area;
- $\mu^{HFHM}_h$ : Home-team advantage for keeping possession in the transition from the “F” area to the “M” area;
- $\mu^{HFHM}_{off}(X)$ : Offensive strength of team X for keeping possession in the transition from the “F” area to the “M” area;
- $\mu^{HFHM}_{def}(X)$ , Defensive strength of team X against keeping possession in the transition from the “F” area to the “M” area;

(3) 20-state model: According to Figure 2(c), we define the following factors:

- $\lambda^I$ : Intercept for scoring goals from the “I” area ( $I=1, \dots, 9$ );
- $\lambda^I_h$ : Home-team advantage for scoring goals from the “I” area ( $I=1, \dots, 9$ );
- $\lambda^I_{off}(X)$ : Offensive strength of team X for scoring goals from the “I” area ( $I=1, \dots, 9$ );
- $\lambda^I_{def}(X)$ : Defensive strength of team X against conceding goals from the “I” area ( $I=1, \dots, 9$ );
  
- $\mu^{AIHJ}$ : Intercept for gaining possession in the transition from the “I” area to the “J” area ( $I, J=1, \dots, 9$ );
- $\mu^{AIHJ}_{home}$ : Home-team advantage for gaining possession in the transition from the “I” area to the “J” area ( $I, J=1, \dots, 9$ );
- $\mu^{AIHJ}_{off}(X)$ : Offensive strength of team X for gaining possession in the transition from the “I” area to the “J” area ( $I, J=1, \dots, 9$ );
- $\mu^{AIHJ}_{def}(X)$ , Defensive strength of team X against losing possession in the transition from the “I” area to the “J” area ( $I, J=1, \dots, 9$ );
  
- $\mu^{HIHJ}$ : Intercept for keeping possession in the transition from the “I” area to the “J” area ( $I, J=1, \dots, 9$ );
- $\mu^{HIHJ}_{home}$ : Home-team advantage for keep possession in the transition from the “I” area to the “J” area ( $I, J=1, \dots, 9$ );
- $\mu^{HIHJ}_{off}(X)$ : Offensive strength of team X for keeping possession in the transition from the “I” area to the “J” area ( $I, J=1, \dots, 9$ );
- $\mu^{HIHJ}_{def}(X)$ , Defensive strength of team X against keeping possession in the transition from the “I” area to the “J” area ( $I, J=1, \dots, 9$ );

We estimate the above factors by means of the maximum likelihood method. Here, we set the possession times such as  $T_H$ ,  $T_{HF}$ , and  $T_{HI}$  as offsets in the models.

### 3 Result and discussions

#### 3.1 Data construction

In order to evaluate the characteristics of teams for goals and possession based on the above models, we need to obtain the total number of transitions between states with the total occupation time in each state for each game. Table 1 shows the total number of transitions between states in a game. We count the number based on the data provided by Data Stadium Inc., in which there are events or actions such as shoot, pass and dribble, together with the time taken by these events or actions with the location of the ball.

Table 1. Total number of transitions between states in a game (V.Sendai vs. M.Yamagata).

		H									A								
		F			M			D			F			M			D		
		1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
H	F	1	45	13		6	2				11	2		1					
		2	4	22	7	1	2	1				14	1			1	1		
		3		10	51		1	15				3	15				2		
	M	4	18	1		77	13		4	3		4	2		6	2			
		5	2	3	2	15	53	19		6		1	2			3			
		6	2	3	23	1	17	##		3	6		1	5		2	14		
D	7	1			9			12	4		1			1			3	1	
	8	2	1	2	6	5	7	7	29	8	1	1		4	1				
	9			3		1	16		4	18		1			3			3	
A	F	1	6			2		1			33	7	1	6	3		2	2	
		2				2	3	1			6	20	6	9	4	8	1		1
		3			7		5			2		1	59		3	21			1
	M	4				10	1		1	2	1	6	3	79	10	1	24	4	
		5			1		6	2		1		3		22	55	11	5	6	3
		6			1		1	10		2	3	3	9	2	16	##	6	18	
D	7							6	5				11	2		66	22	2	
	8				1			1	11	3		1	5	5	1	13	46	11	
	9					1			5	7			1	4	9	12	57		

In Table 1, we can directly see the number such that  $N_{HHH}$  is 45. We can obtain the number of transitions between states in 8-state model from Table 1, such that  $N_{HFHM}$  is  $28=6+2+1+2+1+1+15$ . For 4-state model,  $N_{HA}$  can be obtained as 113, which is the sum in the right upper block.

Table 2 lists part of the data from the games played in the 2015 season. We here look at just 3 teams (V.Sendai, M.Yamagata and S.Hiroshima) with 6 games of them in the 2015 to evaluate the characteristics of teams based on the above models. As the number of parameters are not small in the 8-state and 20 state models, we show a part of the result later in this paper.

Table 2. Part of the statistics of the 2015 season of J1 league.

(a) For 4-state model

		Goal		Transition		Time (min.)	
Home	Away	$N_{HG}$	$N_{AG}$	$N_{HA}$	$N_{AH}$	$T_H$	$T_A$
V.Sendai	M.Yamagata	2	0	113	111	21.2	27.2
M.Yamagata	V.Sendai	1	1	208	208	21.8	22.6
S.Hiroshima	V.Sendai	2	0	152	154	30.0	30.4
V.Sendai	S.Hiroshima	3	4	164	165	33.5	26.0
S.Hiroshima	M.Yamagata	5	1	174	175	28.7	27.2
M.Yamagata	S.Hiroshima	1	3	160	161	28.1	31.1

(b) For 8-state model

Home	Away	Goal		Transition		Time (min.)			
		N <sub>HG</sub>	N <sub>AG</sub>	N <sub>HMHF</sub>	N <sub>AMAD</sub>	T <sub>HF</sub>	T <sub>HM</sub>	T <sub>AM</sub>	T <sub>AD</sub>
V.Sendai	M.Yamagata	2	0	54	66	4.7	10.5	11.5	8.2
M.Yamagata	V.Sendai	1	1	48	46	6.7	9.3	9.0	6.6
S.Hiroshima	V.Sendai	2	0	62	59	7.8	13.0	17.5	6.8
V.Sendai	S.Hiroshima	3	4	79	45	9.8	18.5	10.3	6.3
S.Hiroshima	M.Yamagata	5	1	44	68	6.2	11.1	14.0	8.2
M.Yamagata	S.Hiroshima	1	3	67	47	6.8	15.7	13.2	5.8

(c) For 20-state model

Home	Away	Goal			Transition		Time (min.)				
		N <sub>H2G</sub>	N <sub>A7G</sub>	N <sub>A8G</sub>	N <sub>H5H2</sub>	N <sub>A5A8</sub>	T <sub>H2</sub>	T <sub>H5</sub>	T <sub>A5</sub>	T <sub>A7</sub>	T <sub>A8</sub>
V.Sendai	M.Yamagata	2	0	0	3	6	1.3	2.9	3.1	2.5	3.6
M.Yamagata	V.Sendai	1	0	1	5	3	1.7	2.1	2.1	1.6	2.9
S.Hiroshima	V.Sendai	2	0	0	3	6	1.8	3.5	6.5	3.6	1.5
V.Sendai	S.Hiroshima	3	1	3	6	5	2.6	6.9	3.3	2.0	2.4
S.Hiroshima	M.Yamagata	5	0	1	8	9	2.2	3.3	4.9	2.6	2.9
M.Yamagata	S.Hiroshima	1	1	2	10	3	1.9	5.9	4.2	2.2	1.7

### 3.2 Result of estimation

We now evaluate the characteristics of the three teams using the three models shown in Figure 1. The result of the offensive and defensive strengths of the three teams are represented in Table 3.

Table 3. Offensive and defensive strengths of the three teams.

(a) For 4-state model:

Intercept	Goals from H <sub>p</sub>		Transition from A <sub>p</sub> to H <sub>p</sub>	
	λ	λ <sub>h</sub>	μ	μ <sub>h</sub>
Home advantage	λ <sub>h</sub>	λ <sub>h</sub>	μ <sub>h</sub>	μ <sub>h</sub>
Team	Offence (λ <sub>off</sub> )	Defence (λ <sub>def</sub> )	Offence (μ <sub>off</sub> )	Defence (μ <sub>def</sub> )
S.Hiroshima	1.31	-0.20	-0.28	-0.25
V.Sendai	0.51	-0.33	-0.10	-0.09
M.Yamagata	0	0	0	0

(b) For 8-state model

Intercept	Goals from H <sub>p</sub> <sup>F</sup>		Transition from H <sub>p</sub> <sup>M</sup> to H <sub>p</sub> <sup>F</sup>	
	λ <sup>F</sup>	λ <sub>h</sub> <sup>F</sup>	μ <sup>M</sup>	μ <sub>h</sub> <sup>M</sup>
Home advantage	λ <sub>h</sub> <sup>F</sup>	λ <sub>h</sub> <sup>F</sup>	μ <sub>h</sub> <sup>M</sup>	μ <sub>h</sub> <sup>M</sup>
Team	Offence (λ <sub>off</sub> <sup>F</sup> )	Defence (λ <sub>def</sub> <sup>F</sup> )	Offence (μ <sub>off</sub> <sup>HMHF</sup> )	Defence (μ <sub>def</sub> <sup>HMHF</sup> )
S.Hiroshima	1.51	-0.26	-0.28	-0.19
V.Sendai	0.49	-0.55	-0.09	0.09
M.Yamagata	0	0	0	0

(c) For 20-state model

Intercept	Goals from H <sub>p</sub> <sup>2</sup>		Transition from H <sub>p</sub> <sup>5</sup> to H <sub>p</sub> <sup>2</sup>	
	λ <sup>H2</sup>	λ <sub>h</sub> <sup>H2</sup>	μ <sup>H5</sup>	μ <sub>h</sub> <sup>H5</sup>
Home advantage	λ <sub>h</sub> <sup>H2</sup>	λ <sub>h</sub> <sup>H2</sup>	μ <sub>h</sub> <sup>H5</sup>	μ <sub>h</sub> <sup>H5</sup>
Team	Offence (λ <sub>off</sub> <sup>H2</sup> )	Defence (λ <sub>def</sub> <sup>H2</sup> )	Offence (μ <sub>off</sub> <sup>H5H2</sup> )	Defence (μ <sub>def</sub> <sup>H5H2</sup> )
S.Hiroshima	1.47	-0.05	-0.54	-0.35
V.Sendai	0.61	-0.41	-0.71	-0.20
M.Yamagata	0	0	0	0

In Table 3, we note that in terms of defensive strength, if the value is negatively larger, it means stronger in the sense of conceding fewer goals or losing less possession. According to Table 3, S.Hiroshima has easily the highest goal-scoring parameter - not surprisingly, since S.Hiroshima scored 14 goals in the 6 games. However, S.Hiroshima's defensive parameter is not particularly good, reflecting that it conceded 5 goals.

In terms of gaining possession, Hiroshima is not particularly good in the transition from "M" to "F" as shown in Table 3(b) with the value of -0.28. We can see the detail, that is, Hiroshima is not particularly good in the transition from "5" to "2" as shown in Table 3(c) with the value of -0.54. These results would be interesting since they give insights which could in no way be ascertained simply by examining the data.

## 4 Conclusions and further study

In this paper, we have proposed a statistical model of a soccer game and evaluated the characteristics of teams by means of parameters representing home advantage, offensive and defensive strengths, based on Markov process models considering the location of the ball on the pitch.

As we evaluate the strength of teams considering the location of the ball on the pitch, our approach has produced more insight into the strength of teams than previous work, in which only the data for goals and possession were used in the models. Here, we just use the 6 game data from the J-League Division 1 in 2015. As this study is still in progress, we plan to present more in the conference.

## 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

- [1] Dixon, M.J. and Coles, S.G. (1997) *Modelling association football scores and inefficiencies in the football betting market*. Applied Statistics **46**, 245-280.
- [2] Dixon, M.J. and Robinson, M.E. (1998) *A birth process model for association football matches*. The Statistician **47**, 523-538.
- [3] Hirotzu, N. and 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.
- [4] Hirotzu, N. and Wright, M. (2003a) *An evaluation of characteristics of teams in association football by using a Markov process model*. The Statistician **52**, 591-602.
- [5] Hirotzu, N. and Wright, M. (2003b) *Determining the Best Strategy for Changing the Configuration of a Football Team*. Journal of the Operational Research Society **54**, 878-887.
- [6] Lee, A.J. (1997) *Modeling scores in the Premier League: is Manchester United really the best?* Chance **10**, 15-19.
- [7] Maher, M.J. (1982) *Modelling association football scores*. Statistica Neerlandica **36**, 109-118.

# Sport Strategy Optimization in Beach Volleyball– How to bound direct point probabilities dependent on individual skills

S. Hoffmeister\* and J. Rambau\*\*

University of Bayreuth, Germany, susanne.hoffmeister@uni-bayreuth.de

University of Bayreuth, Germany, joerg.rambau@uni-bayreuth.de

## Abstract

Recently, we presented a two scale approach that uses Markov Decision problems (MDPs) to answer sport strategic questions. We have implemented our method for beach volleyball by developing an appropriate gameplay-MDP and strategic-MDP for a certain strategic benchmark question. Following the two scale approach, the gameplay-MDP is simulated to generate the input probabilities of the strategic-MDP. The strategic-MDP is solved subsequently to answer the sport strategic question. We want to investigate in this paper whether the strategic-MDP probabilities can be directly computed from the gameplay-MDP or whether at least some bounds can be computed.

The derived bounds of this paper are applied to men’s beach volleyball Olympic final 2012 between Germany and Brazil and are part of the presentation *Strategy optimization in beachvolleyball – applying a two scale approach to the olympic games*.

## 1 Introduction

Markov Decision Problems (MDPs) can be used for modelling sport games and answering sport strategic questions. Some examples are: Clarke and Norman (2012) as well as Nadimpalli and Hasenbein (2013) investigate a Markov Decision Problem (MDP) for tennis games to determine when a player should challenge a line call. Hirotsu and Wright (2002) model football as a four state Markov Process and use dynamic programming to determine the optimal timing of a substitution Hirotsu and Wright (2002), the best policy for changing the configuration of a team Hirotsu and Wright (2003) or to determine under which circumstances a team may benefit from a professional foul Wright and Hirotsu (2003). Chan and Singal (2016) use an MDP to compute an optimization-based handicap system for tennis. Clarke and Norman (1998) formulate an MDP for cricket to determine whether the batsman should take an offered run when maximizing the probability that the better batsman is on strike at the start of the next over. Norman (1985) builds a more aggregated MDP for tennis games to tackle the question when to serve fast or when to serve slow at each stage of a game.

All papers mentioned so far investigate MDPs for general rules that are independent of teams and matches. Only a few papers present MDPs that model strategies dependent on special pairings. Most of these team specific MDPs are retrospective (Terroba et al., 2013). This comes from the difficulty to estimate appropriate transition probabilities for matches or pairings that may have not been played before. We have overcome this difficulty by our two scale approach. The idea is that the gameplay-MDP (g-MDP) incorporates only player depended probabilities that constitute the player’s skills and are independent of the opponent. From a simulation of the g-MDP the transition probabilities of a second more aggregated MDP, the strategic-MDP (s-MDP) are generated. The s-MDP and the g-MDP must

be related to each other such that a set of transitions in the g-MDP can be mapped to transitions in the s-MDP. The generated s-MDP transitions depend in contrast to the g-MDP transitions on the opponent. Due to the aggregation, the s-MDP is significantly smaller than the g-MDP and can be solved by dynamic programming.

The question may arise whether the s-MDP transition probabilities can be directly computed from the g-MDP probabilities or whether at least some bounds can be found. This paper considers as a basis the implementation of the two scale approach to beach volleyball, presented in (Hoffmeister and Rambau, 2017).

The main result of this paper is the computation of intervals for the direct point probabilities of the serving situation of the Olympic beach volleyball final in 2012. The computed intervals are

$$P_{estimatedStrat}^{serve} \in [0, 0.2291], \quad \bar{P}_{estimatedStrat}^{serve} \in [0, 0.1936], \quad \hat{P}_{estimatedStrat}^{serve} \in [0.5913, 1],$$

where  $P_{estimatedStrat}^{serve}$  denotes the probability for an ace,  $\bar{P}_{estimatedStrat}^{serve}$  is the probability for a point loss and  $\hat{P}_{estimatedStrat}^{serve}$  is the probability for a subsequent field attack by the opponent team. For the computation of these values we used the skills estimated from all prefinal matches and the estimated strategy of the final.

The paper provides the analytical derivation of the presented bounds on the direct point probabilities of the Olympic beach volleyball final in 2012 in the presentation *Strategy optimization in beachvolleyball – applying a two scale approach to the olympic games* and is organized as follows. In Section 2 the two scale approach implementation is recapped. The analytical bounds for the s-MDP transitions probabilities in terms of the g-MDP transition probabilities are derived in Section 3. Section 4 concludes this paper and gives an outlook to future investigations.

## 2 Two Scale Approach for Beach Volleyball recapped

In our implementation of the two scale approach for beach volleyball we answer the strategic question which attack plan a team should play, depending on the current score and situation. In beach volleyball a team consists of two players and a match of two or three sets. A set is won if one team has gained at least 21 points and is at least two points ahead of the opponent team. A point is scored according to the rally-point system. Let in the following of this paper team  $P$  be the team whose policy should be optimized and team  $Q$  be the opponent team. In both MDPs team  $P$  and team  $Q$  are modelled symmetric. However, team  $Q$  is part of the environment and captured in the transition probabilities.

The s-MDP models a complete beach volleyball set. For the purpose of the benchmark question a state of the s-MDP contains the current score, which team starts the next attack plan and an indicator whether it is a serving state or not. The action set  $A$  is constituted by the set of attack plans of team  $P$ . Attack plans consist of a sequence of hits and moves played in a phase of ball possession, e.g, an attack plan for a field attack after a serve consists of a reception, a set and a smash or shot. In all states where team  $P$  starts the next attack plan it can choose an action  $a \in A$  for its next attack. The reward is modelled such that the expected total reward in a state equals the winning probability of the set starting from the current state. The transition probability  $p_a$  [ $\bar{p}_a$ ] is the probability that team  $P$  playing action  $a$  directly wins [loses] the rally. The probability that none of this happens is denoted by  $\hat{p}_a := 1 - p_a - \bar{p}_a$ . We use  $q, \bar{q}$  and  $\hat{q}$  analogously for the transition probabilities after an serving or field attack of team  $Q$ . Since a serving attack has transition probabilities clearly different from a field attack, we distinguish between them. This is denoted by a superscript *field* or *serve* on the transition probabilities. Thus the evolution of the system is governed by eight probabilities  $p_a^{sit}$ ,  $\bar{p}_a^{sit}$ ,  $q^{sit}$ ,  $\bar{q}^{sit}$ , where  $a \in A$ ,  $sit \in \{serve, field\}$ .

The g-MDP models only a single rally instead of a whole set. A state includes the position of each player, the position of the ball, a boolean variable that indicates the hardness of the last hit and three other parameters that are necessary to track certain beach volleyball rules. A position on the court is defined on basis of the grid presented in Figure 1. The g-MDP is very large and contains around one

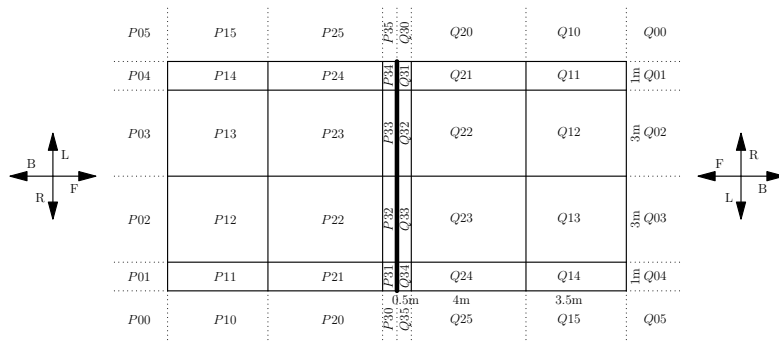


Figure 1: Court grid

billion different states. A state in the g-MDP is observed each time a player hits the ball, or the ball contacts the ground or net. The important point is that all transitions are generated from the individual player skills. The advantage of the player skills is that they can be estimated from training sessions or any match of that player. Furthermore, the player skills are assumed to be more stable in comparison to opponent dependent probabilities.

For each player  $\rho$ , including the opponent team, and each hitting technique  $tech$  with target  $target$  against a ball with hardness  $hardness$ , the probability  $p_{succ, \rho}(tech, [pos(\rho)], [target], [hardness])$  is defined as the probability that the specified target field  $target$  from  $\rho$ 's position is met. Due to modelling issues of the g-MDP implementation, we distinguish skills depending on parameters specified in the curly brackets. The parameters in square brackets are optional parameters. For some hitting techniques, the individual player probabilities are aggregated over certain optional parameters. For example, serving skills are aggregated over all possible serving fields which are  $P01, P02, P03$  and  $P04$ . Since only the relative distance to the ball is important for receives, defences and settings, they are aggregated over all possible target fields. But, we distinguish if a defence or receive was played against a *normal* or *hard* ball which is indicated by the parameter  $hardness \in \{hard, normal\}$ . If in the following some optional parameters are missing, it means that the skill was aggregated over that parameter. The complete tables with the estimated individual player skills and their aggregation can be found in (Hoffmeister and Rambau, 2017, Section 7, Gameplay MDP validation). We want to point out that a successful hit does not mean that a point occurs, since this would include the defending or receiving skills of the opponents. It means that the ball flies toward the target field the player aimed for. The probability of an execution fault is denoted by  $p_{fault, \rho}(tech, [pos(\rho)], [target], [hardness])$  for player  $\rho$  using hitting technique  $tech$ . If neither an execution fault nor a successful hit occurs, the ball will land in a neighbour field of the target field. We call this event a deviation and denote it by the probability  $p_{dev, \rho}(tech, [pos(\rho)], [target], [hardness]) := 1 - p_{succ, \rho}(tech, [pos(\rho)], [target], [hardness]) - p_{fault, \rho}(tech, [pos(\rho)], [target], [hardness])$ . Table 1 summarizes all hitting techniques available in the g-MDP. The possible target fields  $neighbour(pos(\rho)) \setminus (Q, \cdot)$  of a set are all fields that are a neighbouring field of the player's current position  $pos(\rho)$  and not on team  $Q$ 's court. There exist requirements on the state for using a certain technique which we skipped here and are specified in (Hoffmeister and Rambau,

<i>tech</i>	<i>target</i>	description	results in <i>hard</i> ball	skill depends on <i>hardness</i>
Serve				
$S_F$	$Q11 - Q24$	float serve	false	false
$S_J$	$Q11 - Q24$	jump serve (hard)	true	false
Reception				
$r$	$P11 - P34$	receive	false	true
$r_m$	$P11 - P34$	receive with move	false	true
Setting				
$s$	$neighbour(pos(\rho)) \setminus (Q, \cdot)$	set	false	false
Attack-Hit				
$F_{SM}$	$Q11 - Q24$	smash (hard)	true	false
$F_E$	$Q11 - Q24$	emergency shot	false	false
$F_P$	$Q11 - Q34$	planned shot	false	false
Defence				
$d$	$P11 - P34$	defence	false	true
$d_m$	$P11 - P34$	defence with move	false	true

Table 1: Hit specification for player  $\rho$  of team  $P$  and ball  $ball$ ;

2017, see Section 5, A Gameplay MDP for Beach Volleyball). We included the information which hitting technique results in a *hard* ball in the table.

Team actions in the g-MDP are a composition of the players' moves and hits. Each team in the g-MDP plays a team specific g-MDP-policy that is an implementation of an s-MDP action in the g-MDP. We implemented a g-MDP-policy as a variation of a basic policy that guarantees a reasonable match flow. Each team adapts parts of the basic policy according to their preferences. In our implementation modifications of the blocking, serving and attack-hit decisions are possible. All team specific refinements are included in a vector  $\pi$  whose components determine the probability for choosing true in a Boolean decision. In the basic policy all components of  $\pi$  are set to 0.5 which means, that both decision possibilities are equally probable. The blocking policy is specified by  $\pi_b$ , which states with which probability player 1 of a team is the designated blocking player in the next rally. It follows that with probability  $(1 - \pi_b)$  player 2 is the blocking player. The parameter  $\pi_s$  determines the serving policy of a team. With  $\pi_s$  a serve on player 1 of the opponent team is made, i.e., the target field of the serve belongs to the opposing court half that is covered by player 1. Further, a technique and target field decision of the serve and attack-hit are included in  $\pi_h$ . The two parts  $\pi_h^{serve}$  and  $\pi_h^{field}$  of  $\pi_h$  include the policy belonging to the indicated situation. Each part splits up into a technique and target field decision that depend on the hitting player  $\rho$ , i.e.,  $\pi_h^{sit} = (\pi_{h,tech}^{sit}(\rho), \pi_{h,target}^{sit}(\rho))^T$  with  $sit \in \{serve, field\}$ . The subscript term indicates if the decision is related to the technique (*tech*) or target field (*target*) decision. Now we can summarize all parameters that are necessary for defining a g-MDP policy of team  $P$  with players  $P_1$  and  $P_2$ :

$$\pi = \begin{pmatrix} \pi_h \\ \pi_b \\ \pi_s \end{pmatrix}, \quad \pi_h = \begin{pmatrix} \pi_h^{serve} \\ \pi_h^{field} \end{pmatrix}, \quad \pi_h^{sit} = \begin{pmatrix} \pi_{h,tech}^{sit}(\rho) \\ \pi_{h,target}^{sit}(\rho) \end{pmatrix}, \quad sit \in \{serve, field\}, \quad \rho \in \{P_1, P_2\}.$$

For a better memorability we defined the values of the components of  $\pi_h$  always as the probability for the more risky opportunity. In our example, we have two serving techniques available in the g-MDP, namely the float serve  $S_F$  and the jump serve  $S_J$ . The float serve is considered as a safe hit and the jump serve as a



risky hit. So  $\pi_{h,tech}^{serve}(\rho)$  is defined as the probability that  $\rho$  chooses an  $S_J$ . For the attack-hit we have three techniques available the smash  $F_{SM}$ , a planned shot  $F_P$  and an emergency shot  $F_E$ . The emergency shot is normally only played if none of the other attack-hits is possible. The smash is considered as a risky hit and the planned shot as a safe hit. So  $\pi_{h,tech}^{field}(\rho)$  is defined as the probability that  $\rho$  chooses a  $F_{SM}$ . Furthermore, we define all fields that are near the side out of the court as border fields. For example, on court side of team  $Q$  the border fields are  $\partial F := \{Q11 - Q31, Q14 - Q34\}$ . These are more risky target fields than non-border fields. So  $\pi_{h,target}^{serve}(\rho)$  and  $\pi_{h,target}^{field}(\rho)$  are the probabilities with which a border field is chosen as a target field. (Hoffmeister and Rambau, 2017, see Section 6, Gameplay MDP strategy)

Following the two scale approach, the g-MDP is simulated with team  $P$  playing a certain policy. From the simulation the s-MDP transition probabilities are estimated by counting the number of serves and attack plans as well as their outcomes. The outcome, following the definition of the s-MDP, is either a direct point or fault of the attacking team or a subsequent attack by the opponent team.

### 3 Bounds for the Direct Point Probabilities

In the following we derive bounds for the s-MDP probabilities in terms of the g-MDP probabilities. We make these considerations for team  $P$ 's probabilities only. Bounds for the opponent team  $Q$ 's probabilities can be derived analogously. For easier notation, we denote the hitting player throughout this section by  $\rho \in \{P1, P2\}$  and the receiving Player by  $\sigma \in \{Q1, Q2\}$ . Further, let  $S_* [r_*]$  be an unspecified serve [reception]. We assume that  $\rho$  has chosen an reasonable target field, i.e.,  $target$  is inside the court of the opposing team.

We start with an analysis of the serving situation, i.e., we compute bounds for  $p_a^{serve}$ ,  $\bar{p}_a^{serve}$  and  $\hat{p}_a^{serve}$ , where  $a$  is an attack plan of the s-MDP that corresponds to a team specific policy  $\pi$  of the g-MDP. In beach volleyball a direct point after a serve is called an ace. For an ace the following events must all be realized together in a serving situation of the g-MDP:

- the serve is executed without a fault,
- the ball does not land in an outside field,
- the opponent team makes a fault when receiving the ball.

Now we try to calculate the probabilities of these events. Assume first, the hitting player  $\rho$ , the executed serving technique  $S_*$  and the target field  $target$  are known. Then the probability that  $S_*$  is executed by  $\rho$  without a fault is

$$p_{succ, \rho}(S_*, target) + p_{dev, \rho}(S_*, target)$$

Since we know from policy  $\pi$  of team  $P$  with which probability the serves  $S_J$  or  $S_F$  are played we can state that expression more precisely as

$$\begin{aligned} & \pi_{h,tech}^{serve}(\rho) \cdot \left( p_{succ, \rho}(S_J, target) + p_{dev, \rho}(S_J, target) \right) \\ & + (1 - \pi_{h,tech}^{serve}(\rho)) \cdot \left( p_{succ, \rho}(S_F, target) + p_{dev, \rho}(S_F, target) \right) \end{aligned}$$

The hit may only land outside the field if the target field was a border field and a deviation to an outside field occurred. From the policy  $\pi$  the probability  $\pi_{h,target}^{serve}(\rho)$  with which a border field is chosen as a target field is known. In the system dynamic of the g-MDP a deviation to any neighbour-field is equally probable. When we look at the court grid presented in Figure 1, we see that each field has eight neighbour-fields and the number of outside fields that are neighbour-fields range between three and five.

These values depend on the specification of the court grid. In general, let

$$\omega(\text{target}) := \frac{|\text{neighbour}(\text{target}) \cap \text{outside-fields}(\text{target})|}{|\text{neighbour}(\text{target})|}$$

be the probability that the result of the deviation from field *target* is an outside field. Since we do not know the exact target field, we define

$$\begin{aligned} \omega_{\max} &= \max \left\{ \frac{|\text{neighbour}(\text{target}) \cap \text{outside-fields}(\text{target})|}{|\text{neighbour}(\text{target})|} \mid \forall \text{target} \in \text{grid} \right\} \\ \omega_{\min} &= \min \left\{ \frac{|\text{neighbour}(\text{target}) \cap \text{outside-fields}(\text{target})|}{|\text{neighbour}(\text{target})|} \mid \forall \text{target} \in \text{grid} \right\} \end{aligned}$$

For our grid of the g-MDP, we get  $\omega_{\max} = \frac{5}{8}$  and  $\omega_{\min} = \frac{3}{8}$ .

The receiving player  $\sigma$  may use, depending on his position and the position of the ball, a receive  $r$  or a receive with a move  $r_m$  as the receiving technique. For receiving and defending skills, the individual probabilities for that technique depend also on the *hardness* of the ball which may be either *hard* or *normal*. The jump serve is the only serve that leads to a *hard* ball. Since we know from policy  $\pi$  of the attacking team the probability of a jump serve which is  $\pi_{h,tech}^{serve}(\rho)$  we know with which probability the ball to receive is *hard* or *normal*. The absolute position of the receiving player has in the g-MDP no impact on the skills. So  $\sigma$  makes an execution fault with probability

$$\pi_{h,tech}^{serve}(\rho) \cdot p_{\text{fault}, \sigma}(r_*, \text{normal}) + (1 - \pi_{h,tech}^{serve}(\rho)) p_{\text{fault}, \sigma}(r_*, \text{hard}).$$

Since in general a receive with a move should have a higher fault rate than a receive without a move, we can conclude that the probability of a fault lies between

$$\pi_{h,tech}^{serve}(\rho) \cdot p_{\text{fault}, \sigma}(r, \text{hard}) + (1 - \pi_{h,tech}^{serve}(\rho)) p_{\text{fault}, \sigma}(r, \text{normal})$$

and

$$\pi_{h,tech}^{serve}(\rho) \cdot p_{\text{fault}, \sigma}(r_m, \text{hard}) + (1 - \pi_{h,tech}^{serve}(\rho)) p_{\text{fault}, \sigma}(r_m, \text{normal}).$$

It depends on the development of the match which player serves how often. Also we have no stochastic information about the serving distribution. Therefore, we introduce  $\rho_{\max}^{\gamma, tech}$  [ $\rho_{\min}^{\gamma, tech}$ ] as the player that has the maximal [minimal] probability for the specified outcome  $\gamma$  of a skill, i.e., we chose

$$(\rho_{\max}^{\gamma, tech}, \tau_{\max}^{\gamma, tech}) \in \arg \max_{\rho, \text{target}} \left\{ p_{\gamma, \rho}(tech, \text{target}) \mid \rho \in \{P1, P2\} \right\} \forall \gamma \in \{\text{succ}, \text{fault}, \text{dev}\}, tech \in \{S_J, S_F\}$$

and

$$(\rho_{\min}^{\gamma, tech}, \tau_{\min}^{\gamma, tech}) \in \arg \min_{\rho, \text{target}} \left\{ p_{\gamma, \rho}(tech, \text{target}) \mid \rho \in \{P1, P2\} \right\} \forall \gamma \in \{\text{succ}, \text{fault}, \text{dev}\}, tech \in \{S_J, S_F\}.$$

So  $(\rho_{\max}^{\gamma, tech}, \tau_{\max}^{\gamma, tech})$  or  $(\rho_{\min}^{\gamma, tech}, \tau_{\min}^{\gamma, tech})$  can be for the same hit a different player depending on the specified outcome  $\gamma$ . We use an analogous notation for the receiving player  $\sigma$ . The components of the hitting strategy  $\pi_h$  are dependent on the hitting player. Therefore, we define

$$\begin{aligned} \bar{\pi}_{h,tech}^{serve} &= \max_{\rho \in \{P1, P2\}} \{ \pi_{h,tech}^{serve}(\rho) \} \text{ and } \underline{\pi}_{h,tech}^{serve} = \min_{\rho \in \{P1, P2\}} \{ \pi_{h,tech}^{serve}(\rho) \} \\ \bar{\pi}_{h,target}^{serve} &= \max_{\rho \in \{P1, P2\}} \{ \pi_{h,target}^{serve}(\rho) \} \text{ and } \underline{\pi}_{h,target}^{serve} = \min_{\rho \in \{P1, P2\}} \{ \pi_{h,target}^{serve}(\rho) \} \end{aligned}$$

We can summarize the computed bounds for  $p^{serve}$ :

$$\begin{aligned}
 p_a^{serve} \leq & \left( \bar{\pi}_{h,tech}^{serve} \cdot \left( p_{succ, \rho_{max}^{succ, S_J}}(S_J, \tau_{max}^{succ, S_J}) + p_{dev, \rho_{max}^{dev, S_J}}(S_J, \tau_{max}^{succ, S_J}) \right. \right. \\
 & \cdot \left. \left. \left( (1 - \underline{\pi}_{h,target}^{serve}) + \bar{\pi}_{h,target}^{serve}(1 - \omega_{min}) \right) \right) \right. \\
 & + \left. \left( 1 - \underline{\pi}_{h,tech}^{serve} \right) \left( p_{succ, \rho_{max}^{succ, S_F}}(S_F, \tau_{max}^{succ, S_F}) + p_{dev, \rho_{max}^{dev, S_F}}(S_F, \tau_{max}^{succ, S_F}) \right. \right. \\
 & \cdot \left. \left. \left( (1 - \underline{\pi}_{h,target}^{serve}) + \bar{\pi}_{h,target}^{serve}(1 - \omega_{min}) \right) \right) \right) \\
 & \cdot \left( \bar{\pi}_{h,tech}^{serve} \cdot p_{fault, \sigma_{max}^{fault, r_m}}(r_m, hard) + (1 - \underline{\pi}_{h,tech}^{serve}) p_{fault, \sigma_{max}^{fault, r_m}}(r_m, normal) \right)
 \end{aligned}$$

and

$$\begin{aligned}
 p_a^{serve} \geq & \left( \underline{\pi}_{h,tech}^{serve} \cdot \left( p_{succ, \rho_{min}^{succ, S_J}}(S_J, \tau_{min}^{succ, S_J}) + p_{dev, \rho_{min}^{dev, S_J}}(S_J, \tau_{min}^{dev, S_J}) \right) \right. \\
 & \cdot \left. \left( (1 - \bar{\pi}_{h,target}^{serve}) + \underline{\pi}_{h,target}^{serve}(1 - \omega_{max}) \right) \right) \\
 & + \left( 1 - \bar{\pi}_{h,tech}^{serve} \right) \left( p_{succ, \rho_{max}^{succ, S_F}}(S_F, \tau_{min}^{succ, S_F}) + p_{dev, \rho_{min}^{dev, S_F}}(S_F, \tau_{min}^{dev, S_F}) \right. \\
 & \cdot \left. \left( (1 - \bar{\pi}_{h,target}^{serve}) + \underline{\pi}_{h,target}^{serve}(1 - \omega_{max}) \right) \right) \\
 & \cdot \left( \underline{\pi}_{h,tech}^{serve} \cdot p_{fault, \sigma_{min}^{fault, r}}(r, hard) + (1 - \bar{\pi}_{h,tech}^{serve}) p_{fault, \sigma_{min}^{fault, r}}(r, normal) \right).
 \end{aligned}$$

In the next step we consider the probability  $\bar{p}_a^{serve}$  of a serving fault. We make the plausible assumption that the opponent does not try to receive a serve that flies towards an outside field. Each of the following events in the g-MDP lead to a fault after a serve:

- execution fault of the serve, i.e., the ball does not cross the net,
- the ball crosses the net but lands in an outside field.

In the same way as for the direct point probability, we can calculate the probability of an execution fault of the hitting player  $\rho$  when following policy  $\pi$ :

$$\pi_{h,tech}^{serve}(\rho) \cdot p_{fault, \rho}(S_J, target) + (1 - \pi_{h,tech}^{serve}(\rho)) p_{fault, \rho}(S_F, target).$$

Since we assume only reasonable serves, the ball can only land in an outside field if a deviation occurred. Analogously to the analysis of  $p_a^{serve}$ , we can calculate a lower bound for the probability that the ball crosses the net and lands in an outside field:

$$\left( \pi_{h,tech}^{serve}(\rho) \cdot p_{dev, \rho}(S_J, target) + (1 - \pi_{h,tech}^{serve}(\rho)) p_{dev, \rho}(S_F, target) \right) \cdot \pi_{h,target}^{serve}(\rho) \cdot \omega_{min}$$

and an upper bound:

$$\left( \pi_{h,tech}^{serve}(\rho) \cdot p_{dev, \rho}(S_J, target) + (1 - \pi_{h,tech}^{serve}(\rho)) p_{dev, \rho}(S_F, target) \right) \cdot \pi_{h,target}^{serve}(\rho) \cdot \omega_{max}.$$

Bounds for direct point probabilities

Hoffmeister, Rambau

With the same meaning of  $\rho_{max}^{\gamma, tech}$ ,  $\rho_{min}^{\gamma, tech}$ ,  $\tau_{max}^{\gamma, tech}$ ,  $\tau_{min}^{\gamma, tech}$ ,  $\bar{\pi}_{h,tech}^{serve}$ ,  $\underline{\pi}_{h,tech}^{serve}$ ,  $\bar{\pi}_{h,target}^{serve}$  and  $\underline{\pi}_{h,target}^{serve}$  as before, we get an upper bound for  $\bar{p}^{serve}$ :

$$\begin{aligned} \bar{p}_a^{serve} &\leq \bar{\pi}_{h,tech}^{serve} \cdot p_{fault, \rho_{max}^{fault, S_J}}(S_J, \tau_{max}^{fault, S_J}) + (1 - \underline{\pi}_{h,tech}^{serve}) p_{fault, \rho_{max}^{fault, S_F}}(S_F, \tau_{max}^{fault, S_F}) \\ &\quad + \left( \bar{\pi}_{h,tech}^{serve} \cdot p_{dev, \rho_{max}^{dev, S_J}}(S_J, \tau_{max}^{dev, S_J}) + (1 - \underline{\pi}_{h,tech}^{serve}) p_{dev, \rho_{max}^{dev, S_F}}(S_F, \tau_{max}^{dev, S_F}) \right) \cdot \bar{\pi}_{h,target}^{serve} \cdot \omega_{max} \end{aligned}$$

and a lower bound:

$$\begin{aligned} \bar{p}_a^{serve} &\geq \underline{\pi}_{h,tech}^{serve} \cdot p_{fault, \rho_{min}^{fault, S_J}}(S_J, \tau_{min}^{fault, S_J}) + (1 - \bar{\pi}_{h,tech}^{serve}) p_{fault, \rho_{min}^{fault, S_F}}(S_F, \tau_{min}^{fault, S_F}) \\ &\quad + \left( \underline{\pi}_{h,tech}^{serve} \cdot p_{dev, \rho_{min}^{dev, S_J}}(S_J, \tau_{min}^{dev, S_J}) + (1 - \bar{\pi}_{h,tech}^{serve}) p_{dev, \rho_{min}^{dev, S_F}}(S_F, \tau_{min}^{dev, S_F}) \right) \cdot \underline{\pi}_{h,target}^{serve} \cdot \omega_{min}. \end{aligned}$$

Finally, we compute bounds for the case that neither a direct point nor a fault occurs, which happens with probability  $\hat{p}_a^{serve}$ . The following events must occur together after a serving situation in the g-MDP to lead to a subsequent attack by the opponent team:

- the serve is executed without a fault,
- the ball does not land in an outside field,
- the opponent team is receives the ball without a fault.

The first two events are the same as in a direct point scenario. Only the receiving event differs and is the counterpart of the receiving event in the direct point scenario.

$$\begin{aligned} \hat{p}_a^{serve} &\leq \left( \bar{\pi}_{h,tech}^{serve} \cdot \left( p_{succ, \rho_{max}^{succ, S_J}}(S_J, \tau_{max}^{succ, S_J}) + p_{dev, \rho_{max}^{dev, S_J}}(S_J, \tau_{max}^{succ, S_J}) \right) \right. \\ &\quad \cdot \left( (1 - \underline{\pi}_{h,target}^{serve}) + \bar{\pi}_{h,target}^{serve} (1 - \omega_{min}) \right) \\ &\quad + (1 - \underline{\pi}_{h,tech}^{serve}) \cdot \left( p_{succ, \rho_{max}^{succ, S_F}}(S_F, \tau_{max}^{succ, S_F}) + p_{dev, \rho_{max}^{dev, S_F}}(S_F, \tau_{max}^{succ, S_F}) \right) \\ &\quad \cdot \left. \left( (1 - \underline{\pi}_{h,target}^{serve}) + \bar{\pi}_{h,target}^{serve} (1 - \omega_{min}) \right) \right) \\ &\quad \cdot \left( \bar{\pi}_{h,tech}^{serve} \cdot \left( p_{succ, \sigma_{max}^{succ, r}}(r, hard) + p_{dev, \sigma_{max}^{dev, r}}(r, hard) \right) \right. \\ &\quad \left. + (1 - \underline{\pi}_{h,tech}^{serve}) \cdot \left( p_{succ, \sigma_{max}^{succ, r}}(r, normal) + p_{dev, \sigma_{max}^{dev, r}}(r, normal) \right) \right) \end{aligned}$$

and

$$\begin{aligned} \hat{p}_a^{serve} &\geq \left( \underline{\pi}_{h,tech}^{serve} \cdot \left( p_{succ, \rho_{min}^{succ, S_J}}(S_J, \tau_{min}^{succ, S_J}) + p_{dev, \rho_{min}^{dev, S_J}}(S_J, \tau_{min}^{dev, S_J}) \right) \right. \\ &\quad \cdot \left( (1 - \bar{\pi}_{h,target}^{serve}) + \underline{\pi}_{h,target}^{serve} (1 - \omega_{max}) \right) \\ &\quad + (1 - \bar{\pi}_{h,tech}^{serve}) \cdot \left( p_{succ, \rho_{min}^{succ, S_F}}(S_F, \tau_{min}^{succ, S_F}) + p_{dev, \rho_{min}^{dev, S_F}}(S_F, \tau_{min}^{dev, S_F}) \right) \\ &\quad \cdot \left. \left( (1 - \bar{\pi}_{h,target}^{serve}) + \underline{\pi}_{h,target}^{serve} (1 - \omega_{max}) \right) \right) \\ &\quad \cdot \left( \underline{\pi}_{h,tech}^{serve} \cdot \left( p_{succ, \sigma_{min}^{succ, r_m}}(r_m, hard) + p_{dev, \sigma_{min}^{dev, r_m}}(r_m, hard) \right) \right. \\ &\quad \left. + (1 - \bar{\pi}_{h,tech}^{serve}) \cdot \left( p_{succ, \sigma_{min}^{succ, r_m}}(r_m, normal) + p_{dev, \sigma_{min}^{dev, r_m}}(r_m, normal) \right) \right). \end{aligned}$$

Figure 2 summarizes realisation sequences of the g-MDP serving situation and the related s-MDP transitions. Probability of the paths in the g-MDP that end up in a green disk correspond to the path of  $p_a^{serve}$  in the s-MDP, probabilities of paths with an orange circle correspond to  $\hat{p}_a^{serve}$  and red disks to  $\bar{p}_a^{serve}$ .

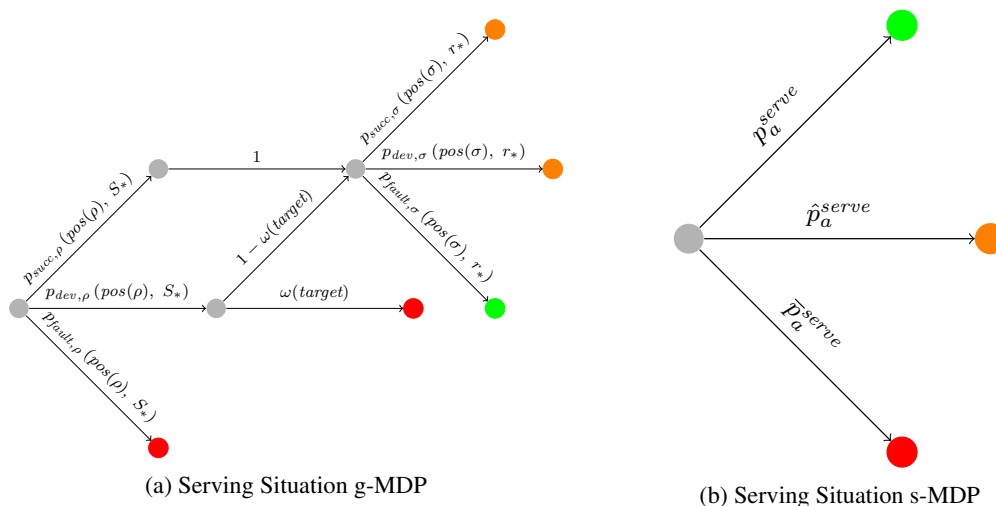


Figure 2: Summary Serving Situation

If we insert the estimated pre-final skills of Brink-Reckermann and Alison-Emanuel as well as the estimated final strategy of Brink-Reckermann in the presented equations we get the intervals for the direct point probabilities as presented in the introduction of that paper. For a comparison, the estimates for the direct point probabilities from the g-MDP simulation, presented in (Hoffmeister and Rambau, 2017), are:

$$p_{estimatedStrat}^{serve} = 0.0769, \quad \bar{p}_{estimatedStrat}^{serve} = 0.1380, \quad \hat{p}_{estimatedStrat}^{serve} = 0.7851.$$

## 4 Conclusion

We conclude that our estimates for the direct point probabilities from the g-MDP simulation lie in the computed intervals of this paper. The derived bounds have a relative large spread because of the large number of possible actions and realisations in the g-MDP. Since it is not possible, e.g., to predict how often a player may serve in a set or which player of the opponent team will receive the ball, we had to make rough assessments in the terms for the bounds. The computed intervals will get smaller if the players in a team have more similar skills.

It is work in progress to analyse the field attack situation in a similar way even if this will be probably even more complicated. However, we conclude that it is possible to compute some bounds of the s-MDP transition probabilities in terms of the g-MDP strategy and skills.

## References

- [1] Chan, T. C. Y. and Singal, R. (2016) *A Markov Decision Process-based handicap system for tennis*. Journal of Quantitative Analysis in Sports, Vol. 12, pp. 179–189.

- [2] Clarke, S. R. and Norman, J. M. (1998) *Dynamic programming in cricket: Protecting the weaker batsman*. Asia Pacific Journal of Operational Research, Vol. **15**.
- [3] Clarke, S. R. and Norman, J. M. (2012) *Optimal challenges in tennis*. Journal of the Operational Research Society, Vol. **63**, pp. 1765–1772.
- [4] Hirotsu, N. and 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, Vol. **53**, pp. 88–96.
- [5] Hirotsu, N. and Wright, M. (2003) *Determining the best strategy for changing the configuration of a football team*. Journal of the Operational Research Society, Vol. **54**, pp. 878–887.
- [6] Hoffmeister, S. and Rambau, J. (2017) *Strategy Optimization in Sports – A Two-Scale Approach via Markov Decision Problems*. URL: [http://www.wm.uni-bayreuth.de/de/download/xcf2d3wd4lkj2/preprint\\_sso\\_bv.pdf](http://www.wm.uni-bayreuth.de/de/download/xcf2d3wd4lkj2/preprint_sso_bv.pdf).
- [7] Nadimpalli, V. K. and Hasenbein, J. J. (2013) *When to challenge a call in tennis: A Markov decision process approach*. Journal of Quantitative Analysis in Sports, Vol. **9**, pp. 229–238.
- [8] Norman, J. M. (1985) *Dynamic Programming in Tennis – When to Use a Fast Serve*. The Journal of the Operational Research Society, Vol. **36**, pp. 75–77.
- [9] Terroba, A., Kusters, W., Varona, J. and Manresa-Yee, C. S. (2013) *Finding optimal strategies in tennis from video sequences*. International Journal of Pattern Recognition and Artificial Intelligence, Vol. **27**, pp. 1–31.
- [10] Wright, M. and Hirotsu, N. (2003) *The professional foul in football: Tactics and deterrents*. Journal of the Operational Research Society, Vol. **54**, pp. 213–221.

# Using Player Quality and Injury Profiles to Simulate Results in Australian Football

Karl Jackson

Champion Data - Melbourne, Australia. karl.jackson@championdata.com.au

## Abstract

Traditionally, prediction algorithms in team sports are focused on the results of single matches. Short-term predictions allow an agnostic approach to team selection to be valid, since gross changes in availability and quality of player personnel are unlikely. When aiming for long-range predictions, both of these potential changes become significant factors. This research aims to use projections of potential injuries (prevalence and severity) and player improvement or decline as a means to allow for more accurate simulations of a full season of matches in the Australian Football League (AFL). Injury distributions will be based on observed injuries to AFL-listed players from seasons 2013-2016. Changes in player quality will be modeled on observed changes of player performance as measured by the official AFL Player Ratings, from seasons 2010-2016. This methodology can then be used to forecast team performance, allowing realistic expectations to be set by professional teams, to assess the impact of an uneven fixture on team expectations.

## 1 Introduction

Australian football is the country's most popular football code which can be seen as a hybrid between association football, basketball, and rugby codes. Its highest level of competition is the Australian Football League (AFL), an 18 team league that runs for a 23 round regular season, followed by a finals series contested by the top-eight teams. Each team is comprised of 22 players, 18 of which are allowed on the team at one time. Minimal restrictions are placed on player positions ("offsides") and each team is allowed 90 player rotations per match.

Research initiated in 2007 by Champion Data, the AFL's official data providers, led to the creation of the Official AFL Player Ratings, as first introduced in Jackson (2008) [6], and fully defined in Jackson (2016) [7]. This rating system places a value on player performance by measuring the change in the equity of team scores as the result of a player's actions on the field. The Official AFL Player Ratings will be used as a proxy for player quality, and by extension, the quality of a selected team of players. Match simulations will be conducted based on this team quality metric.

While match simulations and result forecasting in sports is a well-studied area, with applications in Australian football dating back to Clarke (1993) [2], the vast majority of ratings treat teams as singular entities. Studies generally fit into one of three categories:

1. Adaptive team ratings (ELO-like), such as Ryall & Bedford (2010) [9], where often many years of data are collected to train the model for future matches, and competing teams' ratings are updated based on the result of each match.

2. Transformation of team rankings, such as Glasson, et. al. (2001) [5], where the ordered ranking of teams is used as the main input to the prediction model rather than an explicit rating.
3. Team statistics, such as Dyte & Clarke (2000) [3], where historical performance of teams in key statistical areas, like goals scored and goals conceded, are used to model outcomes.

This research considers the impact of individual players on team success. Some research exists in this realm, such as Robertson, et. al. (2016) [8], but most is limited to discrete sports like baseball - Freeze (1974) [4] - and cricket - Bukiet (2006) [1].

## 2 Preliminary Simulations

The measure of team quality used in this research is the sum of each player's average score in the league's official player rating system - the AFL Player Ratings. A rolling two-year window of player performance contributes to each player's average per game, with exponential smoothing applied to each match in a player's sample to ensure more weight is applied to more recent games.

### 2.1 Prediction Model

This method is similar to ELO-like ratings at team level, in that player averages are updated after each match played to reflect performance in that match. The assumed advantage over ELO-like team models is the ability to adjust the model for team selections, particularly when key players are excluded via injury or suspension. The model used to predict win probabilities for matches is:

$$\text{logit}(P) = 0.124 \times \text{Home} + 0.380 \times \text{Travel} + 0.040 \times \text{Quality} \quad (1)$$

where Home gives an advantage to the home team and Travel gives a further advantage if the opposition team has traveled to the home team's ground. This split in home ground advantage is required in the AFL because 10 of the 18 teams are based in or near Melbourne, and the other eight teams are spread with two each in four states - meaning games are regularly played with both teams in their home city. The Quality variable is the difference in team quality between the two competing teams.

Note that from (1), a difference in team quality of +12.6 points is enough to overcome the disadvantage of being a travelling away team. Only 30% of matches have a difference in team quality of below 12.6 points to either team. The median of the absolute difference in Team Quality is roughly 50 points, and the largest seen from 2012-2016 was 110 points.

To test the viability of the model, we perform 10,000 random permutations of matches where 60% are used as a training set and the remaining 40% (432 matches) are used as a testing set. Stefani and Clarke (1992) [10] predicted 10 individual seasons of Australian football using two different algorithms, with an overall accuracy of 68% for both algorithms. This model correctly predicted results above 68% in 98% of match permutations, and had an overall accuracy of 72%. This indicates that the model described in (1) is more than suitable to use as a prediction tool for Australian Football matches. A distribution of prediction results can be seen in Figure 1



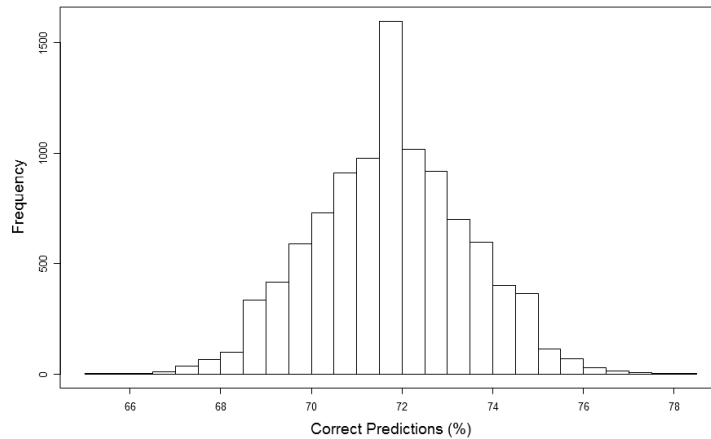


Figure 1: Prediction power of 10,000 random permutations of 1,080 matches using 60% as a training set and 40% as a testing set.

## 2.2 Simulation Method

The main aim of this research is to perform a simulation of the 2017 AFL season. For a preliminary simulation, this was done under the assumption that each club had their best available team ready to play in each match, and that no improvement or decline was present in any players. For each match, the strength of both teams was used to determine a win percentage for the home team, as calculated using the model described in (1). The result of the match was simulated as a Bernoulli variable corresponding to win or lose, and at the conclusion of each season teams were ordered by wins, with ties broken by random sorting. This was repeated 100,000 times.

## 2.3 Simulation Results

Results from these simulations can be seen in Table 1. Competition favourites GWS Giants unsurprisingly sit on top of the list with 46% of simulations resulting in them finishing in first position. At the other end of the table, favourites for the wooden spoon, Brisbane Lions, did finish in last place in 69% of simulations.

In none of the 100,000 did the GWS Giants finish below 12th on the ladder, nor did Brisbane finish above 12th. Half of the competition had no simulation results where they finished on the bottom of the ladder, and seven teams did not appear in the bottom two teams. At the top of the ladder there was more variation, with just six teams failing to finish on top and three teams failing to finish in the top two positions. All teams except Brisbane had at least one simulation finishing in the top-eight teams, and thus qualifying for finals.

## 2.4 Effect of Fixture

As an extension of these results, we can begin to examine the effect of factors such as each team's strength of schedule and the location of matches. Each game was adjusted by firstly removing home ground advantage

Ladder Pos:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
GWS	46	21	13	8	5	3	2	1	1	0	0	0	*	*	*	*	*	*	
WB	15	18	16	13	11	9	7	5	3	2	1	1	0	0	0	*	*	*	
WCE	11	15	15	14	12	10	8	6	4	3	2	1	1	0	0	*	*	*	
HAW	8	13	13	13	12	10	9	7	6	4	3	2	1	1	0	0	*	*	
SYD	7	11	12	13	12	11	10	8	6	4	3	2	1	1	0	0	*	*	
ADEL	5	8	10	11	11	11	10	9	7	6	4	3	2	1	1	0	*	*	
PA	4	8	10	11	12	12	11	10	8	6	4	3	2	1	1	0	*	*	
GEE	1	3	4	6	7	9	10	11	11	10	8	7	5	4	2	1	0	*	
COLL	1	3	4	6	8	9	10	11	11	10	9	7	5	4	2	1	0	*	
MELB	0	1	1	2	3	5	6	8	10	11	11	11	10	9	7	4	1	0	
RICH	0	0	1	1	2	4	5	7	9	10	11	12	12	10	8	5	2	0	
FREM	0	0	1	1	2	3	5	6	8	10	11	12	12	11	9	6	2	0	
GC	*	0	0	1	1	2	3	5	7	9	11	12	13	13	11	8	3	1	
SK	*	0	0	1	1	2	3	4	6	8	9	11	13	14	14	11	4	1	
ESS	*	0	0	0	1	1	2	3	4	6	8	10	12	15	17	14	7	2	
NM	*	*	*	0	0	0	1	1	2	3	4	6	8	12	17	25	16	5	
CAR	*	*	*	*	*	*	*	0	0	0	1	1	2	4	9	19	41	22	
BL	*	*	*	*	*	*	*	*	*	*	*	*	0	0	1	2	6	22	69

Table 1: Simulated ladder position (percentage of simulations) when the best possible team is available for each team in each game. \* = no simulation results.

and the travel advantage of the model, then by playing each game against a neutral opponent (the average quality of the 17 other teams combined). The results of these simulations can be found in Table 2.

Of the two fixture-related factors, the strength of schedule had the greatest effect on season outcomes. Forcing all matches to be played in neutral venues has a net effect of less than  $\pm 0.3$  wins for all teams. This effect was roughly one-third of the effect of the strength of schedule, which display a range of  $\pm 0.9$  wins difference from simulations where all games were played against neutral opponents.

Note that by comparing Table 1 and Table 2 we can see a clear effect between the strength of each team and the difficulty of that team's draw. For example, GWS has 16.7 expected wins under the true fixture, but would be expected to win 17.6 wins if all matches were against neutral opponents. Brisbane, however, has an expected wins count of 3.4 from the actual fixture, 0.5 wins above what would be expected from a neutral draw - the fifth-highest advantage of all teams. This is likely to be by design in the AFL Fixture as a means of competition equalisation. When creating the fixture, the league assigns the top-six teams a harder bracket of repeat opponents than the middle-six, and bottom-six teams from the previous season.

## 2.5 Player Contributions

An additional measure of impact that can be extracted from repeat simulations is the value of a single player. By making a single player unavailable for the entirety of the season and leaving all other factors constant, we can see the change in the number of expected wins for that player's club and use that as a measure of the

Club	Fixture	Venue	Diff	Rank	Opp	Diff	Rank	Neutral	Diff	Rank
GC	9.5	9.3	0.2	3	8.6	0.9	2	8.4	1.1	1
ESS	8.6	8.6	0.0	9	7.7	0.9	1	7.7	0.9	2
BL	3.4	3.1	0.3	1	2.9	0.5	5	2.7	0.7	3
MELB	10.4	10.2	0.2	4	10.0	0.4	6	9.8	0.6	4
RICH	10.1	10.2	-0.1	12	9.5	0.6	3	9.5	0.6	5
CAR	5.2	5.3	-0.1	13	4.7	0.5	4	4.7	0.5	6
NM	7.2	7.2	0.0	8	6.8	0.4	7	6.7	0.5	7
FREM	9.9	9.8	0.1	5	9.6	0.3	8	9.5	0.4	8
SK	9.1	9.1	0.0	7	9.0	0.1	9	8.9	0.1	9
ADEL	13.4	13.2	0.2	2	13.8	-0.4	10	13.7	-0.3	10
PA	13.3	13.3	0.0	10	13.7	-0.4	11	13.7	-0.3	11
COLL	11.9	12.0	-0.1	14	12.3	-0.4	12	12.4	-0.5	12
WB	14.9	14.9	0.0	11	15.6	-0.7	15	15.7	-0.8	13
GEEL	11.9	11.9	0.0	6	12.6	-0.7	14	12.8	-0.8	14
HAW	14.1	14.2	-0.1	15	14.8	-0.7	13	15.0	-0.9	15
SYD	14.0	14.2	-0.2	18	14.9	-0.9	16	15.1	-1.1	16
WCE	14.5	14.6	-0.1	16	15.4	-0.9	18	15.6	-1.1	17
GWS	16.7	16.8	-0.2	17	17.6	-0.9	17	17.8	-1.1	18

Table 2: Expected wins from simulations under the existing fixture, with all games at neutral venues, with all games against neutral opponents, and with both venues and opponents neutral.

player's value.

Under these preliminary simulations the results aren't particularly insightful. The model used is linear on team quality, and the change in team quality is simply the difference between the player excluded and each club's 23rd-best player. In order to gain a more insightful measure, more variation needs to be applied to the model to make it more realistic.

### 3 Injury Profiles

One method of introducing more realistic variability to simulations of the season is to remove the assumption that each team has their best possible team available for each match. This will be done by simulating injuries to players. For the purposes of this research, we will use player injuries as published on the league's website. This information has been published since the 2013 season and recorded by Champion Data in terms of when injuries first occurred and how many weeks the player missed.

Excluding weeks where a player carried over an injury from a previous round, and including weeks where players had no injuries, there were a combined total of 68,219 player weeks in the data set. In that time there were 3,758 new injuries that forced players to miss at least one week of football. This corresponds to one injury in roughly every 18 available matches.

For the length of each injury, there is a competing distribution where season-ending injuries are increas-

ingly likely late in the season, and non-season-ending injuries follow a geometric distribution. A visual representation of these competing distributions can be seen in Figure 2.

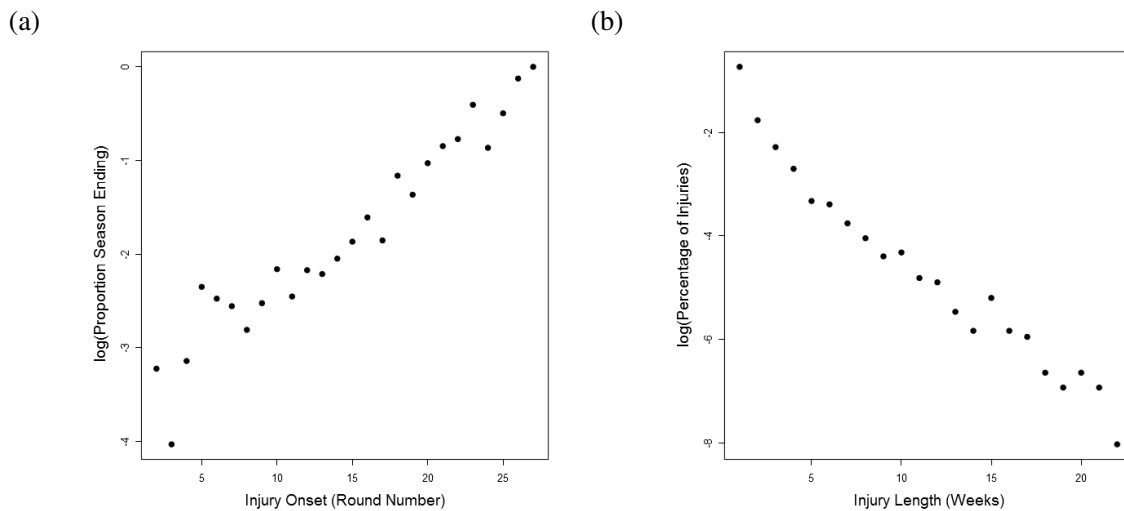


Figure 2: (a) Percentage of injuries that are season-ending by round of first onset. (b) Distribution of injury length (Weeks).

Random injuries for players were introduced as a three-stage process for each of the 23 rounds of matches in the fixture. All non-injured players were randomly assigned an injury at a rate of one-in-18. Of these players, season-ending injuries occurred at a rate relative to the current round number, as shown in (2).

$$\Pr(\text{Season Ending}) = \exp(-3.66 + 0.13 \times \text{Round Number}) \quad (2)$$

Of the injuries that weren't season-ending, the length of the injury was simulated as in (3).

$$\Pr(\text{Injury Length} = n) = \exp(-1.52 + 0.28 \times n) \quad (3)$$

### 3.1 Simulation Results

Results of these updated simulations can be seen in Table 3. It is clear when comparing to the preliminary results in Table 1 that there is a much larger spread of finishing positions for most teams. GWS still has the highest number of expected wins, but just 31% of simulations resulted in them finishing on top, compared to 46% in the initial simulations. Just five teams avoided finishing on the bottom of the ladder, compared to nine beforehand, and only two teams failed to finish on top of the ladder - Brisbane and Carlton - compared to six in the original simulation.

## 4 Player Improvement/Decline

Another method of making simulations more realistic is by randomising the improvement/decline of players based on their age. Observed changes in player performance in the AFL Player Ratings are available from

Ladder Position:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
GWS	31	20	14	10	8	6	4	3	2	1	1	0	0	0	0	0	*	*
WB	25	19	16	11	9	7	5	3	2	1	1	0	0	0	0	0	*	*
WCE	13	14	14	12	11	9	7	6	4	3	2	1	1	1	0	0	0	*
SYD	11	13	13	12	11	10	8	7	5	4	3	2	1	1	0	0	0	*
PA	7	10	11	12	12	10	9	8	6	5	4	2	1	1	0	0	0	0
HAW	5	8	9	10	11	11	10	9	8	6	5	4	2	2	1	0	0	0
ADEL	4	6	8	8	9	10	10	10	8	7	6	5	3	3	2	1	0	*
COLL	2	4	5	7	8	9	9	10	10	9	8	7	5	4	3	1	0	0
GEE	2	3	4	6	7	8	9	9	9	10	9	7	6	5	3	2	1	0
MELB	1	1	2	3	4	5	7	8	10	10	11	10	10	8	6	3	1	0
RICH	0	1	1	2	3	5	6	7	9	9	11	11	11	10	8	5	2	0
FREM	0	1	1	2	2	4	5	7	8	9	10	11	11	10	9	5	2	0
SK	0	0	1	1	2	3	4	6	7	9	11	11	12	12	10	7	2	0
GC	0	0	1	1	1	2	4	5	6	8	10	12	13	14	12	8	4	1
ESS	0	0	0	0	1	1	2	3	4	6	7	9	12	15	17	14	7	2
NM	0	0	0	0	0	0	0	1	1	2	3	5	8	11	17	28	17	6
CAR	*	*	*	*	*	0	0	0	0	0	1	1	2	4	9	19	42	20
BL	*	*	*	*	*	*	*	*	*	0	0	0	0	1	1	6	22	70

Table 3: Simulated ladder position (percentage of simulations) with random injuries added.

\* = no simulation results.

2010 to 2016. Distributions of changes from one year to the next are shown for three age groups in Figure 3. Note that the younger of the three age groups has a much higher rate of improvement than the middle age group, which is essentially random around the zero point, and the older age group, where there is more decline than improvement. 59% of 21-year-olds improved on their performance from the previous season, with an average change of +0.7 points per game. Just 45% of 25-year-olds improved, with an average change of -0.1 points per game, and just 41% of 32-year-olds improving, with an average change of -0.1 points per game.

In each simulation, each player's value is randomly permuted based on the observed distribution at that player's age, and these values are used in the simulation model as in the previous section with random injuries still included.

#### 4.1 Simulation Results

Results of these simulations can be seen in Table 4. This final simulation provides more variation again amongst teams compared to previous simulations. GWS is now the only team that fails to finish on the bottom of the ladder, and Brisbane finishes as high as third overall.

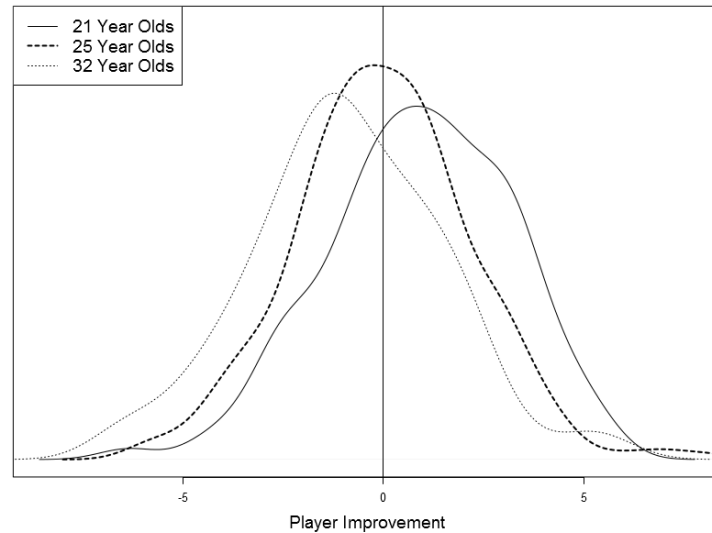


Figure 3: Distribution of improvement/decline for three selected age groups.

## 5 Discussion & Conclusion

This research has made an attempt to provide accurate simulations of an entire season of Australian football. Simulating full seasons of matches allows us to place realistic expectations on teams before a season starts, triggers analysis and discussion in the media and amongst the viewing public, and also allows us to measure the impact of specific inputs to simulation models.

Since the 2017 had started prior to this research being undertaken, a full range of season-start bookmaker odds were not available. One applicable market that was able to be found was for GWS to finish on top of the ladder at odds of \$4.50, which implies a probability of between 19% and 21% depending on the size of the bookmaker's overround. The results of the final simulation has GWS finishing on top 28% of the time, compared to 31% of simulations without player improvement, and 46% of simulations with no injuries.

Melbourne to finish in the top-eight teams had odds of \$2.10, implying a probability of between 42% and 45%. In the final simulation model they finished in the top-eight in 42% of simulations, compared to 31% of simulations with no player improvement, and 26% of simulations with no injuries. Similarly, Gold Coast had odds of \$4.00 to make the top-eight, implying a probability of 22% to 24%. Final simulations had Gold Coast in the top-eight teams 22% of the time, compared to 14% and 12% from the respective previous models.

These results indicate that random changes in player performance, and random injuries throughout a season increase the accuracy of full-season predictions. The tradeoff of such a result is that each team's predicted standings are much more open than under a simplistic model, potentially diluting any information that can be gained from predicting final outcomes, implying that they are highly random around some baseline of team quality.

Each of the adjustments to standard simulations of match results introduced in this paper can be improved with future research. Injury profiles are likely to change based on player age, player role, and injury

Ladder Position:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
GWS	28	18	13	10	8	6	5	4	3	2	1	1	1	0	0	0	0	*
WB	21	17	14	11	9	7	6	5	3	2	2	1	1	0	0	0	0	0
SYD	12	13	12	11	10	9	7	7	5	4	3	3	2	1	1	0	0	0
WCE	11	12	12	10	9	8	8	7	5	5	4	3	2	2	1	1	0	0
PA	6	9	10	10	10	10	9	8	6	6	5	4	3	2	1	1	0	0
ADEL	6	8	9	9	9	9	8	8	8	6	5	5	4	3	2	1	1	0
HAW	4	6	8	8	8	8	9	8	8	7	6	6	5	3	3	1	1	0
COLL	3	5	6	7	8	9	8	8	8	8	7	6	6	4	3	2	1	0
GEE	3	4	5	7	7	7	9	8	8	8	7	7	6	6	4	3	1	0
MELB	2	3	4	5	6	7	7	8	8	9	9	8	8	6	5	3	2	1
RICH	1	2	2	3	4	4	5	7	8	8	9	10	9	10	8	6	4	2
FREM	1	1	2	3	3	4	5	6	7	9	9	9	10	10	9	7	4	1
SK	0	1	2	2	3	5	5	6	7	8	9	9	10	10	9	7	4	1
GC	0	1	1	2	3	4	5	6	7	8	9	10	10	10	9	7	4	2
ESS	0	0	1	1	1	2	3	3	4	6	8	9	10	12	13	12	9	5
NM	0	0	0	0	1	1	1	2	2	3	4	6	8	11	14	19	17	11
CAR	*	0	0	0	0	0	0	1	1	1	2	3	4	7	12	17	28	23
BL	*	*	0	*	0	0	0	0	0	0	0	1	1	2	5	11	25	54

Table 4: Simulated Ladder Position (percentage of simulations) with random injuries added and random changes in player value. \* = no simulation results.

history. These factors, as well as the baseline quality of players, are also likely to affect the distribution of improvement of players. As an example, Marcus Bontempelli of the Western Bulldogs is already performing alongside the elite players of the sport despite being just 21 years of age. Using only his age to determine a rate of improvement is likely to be misleading, since his current performance would suggest that he has a smaller scope for improvement than a player with a lower baseline of past performance.

The model used for simulating games performs very well to predict future matches (at 72% on average) but further research could improve this accuracy. Treating each team as a complex system rather than a linear combination of 22 players may give more information about the quality of each team, and further investigation into shared experience of a selected team, smarter application of home ground advantage, and consideration for the length of breaks before games may all have some impact on the prediction power of the underlying model.

Future research could also focus more on the specific contribution of individual players to assess their value to the team in terms of wins added. This can be done on an individual basis, or as part of a collective assessment of the club's entire playing list to highlight weaknesses or to assess depth of player talent.

## References

- [1] Bukiet, B., and Ovens, M. (2006) *A Mathematical Modelling Approach to One-Day Cricket Batting Orders*. Journal of Sports Science & Medicine **5(4)**, 495-502.

- [2] Clarke, S.R. (1993) *Computer Forecasting of Australian Rules Football for a Daily Newspaper*. The Journal of the Operational Research Society **44 (8)**, 753-759
- [3] Dyte, D., and Clarke, S.R. (2000) *A ratings based Poisson model for World Cup soccer simulation*. Journal of the Operational Research Society **51 (8)**, 993-998.
- [4] Freeze, R.A. (1974) *An Analysis of Baseball Batting Order by Monte Carlo Simulation*. Operations Research **22 (4)**, 728-735.
- [5] Glasson, S., Jeremiejczyk, B., and Clarke, S.R. (2001) *Simulation of Women's Beach Volleyball Tournaments*. ASOR Bulletin 2001.
- [6] Jackson, K.B. (2008) *A Player Rating System for Australian Rules Football Using Field Equity Measures*. Mathematics and Computers in Sport **9**, 68-74.
- [7] Jackson, K.B. (2016) *Assessing Player Performance in Australian Football Using Spatial Data*. PhD Thesis, Swinburne University of Technology.
- [8] Robertson, S., Gupta, R., and McIntosh, S. (2016) *A method to assess the influence of individual player performance distribution on match outcome in team sports*. Journal of Sports Sciences **34 (19)**, 1893-1900.
- [9] Ryall, R. and Bedford, A. (2010) *An optimized ratings-based model for forecasting Australian Rules football*. International Journal of Forecasting **26 (3)**, 511-517.
- [10] Stefani, R and Clarke, S.R. (1992) *Predictions and home advantage for Australian rules football*. Journal of Applied Statistics **19 (2)**, 251-261.



# A statistical rating method for team ball games and its application to prediction in the Rio Olympic Games

Eiji Konaka\*

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

## Abstract

This study presents the prediction results of ball games such as basketball, handball, volleyball, and water polo in the Rio Olympic Games. First, the study proposes a statistical rating method for team ball games. Only one parameter, called a rating, shows the strength and/or skill of each team. We assume that the difference in the rating values explains the scoring rate in a game via a logistic regression model. The rating values are estimated from major international competition results, including world championships, worldwide league competitions, and Olympic continental and world qualifying tournaments held before the Rio Olympic Games. The rating values are calculated using an iterative method. This method is easy to implement and numerically stable. The results of these ball games in the Rio Olympic Games are estimated based on the calculated rating values. The prediction results demonstrate that the proposed method can more accurately predict the result than the official world rankings or world ranking points. The proposed method made 220 correct predictions out of 294 matches in eight events, while the official world rankings made only 202 correct predictions. This result shows a significant difference between the two criteria.

## 1 Introduction

This study presents a unified ability evaluation model for several ball games, and its applies it to predict the results of four ball games (basketball, handball, volleyball, and water polo) in the Rio Olympic Games.

Accurate ranking systems for sports are required because rankings are used as criteria in group draws, player (team) seeding, guest players (teams) selection, and so on.

The number of wins and/or the percentage of victories are the most "fair" ranking criteria if all players are matched at the same time 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, the national teams of major sports cannot all compete in a fair round-robin format. As a result, each team has different opponents and a different number of matches.

In order to rank and order the 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. Ranking points are calculated as the sum of the points attributed to international tournaments and a standings in the tournament. The sum is calculated for a designated period, such as four years. The four ball games here determine their world rankings using the accumulative method[2, 3, 4] (FINA does not disclose the world rankings and ranking points of water polo on their website. The rankings and ranking points of water polo used here is collected from

personal websites and sports news), however, these ranking points have no clear mathematical or statistical background. Therefore, the values and rankings cannot be used as a quantitative measure of the ability of the teams.

A **points exchange** is another possible choice of ranking system. Here, each team has a ranking point, which they are exchanged based on match results. For example, several points are moved from the losing team to the winning team in the match. The most popular points exchange system is the Elo rating[5] used in Chess ranking. Rugby uses a modified Elo-based ranking system[6]. In these systems, calculated ranking points are converged into the real values if the abilities of all teams are constant and an adequately large number of matches are played within a certain period. In general, ranking points in a points exchange system require more calculation than those in accumulative points systems.

## 1.1 Ranking and rating

Here, we clearly define *ranking* and *rating* as follows:

- *ranking*: order of teams.
- *rating*: 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 non-constant elements.

The ranking point in the accumulative method includes both sets of elements. On the other hand, the point exchange system intends to estimate the first set of elements by denoising the effects from the second set. This study defines that the rating is a quantitative value that shows the first set of elements calculated by a statistical method.

Many statistical skill-assessment studies have been reported. However, too detailed a situation analysis increases the number of explanatory variables, while collectable data are bounded. This phenomena is the well-known *curse of dimensionality*[7] in a machine learning context. Simple structure should be tested with simple data first.

Compared to these diverse results on skill assessment, rating methods for volleyball teams are difficult to find. Massey's method[8] is applied to college volleyball[9] and beach volleyball[10]. Some match prediction models on basketball games have been reported and their prediction accuracy has been discussed[11, 12], however, no rating methods for national teams have been developed.

## 1.2 Objective

As mentioned above, few studies on quantitative ability-evaluation methods for national ball game teams are available. Moreover, the conventional studies exploit specific features of each sport. The main objective of this study is to propose a simple and unified rating framework for different ball games.

The unified method should use only commonly recorded values among different sports. All four ball games considered here have a common value — score. This study proposes a unified and statistical rating

method that only uses scores in each match (or sets, in volleyball). Only one parameter, called rating, shows the strength and/or ability of each team. We assume that the difference in rating values explains the scoring rate in a game via a logistic regression model. The rating values are estimated from major international competition results, including world championships, worldwide league competitions, and Olympic continental and world qualifying tournaments held before the Rio Olympic Games. The rating values are calculated by an iterative method. This method is easy to implement and numerically stable.

Results of these ball games in the Rio Olympic Games are estimated based on the calculated rating values. The prediction results demonstrate that the proposed method can more accurately predict the result than the official world rankings or world ranking points can. The proposed method made 220 correct predictions out of 294 matches in eight events, while the official world rankings only made 202 correct predictions. This result shows a significant difference, with  $p = 0.016$ , between the two criteria. The proposed method can also made nine correct predictions out of 24 medals, together with their medal colors (37.5%). Moreover, we made 16 correct predictions on podium finishes (66.7%). These prediction results are clearly better than those provided by Sports Illustrated (25.0%, 58.3%), USA Today (25.0%, 50.0%), and Gracenote (37.5%, 54.2%).

The proposed rating can evaluate the distribution of the competitive strength of national teams, and can be applied to compare the distributions between different sports. Unlike the skill parameter in the Bradley-Terry model, the proposed rating is an interval scale. Therefore, the proposed rating values can be clustered by a distance-based hierarchical clustering method such as the Ward method. The clustering result shows the number of teams that are equally matched.

## 2 Definition and calculation of rating

In this section, we first point out the problems with the accumulative and point exchange methods. Then we propose a novel rating calculation method that can resolve the problems.

### 2.1 Current ranking system

The Fédération Internationale de Volleyball (FIVB), the world governing body for volleyball, regularly reports the ranking of its member nations' teams. The FIVB Board of Administration designs the system of point attribution for selected FIVB world and other official competitions[4]. Table 1 shows the points awarded for four major international competitions.

This table shows huge inconsistency. Why are champions of these competition each awarded 100 points? What is the reason for the point difference among standings? The answers could not be found in FIVB website.

Basketball[2] and handball[3] have a similar accumulative ranking system, and also no explanation on the mathematical fundamentals of the point attribution systems. Fédération Internationale de Natation (FINA) no longer even disclosed the world rankings of water polo in 2016.

### 2.2 Proposed method

As mentioned above, the official ranking points cannot be utilized to estimate the team ability, nor to predict future match results because they have no mathematical fundamentals.

Table 1: FIVB Ranking Point System

Standing	Competition name			
	Olympic	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

We propose a unified statistical estimating method of scoring ratios based on the score in each match, which is always officially recorded and is common among different ball games.

Assume that the scoring ratio of a team  $i$  against a team  $j$  ( $i$  and  $j$  are the team indices), denoted as  $p_{i,j}$ , is given as

$$p_{i,j} = \frac{1}{1 + e^{-(r_i - r_j)}}, \quad (1)$$

where  $r_i$  is defined as the *rating* of team  $i$ . This mathematical structure is called as *logistic regression model*. It is widely used in various areas, such as the winning probability assumption of Elo ratings in Chess games[5], and correct answer probability for questions in item response theory[13, 14].

### 2.2.1 Point exchange system and its limitations

The Elo rating[5] uses

$$p_{i,j} = \frac{1}{1 + 10^{\frac{-(r_i - r_j)}{400}}}, \quad (2)$$

which is a linear transformation of (1), to explain winning rate of player  $i$  against player  $j$ . Therefore, the rating difference can explain and predict the match result if the rating values are accurately calculated. World rankings calculated by World Rugby, the governing body for rugby union, adopts a linear transformation of

(1) with polygonal line approximation[6]. This ranking point system clearly intends to connect the ranking points and the team abilities. Therefore, the point system can, for example, evaluate how large an upset a match result is. For instance, the World Rugby officially reports that "World rankings confirm Japan's victory as biggest shock"[15]. The upset by Japan against South Africa in the Rugby World Cup 2015 is confirmed as the biggest upset, with a ranking point difference 13.09.

In an Elo rating system, the match result

$$s_{i,j} = \begin{cases} 1 & i \text{ win.} \\ 0 & i \text{ lose.} \end{cases} \quad (3)$$

leads to the rating update

$$r'_i = r_i + K(s_{i,j} - p_{i,j}). \quad (4)$$

Here,  $r_j$  is updated in a similar way. By this definition, the calculation of a current rating requires all match results played by all teams.

The constant  $K$  in (4) determines the update value of the rating value. For example,  $K = 16$  is used in professional class chess matches. A large  $K$  implies that the rating values are strongly affected by the latest match results. On the other hand, the rating values dwell on the past match results in the case of a small  $K$ . The accuracy of the rating values depends on this parameter  $K$ . Moreover, the rating value cannot converge to "real" ability when the number of matches is small because the rating can be updated only a few times.

## 2.2.2 Proposed method

The main limitations of the (Elo-based) point exchange method can be summarized as follows:

- It is difficult to determine the appropriate  $K$  in (4).
- The rating values are updated only once after a match. The rating values might not converge to real ability with a few matches.

This study proposes a novel rating method in order to overcome these limitations of the conventional rating method.

The proposed method is summarized as follows:

- The rating values are updated using ALL match results in a predefined interval, e.g., one year.
- $K$  in (4) is a small value.
- A match result is randomly selected and the abovementioned rating update is iterated until all rating values converge.

The detail of the proposed unified algorithm is described in the following. Table 2 lists the notations used throughout in this paper.

### Unified rating calculation algorithm

1. Let  $r^{(0)} = 0$ . Let  $\varepsilon_{th} > 0$  be a small value. Set  $K > 0$ . Several experiments tell that the value should be  $10^{-3}$  or smaller. Set iteration index  $k = 0$ .

Store the results in  $N_S$  matches (or sets) in a database. Here,  $\langle i, j, s_i, s_j \rangle$  denotes one match (or set) result.

Table 2: Notations

$N_T$	Number of teams
$r \in \mathfrak{R}^{N_T}$	Rating vector
$N_S$	Number of sets (in volleyball) Number of games (in others)
$\langle i, j, s_i, s_j \rangle$	Result of one set (or game). Team $i$ and $j$ scored $s_i$ and $s_j$ points in a set (or a game). $N_S$ tuples are stored in database.
$\epsilon_{th}$	Threshold value
$K$	Parameter used in rating update
$k$	Iteration index
$0, 1$	Column vector composed of zeros and ones with suitable dimensions
$\ x\ $	Euclidean norm of vector $x$

2. The results in the database are sorted randomly.
3. Retrieve a result  $\langle i, j, s_i, s_j \rangle$  from the database.
4. The rating values  $r_i$  and  $r_j$  are updated as follows:

$$p_{i,j} = \frac{1}{1 + e^{-\left(\frac{r_i^{(k)} - r_j^{(k)}}{s_i + s_j}\right)}}, \quad s_{i,j} = \frac{s_i}{s_i + s_j}, \quad (5)$$

$$r_i^{(k+1)} = r_i^{(k)} + K(s_{i,j} - p_{i,j}), \quad r_j^{(k+1)} = r_j^{(k)} + K((1 - s_{i,j}) - (1 - p_{i,j})). \quad (6)$$

The above update is performed for all results in the database.

5. If  $\|r^{(k+1)} - r^{(k)}\| < \epsilon_{th}$ , output  $r^{(k+1)}$  and terminate the algorithm. Otherwise,  $k \leftarrow k + 1$ , store the match result to the database, and go to Step 2.  $\square$

A rating value may diverge to  $\infty$  if a team has no conceding scores in the interval. If such a team is included,  $s$  in (5) is modified as

$$s = \frac{s_i + \epsilon_s}{s_i + s_j + 2\epsilon_s}, \quad (7)$$

where  $\epsilon_s > 0$  is a small positive value compared to  $s_i$  and  $s_j$ .

By definition, the rating is an interval scale. Therefore, its origin,  $r = 0$ , can be arbitrarily selected and a constant value can be added to all  $r_i$ . For example,

$$r \leftarrow r - (\max r) \cdot 1 \quad (8)$$

implies that  $r = 0$  always shows the highest rating, and  $r < 0$  shows the distance from the top team.

### 2.2.3 Convert rating on scoring ratio to winning probability

The rating in (1) calculated by the proposed method in Section 2.2.2 explains the scoring ratio. This differs between sports showing how the scoring ratio affects the winning probability. Here, we convert the rating on the scoring ratio to that of a winning probability, as follows:

$$w_{i,j} = 1 \text{ (} i \text{ wins), or } 0 \text{ (} j \text{ wins)} \quad (9)$$

denotes a win or loss for team  $i$  against team  $j$ . Find  $D_k^*$ , where  $k$  is an index of sports, that satisfies

$$\hat{w}_{i,j} = \frac{1}{1 + \exp(-D_k(r_i - r_j))}, \quad (10)$$

$$D_k^* = \arg \min_{D_k} \sum (w_{i,j} - \hat{w}_{i,j})^2. \quad (11)$$

Then,  $r_i$  is converted as follows:

$$\bar{r}_i = D_k^* r_i, \quad i = 1, 2, \dots, N_T. \quad (12)$$

Therefore,  $\bar{r}_i$  is a rating that explains the winning probability, and  $\bar{r}$  can be utilized in match result predictions.

## 3 Rating calculation for four ball games and its application to match result predictions in Rio Olympic Games

### 3.1 Data set

We calculate the rating values of national teams on the following four ball games: basketball, handball, volleyball, and water polo. The match results used in the rating calculation consist of the following:

- Rio Olympic qualifying tournaments, including continental championships.
- Worldwide tournaments, for example, world championships, and World League (men's volleyball), held from 2014 to 2016/8 (just before Rio 2016).

The number of teams participating in at least one tournament and the number of matches in the data set are listed in Table 3.

Table 3: Number of teams and matches

Sport	Gender	Teams	Matches	Sport	Gender	Teams	Matches
Basketball	M	69	334	Basketball	W	57	238
Handball	M	69	375	Handball	W	44	311
Volleyball	M	43	466	Volleyball	W	36	337
Water polo	M	31	346	Water polo	W	26	294

### 3.2 Results

Figure 1 shows the results of all 38 matches (30 group round-robin matches, four quarterfinals, two semifinals, and two medal matches) of the men's basketball in Rio 2016. The horizontal and vertical axes are the predicted scoring ratio from the calculated rating values and real scoring ratio, respectively.

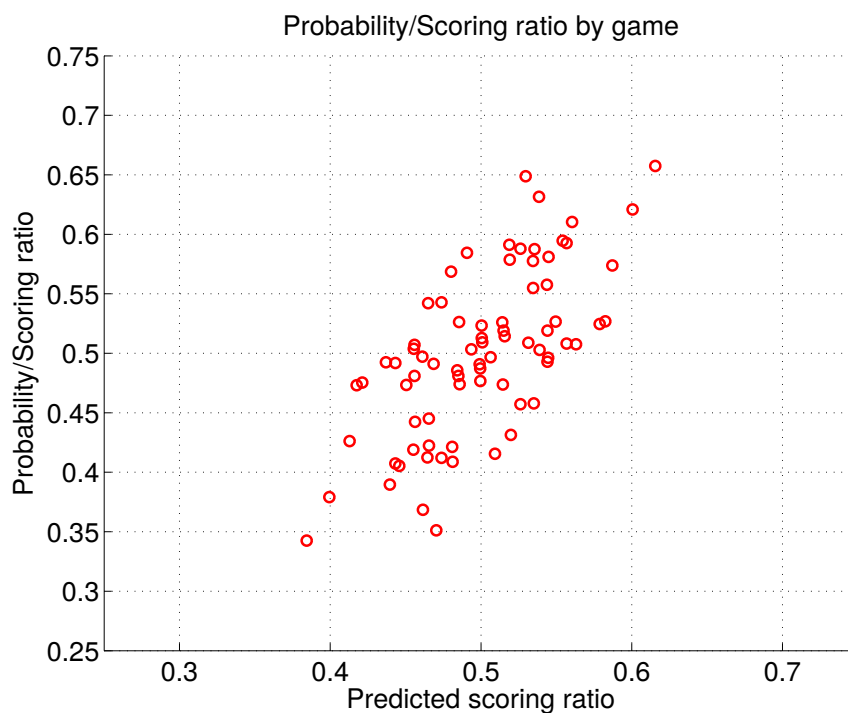


Figure 1: Predicted and real scoring rate in each game (Rio 2016, Basketball, Men)

As a comparison, in Figure 2, the horizontal axis now shows the difference in the official world rankings.

Table 4 shows the prediction accuracy comparison between the proposed method and the official world rankings. The prediction law is very simple: "a team with a higher rating (ranking) scores more." Draws are judged as incorrect in both methods.

Table 5 lists the normalization parameters  $D_k^*$ .

Table 6 lists the detailed prediction for men's basketball. The rating values are normalized using  $D_k^*$ , and are shifted so that the lowest rating is zero. All 38 matches are simulated  $10^6$  times. The table lists the average. The underline and bold numbers denote the prediction and the result, respectively.

The teams winning medals are predicted for eight events in four sports. The prediction is evaluated from two different viewpoints, "Medal with color" and "Podium." For example, the prediction in Table 6 tells us that gold, silver, and bronze medals would have been awarded to USA, ESP, and SRB. The actual result is



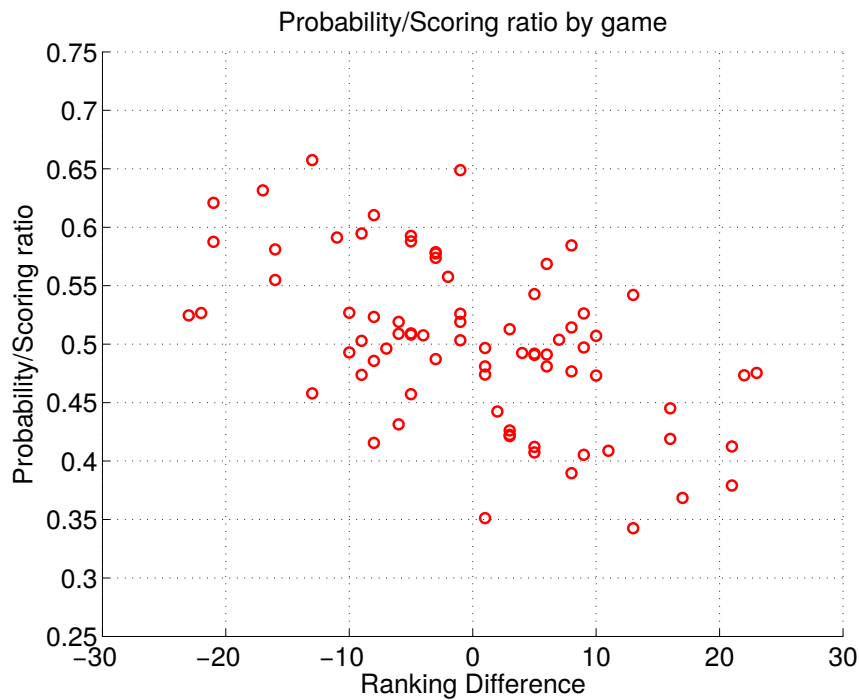


Figure 2: Ranking gap and real scoring rate in each game (Rio 2016, Basketball, Men)

USA, SRB, and ESP. In this case, the proposed method predicts one medal with color, and three podium finishes.

The proposed prediction result is compared to the predictions by Sports Illustrated (SI)[16], USA Today[17], and Gracenote[18]. Table 7 shows the result. Bold number show the most accurate prediction.

By definition, the proposed rating is an interval scale. Therefore, the distribution can be clustered based on its difference, namely, distance. The Ward method[19] with a threshold of 0.5 is used for the teams that qualified for Rio 2016. The results are shown in Figure 3. In this result, the rating values are normalized by  $D_k^*$  and are shifted so that the top-rated team is zero.

### 3.3 Discussion

Table 4 shows that the proposed rating method can realize an accurate prediction (220 correct out of 294 matches) compared to the official (accumulative) world ranking system (202 correct out of 294 matches). The null hypothesis that "the prediction accuracy of the proposed method is the same as that of the world ranking system" is rejected by Pearson's  $\chi^2$  test with  $p = 0.016 < 0.05$ . Moreover, the correlation between the predicted and the real scoring ratio is stronger than that between the ranking gap and the ratio. This result implies that the proposed rating value is a better quantitative measure of the ability of national teams of these

Rating and prediction of team ball games in the Rio Olympic Games

E. Konaka

Table 4: Prediction accuracy in Rio 2016

		Matches	Corrects		Corr. Coeff.	
			Rating	Ranking	Rating	Ranking
Basketball	M	38	30	29	0.679	-0.542
Handball	M	38	25	20	0.592	-0.492
Volleyball	M	38	30	27	0.731	-0.790
Water polo	M	42	27	21	0.560	-0.438
Basketball	W	38	34	29	0.818	-0.698
Handball	W	38	22	31	0.579	-0.572
Volleyball	W	38	33	31	0.731	-0.663
Water polo	W	24	19	14	0.905	-0.697
All	M	156	112	97		
All	W	138	108	105		
All		294	220	202		

Table 5: Normalization parameters  $D_k^*$ 

		Gender	$D_k^*$			Gender	$D_k^*$
Basketball	M	M	11.660	Basketball	W	W	9.193
Handball	M	M	12.299	Handball	W	W	9.090
Volleyball	M	M	15.019	Volleyball	W	W	9.868
Water polo	M	M	5.288	Water polo	W	W	4.055

four ball games than is the official world ranking.

Table 5 shows that  $D_k^*$  is larger in men's events than it is in women's event in the same sport.  $D_k^*$  is a parameter used to convert the rating on scoring ratio to the rating on winning probability. Large  $D_k^*$  implies that many men's teams are equally matched and many matches are hotly contested, that is, the scoring ratio is around 0.5. Table 4 also shows that the official ranking system does not provide accurate ability evaluations, especially for men's competitions.

Table 7 shows that the proposed method can realize better predictions than can a famous sports magazine and a nationwide newspaper. These are comparative to a company providing statistics. However, the advantage of the proposed method cannot be tested statistically because there are only small samples.

Surprisingly, the proposed method can achieve better prediction results than those of the official ranking system and professional sports journalists, even though the proposed method uses only one unified model and does not include specific features in each sport and event.

Figure 3 shows the normalized rating values on winning probability. The rating values can be compared because they are normalized. These figures and the prediction results imply the following:

- In handball, no outstanding strongest team exists. Six teams with  $\bar{r} > -1$  qualified for Rio 2016 in both the men's and women's events. Therefore, it is difficult to predict the match results ( $\bar{r} = -1$  implies

Table 6: Medal prediction (basketball, men)

Team	Rating (normalized)	Group	Gold	Silver	Bronze	4th
FRA	3.3714	A	0.0236	0.1281	0.1692	0.1849
USA	5.9376	A	<b>0.7933</b>	0.1077	0.0695	0.0060
VEN	0.7395	A	0.0000	0.0001	0.0003	0.0019
SRB	3.6863	A	0.0453	<b>0.2225</b>	<u>0.2658</u>	0.1836
CHN	0.0000	A	0.0000	0.0000	0.0000	0.0004
AUS	3.3972	A	0.0248	0.1272	0.2097	<b>0.2392</b>
ARG	2.1873	B	0.0008	0.0112	0.0143	0.0522
ESP	4.3037	B	0.1081	<u>0.3611</u>	<b>0.1796</b>	0.0656
BRA	2.2445	B	0.0007	0.0122	0.0266	0.0851
LTU	2.3811	B	0.0024	0.0173	0.0419	0.1135
CRO	2.1984	B	0.0010	0.0125	0.0229	0.0659
NGR	0.9001	B	0.0000	0.0001	0.0002	0.0017

underline: prediction, **bold**: result

Table 7: Medal predictions

	All medals	Medal with color	Podium
Proposed	24	<b>9</b>	<b>16</b>
SI	24	6	14
USA Today	24	6	12
Gracernote	24	<b>9</b>	13

**bold**: best prediction

that the team beats the top-rated team with probability  $1/(1 + e^1) = 0.2689$ ).

- The other three sports have one to three outstanding teams, i.e.,  $\bar{r} > -1$ .
- Except for the abovementioned outstanding teams, the slope of the plot of the men's rating is more moderate than that of the women's rating. This implies that the many equally matched teams are in the men's event. Many match results follows the match previews in the women's events because there are clear differences in the ability of the teams.

## 4 Conclusion

This study has presented the prediction results of four ball games, basketball, handball, volleyball, and water polo in the Rio Olympic Games based on a unified statistical rating method. Both the unified rating method and its calculation method have been proposed. The rating values for all teams participating in Olympic qualification tournaments within a year are calculated.

Surprisingly, the proposed method has been able to achieve better prediction results than the official ranking system and professional sports journalists, even though the proposed method uses only one unified model and does not include specific features in each sport and event.

## References

- [1] Stefani Ray. The methodology of officially recognized international sports rating systems. *Journal of Quantitative Analysis in Sports*, 7(4), 2011.
- [2] FIBA. FIBA world ranking. <http://www.fiba.com/rankingmen>, 2016. referred in 2016/12/22.
- [3] IHF. Ranking table. <http://www.ihf.info/en-us/thegame/rankingtable.aspx>, 2016. referred in 2016/12/22.
- [4] FIVB. FIVB volleyball world rankings. <http://www.fivb.org/en/volleyball/Rankings.asp>, 2016. referred in 2016/6/14.
- [5] Arpad E. Elo. *Ratings of Chess Players Past and Present*. Harper Collins Distribution Services, hardcover edition, 1979.
- [6] World Rugby. Rankings explanation. <http://www.worldrugby.org/rankings/explanation>, 2014. referred in 2016/6/14.
- [7] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613. ACM, 1998.
- [8] Ken Massey. Massey rating. <http://www.masseyratings.com/>, 1997. referred in 2016/6/14.
- [9] Elizabeth Knapper and Hope McIlwain. Predicting wins and losses: A volleyball case study. *The College Mathematics Journal*, 46(5):352–358, 2015.
- [10] Sam Glasson, Brian Jeremiejczyk, and Stephen R. Clarke. Simulation of women’s beach volleyball tournaments. *Australian Society for Operations Research*, 20(2):2–7, 2001.
- [11] Kubatko Justin, Oliver Dean, Pelton Kevin, and Rosenbaum Dan T. A starting point for analyzing basketball statistics. *Journal of Quantitative Analysis in Sports*, 3(3), 7 2007.
- [12] Michael J. Lopez and Gregory J. Matthews. Building an NCAA men’s basketball predictive model and quantifying its success. *Journal of Quantitative Analysis in Sports*, 11(1), 5 2015.
- [13] R. Hambleton. *Fundamentals of Item Response Theory (Measurement Methods for the Social Science)*. Sage Publications, Incorporated, new. edition, 9 1991.
- [14] R. J. de Ayala. *The Theory and Practice of Item Response Theory (Methodology in the Social Sciences)*. Guilford Pr, 1 edition, 12 2008.
- [15] World Rugby. World rankings confirm Japan’s victory as biggest shock. <http://www.rugbyworldcup.com/news/111746>, 10 2015. referred in 2016/6/14.
- [16] Brian Cazeneuve. Olympic medal predictions: Picking gold, silver, bronze in all 306 events. <http://www.si.com/olympics/2016/08/01/rio-2016-olympics-medal-picks-predictions-projected-medal-count>, 6 2016. referred in 2016/8/1.
- [17] USA Today. 2016 rio olympics medal projections. <http://www.usatoday.com/story/sports/olympics/2016/07/30/2016-rio-olympics-medal-projections/87779154/>, 7 2016. referred in 2016/8/1.
- [18] Gracernote. Gracernote’s data analytics predicts winners and losers of 2016 rio olympics. <http://www.gracernote.com/gracenotes-data-analytics-predicts-winners-losers-2016-rio-olympics/>, 4 2016. referred in 2016/8/1.
- [19] Joe H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

Rating and prediction of team ball games in the Rio Olympic Games

E. Konaka

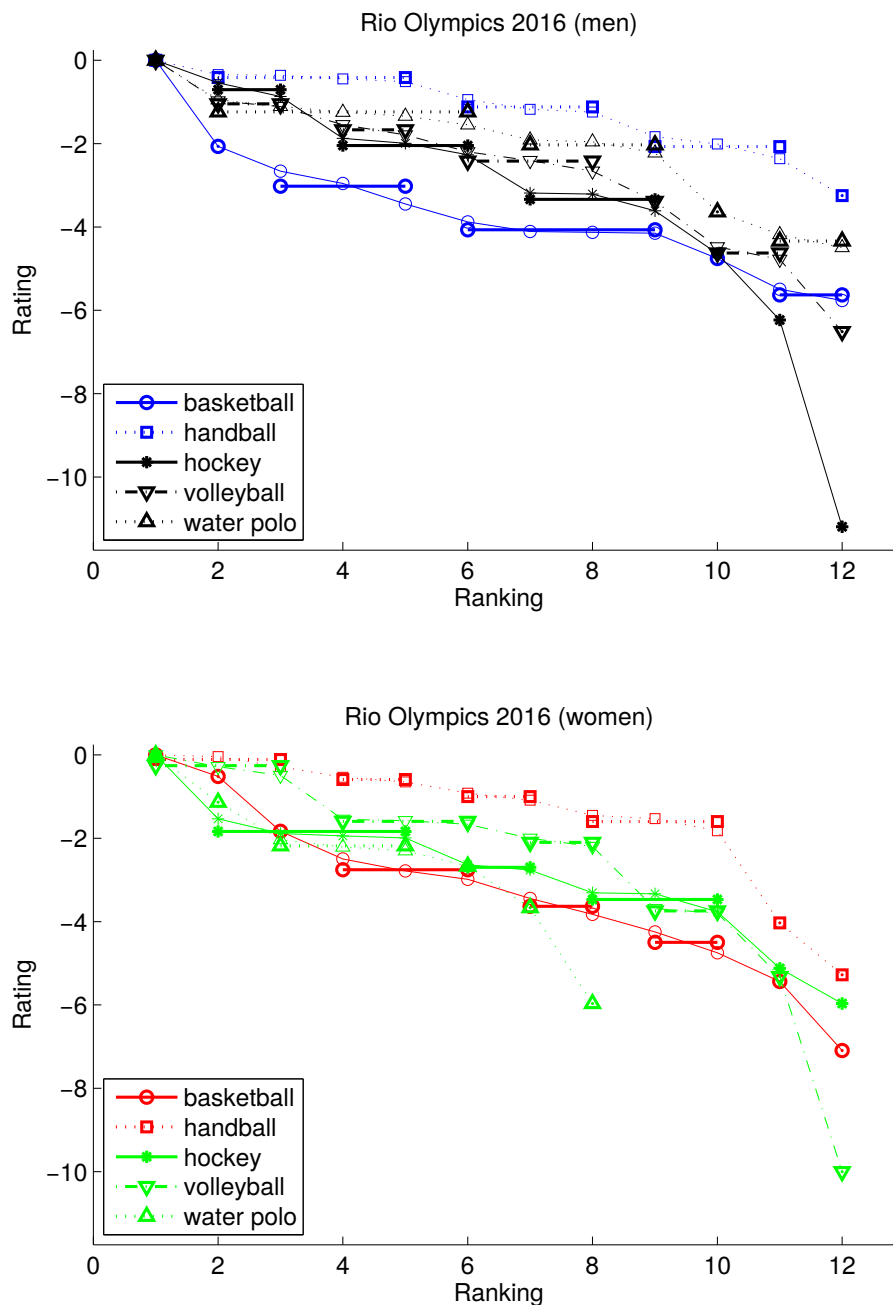


Figure 3: Clustering result of normalized rating of four sports in Rio 2016 for qualified teams

# Estimating the Duration of Professional Tennis Matches for Varying Formats

Stephanie Kovalchik\* and Martin Ingram\*\*

\*Institute of Sport, Exercise and Active Living, Victoria University,

PO Box 14428, Melbourne 8001, VIC, Australia, email: stephanie.kovalchik@vu.edu.au

\*\* Silverpond, 382 Little Collins Street Melbourne VIC, 3000, email: martin.ingram@gmail.com

## Abstract

There is growing concern in professional tennis over the duration of tennis matches. Yet as governing bodies begin to consider introducing faster match formats like Fast4, the impact these formats could have on the professional game remains poorly understood. In this paper, we develop a shot-level Monte Carlo match simulation approach for estimating the duration, points played, and upset probability given a specific match format. Our model is built on studies of predictors of the in-play and between-play time of matches using Hawk-eye tracking data and publicly available shot-level tennis data. When we applied our models to a variety of match formats with serve characteristics that were representative of current elite players, we found that Fast4 formats had an expected duration of 1 hour, best of 3 of 90 minutes, and best of 5 of two hours. Our results also show that longer matches tend to favor the better player and make match outcomes more predictable. Fast4 formats have a typical upset frequency of 20% compared to 13% for best of 3 matches and 10% for best of 5 matches. The modeling approach we have developed can be a useful resource for tennis governing bodies in assessing the impact of new match formats.

## 1 Introduction

There is growing concern in professional tennis over the duration of tennis matches. Since the serve-and-volley era of the 1990s, tennis has shifted toward a slower style of play characterized by fewer points at net and longer rallies from the baseline. The depth of the men's and women's tours have also grown over time, resulting in more tightly contested matches and more total points played. Each of these factors have contributed to a steady increase in the typical length of matches and a greater likelihood of epic matches like the 11-hour Isner-Mahut marathon at the 2010 Wimbledon Championships.

Extended matches raise a number of possible problems for the sport. Increases in match lengths can make it difficult for tournaments to complete matches on schedule, particularly when tournaments have to also contend with weather or other delays. There is also concern that the sport will lose fans if four-hour match lengths were to become the norm. A less well-understood consequence of longer periods of competitive play is the potential increased risk of injury to players.

In response to these trends and concerns, tennis stakeholders have begun to consider ways to modify the match format of today's game in order to curb long match lengths. The alternative formats involve a variety of changes to the way games, sets and matches are won. Common strategies that have been proposed include

the exclusion of advantage points, reductions to the number of games played in a set, and changes to how the final set of close matches are decided.

Only one previous study has examined the question of match duration and match format in detail. In this work, Barnett and colleagues considered the change in total points played for a variety of formats (Barnett, 2014). Although an important contribution, there are two major questions that were not addressed which limit the usefulness of this prior work. First, the paper did not estimate the match duration implied by the number of points played, which is the most relevant unit of duration for tournament organizers and fans. Secondly, the paper did not consider some of the more recently proposed match formats of most interest to tennis stakeholders, such as the Fast4 format or format used by International Premier Tennis League (IPTL).

In this paper we present a methodology for estimating the time of professional tennis matches. Our approach separates match duration into time in play and time between play. To improve the accuracy of the estimates time in play, we simulate the number of shots played per point and points played per match given the service strength of the two competitors. The time between play, on the other hand, is a by-product of the time between points, changes of ends and number of sets played and how much time the tournament allocates to each. The advantage of the methodology we present is that, given information about the expected service performance of each competitor, it can be used to estimate the expected time of numerous possible match formats. This paper describes the method in detail and uses it to examine the duration and randomness of match outcomes for nine current and proposed formats for matches in elite tennis.

## 2 Model of Match Duration

Estimating the duration of a tennis match requires an understanding of the factors that influence the length of a point. In this work, we divide the duration of a match into *time in play* and *time between play*.

### 2.1 Time In Play

The time in play is the time from the start of the serve up until a linesman makes an out call, the ball goes into the net, or the player wins the point. The duration of time within the point is determined by the number of shots, the speed of shots, and the time it takes player to get to make contact with the ball. This process repeats for every point until the game is won.

We model the time in play by first determining the number of points for the match, denoted as  $\eta$  in Figure 1. This is simulated using Monte Carlo methods of a tennis match under a specific format, as has been previously described (Newton and Aslam, 2006). The input parameters for the point simulation are the serve bonus,  $\delta$ , and malus,  $\Delta$ . Using the terminology of Klaassen and Magnus, 2014, the bonus is the sum of the proportion of points on serve by both competitors. The malus is the absolute difference in the proportion of points won on serve.

To estimate the time within a point, we first estimate the number of shots,  $\mathcal{S}$ . Given the number of shots, we then estimate the expected time to play that many shots in a rally,  $\mathcal{I}\mathcal{P}$ . The sum of the time in play across all points gives the estimate of the total play time for the match.

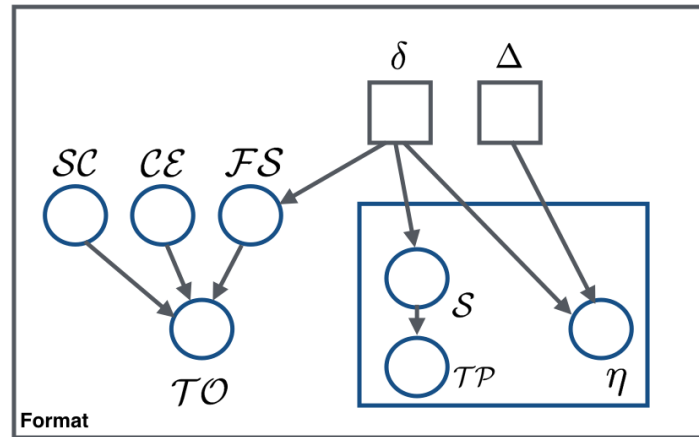


Figure 1: Plate diagram of Monte Carlo simulation for match duration

## 2.2 Time Between Play

The time outside of play includes the number and time a player takes to prepare to serve within a game, the number and time of changes of ends, and the number and time of set changes. For simplicity, we will exclude the additional time due to unusual match interruptions such as a medical timeout or bathroom break.

## 3 Methods

### 3.1 Model

Having now described the broad approach for the Monte Carlo model, in this section we detail the fixed and stochastic components in detail.

The parameters of the model are listed in Table 1 and their role in the model is depicted in Figure 1. Of the stochastic components,  $\eta$ , the number of points, was dictated by the match format and service characteristics, which were all fixed parameters of the model. The remaining stochastic components required a distributional assumption.

To determine a reasonable distribution for the shots per point, we examined the relationship between the number of shots per rally and the service bonus, service malus, and surface of the match using data from the Match Charting Project ([www.tennnisabstract.com](http://www.tennnisabstract.com)) for 1,582 men's matches and 966 women's matches from 2010 to the present. We found that the expected shot count and variance could be accurately approximated with a quasi-Poisson distribution conditional on the service bonus. Figure 2 shows the observed distribution of shots against an equal number of simulated shots from the quasi-Poisson distribution and indicates an overall good fit on all surfaces.

Determining a model for the time taken to complete a rally of a specific shot length required accurate information of the start and end of points of varying rally lengths. For this purpose, we used tracking data from the Australian Open obtained by Hawk-eye Technologies. Using data for 33,788 points across 161 men's matches and 21,450 points across 170 women's matches at the 2015 and 2016 Australian Opens, we



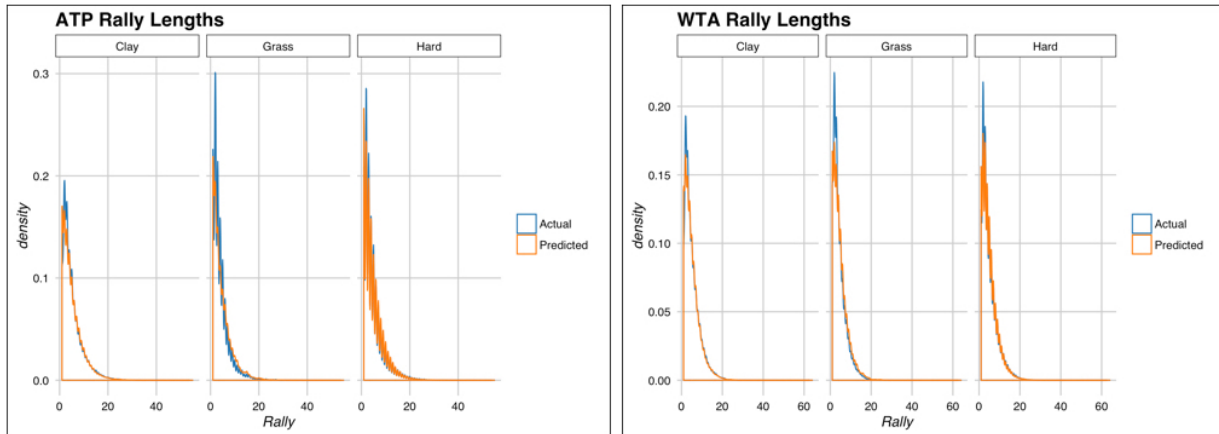


Figure 2: Observed shot distributions against an equal number of simulated shot distributions from the quasi-poisson model use in the match simulation for 1,582 ATP matches (left) and 966 WTA matches (right)

found a strong log-log linear relationship between the seconds played and number of shots played in the rally. Figure 3 compares the observed and simulated time distributions for the same distribution of rally lengths and show a good representation of the actual times in play.

We were concerned that conditioning on shot number and serve characteristics might not be sufficient to account for the influence of surface on the lengths of points. Because the authors only had access to tracking data at the Australian Open, we could not examine the time in play differences by surface directly at the point-level. However, we did have the total match duration and total shots played for the sample of matches from the Match Charting Project. Using this aggregate data, we performed a regression analysis to see whether surface explained differences in match times after accounting for shots played and the service characteristics of the players through the bonus and malus for the match. No statistically significant differences for surface were found, which indicated that surface effects primarily act through the bonus and malus conditions of the match.

The time between play was derived from the total points, percentage of first serves in, total changes of ends, and total set changes. The changes of ends and set changes are byproducts of the match format and serve characteristics. We assume that a match format will allocate a fixed time for every change of end and set change. Current professional matches implement a change of ends of 90 seconds in length and a set change of 120 seconds, what we will call a *long* break scenario. Some alternative formats like Fast4 use shorter breaks; 60 seconds for changes of ends and 90 seconds for set changes, what we will call a *short* break scenario.

The remaining component of the time between play is the time to prepare for serve. This is determined by a player's pace and their first serve percentage. For simplicity, we assume that a player will use the maximum allocated time to prepare for their first and second serves, which is 20 seconds and 10 seconds at Grand Slam events.

Our exploratory analysis of the Match Charting Project found that the first serve percentage was associated with the service characteristics of the match. For men, we found that the percentage of first serves in was best modelled as a function of the service bonus and surface (Figure 4, left panel). Women's first serve

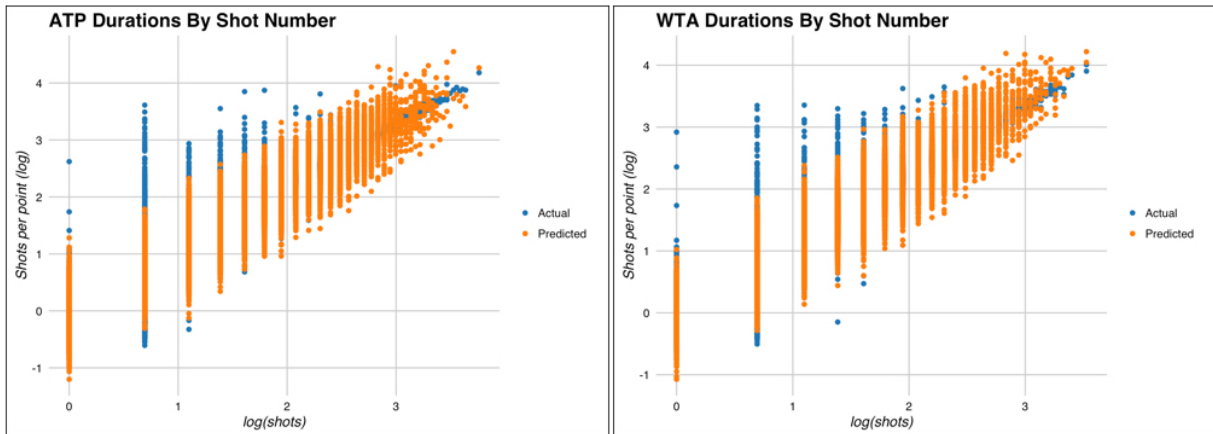


Figure 3: Observed and simulated time durations by observed rally length from a linear model for the ATP (left) and WTA (right)

percentage in was more random and appeared independent of serve characteristics and surface (Figure 4, right panel).

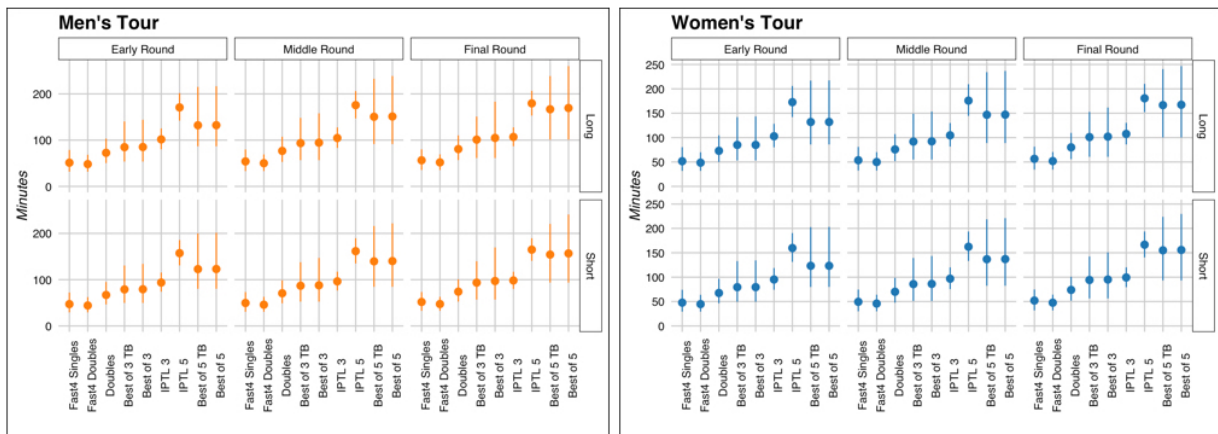


Figure 4: Observed and simulated percentage of first serves in by serve bonus from a linear model for the ATP (left) and WTA (right)

All of the above components define the necessary information for simulating the points and time duration of a match with a specific format and service characteristics.

### 3.2 Formats

We applied the match simulation model to nine different match formats. These are described in Table 2. The Best of 3 format with a tiebreak is the most commonly used at the professional level, being used for all WTA

Parameter	Description	Distribution
$\delta$	Serve bonus	Fixed
$\Delta$	Serve malus	Fixed
$\eta$	Number of points	Determined by format given $\delta, \Delta$
$\mathcal{S}$	Shots per point	Quasi-poisson $\sim \lambda = 2.89 - 1\delta, \phi = 3.3$ [ATP] Quasi-poisson $\sim \lambda = 2.33 - 0.7\delta, \phi = 2.7$ [WTA]
$\mathcal{I}\mathcal{P}$	Time-in-play per point	Gaussian (log) $\sim \mu = 0.1\log(\mathcal{S}), \sigma = 0.3$
$\mathcal{C}\mathcal{E}$	Changes of ends	Determined by format given $\delta, \Delta$
$\mathcal{I}\mathcal{C}\mathcal{E}$	Time of change of end	Determined by format given $\delta, \Delta$
$\mathcal{S}\mathcal{C}$	Set changes	Determined by format given $\delta, \Delta$
$\mathcal{I}\mathcal{S}\mathcal{C}$	Time of set change	Determined by format given $\delta, \Delta$
$\mathcal{F}\mathcal{S}$	Percent first serves in	Gaussian $\sim \mu = 48 + 11\delta + 3GRASS + 2CLAY, \sigma = 5$ [ATP] Gaussian $\sim \mu = 62, \sigma = 6$ [WTA]
$\mathcal{I}\mathcal{O}$	Time-between-play	$CE \times \mathcal{I}\mathcal{C}\mathcal{E} + SC \times \mathcal{I}\mathcal{S}\mathcal{C} + v[20 + 10 \times (1 - \mathcal{F}\mathcal{S})]$

Table 1: Description of simulation parameters and distributions

tour events and the US Open. And the other women's Grand Slams an advantage set is played in the final set of the Best of 3 match. The Best of 5 format, although less common, is of particular importance because it is used for men's Grand Slams and Davis Cup. As on the women's tour the US Open is the only Grand Slam that uses a final set tiebreak.

In addition to the current formats used for singles matches, we considered using the doubles match format for singles, which can be thought of as an abbreviated Best of 3. What we call the 'IPTL' format is inspired by the new exhibition tour. The IPTL is a team format, where singles and doubles matches are played over the course of five sets and points are earned based on the games won in each set. In this paper, we consider a 3 and 5-set singles format of the IPTL, including its features of dropping advantage points and lets as well as the use of the first-to-seven tiebreak.

The final formats considered are the Fast4, the most extreme variant on current match formats. Sometimes referred to as the T20 of tennis, the Fast4 was introduced by Tennis Australia to provide a match format that could be played within 60 or 90 minutes. It's main features are the 4-game set, the absence of ads and lets, and the use of a first-to-five point tiebreak; features which makes this format most similar to the IPTL.

### 3.3 Analysis

We simulated professional singles matches for each of the nine match formats under a variety of service conditions and between-play break lengths. The service conditions were selected to capture differences by surface and tournament round. Using the sample of matches from the Match Charting Project we found that the range of bonus and malus shown in Figure 5 captured 95% of match conditions.

For each combination of bonus and malus and for both a long and short break condition, we simulated

Name	Description
Best of 3	Sets are played to six games. Winner is first to win two sets. Tiebreak is played at 6-6 except in third set when an advantage set is played.
Best of 3 TB	This format is the same as the Best of 3 except that third set uses a tiebreak at 6-6.
Best of 5	Same as Best of 3 but the winner is the first to win three sets.
Best of 5 TB	This format is the same as the Best of 5 except that third set uses a tiebreak at 6-6.
Doubles	First two sets are played as a Best of 3 but there is no ad and tiebreaks are played at 6-6. Third sets are decided by a ten-point tiebreak.
IPTL 3	Three sets with no ad and no lets. At 5-5, players play a first to 7-point tiebreak. <a href="http://www.iptlworld.com/format">http://www.iptlworld.com/format</a>
IPTL 5	Same as IPTL but five sets.
Fast4 Singles	Best of 3 sets with sets played to four games. There is no ad and no lets. At 3-3, players play a first to five-point tiebreak.
Fast4 Doubles	Same format as Fast4 singles for first 2 sets. Third set is decided by ten-point tiebreak.

Table 2: Tennis Match Formats

100,000 matches for each format. When summarising the matches, we grouped the bonus and malus to represent the different tours and different round conditions. In doing this, we tried to reflect the tendency for early round matches to have a greater disparity in serve performance than later round matches. For the men's tour, we defined an early round match as one with a bonus of 1.25 or less and malus of 0.10 or greater, final round matches were conditions with a bonus greater than 1.25 and malus less than 0.10, all other conditions were reported as middle round matches. For women, an early round match as one with a bonus of 1.10 or less and malus of 0.10 or greater, final round matches were conditions with a bonus greater than 1.10 and malus less than 0.10, all other conditions were reported as middle round matches.

The results include the distribution of minutes played, the distribution of points played, the frequency of matches over 3 hours, and the frequency of upsets. Upsets were defined as the event of a player with the lower serve percentage in the match losing the match. Distributional summaries used the mean and 95% uncertainty interval. The simulation of points played were performed in Python and the time simulation given the number of points was performed in R using programs written by the authors.

	<b>WTA Grass</b>	<b>ATP Grass</b>					
	<b>WTA Hard/Clay</b>	<b>ATP Hard/Clay</b>					
<b>Bonus <math>\delta</math></b>	1.05	1.10	1.15	1.20	1.25	1.30	1.35
<b>Malus <math>\Delta</math></b>	0.00	0.05	0.10	0.15			

Figure 5: Serve conditions evaluated in simulation study

## 4 Results

Across the nine match formats, the expected points played varied from 72 (Fast4 Doubles) to 277 (IPTL 5) points for the men’s tour and 71 (Fast4 Doubles) to 276 (IPTL 5) points for the women’s tour (Figures 6). Fast4 matches had the fewest points played on average, having half the points played of Best of 3 matches and one-third the points played of best of 5 matches, approximately. The Doubles format was the next fastest, after Fast4, having about 20 points fewer points played on average than a corresponding Best of 3 match.

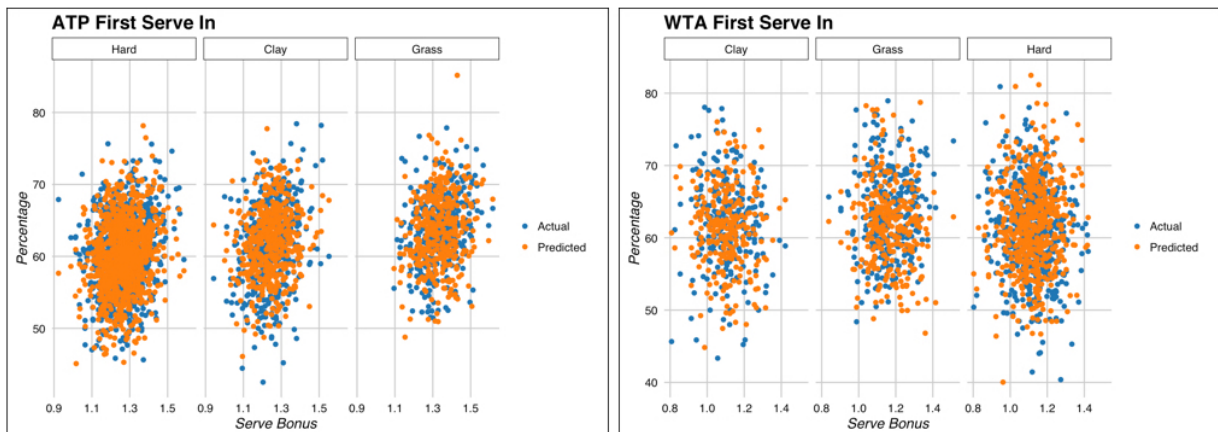


Figure 6: Point distribution by format and round for the men’s tour (left) and women’s tour (right). Points show the mean over the match simulations and lines denote the 95% uncertainty interval.

In contrast to Best of 3 and Best of 5 formats the IPTL format that play all sets, regardless of score, had a higher average number of points. However, these formats also had much less variance around the expected number of points. Similarly, when we contrast tiebreak final set formats to advantage set formats, we see that they are very similar in most respects but advantage sets have a higher upper bound (especially in final

rounds where the service malus is small), indicating a higher risk of extended matches.

All formats saw an increase in the expected points played from early to later rounds, as the service bonus increased and service malus decreased (Figure 6). However, the relative increase varied across formats. The formats that change the least with the strength of the matchup were the Fast4 and IPTL formats, where we found an average increase of under 3% from the early to final rounds. The formats that were the most sensitive to the strength of competition were the Best of 3 and Best of 5 formats with a final set tiebreak, where the average points played increased by 30% from the early to final rounds.

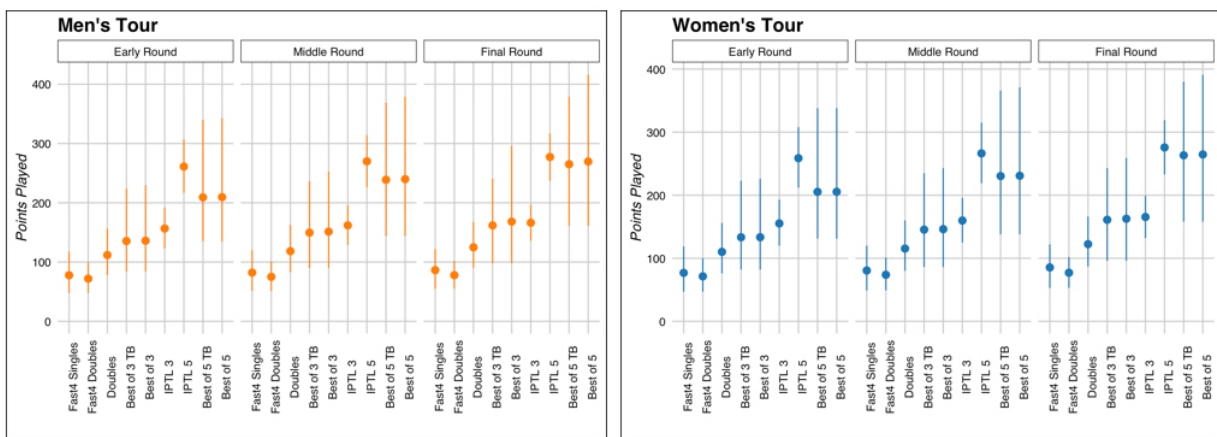


Figure 7: Time distribution by format, round and break length for the men's tour (left) and women's tour (right). Points show the mean over the match simulations and lines denote the 95% uncertainty interval.

In terms of the duration of matches, the Fast4 format is estimated to have an average time under 60 minutes, the quickest format considered. The longest format, the IPTL 5, had an average length of 170 minutes, slightly less than 3 hours on average (Figure 7). As we observed for the point distribution, 'best of' formats had quicker average times compared to the IPTL but also greater variance, meaning a higher probability of extremely long matches.

Best of 3 matches had an expected duration of approximately 90 minutes. By contrast, the Doubles format was about 15 minutes faster on average. This format also had less variability and would have minimal risk of exceeding 2 hours.

The relative increases in duration across round paralleled what was found for the point distribution. The use of long breaks added 5 to 10 minutes to the typical match length compared to short breaks (Figure 7).

We can see the relative differences in the risk of extended matches by looking at the frequency of matches over 3 hours across the different match formats and rounds. We found that only a handful of the formats had a positive probability of long matches: Best of 5 matches, the IPTL 5 and Best of 3 matches with a final advantage set (Figure 8). The formats with the highest risk were final round matches using a IPTL 5 format, where the frequency was 50% for a match with long breaks. The risk of a long match was very small though measurable for Best of 3 advantage set matches.

The frequency of 3 hour matches show the impact of the serve conditions and breaks most strongly. The frequency increased by about 10 percentage points among the 5-set formats for each round (Figure 8). Long breaks within the same round doubled the risk of an extended match for these formats.

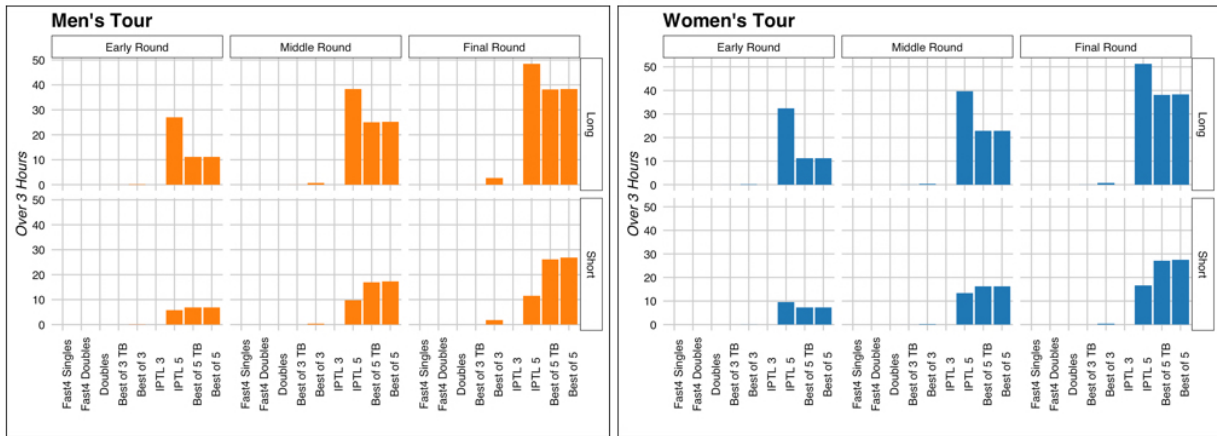


Figure 8: Frequency of matches over 3 hours for the men's tour (left) and women's tour (right).

The final outcome we report are the frequency of upsets expected with each format. In general, the fastest formats tended to have the highest frequency of upsets (Figure 9). In early rounds, for example, Fast4 formats are expected to have 1 upset for every 6 matches. In contrast, Best of 5 matches are expected to have an upset in early rounds only 3% of the time. Tiebreaks tend to increase the risk of an upset, though the difference is very small. A larger effect was found with the use of the IPTL format which had a higher upset frequency than the corresponding 'best of' formats across all rounds.

## 5 Conclusions

We have developed a shot-level Monte Carlo match simulation approach for estimating the duration, points played, and upset probability given a specified match format. This can be a useful resource for tennis stakeholders interested in changing the dynamics of the sport. It can also be helpful for tournament directors when formulating a schedule and making an assessment about the risk that some matches won't be played at their scheduled time.

Studying a number of currently used and currently proposed match formats we found a wide range in the duration and upset characteristics. Fast4 formats have an expected duration of 1 hour, Best of 3 90 minutes, and Best of 5 two hours. Matches that can play to 5 sets were the most vulnerable to extended matches, especially in the later rounds of a tournament when the strengths of the opponents are likely to be close. We did find that tournament directors could substantially decrease this length without impacting play if they were to reduce the time between play.

We have provided a description of the expected characteristics of match durations for a typical ATP and WTA tournament. At the same time, we have seen that these conditions are heavily dependent on the particular strength of the opponents and how they are seeded in the draw. These characteristics would change, for example, if the depth of the tournament underwent a shift. We can see this most dramatically by considering the characteristics of the Inzer-Mahut match of the 2010 Wimbledon Championships where the high percentage on serve of each player equated to a service bonus of 1.6 (Figure 10), conditions that

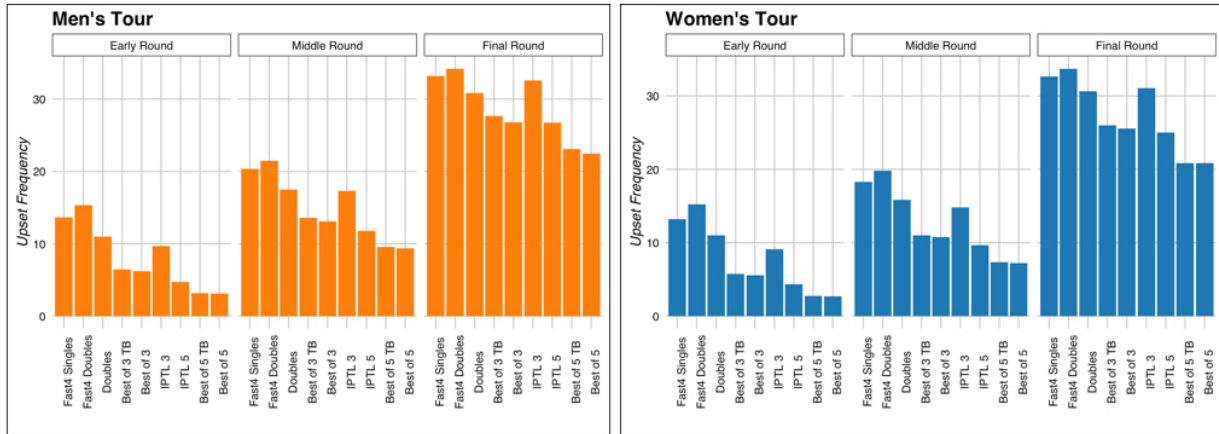


Figure 9: Frequency of upsets for the men's tour (left) and women's tour (right).

increase times across all formats but exaggerate the vulnerabilities of the advantage set formats in particular.

While there has been a growing momentum to introduce a faster format into the professional game, these discussions often overlook the impact these changes would have on the randomness of match outcomes. Our results show that longer matches tend to favor the better player and would make match outcomes less predictable. The difference in predictability due to the Best of 5 format played at men's Grand Slams compared to the Best of 3 format at women's Grand Slams is a matter of a few percentage points, yet it is a major cause behind the perceived *inconsistency* of the women's game (Kovalchik, 2015). This suggests that, without a corresponding change in tournament format, the wide adoption of formats like Fast4 would do more than impact match lengths; it would profoundly alter the outcomes and stars on the professional tours.

## References

- [1] Tristan Barnett. "A recursive approach to modelling the amount of time played in a tennis match". In: (2014).
- [2] Franc Klaassen and Jan R Magnus. *Analyzing Wimbledon: The power of statistics*. Oxford University Press, USA, 2014.
- [3] Stephanie Kovalchik. "Grand Slams are short-changing women's tennis". In: *Significance* 12.5 (2015), pp. 12–17.
- [4] Paul K Newton and Kamran Aslam. "Monte carlo tennis". In: *SIAM review* 48.4 (2006), pp. 722–742.



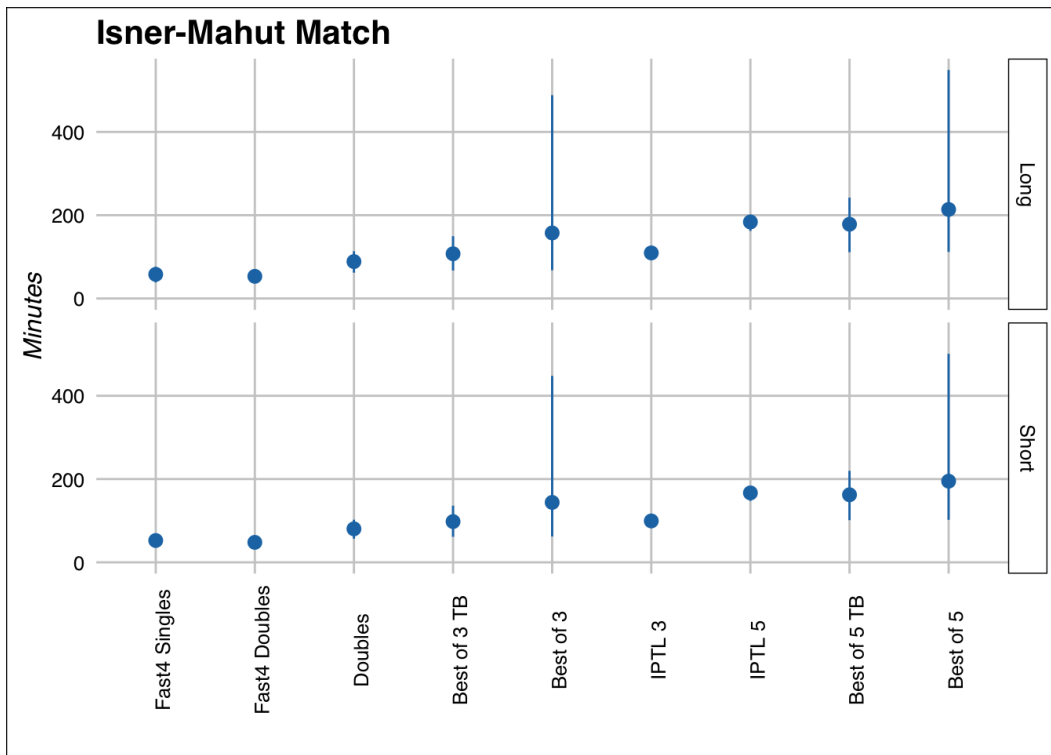


Figure 10: Simulated duration for the Isner-Mahut 2010 Wimbledon match under varying match formats

# Predicting soccer match outcome using machine learning algorithms

C. Liti\*, V. Piccialli\* and M. Sciandrone\*\*

\*DICII University of Rome Tor Vergata Via del Politecnico 1 00133 Rome:

chiara.liti@uniroma2.it

veronica.piccialli@uniroma2.it

\*\*Dipartimento di Ingegneria dell'Informazione, University of Florence Via di Santa Marta 3, 50139 Florence:

marco.sciandrone@unifi.it

## Abstract

The aim of this study is to predict the results of soccer matches of the Italian Serie A TIM championship finished with a draw at the end of the first half using mainly descriptive match-statistics collected during the first half. Moreover we aim to showing the usefulness of these match-statistics. The analysed dataset contains the results of matches played during the second half of season 2014-2015 and the first half of season 2015-2016. Due to the randomness of data and the existence of complex interacting factors, the prediction of soccer match outcomes translates into a hard three-class classification problem (i.e. the Home Win, the Draw or the Away Win). Indeed, the construction of a reliable predictive model is complicated by the limited number of overall available examples and by the even smaller number of instances of the Away win class which makes the dataset unbalanced. We manipulated the data in order to better represent the phenomenon and at the same time reduce the number of features given the small number of instances. We tested different classifiers. The default version of all classifiers gave poor results due to the entropy of the dataset. The adoption of suitable feature selection techniques allowed us to attain promising results compared with those of a reasonable baseline.

## 1 Introduction

Over the last few years there has been an increased interest in the study of numerical predictive models related to sport matches. In particular, outcome prediction in football matches has received a considerable attention in the literature, with different approaches derived both from statistics, and data mining.

In [8] a sequence of goals scored during sport match is modelled as a realization of two dependent random point processes assuming that the scoring intensity of each team has several components depending on time or on factors describing the teams and other conditions of the match. This dependence is modelled with the aid of a semi-parametric multiplicative regression model of intensity.

In [10], the authors suggest a Bayesian dynamic generalized linear model to estimate the time-dependent skills of all teams in a league, and to predict the next week-end's soccer matches. In order to estimate the skills of all teams simultaneously they use the Markov chain Monte Carlo iterative simulation technique, and apply the proposed model the English Premier League and division 1 in 1997-1998.

Predicting soccer match outcome using machine learning algorithm C.Liti, V. Piccialli, M. Sciandrone

In [2], the outcome of a match is forecasted on the basis only of data coming from previous match results. A goal score prediction model is built that uses latent features obtained from matrix factorization process and a Naive Bayes Classifier is used to predict the outcome of the match. The algorithm has been tested on results of the FIFA World Cup 2014.

In [1] a Bayesian network model is built for forecasting Association Football matches in which the subjective variables represent the factors that are important for prediction but which historical data fail to capture. In particular, the authors use the historical data to generate prior forecasts that are 'anonymous' by using predetermined levels of team-strength, rather than distinct team-names. They replace each team-name in each match in the database with a ranked number that represents the strength of that particular team for a particular season. The team-strength number was derived from the total number of points that the particular team achieved during that particular season. The model (pi-football) was used to generate forecasts about the outcomes of the English Premier League (EPL) matches during season 2010/11, and it proved to be competitive with bookmakers' performances.

In [5], a Bivariate Poisson regression is used to estimate forecasting models for goals scored and conceded. Ordered probit regression is used to estimate forecasting models for match results. Both types of models are estimated using the same 25-year data set on English league football match outcomes. The best forecasting performance is achieved using a hybrid specification, in which goals-based team performance covariates are used to forecast win, draw, lose match results.

In [7] the authors compare the performance of a Bayesian Networks (BNs) and other machine learning techniques (MC4, decision tree, Naive Bayesian, Data Driven Bayesian and a K-nearest neighbor) on predicted results (win, lose, or draw) of matches played by Tottenham Hotspur Football Club during the seasons 1995-1997. Two different sets of feature were used in this work: the first one, called the expert model, contains the following attributes: the presence or absence of three players (true, false), the playing position of Wilson, the quality of the opposing team (high, medium, and low), the venue (whether the game is played at Spurs' home ground or away), the quality of the Spurs' attacking force (low, medium, and high), the overall quality of the Spurs' team (low, medium, and high) and the Spurs' performance. The second one, called the general model, includes 30 attributes (28 players, venue, and opponent quality). The data for each season was divided into three groups of ten matches (period 1,2,3) and one group of eight matches (period 4), organised chronologically. The ordering of games was always maintained, this means that the training data set are chronologically earlier than the test data set. Every algorithm was used with both models and with different training and test sets in according to the chronology of the data (e.g., training period 1 test period 2,3,4 season 95-96, etc). The expert BN outperforms the other algorithms with mean value of accuracy of 59% (average of all disjoint training/data sets). However this model falls short in several regards: self-admittedly, it relies upon trends from a specific time period and is not extensible to later seasons.

In this work we consider the problem of predicting the outcome of a soccer match finished with a draw at the end of first half using mainly the information stored during the first half of the game. This was done with the aim of analysing "equilibrate" matches for which the prediction task may be particularly difficult. The outcome of the game (Home Win, the Draw or the Away Win) is predicted using several match-statistics (for instance, number of crosses, medium value of all the players' barycentre, number of completed passes, and so on). Some features related to the teams ( for instance, sum of points collected during previous matches, rank positions, and so on) were also embedded in the model. We built several predictive models using different machine learning techniques (neural networks, support vector machines, etc.) and feature selection

strategies. The computational experiments were performed on a limited number (166) soccer matches of the Italian Serie A TIM championship. The results obtained with some of the built classifiers (taking into account the small number of available instances) and the comparison with a reasonable baseline seem to show that the recorded match-statistics of the first half may contain information useful to predict the outcome of a match.

The paper is organized as follows. In Section 2 the available data of the soccer matches are described. In Section 3 we present the machine learning models and the feature selection strategies used in our computational experiments. The obtained numerical results are shown in Section 4. Finally, Section 5 contains some concluding remarks.

## 2 The dataset

The analysed dataset contains the results of matches finished with a draw at the end of the first half during the second half of season 2014-2015 and the first half of season 2015-2016 of Italian Serie A TIM, thus it contains 166 matches. The initial set of features is composed by 50 attributes representing the most relevant match-statistics (shown in Table 1) collected for each team during the first half.

Table 1: Selected descriptive match statistics

Attribute	Description
Goals	Number of Half Time Goals.
Score	Sum of points collected during previous matches.
Minutes	Number of minutes spent on the field during the first halve.
Yellow Cards	Number of Yellow Cards.
Red Cards	Number of Red Cards.
Possession	Possession is measured in minutes of effective playing time.
Medium Position	Medium value of all the players' barycentre
Corners	Number of kick taken from either end of the goal line
Free Cross Kicks	Number of free cross kicks.
Free Kicks	Number of free kicks.
Fouls Committed	Number of fouls committed by all players leading to a free kick for the opposite team.
Offsides	Number of offsides.
Completed Passes	Number of completed passes.
Long completed passes	Number of long completed passed
Killer Passes	Number of penetrative passes.
Counter-attacks	Number of counter-attacks.
Shots on Goal	Number of shots on target.
Crosses	Number of crosses.

Table 1 Selected descriptive match-statistics

Name of attribute	Description
Saves	Number of times when a goalkeeper catches or punches away a shot on goal.
goalkeeper out	number of times the goalkeeper goes out of the goal post
ball steals	number of ball steals at the end of a play
actual ball steals	number of real ball steals
temporary ball steals	number of temporary ball steals
points	Sum of points collected in the last two matches.
Team market Value	official team market value at the time of the match
Effective Team market Value	team market value computed on the team members that are playing the match
Full time result	H=Home Win, D=Draw, A=Away Win

### 3 Feature selection and machine learning classifiers

Due to the randomness of data and the existence of complex interacting factors, the prediction of soccer match outcomes could be translated into a hard three-class classification problem. Due to the dimension of the sample analysed the number of attributes was reduced computing the difference between the Home and Away descriptive match-statistics. Moreover we decided to remove the attribute *Red Cards* since it contained few values different from zero. As a result of these action the dataset was composed of 27 attributes.

In order to ensure robustness of the results, ten different pairs of training (containing 116 instances corresponding to about 70% of the overall set) and testing sets (with the remaining 50 instances) were analysed. Each pair is disjoint and the number of instances for class is distributed proportionally among the training and testing sets.

As a first step, we applied four different classifiers, in their default versions:

1. Naive Bayes (already used in [7])
2. LibSVM with C-classification [6]
3. LibSVM with  $\nu$ -classification [6]
4. RBFClassifier implemented in WEKA [4] [3]

In Tables 2 and 3, the (average) results are reported in terms of Accuracy and True Positive Rate for the three classes both on training and test respectively. As we can see, the best classifiers are Naive Bayes and RBFClassifier.

In order to improve the results, we defined four different feature selection strategies making use of the filter ReliefF [9]. This method is a robust feature selection technique that randomly selects an instance  $R_i$ , then searches for  $k$  of its nearest neighbours from the same class called nearest hits  $H_j$ , and also  $k$  nearest neighbours from each of the different classes, called nearest misses  $M_j(C)$ . It updates the quality estimation

Predicting soccer match outcome using machine learning algorithm C.Liti, V. Piccialli, M. Sciandrone

Classifier	Accuracy	TPR Home Win	TPR Away Win	TPR Draw
Naive Bayes	0.5517	0.6884	0.3679	0.5356
LibSVM C-classification	0.8207	0.9279	0.5607	0.8800
LibSVM $\nu$ -classification	0.9940	1.0000	0.9857	0.9933
RBFClassifier	0.5852	0.6256	0.3071	0.7201

Table 2: Average TP rates on Training Set employing the classifiers in their default version

Classifier	Accuracy	TPR Home Win	TPR Away Win	TPR Draw
Naive Bayes	0.4340	0.6278	0.2167	0.3900
LibSVM C-classification	0.3920	0.5222	0.0333	0.4900
LibSVM $\nu$ -classification	0.3640	0.4167	0.2083	0.4100
RBFClassifier	0.4360	0.4833	0.2501	0.5200

Table 3: Average TP rates on Test Set employing the classifiers in their default version

$W[A]$  for all attributes  $A$  according to their values for  $R_i$ , hits  $H_j$  and misses  $M_j$ . Note that before performing all the feature selection strategies, we normalized all the training sets. Due to the noise of the data we have decided to weight nearest neighbours by their distance.

**First Strategy** We applied the filter ReliefF on each training set, and we deleted only those attributes having a negative score in all training sets.

**Second Strategy** We applied the filter ReliefF on each training set, and we deleted only those attributes having a negative mean score (the mean was computed among the ten scores for each attribute).

**Third Strategy** We applied ReliefF again on the reduced dataset produced by the first strategy maintaining all the attributes having positive score in at least six training sets.

**Fourth Strategy** We applied ReliefF again on the reduced dataset produced by the first strategy removing all the attributes with negative scores in at least eight training sets.

## 4 Numerical Results

The following tables show the classification results in terms of average accuracy and true positive rate for each class of all training sets using the algorithms described above. We omitted the detailed results for each split, but it emerges clearly that the dataset is highly unstable (the performance may vary a lot depending on the considered split) and we believe this is due both to the complexity of the prediction task and to the too small number of instances.

In order to evaluate the quality of the achieved results, we compare our true positive rates with the ones obtained by a suitable “thumb rule” that we call “baseline”. The baseline determines the outcome of each match based on the following simple rule: if the absolute value of the difference between scores at the actual

Classifier	Accuracy	TPR Home Win	TPR Away Win	TPR Draw
Naive Bayes	0.5793	0.6768	0.4679	0.5556
LibSVM C-classification	0.8147	0.9186	0.5714	0.8667
LibSVM $\nu$ -classification	0.9940	1.0000	0.9857	0.9933
RBFClassifier	0.5811	0.6279	0.2750	0.7267

Table 4: Average TPR on the Training Set employing the Feature Selection 1

Classifier	Accuracy	TPR Home Win	TPR Away Win	TPR Draw
Naive Bayes	0.4280	0.6167	0.2583	0.3600
LibSVM C-classification	0.3960	0.5389	0.0167	0.4950
LibSVM $\nu$ -classification	0.3860	0.4556	0.2000	0.4350
RBFClassifier	0.4520	0.5000	0.2582	0.5256

Table 5: Average TPR on the Test Set employing the Feature Selection 1

Classifier	Accuracy	TPR Home Win	TPR Away Win	TPR Draw
Naive Bayes	0.5603	0.6674	0.4286	0.5400
LibSVM C-classification	0.7810	0.9047	0.5036	0.8355
LibSVM $\nu$ -classification	0.9940	1.0000	0.9857	0.9933
RBFClassifier	0.5242	0.6140	0.2322	0.6978

Table 6: Average TPR on the Training Set employing the Feature Selection 2

Classifier	Accuracy	TPR Home Win	TPR Away Win	TPR Draw
Naive Bayes	0.4340	0.5944	0.3000	0.3700
LibSVM C-classification	0.3860	0.5000	0.0667	0.4750
LibSVM $\nu$ -classification	0.3860	0.4556	0.2000	0.4350
RBFClassifier	0.4220	0.5056	0.2083	0.4750

Table 7: Average TPR on the Test Set employing the Feature Selection 2

Classifier	Accuracy	TPR Home Win	TPR Away Win	TPR Draw
Naive Bayes	0.5560	0.6349	0.4500	0.5467
LibSVM C-classification	0.7440	0.8512	0.4143	0.8467
LibSVM $\nu$ -classification	0.9673	0.9837	0.9322	0.9734
RBFClassifier	0.5386	0.6115	0.2536	0.6466

Table 8: Average TPR on the Training Set employing the Feature Selection 3

Classifier	Accuracy	TPR Home Win	TPR Away Win	TPR Draw
Naive Bayes	0.4440	0.5945	0.3333	0.3750
LibSVM C-classification	0.3740	0.4778	0.0167	0.4950
LibSVM $\nu$ -classification	0.3680	0.4056	0.2250	0.4200
RBFClassifier	0.4260	0.5057	0.2084	0.4850

Table 9: Average TPR on the Test Set employing the Feature Selection 3

Classifier	Accuracy	TPR Home Win	TPR Away Win	TPR Draw
Naive Bayes	0.5690	0.6721	0.4464	0.5467
LibSVM C-classification	0.7931	0.9139	0.5072	0.8555
LibSVM $\nu$ -classification	0.9888	0.9953	0.9822	0.9867
RBFClassifier	0.5904	0.6233	0.3134	0.7198

Table 10: Average TPR on the Training Set employing the Feature Selection 4

Classifier	Accuracy	TPR Home Win	TPR Away Win	TPR Draw
Naive Bayes	0.4320	0.6000	0.2667	0.3800
LibSVM C-classification	0.3920	0.5333	0.0333	0.4800
LibSVM $\nu$ -classification	0.3880	0.4556	0.2250	0.4250
RBFClassifier	0.4810	0.4778	0.2415	0.4700

Table 11: Average TPR on the Test Set employing the Feature Selection 4



Classifier	Accuracy	TPR Home Win	TPR Away Win	TPR Draw
Baseline		0.3220	0.2000	0.5100
Naive Bayes	0.4440	0.5945	0.3333	0.3750
LibSVM C-classification	0.3860	0.5000	0.0667	0.4750
LibSVM nu-classification	0.3880	0.4556	0.2250	0.4250
RBFClassifier	0.4520	0.5000	0.2582	0.5256

Table 12: Summary of the results and comparison with the baseline

moment of the championship is less than a threshold, the baseline will predict a Draw. If this difference is positive and higher than the threshold for the Home team, then the baseline will predict a Home Win. While if this value is higher than the threshold in favour of the away team it will predict an Away win. We decided to use five as threshold for decision making.

In order to summarize the results we report in the following table the best predictions for each classifier comparing with the outcome predicted using the baseline.

It emerges that despite the small size of the dataset, we get results that are better than those predicted by the baseline, and this shows that the selected attributes contain some information useful to predict the outcome, and we expect much better results on larger datasets.

Table 13 and 14 show the results obtained by choosing an appropriate parameter setting for the RBF-Classifier increasing the number of neurons in the hidden layer. The dataset used to employing this strategy is the one defined using the first feature selection rule. We decided to use this attribute reduction because it turned out to be the best feature configuration for RBFClassifier.

Classifier	Accuracy	TPR Home Win	TPR Away Win	TPR Draw
RBFClassifier with two basis functions	0.6101	0.6768	0.2534	0.7733
RBFClassifier with five basis functions	0.6387	0.6950	0.3355	0.7734
RBFClassifier with seven basis functions	0.6664	0.7255	0.3892	0.7824

Table 13: Average TPR on Training Set employing the Feature Selection 1 and optimizing the RBFClassifier

Classifier	Accuracy	TPR Home Win	TPR Away Win	TPR Draw
RBFClassifier with two basis functions	0.4100	0.4111	0.1572	0.5600
RBFClassifier with five basis functions	0.4200	0.4223	0.2082	0.5450
RBFClassifier with seven basis functions	0.3780	0.4055	0.2084	0.4550

Table 14: Average TPR on Test Set employing the Feature Selection 1 and optimizing the RBFClassifier

## 5 Conclusion and future work

In this work we aimed to predict the outcome of a soccer match using mainly match-statistics collected during the first half in order to evaluating also the usefulness of these statistics. The considered prediction task is very hard taking into account the existence of complex interacting factors and the presence of unavoidable randomness in the phenomenon. The results obtained by coupling machine learning models with standard feature selection techniques, and the comparison with a baseline seem to show that match-statistics collected during the first half of a match may contain potentially useful information to predict the outcome of a match. We believe that significant improvements in the prediction performance could be obtained using a larger amount of training data. Future work will be devoted to build a homogeneous dataset with thousand of matches (of the same championship) and without outliers that could be detected by experts taking into account the peculiarities of this phenomenon. Furthermore, ad hoc feature selection techniques will be designed and extensive computational experiments will be performed in order to draw sound conclusions about the reliability of machine learning models in predicting the outcome of a soccer match on the basis of match-statistics.

## References

- [1] Constantinou, A. C., Fenton, N. E., and Neil, M. (2012) *pi-football: A Bayesian network model for forecasting Association Football match outcomes*. Knowledge-Based Systems, **36**, 322-339.
- [2] Dobravec S. (2015). *Predicting sports results using latent features: A case study*, In 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, 2015, pp. 1267-1272.
- [3] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, Morgan Kaufmann, Fourth Edition, 2016.
- [4] Frank, E. (2014). Fully supervised training of Gaussian radial basis function networks in WEKA.
- [5] Goddard, J. (2005). *Regression models for forecasting goals and match results in association football*. International Journal of forecasting, **21**(2), 331-340.
- [6] Hsu, C. W., Chang, C. C., and Lin, C. J. (2003). A practical guide to support vector classification
- [7] Joseph, A., Fenton, N. E., and Neil, M. (2006). *Predicting football results using Bayesian nets and other machine learning techniques*. Knowledge-Based Systems, **19**(7), 544-553.
- [8] Owen A. (2011). *Dynamic Bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter*. IMA J Management Math **22** (2): 99-113.
- [9] Robnik-Sikonja, M., and Kononenko, I. (2003). *Theoretical and empirical analysis of ReliefF and RReliefF*. Machine learning **53** (1-2) : 23-69.
- [10] Rue, H., and Salvesen, O. (2000). *Prediction and retrospective analysis of soccer matches in a league*. Journal of the Royal Statistical Society: Series D (The Statistician), **49**(3), 399-418.

# Draw importance in football

Michael de Lorenzo\*, Stella Stylianou\*, Ian Grundy\* and Bradley O’Bree\*

\*RMIT University, Melbourne, Australia

Corresponding author email: michael.delorenzo@rmit.edu.au

## Abstract

Probabilistic models have been frequently applied in professional sports to quantify the importance of winning a match. However, existing models fail to account for the possibility of a draw outcome, a frequent occurrence in multi-result sports like football. In this paper, we calculate match importance using a trinomial distribution model that accounts for the possibility of a drawn result. Using German Bundesliga football, we demonstrate through case studies that the importance of a match can be evaluated with respect to win and draw results separately.

## 1 Introduction

The importance of matches in professional sports has direct relevance to in statistical modeling of match outcomes and attendance figures. Probabilistic models have frequently been applied to quantify match importance with respect to a team achieving some end-of-season outcome, where importance is defined as the difference in success probabilities conditional on the result of a match (Schilling, 1994). This definition has been applied to English Premier League football (Scarf and Shi, 2008) and Australian Rules football (Bedford and Schembri, 2006), as well as in live tennis gameplay (González-Díaz et al., 2012).

A drawback of applying this probabilistic definition to football (soccer) is that the possibility of the draw outcome is not accounted for. To date, “extending the definition to include draws is not straightforward” (Goossens et al., 2012). In this paper, we further this probabilistic definition to account for the draw outcome, demonstrating that the match importance can be evaluated with respect to win and draw results separately. We complete this by applying a trinomial distribution model to German Bundesliga football.

## 2 Methods

The conditional probabilities definition of importance by Schilling (1994) is calculated by assessing the difference between the probabilities of achieving some end-of-season outcome given a team achieves a win or loss in their next match. However, since there are three match outcomes (win/draw/loss) in football, we extend the definition to include the draw as a non-negative result. Therefore, in this paper, match importance is defined as the difference between two success probabilities: the probability that a team achieves some end-of-season outcome given a non-negative (win/draw) result in their next match; minus the probability that a team achieves some end-of-season outcome given a negative (loss) result in their next match. This is presented in (1).

$$IMP(r + 1) = P(Outcome X|W \vee D r + 1) - P(Outcome X|L r + 1) \quad (1)$$

Note here that this adjusted definition assumes that a draw outcome for a team is a non-negative result. This may not be the case when a team requires nothing less than a win to achieve a specific end-of-season outcome. However, we include it as a non-negative result as a draw still awards a team one point in football, which is more beneficial than a loss, which awards zero points. We discuss this further in the Discussion section of this paper.

To calculate the probabilities, a trinomial distribution model is applied. The model requires information about the total points of the current team of interest and the current team in position  $s$  after the completion of round  $r$ . By identifying success and failure scenarios such that a lower-ranked team can overtake, but not equal, a higher-ranked team, the trinomial distribution model calculates the probability that a team achieves some end-of-season outcome given they win, draw, or lose their next match. This is completed to calculate both the win and draw importance, which can be summed together to form an overall 'result importance'.

In order to complete calculations, match outcome probabilities are required for all teams throughout the season. To demonstrate the functionality of the model, static match outcome probabilities are applied to all teams. While this does not account for critical factors such as home ground advantage or team strength, the static probabilities allow a baseline model to be established and tested. To determine the probabilities, match results from seasons 2005/2006 through 2014/2015 for first and second division Bundesliga were collected from [www.football-data.co.uk](http://www.football-data.co.uk), totaling 5,508 matches. The proportion of drawn matches within the first and second division Bundesliga was found to be 25% and 28%, respectively. Therefore, by splitting the remaining percentage into win and loss, the win/draw/loss probabilities for first division Bundesliga are equal to 37.5%-25%-37.5%, while the probabilities are set to 36%-28%-36% for the second division.

### 3 Results

To demonstrate the trinomial distribution model, case studies from past Bundesliga seasons are assessed. While the model is designed to explore difference end-of-season outcomes in football (e.g. qualification for Champions League, promotion), we elect to focus on winning the league championship as this is the primary objective of any professional football team.

#### 3.1 First case study

The first case study is from the 2008/09 Bundesliga season, where VfL Wolfsburg won the league championship by two points over FC Bayern Munich. The result importance, as well as the win and draw components, are presented in Figure 1.

At the halfway mark of the season, VfL Wolfsburg was in position nine in the standings while FC Bayern Munich was ranked second. This is reflected in Figure 1 where the overall result importance for VfL Wolfsburg is lower than FC Bayern Munich, who was only ranked second due to total goal difference. VfL Wolfsburg finished the season winning 14 of their final 17 matches, including a 5-1 final round home victory against Werder Bremen to secure the championship. This increase in form is observed in Figure 1, where the importance steadily increases as the team moves up the standings.

Due to the nature of the trinomial distribution model, the win and draw importance for both teams follow a similar trend to the overall result importance, where the win outcome always has a higher importance than the draw outcome. However, the two components do not follow an identical distribution to each other. Focusing on VfL Wolfsburg, it can be seen at round 31 that their draw

importance decreased while their win importance increased. The draw importance remained low until the final round of the season, where VfL Wolfsburg (two points ahead of FC Bayern Munich) required only a draw in their result match to secure the league championship.

During the final round, the draw importance for FC Bayern Munich decreased to zero, meaning that they require nothing less than a win in their final match. In this situation, a draw result can be seen as a negative result for the team. Despite the draw being included as a non-negative result during calculations, this example demonstrates that the trinomial distribution model is still correctly identifying when a draw outcome is sufficient to a team. Like VfL Wolfsburg, FC Bayern Munich would go on to win their final match of the season.

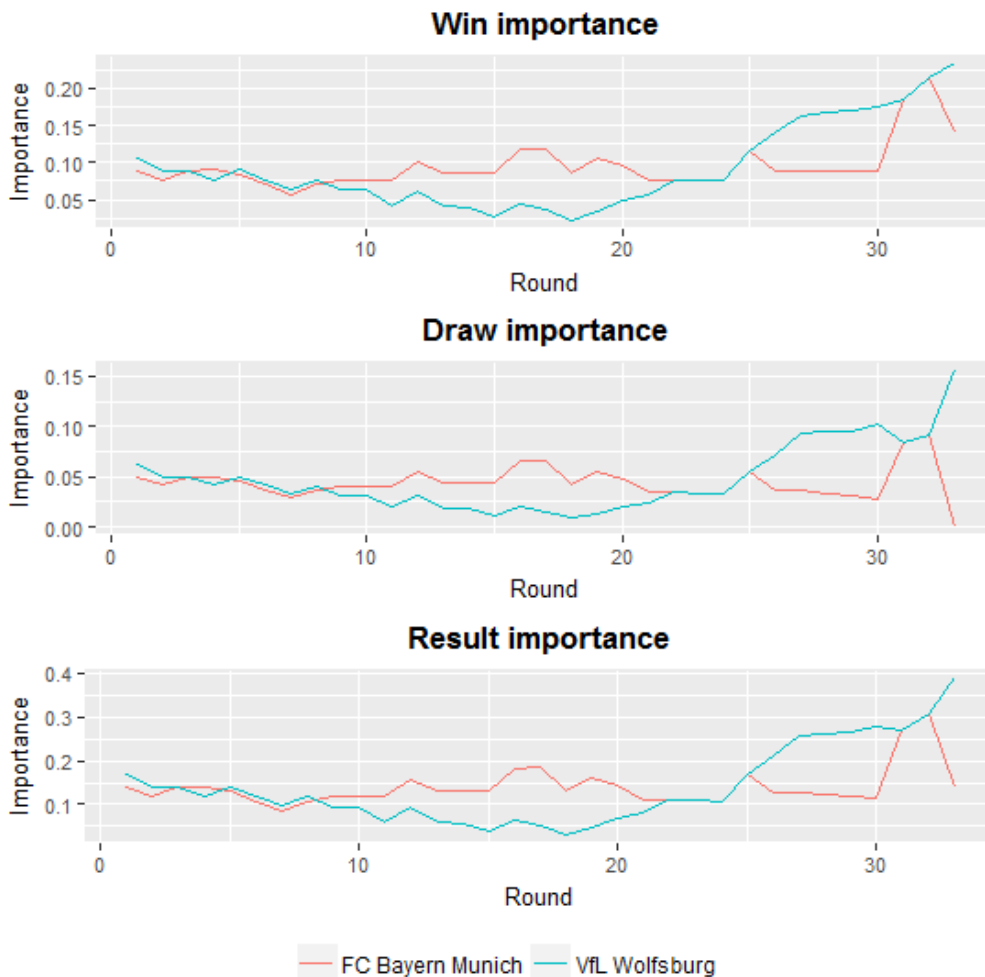


Figure 1: Win/draw/result Importance for VfL Wolfsburg and FC Bayern Munich from 2008/09 Bundesliga season

### 3.2 Second case study

The second case study explores the top three teams from the 2011/2012 2. Bundesliga season. In this season, SpVgg Greuther Fürth won the league championship by two points over Eintracht Frankfurt, who in turn finished six points ahead of Fortuna Düsseldorf. The win/draw/result importance for all three teams to finish in first position is presented in Figure 2.

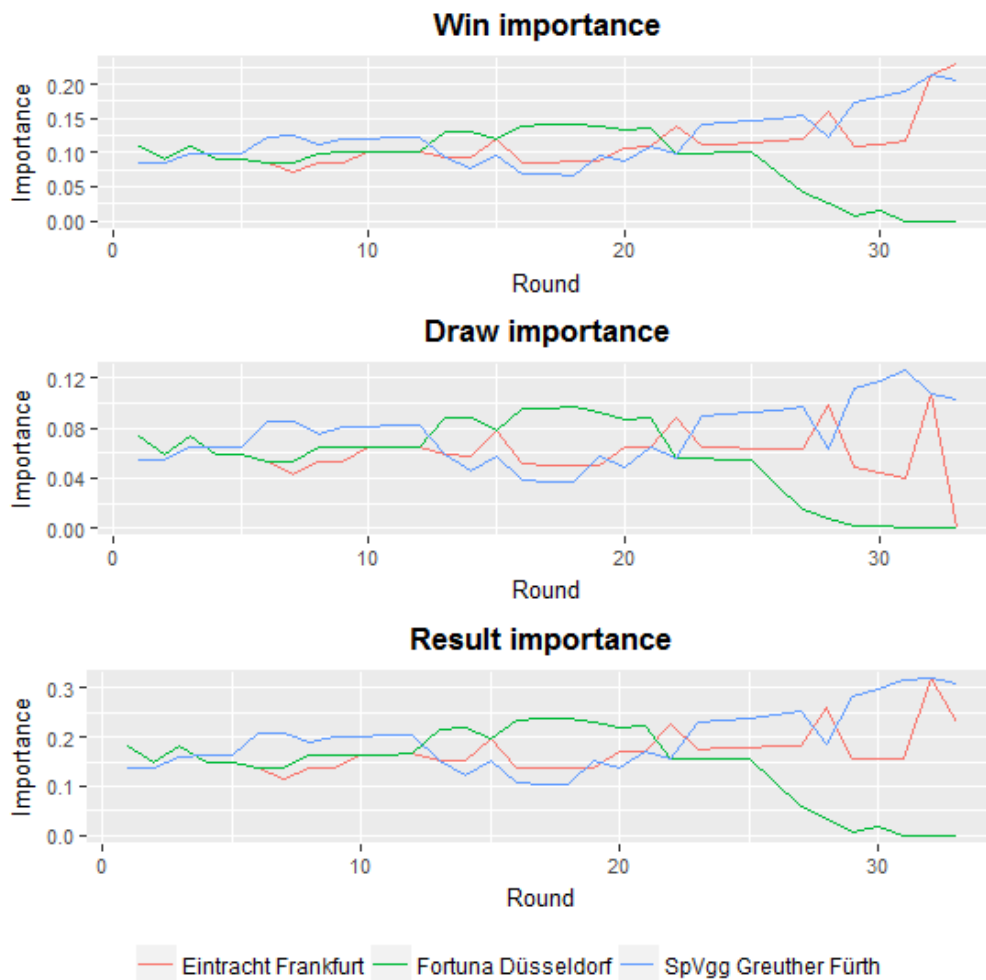


Figure 2: Win/draw/result Importance for SpVgg Greuther Fürth, Eintracht Frankfurt and Fortuna Düsseldorf from 2011/12 2. Bundesliga season

During the season, the three teams held the top three positions in the standings for 20 of the 34 rounds. Over the final nine rounds, Fortuna Düsseldorf recorded only two victories while drawing another five matches. The high number of draw outcomes resulted in their overall result importance steadily decreasing toward the conclusion of the season, where it eventually equaled zero when the

team was mathematically eliminated from championship contention. During these same final rounds, SpVgg Greuther Fürth recorded four wins and five draws while Eintracht Frankfurt won five matches and drew only one.

After round 32 of the season, SpVgg Greuther Fürth and Eintracht Frankfurt were equal on the standings with only total goal difference separating the teams. SpVgg Greuther Fürth would go on to draw their final two matches while Eintracht Frankfurt would lose their final two. Despite the two draw outcomes leading to the league championship for SpVgg Greuther Fürth, the importance in Figure 2 indicates that the draw result was becoming increasingly unfavourable compared to a win outcome. This result is a reflection of the close proximity of the teams in the standings, where a win outcome would increase either team's chances of winning the championship greater than a draw outcome. This observed result further demonstrates that the trinomial distribution model is correctly identifying when a draw outcome can be sufficient to a team.

## 4 Discussion

The results presented in this paper demonstrate that the importance of a draw outcome in football can be quantified by using a trinomial distribution model. While the draw importance is relatively low throughout the season compared to the win importance, there are circumstances in which the draw outcome can be beneficial to a team, including late in the season when a team is attempting to maintain first position in the standings. While the work presented in the previous section of this paper focused on winning the league championship, the work can easily be applied to other end-of-season outcomes, such as qualifying for Champions League or avoiding relegation.

During calculations using the trinomial distribution model, the draw outcome was included as a non-negative result. As mentioned in the previous section of this paper, a draw outcome may not always be non-negative to a team, particularly if a team requires nothing less than a win to achieve some end-of-season outcome. However, as shown through case studies from both first and second division Bundesliga football, the trinomial distribution model is correctly identifying when a draw outcome can be sufficient for a team. Nevertheless, future research and development of the model should assess varying the draw outcome between a negative and a non-negative result for individual teams.

To demonstrate the functionality of the trinomial distribution model, static match outcome probabilities were applied across all teams during calculations. However, use of constant probabilities across all teams does not account for factors such as team strength, home ground advantage and current team form. For example, over the span of the data, FC Bayern Munich had a win probability equal to 67%, which increased to 76% when playing at home. Despite this, the application of static match outcome probabilities provided positive results that the trinomial distribution model is reasonably quantifying importance in football. Nevertheless, future research on varying the match outcome probabilities to take into account these factors would likely further enhanced the observed match importance results.

While the model presented in this paper was applied to German Bundesliga football, the nature of the model allows it to be easily applied to other football leagues, including the English Premier League. Furthermore, by replacing the trinomial distribution with a binomial distribution, the model can easily be applied to sports where there are only two match outcomes (win/loss), such as NBA basketball or American football. Further research can explore the application of the model in other

Draw importance in football

M. de Lorenzo, S. Stylianou, I. Grundy & B. O'Bree

professional sports leagues, which could lead to an interesting league-to-league comparison of when the draw outcome in football is important.

## 5 Conclusion

In this paper, a trinomial distribution model was applied to German Bundesliga football to quantify match importance while accounting for the draw outcome. Results from both the first and second division indicate that the draw result in football can be important to a team given certain season circumstances. The work completed in this paper helps further the knowledge of quantifying match importance in multi-result sports.

## References

- BEDFORD, A. & SCHEMBRI, A. 2006. A Probability Based Approach for the Allocation of Player Draft Selections in Australian Rules Football. *Journal of Sports Science and Medicine*, 5, 509-516.
- GONZÁLEZ-DÍAZ, J., GOSSNER, O. & ROGERS, B. W. 2012. Performing best when it matters most: Evidence from professional tennis. *Journal of Economic Behavior & Organization*, 84, 767-781.
- GOOSSENS, D., BELIÉN, J. & SPIEKSMAN, F. R. 2012. Comparing league formats with respect to match importance in Belgian football. *Annals of Operations Research*, 194, 223-240.
- SCARF, P. A. & SHI, X. 2008. The importance of a match in a tournament. *Computers & Operations Research*, 35, 2406-2418.
- SCHILLING, M. F. 1994. The Importance of a Game. *Mathematics Magazine*, 67, 282-288.



# Home Team Advantage in English Premier League

Patrice Marek\* and František Vávra\*\*

\*European Centre of Excellence NTIS – New Technologies for the Information Society,  
Faculty of Applied Sciences, University of West Bohemia, Czech Republic: patrke@kma.zcu.cz

\*\* Department of Mathematics, Faculty of Applied Sciences,  
University of West Bohemia, Czech Republic: vavra@kma.zcu.cz

## Abstract

The home team advantage in association football is a well known phenomenon. The aim of this study is to offer a different view on the home team advantage. Usually, in association football, each two teams – team A and team B – play twice in a season. Once as a home team and once as a visiting team. This offers two results between teams A and B which are combined together to evaluate whether the team A against its opponent B recorded a result at the home field – in comparison to the away field – that is better, even, or worse. This leads to a random variable with three possible outcomes, i.e. trinomial distribution. Combination and comparison of home and away results of the same two teams is the key to eliminate problems with different strength of teams in the league. Using a uniform distribution as a prior we obtain a Dirichlet distribution as a posterior. This is later used to determine point and interval estimates of unknown parameters of the source trinomial distribution, i.e. the probability that the result at home will be better, even, or worse. Moreover, it is possible to test a hypothesis that the home team advantage for a selected team is statistically significant. This approach can be used to construct a measure of the home team advantage for a single team. Described procedure is demonstrated on English Premier League results from the 1992/1993 season to the 2015/2016 season.

## 1 Introduction

Home team advantage is phenomenon that is well known. It is used in models that estimate probability of win, draw and loss in a match. Usage of home team advantage in modelling and predicting sports results can be traced back to Maher (1982) who used one parameter to adjust strength of team's attack and weakness of team's defence for matches played on away field. Home team advantage was later used in many papers that studied different sports, e.g. in association football by Dixon & Coles (1997), in water polo by Karlis & Ntzoufras (2003) and in ice hockey by Marek et al. (2014).

Home team advantage as a self-standing phenomenon was deeply studied by Pollard & Pollard (2005). Their paper offers nice summary of previous research on this phenomenon and analysis of more than 400,000 matches in many sports played between years 1876 and 2003. They quantified home team advantage in association football as "*the number of points obtained by the home team expressed as a percentage of all points obtained in all games played*". The same definition of home team advantage was used by Allen & Jones (2014) in analysis of the English Premier League in the seasons 1992/1993–2011/2012. Their results showed that 60.77% ( $\pm 8.30$ ) of total points was won in home games.

This paper offers a slightly different view on home team advantage and – instead of points – home team advantage is based on number of goals scored and their differences. The advantage of using goals can be demonstrated on results of a team that played the same opponent at the home and away field. Let us assume, that the result at home field was 3–0 win, and the result at away field was 2–1 win. Obviously, better result was recorded at the home field; however, based on points obtained, it is not possible to distinguish between these results as the team is always awarded by 3 points. Method described in the following part will allow to distinguish between these results, and it will offer to measure the home team advantage for individual teams and observe changes during the time.

## 2 Data and Methods

English Premier League results from the 1992/1993 season to the 2015/2016 season were obtained from England Football Results and Betting Odds (2017). Data for the first English Premier League season (1992/1993) were obtained from official website Premier League Football News, Fixtures, Scores & Results (2017). This website was also used for basic control of all data, e.g. total number of scored goals by team in the whole season.

Premier League consisted of 22 teams in the first 3 seasons and of 20 teams in the rest of seasons. Balanced schedule was used in all seasons, i.e. each team played each other team exactly two times, once as a home team and once as a visiting team. This means that for each team there are 19 opponents (21 in the first three seasons) with two results in a season. These two results are combined together and used to measure home team advantage which is evaluated according to Definitions 1, 2 or 3. Naturally, each season is analysed separately to eliminate changes in teams that form the league and to eliminate changes in rosters that are usually bigger between seasons.

**Definition 1.** Active measure of home team advantage is a random variable  $A$  that can take values  $-1, 0$ , and  $1$ .  $A = -1$  for team  $T_1$  if two matches between teams  $T_1$  and  $T_2$  in a season ended with a result where team  $T_1$  scored more goals on a field of team  $T_2$  than on its own field.  $A = 0$  for team  $T_1$  if this team scored exactly the same number of goals on a home field and away field and  $A = 1$  for team  $T_1$  if this team scored more goals on its own field than on a field of team  $T_2$ . With results  $h_{T_1} : a_{T_2}$  on a home field of team  $T_1$  and  $h_{T_2} : a_{T_1}$  on a home field of team  $T_2$  the value of random variable  $A$  is determined as

$$A = \text{sgn}(h_{T_1} - a_{T_1}). \quad (1)$$

**Definition 2.** Passive measure of home team advantage is a random variable  $P$  that can take values  $-1, 0$ , and  $1$ .  $P = -1$  for team  $T_1$  if two matches between teams  $T_1$  and  $T_2$  in a season ended with a result where team  $T_1$  conceded more goals on a home field than on a field of team  $T_2$ .  $P = 0$  for team  $T_1$  if this team conceded exactly the same number of goals on a home field and away field and  $P = 1$  for team  $T_1$  if this team conceded more goals on a field of team  $T_2$  than on its own field. With results  $h_{T_1} : a_{T_2}$  on a home field of team  $T_1$  and  $h_{T_2} : a_{T_1}$  on a home field of team  $T_2$  the value of random variable  $P$  is determined as

$$P = \text{sgn}(h_{T_2} - a_{T_2}). \quad (2)$$

**Definition 3.** Combined measure of home team advantage is a random variable  $C$  that can take values  $-1, 0$ , and  $1$ .  $C = -1$  for team  $T_1$  if two matches between teams  $T_1$  and  $T_2$  in a season ended with a better result –

measured by a goal difference in matches – for team  $T_1$  on an away field.  $C = 0$  for team  $T_1$  if goal difference in both matches was exactly the same from  $T_1$ 's point of view and  $C = 1$  for team  $T_1$  if this team recorded better result – measured by a goal difference in matches – on its own field. With results  $h_{T_1} : a_{T_2}$  on a home field of team  $T_1$  and  $h_{T_2} : a_{T_1}$  on a home field of team  $T_2$  the value of random variable  $C$  is determined as

$$C = \text{sgn}((h_{T_1} - a_{T_2}) - (a_{T_1} - h_{T_2})). \quad (3)$$

All three measures are defined so that value 1 means that a result was better on a home field, 0 means that there was no difference and  $-1$  means that better result was recorded on an away field. Obviously, active measure for team  $T_1$  is passive measure for team  $T_2$ . More or less, combination of results between two same teams – as used in Definitions 1, 2 or 3 – eliminates the fact that teams in league are of different quality. All three random variables can take same values with same interpretation; therefore, in following parts the combined measure  $C$  is used and it can be easily substituted by  $A$  or  $P$  to obtain results for other two measures.

English Premier League used balanced schedule in all seasons with exactly two matches between each two teams. Let  $L$  denote number of teams in a league (for our data  $L = 22$  or  $L = 20$ ) then for each team in a season, there are  $K$ ,  $K = L - 1$ , opponents. Random sample  $C_1, C_2, \dots, C_K$  is obtained as one season's results of given team and its opponents.  $C_i$ 's are considered to be identically distributed because there are no big changes in a team during one season. Therefore, probabilities  $p_{-1}, p_0$  and  $p_1$  of possible outcomes  $-1, 0$  and  $1$  are considered constant in a season. The meaning is that during a season the home team advantage of a team is stationary. The second assumption is that  $C_i$ 's are independent. The interpretation is that matches with one opponent does not influence matches with other opponents.

*Remark 1.* Assumption that  $C_i$ ,  $i = 1, 2, \dots, K$ , are i.i.d. may not be true in reality. However, it can be expected that violation of this assumption is not strong, and therefore, it is used in the same sense in majority of studies that deal with sports. Without this simplification it would be impossible to use statistics for sports as every single match could be played under slightly different conditions (for example, in different weather conditions). Moreover, undermentioned methods will be robust, and this simplification should not result in any problems with interpretation of obtained findings.

Let  $Z_r$ ,  $r = -1, 0, 1$ , is random variable which describes number of cases in a season where it is possible to observe home team advantage ( $r = 1$ ), away team advantage ( $r = -1$ ) and no advantage ( $r = 0$ ). Obviously, for  $K$  matches in a season  $Z_1 + Z_0 = K - Z_{-1}$ . Vector  $(Z_{-1}, Z_0, Z_1)$  follows trinomial distribution with parameters  $K$  and  $p_{-1}, p_0, p_1$ . Probability mass function under this notation is given by

$$P(k_{-1}, k_0, k_1) = \frac{K!}{k_{-1}!k_0!k_1!} p_{-1}^{k_{-1}} p_0^{k_0} p_1^{k_1}, \quad (4)$$

where  $K$  is total number of opponents in a season for one team,  $p_{-1}, p_0, p_1$  are probabilities of occurring a home team advantage ( $r = 1$ ), an away team advantage ( $r = -1$ ) and no advantage ( $r = 0$ ).  $k_{-1}, k_0, k_1$ ,  $k_{-1} + k_0 + k_1 = K$ , are observations of appropriate advantage.

Bayesian inference is used to estimate unknown parameters and consequently confidence intervals. Prior distribution of parameters  $p_{-1}, p_0$  and  $p_1$  is set to be uniform, i.e. it does not matter where a team plays a match and probability in Equation 4 is used as conditional probability of observation under given parameters, i.e.  $P(k_{-1}, k_0, k_1 | p_{-1}, p_0, p_1)$ . This leads to posterior probability density of parameters  $p_{-1}, p_0, p_1$  given by

$$P(p_{-1}, p_0, p_1 | k_{-1}, k_0, k_1) = \frac{\Gamma(K+3)}{\Gamma(k_{-1}+1)\Gamma(k_0+1)\Gamma(k_1+1)} p_{-1}^{k_{-1}} p_0^{k_0} p_1^{k_1}, \quad p_{-1}, p_0, p_1 \geq 0, \quad \sum_{r=-1}^1 p_r = 1, \quad (5)$$

where  $K$  is total number of opponents in a season for one team and  $k_{-1}, k_0, k_1$ ,  $k_{-1} + k_0 + k_1 = K$ , are observations of given advantage. Equation 5 is probability density function of a Dirichlet distribution  $\text{Dir}(\alpha_1 = k_{-1} + 1, \alpha_2 = k_0 + 1, \alpha_3 = k_1 + 1)$ . Bayesian estimator of probabilities in 4 is given (using squared-error loss function) as mean value of this Dirichlet distribution, i.e.

$$\hat{p}_r = \frac{n_r + 1}{K + 3}, \quad r = -1, 0, 1. \quad (6)$$

If  $p_{-1}, p_0, p_1$  follows Dirichlet distribution  $\text{Dir}(\alpha_1 = k_{-1} + 1, \alpha_2 = k_0 + 1, \alpha_3 = k_1 + 1)$ ,  $k_{-1} + k_0 + k_1 = K$ , then marginal distribution of  $p_r$ ,  $r = -1, 0, 1$ , is  $\text{Beta}(\alpha = k_r + 1, \beta = K - k_r + 2)$  (see (Pitman 1993, p. 473)). This can be used to find individual  $(1 - \alpha_l - \alpha_u)$ -confidence intervals  $(\hat{p}_{r,l}, \hat{p}_{r,u})$  for each  $p_r$  which are given by

$$\hat{p}_{r,l} = \text{Beta}^{-1}(\alpha_l, k_r + 1, K - k_r + 2) \quad (7)$$

and

$$\hat{p}_{r,u} = \text{Beta}^{-1}(\alpha_u, k_r + 1, K - k_r + 2) \quad (8)$$

*Remark 2.* These individual confidence intervals can be used for simultaneous confidence interval of all three parameters. Based on Bonferroni inequality, they form together a  $(1 - 3(\alpha_l + \alpha_u))$ -simultaneous confidence interval.

For testing hypothesis it is necessary to obtain  $P(p_1 > p_{-1})$  from Equation 5. Using results of (Omar & Joarder 2012, p. 932) and observed values of  $k_1$  and  $k_{-1}$  this probability is estimated as

$$P(p_1 > p_{-1}) = 1 - I_{1/2}(k_1 + 1, k_{-1} + 1), \quad (9)$$

where  $I_{1/2}(k_1 + 1, k_{-1} + 1)$  is regularized incomplete beta function or cumulative distribution function of Beta distribution.

*Remark 3.*  $P(p_1 > p_{-1})$  in this paper is an estimate based on observed values of  $k_1$  and  $k_{-1}$ . However, for better readability, the word *estimate* is omitted in the following text.

$P(p_1 > p_{-1})$  is the probability of occurrence of home team advantage, i.e. it can be used as a measure of home team advantage (the higher value of  $P(p_1 > p_{-1})$ , the higher home team advantage). Hypothesis that the home team advantage is real can be accepted if  $P(p_1 > p_{-1}) \geq 1 - \alpha$ .

### 3 Results

As mentioned before, we analysed English Premier League from the 1992/1993 season to the 2015/2016 season. Totally, 9,366 matches were played in these seasons, and, thanks to promotion and relegation, there are 47 teams that played at least one season in the English Premier League. Out of these teams, only

seven played in each season (Arsenal, Aston Villa, Chelsea, Everton, Liverpool, Manchester United, and Tottenham). We also remind that in the first three seasons English Premier League consisted of 22 teams and of 20 teams in the following seasons.

For each team in each season the hypothesis that home team advantage is real was tested (see Equation 9). The hypothesis is accepted in the case where  $P(p_1 > p_{-1}) \geq 0.95$ . These tests were performed for the combined measure of home team advantage that was described in Definition 3. Numbers of teams for which the hypothesis about home team advantage was accepted are presented in Table 1. The highest number was recorded in the 2009/2010 season (17 teams out of 20), and the lowest number was recorded in the 2015/2016 season (2 teams out of 20).

Season	Teams	Season	Teams	Season	Teams
1992/93	11	2000/01	9	2008/09	5
1993/94	5	2001/02	8	2009/10	17
1994/95	12	2002/03	8	2010/11	10
1995/96	8	2003/04	7	2011/12	9
1996/97	4	2004/05	10	2012/13	4
1997/98	9	2005/06	10	2013/14	5
1998/99	6	2006/07	8	2014/15	5
1999/00	13	2007/08	12	2015/16	2

Table 1: Numbers of teams for which the hypothesis about home team advantage was accepted.

Table 2 contains numbers of cases where combined measure of home team advantage ( $C_i$ ) took value of  $-1, 0$ , or  $1$  in the 2015/2016 season. Each team played with 19 opponents, and therefore 19 observations (samples) are obtained for each team. This table also contains  $P(p_1 > p_{-1})$  (based on  $C_i$ 's), and two teams – Newcastle and Swansea – where it is possible to accept the hypothesis that home team advantage exists are marked with an asterisk.

Now, we will present evolution of  $P(p_1 > p_{-1})$ , estimate  $\hat{p}_1$ , and 95% confidence interval  $(\hat{p}_{1,l}, \hat{p}_{1,u})$  during the time. These results are presented for two selected teams (we choose among the previously mentioned seven teams that played in each season of English Premier League). The first presented team – Liverpool – is the team with the highest home team advantage (measured simply as an average of obtained probabilities  $P(p_1 > p_{-1})$  in all seasons). Liverpool is also the team with the lowest changes in  $P(p_1 > p_{-1})$ . These changes were measured using two criteria; the first was sample standard deviation of  $P(p_1 > p_{-1})$ , and the second was sum of absolute differences in  $P(p_1 > p_{-1})$  between two consecutive seasons. In both criteria, Liverpool recorded the lowest value out of the seven mentioned teams. Results of Liverpool are in Figure 1 and Figure 2; the first figure contains evolution of  $P(p_1 > p_{-1})$  and the second figure contains evolution of  $\hat{p}_1$ ,  $\hat{p}_{1,l}$ , and  $\hat{p}_{1,u}$ . Seasons where it is possible to accept hypothesis that home team advantage exists, i.e. where  $P(p_1 > p_{-1}) \geq 0.95$ , are denoted by full bullets (●) in Figure 1.

The team with highest changes in  $P(p_1 > p_{-1})$  was Arsenal (this holds for both used criteria). Arsenal also had the second lowest home team advantage (i.e. average value of  $P(p_1 > p_{-1})$ ). The lowest home team advantage among the seven mentioned teams was recorded by Chelsea with average value of  $P(p_1 > p_{-1})$  equalling to 0.818. For comparison, the average value of this probability for Arsenal was 0.833 and for Liverpool 0.892. Evolution of parameters for Arsenal are presented in Figure 3 and Figure 4.

Team	$C_i = -1$	$C_i = 0$	$C_i = 1$	Sum	$P(p_1 > p_{-1})$
Arsenal	5	4	10	19	0.895
Aston Villa	5	6	8	19	0.788
Bournemouth	8	5	6	19	0.304
Crystal Palace	9	4	6	19	0.227
Everton	8	3	8	19	0.500
Chelsea	6	5	8	19	0.696
Leicester	6	6	7	19	0.605
Liverpool	6	5	8	19	0.696
Man City	5	3	11	19	0.928
Man United	4	5	10	19	0.941
Newcastle*	2	4	13	19	0.998
Norwich	5	4	10	19	0.895
Southampton	7	0	12	19	0.868
Stoke	6	3	10	19	0.834
Sunderland	6	1	12	19	0.916
Swansea*	5	2	12	19	0.952
Tottenham	5	8	6	19	0.613
Watford	4	8	7	19	0.806
West Brom	9	3	7	19	0.315
West Ham	9	1	9	19	0.500

Table 2: Results for the 2015/2016 season

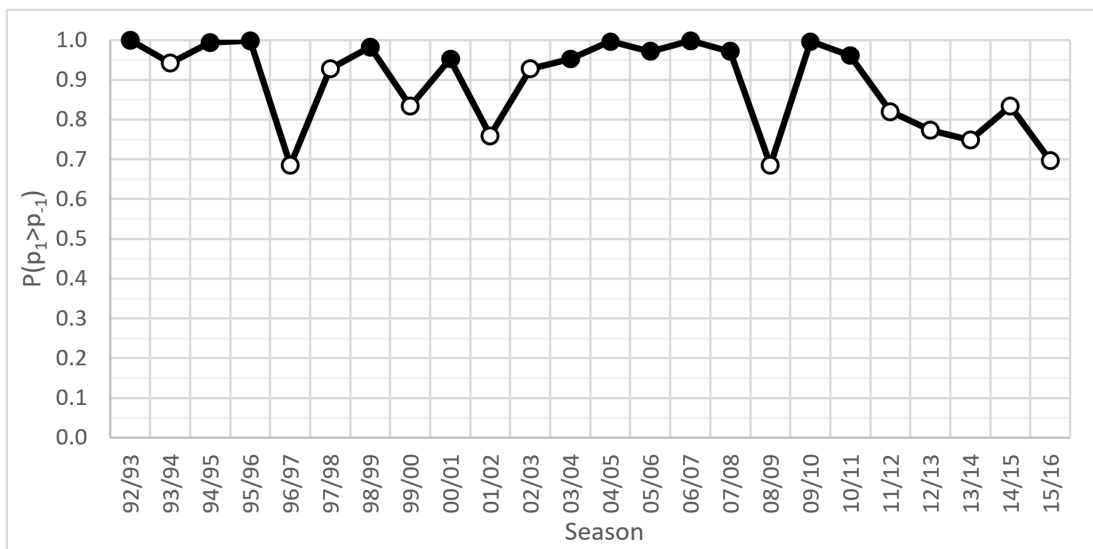


Figure 1: Evolution of  $P(p_1 > p_{-1})$  for Liverpool.

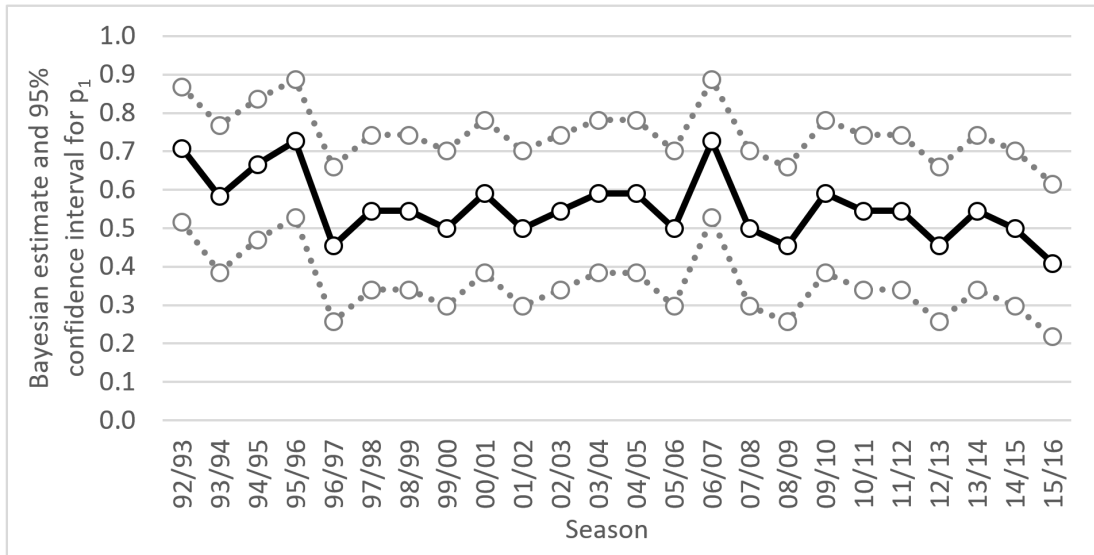


Figure 2: Evolution of Bayesian estimate and symmetric 95% confidence interval for  $p_1$  for Liverpool.

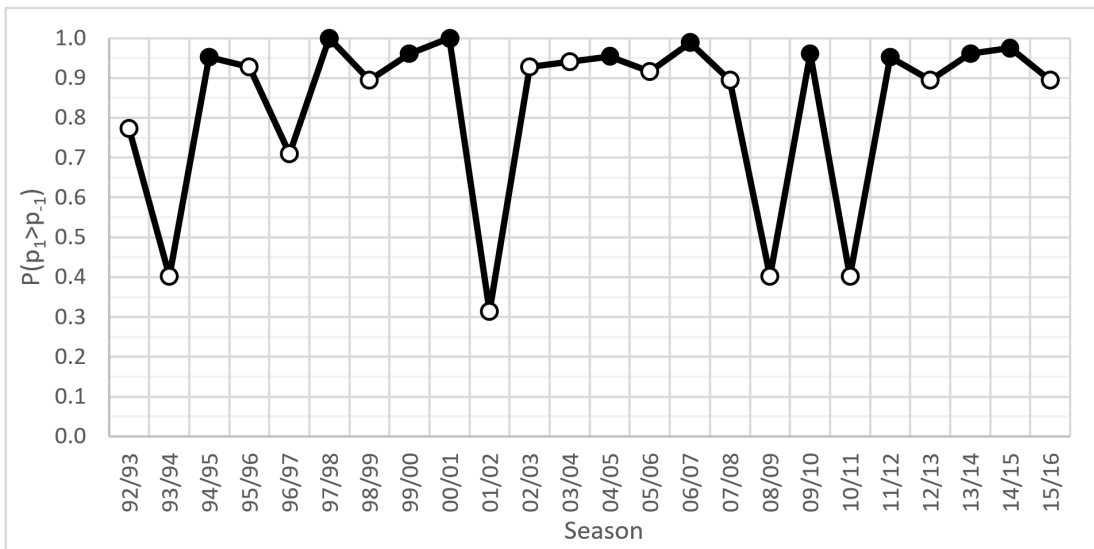


Figure 3: Evolution of  $P(p_1 > p_{-1})$  for Arsenal.

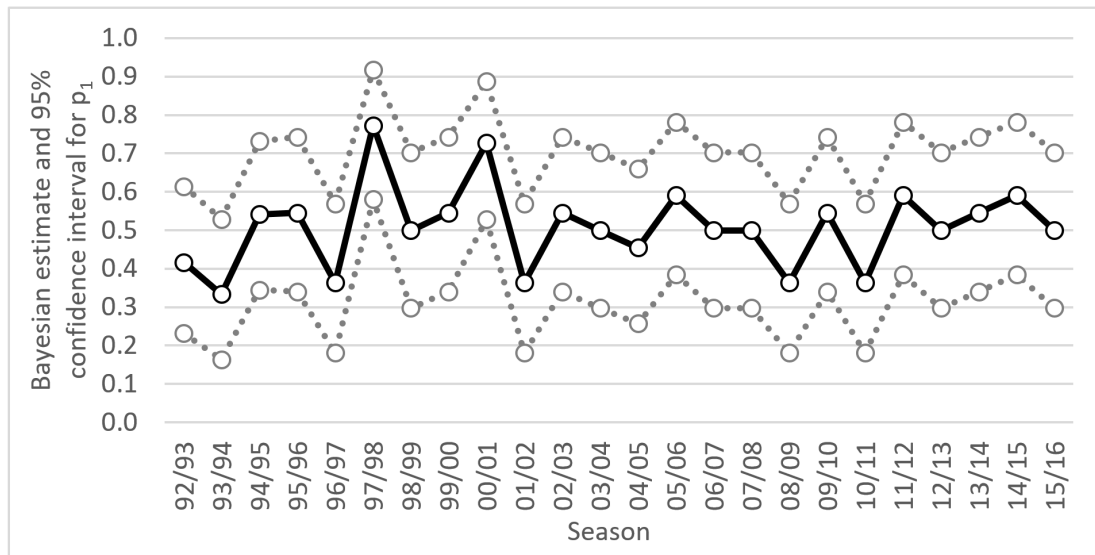


Figure 4: Evolution of Bayesian estimate and symmetric 95% confidence interval for  $p_1$  for Arsenal.

Evolution of  $P(p_1 > p_{-1})$  for all teams that played at least once between the 2012/13 season and the 2015/2016 season is presented in Table 3. Bold font is used for those results where it is possible to accept hypothesis that home team advantage exists. Norwich in the 2013/2014 season is nice example that the home team advantage does not ensure good results. It only ensures that results on a home field are better than on an away field but both can mean loss. Norwich in the 2013/2014 season recorded three times  $C_i = -1$ , once  $C_i = 0$ , and 15 times  $C_i = 1$ . For example, Norwich lost 0–1 to Manchester United at home field and 0–4 in Manchester. Obviously, 0–1 is better results than 0–4, and therefore  $C_i = 1$  in this case, as described in Definition 3. In fact, home team advantage can be, in this sense, called away field disadvantage.

The last presented results are extreme values obtained in all seasons. Five lowest values of  $P(p_1 > p_{-1})$  are presented in Table 4 and five highest values in Table 5. These tables also contain numbers of cases where combined measure of home team advantage ( $C_i$ ) took value of  $-1, 0$ , or  $1$  in the referred season. It can be seen that  $P(p_1 > p_{-1})$  is in many cases close to 1 but it is usually far from 0.

## 4 Discussion

Methods were presented on English Premier League data between 1992/1993 season and 2015/2016 season. Each team was tested in each season to identify whether it is possible to accept hypothesis about the home team advantage. Results are diverse – from two teams with the home team advantage in the 2015/16 season to 17 teams in the 2009/2010 season – and with no clear trend. Full results for the 2015/2016 season were presented along with  $P(p_1 > p_{-1})$  (i.e. probability that probability of home team advantage is higher than probability of away team advantage) that can be used as a measure of the home team advantage; the higher value, the higher home team advantage. In the 2015/2016 season only Swansea and Newcastle had this probability over 0.95, and hypothesis about existing home team advantage can be accepted for them.



Team	Season			
	12/13	13/14	14/15	15/16
Arsenal	0.895	<b>0.962</b>	<b>0.975</b>	0.895
Aston Villa	0.760	0.685	0.820	0.788
Bournemouth	—	—	—	0.304
Burnley	—	—	0.685	—
Cardiff	—	0.928	—	—
Chelsea	0.834	0.849	0.849	0.696
Crystal Palace	—	0.696	0.212	0.227
Everton	<b>0.994</b>	0.928	0.849	0.500
Fulham	0.760	<b>0.962</b>	—	—
Hull	—	0.928	0.928	—
Leicester	—	—	0.895	0.605
Liverpool	0.773	0.748	0.834	0.696
Man City	<b>0.952</b>	<b>1.000</b>	0.820	0.928
Man United	0.867	0.500	<b>0.996</b>	0.941
Newcastle	0.760	0.820	<b>0.975</b>	<b>0.998</b>
Norwich	<b>0.994</b>	<b>0.998</b>	—	0.895
QPR	0.500	—	<b>0.996</b>	—
Reading	0.788	—	—	—
Southampton	0.788	0.849	0.881	0.868
Stoke	0.773	<b>0.975</b>	0.916	0.834
Sunderland	0.773	0.500	0.402	0.916
Swansea	0.788	0.928	0.867	<b>0.952</b>
Tottenham	0.500	0.895	0.928	0.613
Watford	—	—	—	0.806
West Brom	0.928	0.941	0.676	0.315
West Ham	<b>0.994</b>	0.788	<b>0.952</b>	0.500
Wigan	0.304	—	—	—

Table 3: Evolution of  $P(p_1 > p_{-1})$  for all teams in the seasons 2012/13–2015/16.

Team	Season	$P(p_1 > p_{-1})$	$C_i = -1$	$C_i = 0$	$C_i = 1$
Hull	2008/09	0.038	11	4	4
Norwich	1993/94	0.072	11	5	5
Blackburn	2003/04	0.166	10	3	6
Wolves	2011/12	0.166	10	3	6
Crystal Palace	1997/98	0.180	11	1	7

Table 4: Five lowest obtained values of  $P(p_1 > p_{-1})$ .

Team	Season	$P(p_1 > p_{-1})$	$C_i = -1$	$C_i = 0$	$C_i = 1$
Blackburn	2009/10	0.99999	0	4	15
Leeds	1992/93	0.99998	1	2	18
West Ham	1997/98	0.99998	1	0	18
Arsenal	1997/98	0.99993	1	2	16
Bolton	2005/06	0.99993	1	2	16

Table 5: Five highest obtained values of  $P(p_1 > p_{-1})$  (more decimal places of estimates are shown only for illustration, all results can be considered as equivalent).

Since the 1992/1993 season, only seven teams played all seasons of English Premier League. Among these teams, Liverpool had the highest home team advantage and Chelsea had the lowest. It is necessary to remind that the home team advantage means that a result at a home field is better than on an away field, and both results can be loss. Therefore, the home team advantage does not imply good results. In fact, home team advantage can be also named away field disadvantage.

In results for all teams and all seasons, the lowest value of  $P(p_1 > p_{-1})$  was obtained for Hull in the 2008/2009 season. This probability was 0.038, and it is based on observation that out of 19 opponents Hull recorded better result on away field for 11 of them. On the other side is Blackburn in the 2009/2010 season with the highest recorded value of  $P(p_1 > p_{-1})$ . Out of 19 opponents, Blackburn played better on a home field in 15 cases, and in 4 cases there was no advantage on either side.

## 5 Conclusion

This paper offers alternative approach for identification of home team advantage in results. The new method is based on goals scored rather than on points awarded. This allows to distinguish matches that looks identical when points are used; for example, a 0–2 loss is not as bad as a 1–5 loss. Three measures of home team advantage were defined: active, passive, and their combination. Later, the Bayesian estimator and confidence intervals for probabilities of appropriate states – home team advantage, no advantage, and away team advantage – were found. The last theoretical part contains test of the home team advantage. The new method was presented on English Premier League, and results suggest that home team advantage is real; however, it cannot be taken for granted.

## Acknowledgement

This publication was supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports.

## References

- Allen, M. S. & Jones, M. V. (2014), ‘The home advantage over the first 20 seasons of the English Premier League: Effects of shirt colour, team ability and time trends.’, *International Journal Of Sport And Exercise Psychology* **12**(1), 10–18.

- Dixon, M. J. & Coles, S. G. (1997), 'Modelling Association Football Scores and Inefficiencies in the Football Betting Market', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **46**(2), 265–280.
- England Football Results and Betting Odds (2017), 'Premiership Results & Betting Odds'. <http://www.football-data.co.uk/englandm.php>.
- Karlis, D. & Ntzoufras, I. (2003), 'Analysis of sports data by using bivariate Poisson models', *The Statistician* **52**(3), 381–393.
- Maher, M. J. (1982), 'Modelling association football scores', *Statistica Neerlandica* **36**(3), 109–118.
- Marek, P., Šedivá, B. & ěoupal, T. (2014), 'Modeling and prediction of ice hockey match results', *Journal of Quantitative Analysis in Sports* **10**(3), 357–365.
- Omar, M. H. & Joarder, A. H. (2012), 'Some Mathematical Characteristics of the Beta Density Function of Two Variables', *Bulletin of the Malaysian Mathematical Sciences Society* **35**(4), 923–933.
- Pitman, J. (1993), *Probability*, 1 edn, Springer.
- Pollard, R. & Pollard, G. (2005), 'Long-term trends in home advantage in professional team sports in North America and England (1876–2003)', *Journal of Sports Sciences* **23**(4), 337–350.
- Premier League Football News, Fixtures, Scores & Results (2017), 'Premier League Football Scores, Results & Season Archives'. <https://www.premierleague.com/results?co=1&se=1&cl=-1>.

# The effect of leadership on AFL team performance

K. Marshall\*

\* Department of Statistics, Data Science & Epidemiology, Swinburne University of Technology, email address: [kmmarshall@swin.edu.au](mailto:kmmarshall@swin.edu.au)

## Abstract

There has been a lot of evidence for the importance of leaders to a team's functioning, across a wide range of domains including sport. In Australian rules football, 2016 saw a number of high profile team leaders missing a substantial proportion of the season due to injury. For example Carlton's captain Marc Murphy, Gold Coast's captain Gary Ablett Jr, and Fremantle's leadership group member Nat Fyfe all missed a large number of matches. These three teams went on to have a poor season performance. In this paper I will investigate the impact the absence of these leaders had on their team's team poor performance, by using a variety of techniques to examine the effect on other team members.

## 1 Introduction

There has been a great deal of research and discussion about the concept of leadership, especially as it relates to team performance. Studies have confirmed the link between leadership and team outcomes, particularly sporting teams (Price & Weiss, 2011).

Historically consideration of sporting leadership emphasised the role of the coach, however recently focus has turned to leadership from the athletes themselves (Fransen et al., 2015), including the captain and vice-captain. Such leaders are typically highly skilful (Loughead, Hardy, & Eys, 2006; Moran & Weiss, 2006; Price & Weiss, 2011), more senior (Rees & Segal, 1984; Tropp & Landers, 1979) and play in more central positions (Klonsky, 1991). These leaders ideally fulfil a number of roles as described by Loughead et al. (2006) such as task, social and external leadership. More recently a fourth role, that of the motivational leader, has been proposed (Fransen, Vanbeselaere, De Cuyper, Vande Broek, & Boen, 2014).

Motivational leaders inspire and encourage higher performance on-field. This is closely related to the concept of transformational leadership (Bass & Riggio, 2006), whereby teammates are motivated to perform at higher levels than they would have otherwise achieved. Studies have confirmed that transformational leadership is strongly linked with increased sporting team performance (Wang, Waldman, & Zhang, 2014).

Evidence from the literature also suggests that any detrimental change to the leadership structure will negatively affect the team's performance. Flores and colleagues (2012) found a decrease in performance following the dismissal of a team's coach (although other studies have reported no effect, for example De Paola and Scoppa (2012), Ter Weel (2011)). Buccioli et al (Buccioli, Foss, & Piovesan, 2014) also found the duration of a coach's tenure to be a predictor of team performance, with consistency in coaching structure associated with higher levels of performance. One would expect that, given the importance of players' leadership, a similar (although potentially weaker) effect would be present for

captains and other team leaders. Additionally, Bass and Riggio (2006) suggest that the effect of leadership is weakened by physical distance, which indicates that if the leader is absent from the team, their ability to motivate and encourage higher performance will be impaired, and therefore a poorer performance will result.

There may be an alternative explanation for any observed deterioration in team performance when a leader is absent. As mentioned before, leaders are often amongst the most highly-skilled players in the team (Loughead et al., 2006; Moran & Weiss, 2006; Price & Weiss, 2011). Poorer performance may simply be a result of the team missing their captain's elite skills on the field. However there is some suggestion that transformational leadership acts above and beyond the skill level of the leader. Zacharatos and colleagues (2000) studied 112 members of adolescent sporting teams, assessing their own perceptions of transformational leadership and team outcomes, as well as obtaining ratings of the players and team by their coach and peers. Results showed the players' level of transformational leadership was related to team performance outcomes even after controlling for the athlete's skill level. This suggests that the leader contributes something extra to the team's performance above and beyond that contributed by their skill. Searching for this effect using objective performance data is the focus of this paper. Specifically, this paper will use a case study design to compare performance when the leader is in the team to when they are absent, controlling for the effects of their skill level. If there is a leadership effect, one would expect a deterioration in team performance when the captain is missing, above and beyond the lower performance expected when the team is missing a highly-skilled player.

This paper will explore this issue in Australian football, a fast-paced invasion team sport. Each team consists of 18 on-field and four additional players who may be substituted on and off throughout the match. The object is to kick, handpass and run (with periodic bounces) an oval-shaped ball down a playing surface (of cricket field dimensions), to kick the ball through the two centre posts for six points or between the outer posts for one point. There are currently 18 teams in the national competition (Australian Rules Football; AFL). Teams play 22 matches per home-and-away season, with the highest-ranked eight teams at the end of the season competing in the finals series (playoffs).

Australian football gives us scope to investigate the impact of absent leadership, as 2016 saw an unusually high number of AFL captains and leadership group members sidelined due to long-term injury, including the captains for Carlton, the Western Bulldogs and the Gold Coast Suns.

This paper will examine Carlton in detail. Expectations for Carlton were low in 2016, with some experts predicting they would finish towards the bottom of the league ("Crystal ball: AFL.com.au's predictions for 2016," 2016). Despite this, they won five of their first ten matches and were on track to finish in the middle of the ladder, before their captain Marc Murphy suffered a serious ankle injury (Balales, 2016). Without Marc Murphy, they won only two of their remaining twelve matches and finished 14th.

It is possible that Carlton was simply missing the skills and experience brought by Murphy, and their performance suffered as a result. However, as suggested by Zacharatos et al. (2000), there may be evidence for the loss of a captain having an effect on team performance above and beyond their skill level. The first objective of this paper is to explore whether a team's performing more poorly without their captain is due to the absence of their skills, or whether the team is missing the leadership normally provided by the captain. This will be investigated using Carlton as a case study. Given the injury occurred mid-season, it is possible to compare their performance before and after the injury, controlling for season effects.

Secondly, this paper will compare the differing fortunes of Carlton and the Western Bulldogs in 2016. Both captains suffered long-term injuries but these teams went on to obtain vastly different final

results. In the third match of the season, Western Bulldogs captain Robert Murphy injured his knee, ruling him out for the remainder of the year. The Bulldogs won 15 from 22 for the season, finishing seventh. However, despite the absence of one of their key players, they went on to win all their finals matches and ultimately claim the Premiership. Because the injury occurred early in the season, we are unable to reliably compare the team's performance before and after. However, we can predict how they would have performed in the home and away season if their captain had been replaced by a team member of equivalent skill; it may be that they would have earned an easier pathway to the Grand Final. More interestingly, this paper will discuss the characteristics of the two teams and their respective captains that may be responsible for the different levels of performance.

This paper will present introductory analyses. It is expected that this will form the basis for further exploration, ideas for which will be presented at the end of the paper.

## 2 Method

This paper modifies the approach taken by Sargent and Bedford (2013). They used network analysis to identify the most highly-connected players in a team (quantified by a centrality measure of the interactions between teammates), then simulated the effect of replacing the highly-connected player with a less-connected one. Their simulation showed that replacing Jimmy Bartel with the less-connected player Shannon Byrnes in Geelong's 2011 Grand Final team would have resulted in a reduction in winning margin of 15 points.

This paper will use a similar methodology, however using Supercoach scores as a measure of player impact rather than network centrality. Supercoach score is a weighted aggregate of player actions within the match, taking into account the efficiency and usefulness of the actions. I will model replacing the injured captain with a player of similar skill and determine whether the team would have won additional matches. In the case where the captain was injured midway through the season (Carlton), the rate of wins will be compared for the two halves of the season. If the team would have been able to win at the same rate with the hypothetical highly-skilled player as with their captain, this would demonstrate the poorer performance was primarily due to the team missing a player with superior skills. However, if by accounting for the missing skills the team would have not been able to win at the same rate, it suggests that the team was missing something else contributed by the captain. Although not providing proof that their leadership was impaired, that this is one potential explanation.

### 2.1 Supercoach scores

Supercoach scores are a weighted aggregation of player actions within a match. Points are awarded or deducted for certain actions, for example a goal earns the player 8 points, an intercept possession is worth 4.5 and a free (penalty) against loses them 4 points ("SuperCoach 2016: Champion Data explains key stats in scoring formula", 2016). This method was developed to value player contributions in a match for a fantasy football league, but demonstrates high utility for modelling overall match performance. For all matches 2013 to 2016, the correlation between average team Supercoach score and final margin is  $r = 0.88$ ,  $df = 1580$ ,  $p < .001$ .

Supercoach scores for each player and each match were obtained from the website Footy Wire ([http://www.footywire.com/afl/footy/supercoach\\_round](http://www.footywire.com/afl/footy/supercoach_round)).

## 2.2 Modelling match outcome

Following the lead of the majority of other papers on AFL (e.g. Clarke, 1993; Ryall & Bedford, 2010; Stefani & Clarke, 1992), I will model each team's final margin for the match, rather than win/loss. The linear regression model uses a team's short- and long-term form (average margin over their past five and 20 matches respectively), and the team's average Supercoach score for that match. The model is given by:

$$M = -405 + 5.40SC + 0.20F_{20} + 0.09F_5 \quad (1)$$

where  $M$  is the final margin,  $SC$  the average Supercoach scores for the team,  $F_{20}$  and  $F_5$  the average margin over their past 20 and 5 matches respectively.

For all home-and-away matches for all teams from 2013 to 2016, the model had an adjusted  $R^2$  of 0.79. 88% of team results were correctly classified as a win or loss using this model. As such, the model was deemed acceptable for this purpose.

## 2.1 Generating a replacement player with equivalent skill to an injured captain

When a player is injured or otherwise excluded from team selection, there is often no easy way to identify the specific player who replaced them. As such, a few assumptions need to be made in the model. One would assume that the captain is amongst the top 22 players in the squad, and that they would be replaced by a player with lesser skill. In order to capture this, a player from the bottom half of the squad (in terms of Supercoach scores over their past 20 matches) is randomly selected. This player is removed from the selected team and replaced with a hypothetical player of the same skill level as the injured captain.

This hypothetical player's contribution to the match is modelled by randomly generating a Supercoach score based on an empirical cumulative distribution of the captain's past 20 matches' Supercoach scores.

The Supercoach scores for the entire team are then inserted into the equation at (1) to model the team's expected performance for that match.

## 3 Case Study 1: Carlton

With Marc Murphy in the team, Carlton won 27 from 43 matches (39%) during the 2013 to 2016 seasons (his tenure as captain to date). However with him missing, they won only two from 18 (11%),  $\chi^2 = 3.7229$ ,  $df = 1$ ,  $p = 0.05$

There was no difference in quality of opposition when Murphy was playing compared to when he was injured, for either long-term form (average margin across the past 20 matches),  $t = 0.03$ ,  $df = 21.97$ ,  $p = 0.98$ , or short-term form (past 5 matches),  $t = 0.13$ ,  $df = 25.89$ ,  $p = 0.90$ . Therefore any changes can reasonably be attributed to Carlton's performance, rather than the opposition.

Results were simulated for each match in which Murphy was absent throughout the 2016 season (rounds 11 through 23; 12 matches excluding one bye). The process described in section 2 was repeated 1000 times for each match. The pertinent results are in Figure 1, which displays a distribution of predicted match results with the captain replaced by a player of equivalent skill. For comparison, each graph also shows the actual match result and the predicted match result given the actual skill level of the selected side (i.e. without the captain or equivalent highly-skilled player).

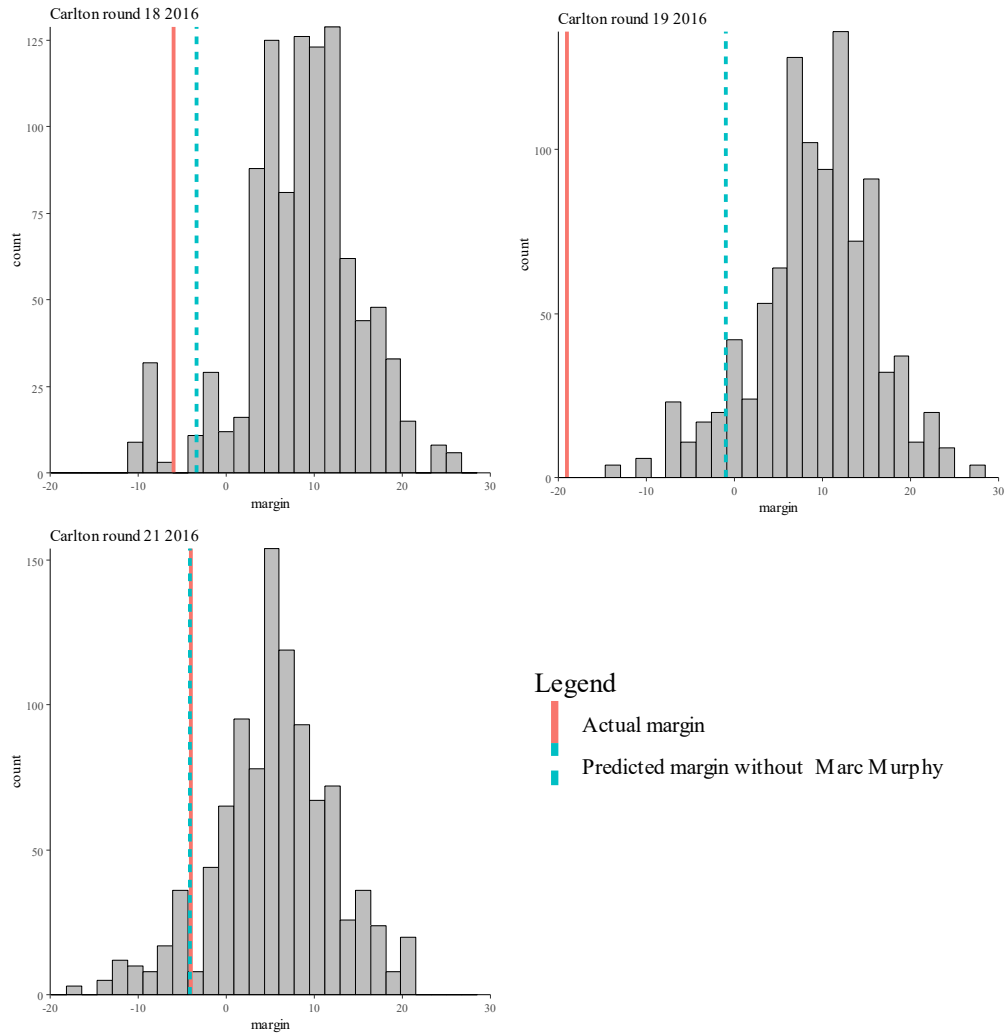


Figure 1. Simulated match results for Carlton in rounds 18, 19 and 21 in season 2016 with the captain replaced by a player with equivalent skill level.

Simulations indicate that had Carlton been able to field a player of equivalent skill to Marc Murphy, they were likely to win their matches in rounds 18 and 21. The model suggests they would also have won their match in round 19, however in reality they performed poorer than predicted by the model. A further two matches would have come within a single goal. This appeared to be due to Carlton’s performance during close matches. With their captain in the team, they won one and lost one close match (final margin under two goals/12 points). Without their captain, they lost all three close matches. As such, the increased overall skill level generated by the simulation was enough to get them over the line in those three close matches.



## 4 Case Study 2: Western Bulldogs

With Robert Murphy in the team, the Western Bulldogs won 65% of matches across the 2015 and 2016 seasons (while he was captain), compared to 66% when he was absent, a non-significant difference ( $p = 1$ ).

Robert Murphy only played two complete matches in 2016 however, the remaining match results were simulated in order to determine how they might have gone had the team been able to replace Murphy with a player of equivalent skill. Selected results are shown in Figure 2.

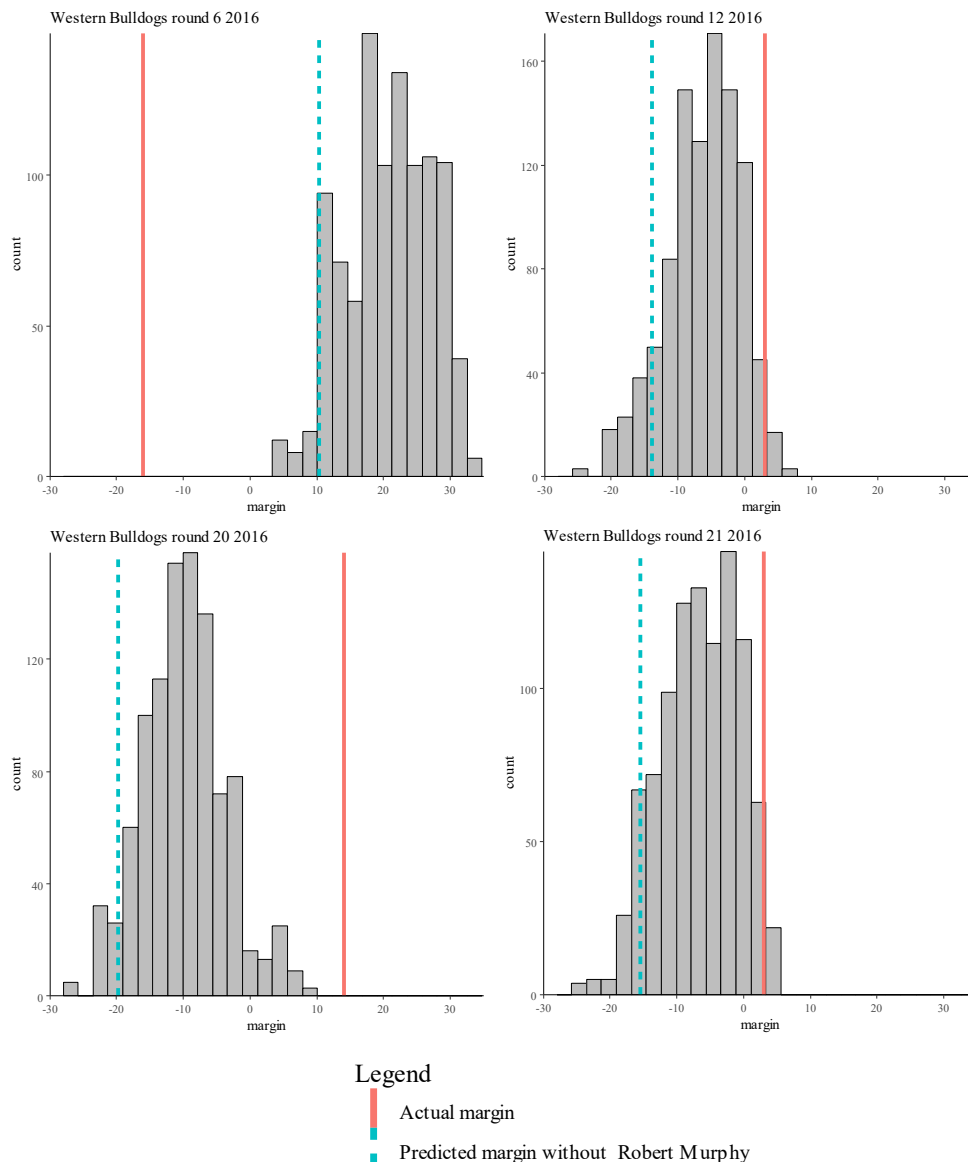


Figure 2. Simulated match results for Western Bulldogs in rounds 6, 12, 20 and 21 in season 2016 with the captain replaced by a player with equivalent skill level.

Simulation results indicate that had the Western Bulldogs been able to field a player of Robert Murphy's skill, they were predicted to win their round 6 match, which they in reality lost. However they were also predicted to win this match given their selected team line-up. The predictive model seemed a poorer fit for a number of Western Bulldogs' matches; for three matches (rounds 12, 20 and 21), they were predicted to lose (with both actual and simulated team line-ups), but managed to win.

Unlike Carlton, the Western Bulldogs were able to win close matches without their captain. With Murphy in the team, they lost the single close match (the match he was injured). Without, they won all five close matches.

## 5 Comparison of the two teams

The average Supercoach scores for each player in 2015 was graphed, see Figure 3. 2015 was selected to provide a more complete picture of the captains' performances. For each team, the captain's average rating is shown as a darker colour.

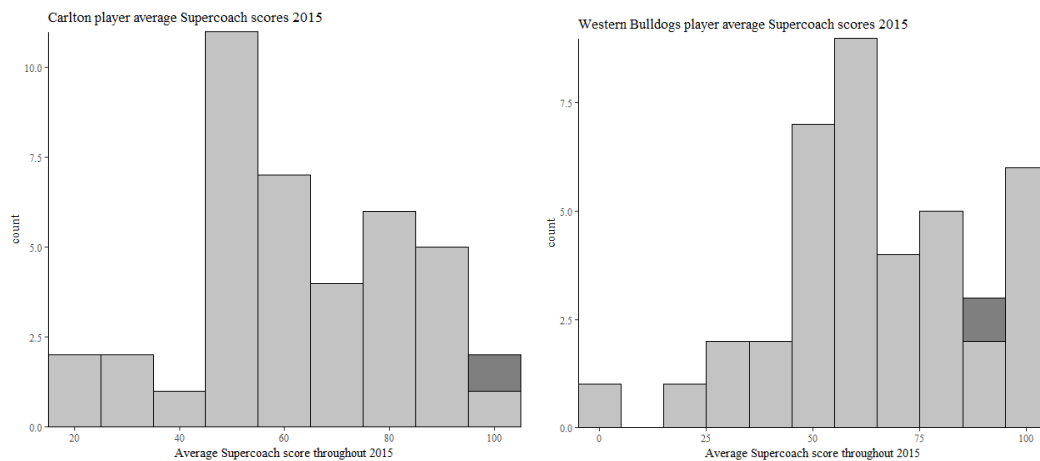


Figure 3. Average Supercoach scores for Carlton and Western Bulldogs players in 2015, with the respective captains highlighted.

Marc Murphy's average Supercoach score for 2015 was 103.16, the highest for Carlton. By comparison, Robert Murphy's average was 91.85, ranking him seventh for the Bulldogs. The average Supercoach scores for Carlton and the Bulldogs in 2015 were 63.07 (SD = 20.74) and 65.96 (SD = 23.36) respectively,  $t = -0.6$ ,  $df = 76.92$ ,  $p = 0.561$ .

## 6 Discussion

This study represents an early attempt to quantify the effect of leadership based on objective performance data. The primary aim was to investigate the impact of a captain's absence on their team's performance. For one Australian football team whose captain was injured mid-season, I simulated results had they been able to replace the captain with a player of similar skill, to see if their seemingly worse performance in the latter half of the season was due to their lacking a player with elite skills, or due to something beyond that, a potential leadership effect.

Results suggest that much of Marc Murphy's influence on Carlton's performance was due to his skill. Had he been replaced with a player of equivalent skill level, I estimate Carlton would have won an additional three matches in the latter half of the season. This would have brought them up to a winning percentage of 42% for the second half of the season, a vast improvement over their actual rate of 17% but slightly less than their winning record while Marc Murphy was in the team in 2016 (50%). Overall, the results primarily indicate that Carlton was relatively resilient in the face of on-field leadership changes in 2016, with the bulk of the effect coming from the absence of their most highly-skilled player. There may have been a slight deterioration in team performance above and beyond skill level. One potential explanation for this may be due to lower leadership on-field, however this cannot be proven by this study.

On the other hand, the results showed the Western Bulldogs may have been able to win only one additional match. This was because the Bulldogs were able to win close matches without their captain, making the skill differential with an injured (highly-skilled) captain less significant. The reasons for the Bulldogs being able to cope better with the pressure of close matches are unclear and require further investigation.

Comparison of the previous year's data revealed Marc Murphy to be Carlton's best player. On the other hand Robert Murphy, while still highly-skilled, ranked seventh for the Bulldogs. (Note this may be due to their different on-field roles being tending to accumulate Supercoach rating points at different rates.) It may be the case that the Bulldogs were better able to replace the skills lost due to their captain's injury – possibly due to restructuring the team and adjusting strategy. Robert Murphy also plays in a less central role compared to Marc Murphy, so has fewer interactions with other players and therefore may require less adjustment to team strategy when absent.

While the literature suggests disruption to a team's leadership is likely to lead to lower performance (e.g. Bass & Riggio, 2006), AFL teams' leadership structures appear relatively resilient. There has been a trend in recent years for teams to have leadership groups, consisting of an additional two to six players beyond the captain and vice-captain, who provide leadership on and off the field. As such, the distributed leadership responsibilities across team members makes teams better positioned to cope with the loss of captains and vice-captains. A recent study found leadership qualities to be distributed across several players in sporting teams, and also found that players other than the captain are often rated by teammates as the strongest leader across a range of roles (Fransen et al., 2014), offering support for the shared leadership model.

One suggestion for future research is to compare the impact of an injured captain on the performance of teams with shared versus traditional leadership structures. If the captain is the sole on-field leadership present, their absence is more likely to be felt by their teammates.

One limitation of this research is only using Supercoach scores as a general measure of player skill. This fails account for strategic changes that would have had to happen to cover for the player absence. Future research may consider within-match plays to get a better assessment of individual player impact.

Overall, it appears that AFL teams are primarily resilient to disruptions to their leadership structures, such as the long-term absence of their captain. The bulk of the performance deterioration seen is when the team is unable to cover for the loss of one of their most highly-skilled players. This finding provides support for AFL teams having larger leadership groups rather than just a captain and vice-captain, so they are better able to manage long-term absences of their key players.

## Acknowledgements

The author would like to thank Denny Meyer, Karl Jackson, Minh Huynh and Maria Gardiner for their comments on an earlier draft of this paper.

## References

- [1] Balales, D. (2016, 16th September 2016). Season review: Marc Murphy. Retrieved from <http://www.afl.com.au/news/2016-03-24/crystal-ball-predictions-for-2016>
- [2] Bass, B. M., & Riggio, R. E. (2006). *Transformational Leadership*. Taylor and Francis: Hoboken NJ.
- [3] Buccioli, A., Foss, N. J., & Piovesan, M. (2014). Pay dispersion and performance in teams. *PLoS ONE*, 9(11).
- [4] Clarke, S. R. (1993). Computer forecasting of Australian Rules Football for a daily newspaper. *Journal of Operational Research Society*, 44, 753-759.
- [5] Crystal ball: AFL.com.au's predictions for 2016. (2016, 24th March 2016). Retrieved from <http://www.afl.com.au/news/2016-03-24/crystal-ball-predictions-for-2016>
- [6] De Paola, M., & Scoppa, V. (2012). The effects of managerial turnover: evidence from coach dismissals in Italian soccer teams. *Journal of Sports Economics*, 13(2), 152-168.
- [7] Flores, R., Forrest, D., & Tena, J. D. (2012). Decision taking under pressure: Evidence on football manager dismissals in Argentina and their consequences. *European Journal of Operational Research*, 222(3), 653-662.
- [8] Fransen, K., Van Puyenbroeck, S., Loughhead, T. M., De Cuyper, B., Vanbeselaere, N., Vande Broek, G., & Boen, F. (2015). The art of athlete leadership: Identifying high-quality athlete leadership at the individual and team level through Social Network Analysis. *Journal of Sport & Exercise Psychology*, 37(3), 274-290.
- [9] Fransen, K., Vanbeselaere, N., De Cuyper, B., Vande Broek, G., & Boen, F. (2014). The myth of the team captain as principal leader: extending the athlete leadership classification within sport teams. *Journal of Sports Sciences*, 32(14), 1389-1397.
- [10] Klonsky, B. G. (1991). Leaders' characteristics in same-sex sport groups: A study of interscholastic baseball and softball teams. *Perceptual and Motor Skills*, 72(3), 943-946.
- [11] Loughhead, T. M., Hardy, J., & Eys, M. A. (2006). The nature of athlete leadership. *Journal of Sport Behavior*, 29(2), 142-158.
- [12] Moran, M. M., & Weiss, M. R. (2006). Peer leadership in sport: Links with friendship, peer acceptance, psychological characteristics, and athletic ability. *Journal of Applied Sport Psychology*, 18(2), 97-113.
- [13] Price, M. S., & Weiss, M. R. (2011). Peer Leadership in Sport: Relationships among Personal Characteristics, Leader Behaviors, and Team Outcomes. *Journal of Applied Sport Psychology*, 23(1), 49-64.
- [14] Rees, R. C., & Segal, M. W. (1984). Role differentiation in groups: The relationship between instrumental and expressive leadership. *Small Group Behavior*, 15(1), 109-123.
- [15] Ryall, R., & Bedford, A. (2010). An optimized ratings-based model for forecasting Australian Rules football. *International Journal of Forecasting*, 26, 511-517.
- [16] Sargent, J., & Bedford, A. (2013). Evaluating Australian Football League player contributions using interactive network simulation. *Journal of Sports Science and Medicine*, 12, 116-121.

- [17] Stefani, R. T., & Clarke, S. R. (1992). Predictions and home advantage for Australian rules football. *Journal of Applied Statistics*, 19(2), 251-261.
- [18] Supercoach 2016: Champion Data explains key stats in scoring system. (2016, 22nd March). Herald Sun. Retrieved from <http://www.heraldsun.com.au/sport/afl/supercoach-news/supercoach-2016-champion-data-explains-key-stats-in-scoring-formula/news-story/7385b9a5cdd20375bf2ed606a02fe1d0>
- [19] Ter Weel, B. (2011). Does manager turnover improve firm performance? Evidence from Dutch Soccer 1986-2004. *De Economist*, 159(3), 279-303.
- [20] Tropp, K. J., & Landers, D. M. (1979). Team Interaction and the Emergence of Leadership and Interpersonal Attraction in Field Hockey. *Journal of Sport Psychology*, 1(3), 228-240.
- [21] Wang, D., Waldman, D. A., & Zhang, Z. (2014). A meta-analysis of shared leadership and team effectiveness. *Journal of Applied Psychology*, 99(2), 181-198.
- [22] Zacharatos, A., Barling, J., & Kelloway, E. K. (2000). Development and effects of transformational leadership in adolescents. *The Leadership Quarterly*, 11(2), 211-226.

# Sensor Analytics in Basketball

R. Metulini\* and M. Manisera\*\* and P. Zuccolotto\*\*\*

\* Department of Economics and Management - University of Brescia, email address: rodolfo.metulini@unibs.it

\*\* Department of Economics and Management - University of Brescia, email address: marica.manisera@unibs.it

\*\*\* Department of Economics and Management - University of Brescia, email address: paola.zuccolotto@unibs.it

## Abstract

A new approach in team sports analysis consists in studying positioning and movements of players during the game in relation to team performance. State of the art tracking systems produce spatio-temporal traces of players that have facilitated a variety of research aimed to extract insights from trajectories. Several methods borrowed from machine learning, network and complex systems, geographic information system, computer vision and statistics have been proposed. After having reviewed the state of the art in those niches of literature aiming to extract useful information to analysts and experts in terms of relation between players' trajectories and team performance, this paper presents preliminary results from analysing trajectories data and sheds light on potential future research in this field of study. In particular, using convex hulls, we find interesting regularities in players' movement patterns.

## 1 Introduction

Studying the interaction between players in the court, in relation to team performance, is one of the most important issues in sport science. Coaches and experts want to explain why, when and how specific movement is expressed because of tactical behaviour. Analysts want to study how cooperative movement patterns react to a variety of factors. To answer these questions, sports analysts borrowed methods from many disciplines, such as machine learning, network and complex systems, geographic information systems, computational geometry, computer vision and statistics. In recent years, the advent of information technology systems made it possible to collect a large amount of different types of big data, which are, basically, of two kinds. On the one hand, play-by-play data report a sequence of significant events, such as passes and shots, that occur during a match. A wide range of academic studies uses these data, with the aim of analysing team's performance: events collected in the play-by-play are used to identify the drivers that affect the probability to win a match, both from a data-mining perspective (Carpita et al. (2013, 2015)), and adopting social network analysis (Wasserman and Katherine (1994), Passos et al. (2011), Cintia et al. (2015)). On the other hand, object trajectories capture the movement of players using optical- or device-tracking and processing systems. These systems are based on Global Positioning Systems (GPS). The trajectory of a single player depends on the trajectories of all other players in the court, both team-mates and opponents, and on a large amount of external factors. A candidate method to approach with this complexity consists in segmenting a match into phases, to facilitate the retrieval of significant moments of the game (Perin et al. (2013)). Furthermore, a promising niche of literature called *ecological dynamics* expresses players in the court as agents who face with external factors (Travassos et al. (2013), Passos, Araujo and Volossovitch (2016)). Visualization tools are required in order to communicate the information extracted from trajectories. The growing interest in applying novel visual tools

to a range of sports contexts has been highlighted by Basole and Saupe (2016). Visualization tools are used on leading outlets such as the *New York Times* (Aisch and Quealy (2016), Goldsberry (2013)), and on academic journals (Perin, Vuillemot and Fekete (2013), Sacha et al. (2014), Polk et al. (2014)).

We start by presenting the state of the art in those niches of sport science literature where the use of play-by-play and trajectories data is helpful to the aim of extracting information for analysts and experts (Section 2). We position among these fields of literature with the final aim to study players' trajectories in basketball, in particular to visualize and characterize the movements of players around the court to find relevant types of movement patterns that could affect the team performance. To this aim, in Section 3 we first present data and methods, while in Section 4 we discuss in detail our research questions. We then present empirical analysis in Section 5. In Section 6 we conclude and discuss future research developments.

## 2 State of the art

### 2.1 Data-driven approach

Data-driven science, an interdisciplinary field about scientific methods, processes and systems, aims to extract insights from data in various forms. It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the sub-domains of machine learning, classification and data-mining. Data-driven approach benefits from the availability of the play-by-play and academic research is proliferating. Carpita et al. (2013, 2015) used cluster analysis and principal component analysis in order to identify the drivers that affect the probability to win a football match. Social network analysis has also been used to capture the interactions between players. Wasserman and Katherine (1994) mainly focus on passing networks and transition networks. A passing network is a graph where each player is modelled as a vertex and successful passes to the player represent links among vertices. Transition networks can be constructed directly from event logs, and correspond to a passing network where play-by-play data are attached. Passos et al. (2011) used centrality measures with the aim to identify central (or key) players, and to estimate the interaction and the cooperation between team members in water polo. In soccer, Cintia et al. (2015) observed players' behaviour on the pitch. They predict the outcome of a long-running tournament such as Italian major league using simple network measures. Moreover, Cintia, Rinzivillo and Pappalardo (2015) proposed and computed a pass-based performance indicator that strongly correlates with the success of the team.

### 2.2 Synchronized movements analysis

The trajectory of a single player depends on a large amount of factors and on the trajectories of all other players in the court, both team-mates and opponents. Because of these interdependencies, in every single moment, a player action causes a reaction. A promising niche of sport science literature, borrowing from the concept of physical psychology, Turvey and Shaw (1995) expresses players in the court as agents who face with external factors (Travassos et al. (2013), Passos, Araujo and Volossovitch (2016)). In addition, typically, players' movements are determined by their role in the game. Predefined plays

are used in many team sports to achieve some specific objective; moreover, team-mates who are familiar with each other's playing style may develop ad-hoc productive interactions that are used repeatedly. On the one hand, experts want to explain why, when and how specific movement behaviour is expressed because of tactical behaviour. Brillinger (2007) addressed the question of how to analytically describe the spatio-temporal movement of particular sequences of passes (i.e. the last 25 passes before a score). On the other hand, analysts want to explain and observe cooperative movement patterns in reaction to a variety of factors, such as coach advices and the corresponding team reactions, characteristics of the stadium (capacity, open/closed roof) and historical and current weather records (high/low temperatures, air humidity). A useful method to approach with complexity in team sport analysis consists in segmenting a match into phases, as it facilitates the retrieval of significant moments of the game. Perin, Vuillemot and Fekete (2013) visually segmented a football match into different phases while Metulini, Manisera and Zuccolotto (2017) segmented a basketball game into phases using a cluster analysis. A key factor in relation to teams' performance is how players control space. Many works are devoted to analyse how the space is occupied by players - when attacking and when defending - or in crucial moments of the match. Examples can be found in football (Couceiro et al. (2014), Moura et al. (2012)) or in futsal (Fonseca et al. (2012), Travassos et al. (2012)).

### 2.3 Visualization tools

In order to communicate the information extracted from the spatio-temporal data, visualization tools are required. Basole and Saupe (2016) highlight the growing interest in applying novel visual tools to a range of sports contexts. Analysts from leading outlets such as the *New York Times* use visualization to tell basketball and football stories (Aisch and Quealy (2016), Goldsberry (2013)). The increasing popularity of sports data visualization is also reflected in a greater academic interest. Perin, Vuillemot and Fekete (2013) developed a system for visual exploration of phases in football, Sacha et al. (2014) present a visual analysis system for interactive recognition of football patterns and situations. Notable works include data visualization in ice hockey (Pileggi et al. (2012)) and tennis (Polk et al. (2014)). In basketball, Losada, Theron and Benito (2016) developed 'BKViz', a visual analytics system to reveal how players perform together and as individuals. For visualizing aggregated information the most common approach is to use heat maps, simple and intuitive tools that can be used to visualize various types of data. Typical examples in the literature show the spread and range of a shooter (Goldsberry (2012)) or count how many times a player lies in specific court zones. More recently, dynamic approaches have been proposed to visualize aggregated information displaying the time dimension: Theron and Casares (2010) employed tools for the analysis of players' movements. Metulini (2017) proposed the use of motion charts for visualizing movements of basketball players' in the court. There are several softwares providing the possibility to reproduce motion charts, more or less intuitive, open source or requiring a license (*Gapminder world*, *Google docs gadget*, *Trend compass*, and *JMP* from SAS institute). In addition, motion charts can be created through web programming languages using *Google application programming interface*, *Google API*, *Flash* or *HTML5*. Applications of motion charts in other academic fields cover the aspects of students learning processes (Santos et al. (2012)) and linguistic changes (Hilpert (2011)), insurance (Heinz (2014)) and development economics (Saka and Jimichi (2015)), medicine (Santori (2014)) and hydrology (Bolt (2015)).



## 3 Data & Methods

### 3.1 Global Positioning Systems (GPS)

Object trajectories capture the movement of players (with or without data about the ball). Players' trajectories are retrieved using optical- or device-tracking and processing systems. Optical tracking systems use fixed cameras to collect the player movement, and the images are then processed to compute the trajectories (Bradley et al. (2007)). There are several commercial vendors who supply tracking services to professional sport teams and leagues (Tracab (2015), Impire (2015)). Tracking systems rely on devices that infer their location, and are attached to the players' clothing or embedded in the ball or puck. These systems are based on Global Positioning Systems (GPS) (Catapult (2015)). The resulting dataset is dense, because GPS collects data at very close instants. The adoption of this technology and the availability to researchers of the resulting data depends on various factors, particularly commercial and technical, such as, for example, the costs of installation and maintenance and the legislation adopted by the sport associations. This data acquisition may be partially restricted in some diffused team sports (as it was for example in soccer until 2015) while allowed for others.

### 3.2 Play-by-play

Play-by-play is a sequence of significant events that occur during a match. Events can be broadly categorised as player events such as passes and shots; and technical events, for example fouls, time-outs, and start/end of period. Event logs are qualitatively different from the player trajectories in that they are not dense, as samples are only captured when an event occurs; however they can be semantically richer as they include details like the type of event and the players involved. Typically, in basketball, the play-by-play consists of a collection of about five hundreds events per game. The collection includes events such as made shots, missed shots, rebounds, fouls, start/end of the period, etc.. .

## 4 Research questions

The overall objective of our research is to visualize and characterize the movement of basketball players around the court by finding relevant types of movement patterns that could affect the team performance.

Going into detail, the first specific objective is to find and demonstrate the usefulness of a visual tool approach in order to extract preliminary insights from trajectories. In this respect, we aim to visualize the synchronized movement of players and to characterize the spatial pattern of them around the court in order to supply experts and analysts with useful tool in addition to traditional statistics, and to corroborate the interpretation of evidence from other methods of analysis. Some preliminary results in this respect have been found in Metulini (2017), which suggested the use of motion charts to visualize the movements of players around the court and found, using sensor data from one game, differences in spacing structure among offensive and defensive plays. However, developments need to be carried on: analysing both team-mate and opponents trajectories together with the ball is fundamental in order to study how the movement of the players of one team reacts to the movement of the opposing team.

Another aspect of research lies on segmenting the match into phases. Specifically, our idea is to find, through a cluster analysis, a number of groups each identifying a specific spatial pattern, in order to: i) characterize the synchronized movement of players around the court, ii) find any regularities and synchronizations in players' trajectories, by decomposing the game into homogeneous phases in terms of spatial relations. Some related exploratory analysis using one game data have been carried on by Metulini, Manisera and Zuccolotto (2017). We plan to extend the analysis to multiple matches. Moreover, we aim to match play-by-play data with trajectories, to extract insights on the relations between particular spatial pattern and the team performance. The effect of the ball's position in determining players' movement also deserves to be studied.

## 5 Preliminary results

GPS trajectories data used in empirical analysis refers to a friendly match played on March 22th, 2016 by a team based in the city of Pavia (Italy). This team played the 2015-2016 season in the C-gold league, the fourth league in Italy. Totally, six players took part to the friendly match. All those players worn a microchip in their clothings. The microchip collected the position in both the  $x$ -axis and the  $y$ -axis. The positioning of the players has been detected at a resolution of milliseconds. Considering all the six players recovered, the system recorded a total of 133,662 space-time observations ordered in time. On average, the system collects positions about 37 times every second. Considering all the six players, the position of each single player is collected, on average, every 162 milliseconds. The dataset is structured such that *tagid* variable uniquely identifies the player, *timestamp\_ms* variable reports the exact millisecond in which the player position has been observed, *klm\_x* and *klm\_y* represent the  $x$ -axis (length), and the  $y$ -axis (width) coordinates, filtered with a Kalman approach. The Kalman filtering is an algorithm that uses a series of measurements observed over time, in order to produce more precise estimates than those based on a single measurement alone.

We first use motion charts to visualize the synchronized spatio-temporal movements of players around the court. In our application, *gvisMotionChart* in *R* is used (Gesmann and de Castillo (2013)) because it outperforms alternatives in terms of open source, friendliness, and because it allows to import data. A video tutorial showing players' trajectories using motion charts can be found at: <http://bodai.unibs.it/BDSports/Ricerca%20-%20DataInn.htm> To the best of our knowledge, motion charts have never been applied to basketball. Motion charts applied to our data show differences in the spacing structure of players among offensive and defensive plays. We defined whether each time instant corresponds to an offensive or a defensive play looking to the average coordinate of the five players in the court. The evidence is confirmed by the statistics describing the convex hull areas and the average distances, which are reported in Table 1. Results clearly highlight larger average distances among players in offensive plays, as well as larger convex hull areas.

To corroborate the previous evidence, Figures 1 and 2 report the convex hulls for selected snapshots from, respectively, the first offensive play and the first defensive play of the match. Once again, players are more spread around the court in offensive plays.

We then perform a cluster analysis, with the aim of characterizing the spatial pattern of the players in the court. We define different game phases, each considering moments being homogeneous in terms of spacings among players. We apply a  $k$ -means cluster analysis to group objects. Objects are the time instants and the similarity is expressed in terms of players' distance. We choose  $k = 8$  based on the value of the between deviance (BD) / total deviance (TD) ratio for different number of clusters. First, we characterize each cluster in terms of players' position in the court. For each cluster, a multidimensional scaling (MDS) is used to plot each player in a 2-dimensional space such that the between player average

Table 1: Average distances among players (in meters) and convex hulls areas (in squared meters) for the full match, for defensive and offensive plays.

	Average distances		Convex hull area	
	attack	defence	attack	defence
Min	2.296	0.400	1.000	1.000
1st Qu.	6.372	4.309	30.000	14.000
Median	7.235	5.086	41.000	20.500
Mean	7.250	5.680	42.590	28.550
3rd Qu.	8.132	6.523	53.000	33.500
Max	13.947	14.260	138.500	180.000

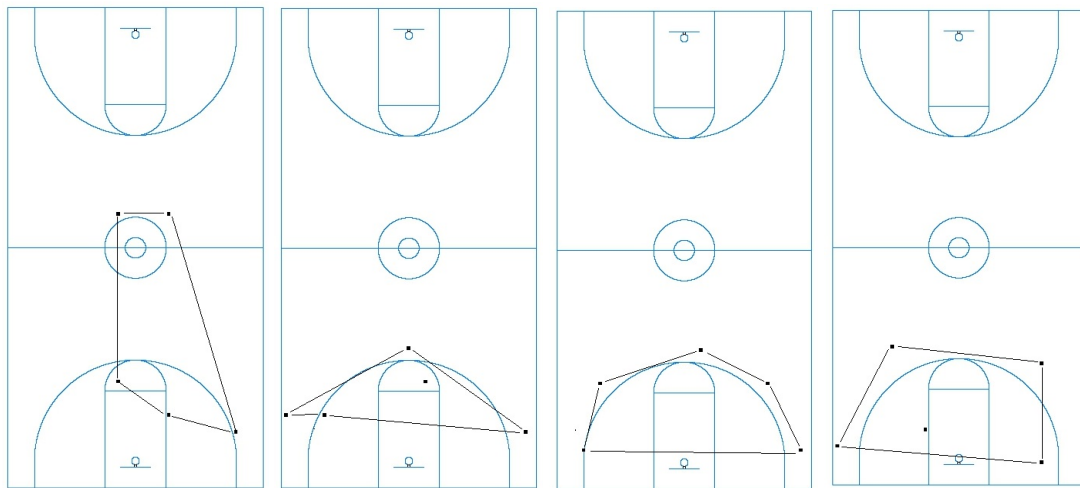


Figure 1: Convex hull for selected snapshots related to the first offensive play of the game

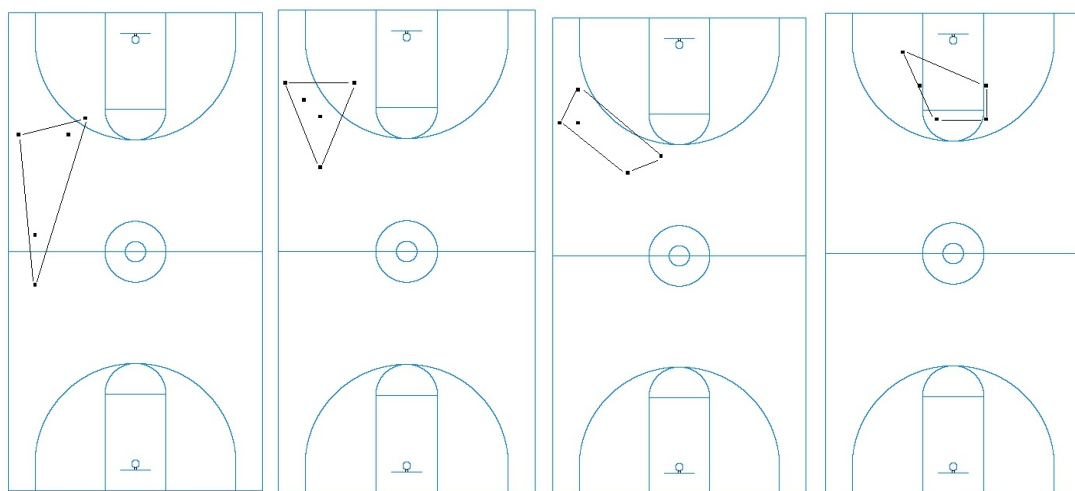


Figure 2: Convex hull for selected snapshots related to the first defensive play of the game

distances are preserved. We find substantial differences among different phases. Results of MDS are presented in Figure 3 and highlight remarkable differences in the positioning structure of players in the court.

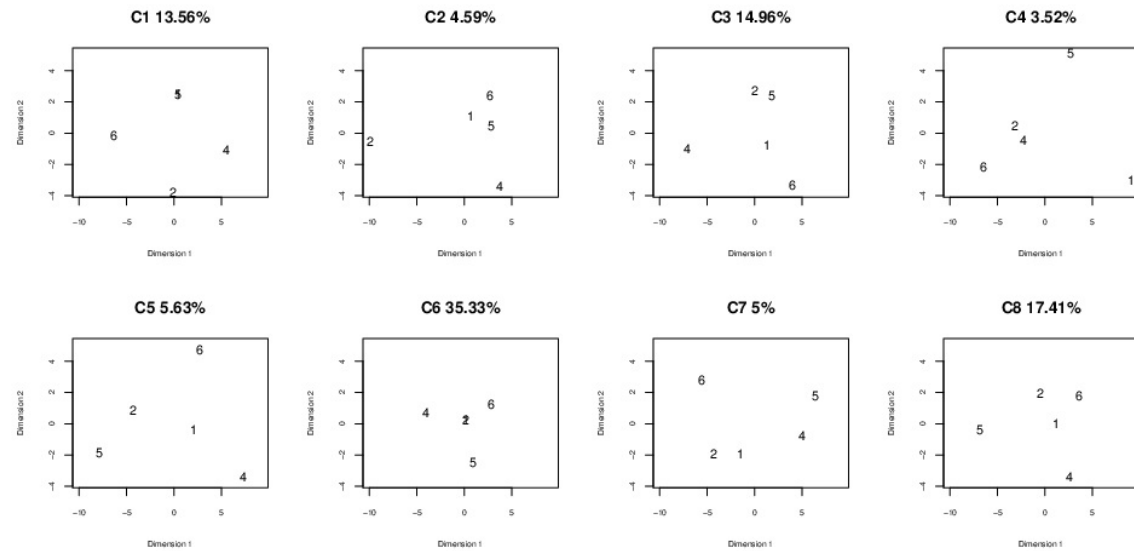


Figure 3: Map representing, for each of the 8 clusters (C1: Cluster 1, C2: Cluster 2, ...), the average position in the  $x - y$  axes of the five players, using MDS. Percentages report the proportion of instants in the dataset belonging to each cluster.

After having defined whether each moment corresponds to an offensive or a defensive play looking to the average coordinate of the five players in the court, we also found that some clusters represent offensive actions rather than defensive. More precisely, we found that cluster 1 (C1), Cluster 2 (C2), Cluster 3 (C3) and Cluster 4 (C4) mainly correspond to offensive plays (respectively, for the 85.88%, 85.91%, 73.93% and 84.62% of the time instants in each cluster) and Cluster 6 (C6) strongly corresponds to defensive plays (85.07%). *Offensive* clusters show larger players' spacings than in the *defensive* cluster. A motivation for this behaviour could be that players in defence have the objective to narrow the opponents' spacings in order to limit their play, while the aim of the offensive team is to maintain large distances among team-mates, to increase the propensity to shot with good scoring percentages. Anyhow, these findings go on the same direction of those of the convex hulls.

## 6 Conclusions and future developments

In recent years, coaches, experts and analysts have received benefits from the availability of large amounts of data to use in team sports analysis, which increased the possibility to extract insights from matches in relation to teams' performance. The advent of information technology systems permits to match play-by-play with players' trajectories and to analyse teams' performance with a variety of approaches. Having the trajectories of the players and the play-by-play available, and inspired by the literature based on the data-driven methods as well as by the increasing interest in visualizing data, we analysed the movement and the positioning of players using visual tools and data-mining techniques,

with the aim of finding regularities and patterns.

The most promising result relates to convex hulls' analysis. We found that players are more spread around the court in offensive plays rather than in defensive plays. Further analysis should be carried out in order to better understand the logic underpinning this regularity. A potential approach could be to examine the time series of the convex hull areas of both teams together. This will answer the question whether the defensive team has success in limiting the spacing of the offensive team. An analysis that aims to assess whether and how the two teams pursue their strategies, and how the achievement of their strategy affects their performance, may be of interest for coaches and experts.

Further research can be carried out with the aim of finding regularities between trajectories and players' (and team) performance by increasing the availability of trajectories data to both team-mates, opponents, and the ball, for multiple games, that sounds essential to better explore the multivariate and complex structure of trajectories in association with teams' performance. Future challenges also aim to experiment the potential of spatial statistics and spatial econometrics techniques applied to trajectory analysis (Brillinger (2010)), also in view of the similarities between sport players and economic agents in terms of endogenous and exogenous factors that impact on locational choices, as illustrated in Arbia (2016).

## Acknowledgements

Research carried out in collaboration with the Big&Open Data Innovation Laboratory (BODaI-Lab), University of Brescia (project nr. 03-2016, title Big Data Analytics in Sports, [www.bodai.unibs.it/BDSports/](http://www.bodai.unibs.it/BDSports/)), granted by Fondazione Cariplo and Regione Lombardia. Authors would like to thank MYagonism (<https://www.myagonism.com/>) for having provided the data. A special thank goes to Raffaele Imbrogno ("Foro Italico" University, Roma IV) and Paolo Raineri (MYagonism) for fruitful discussions.

## References

- [1] Aisch, G., Quealy, K. (2016) *Stephen Curry 3-Point Record in Context: Off the Charts*. New York Times.
- [2] Arbia, G. (2016) *Spatial Econometrics: A Broad View*. Foundations and Trends in Econometrics **8** (3-4), 145-265.
- [3] Basole, R. C., Saupe, D. (2016) *Sports Data Visualization* [Guest editors' introduction]. IEEE Computer Graphics and Applications **36** (5), 24-26.
- [4] Bolt, M. D. (2015) *Visualizing Water Quality Sampling-Events in Florida*. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **2** (4), 73.
- [5] Bradley, P., O'Donoghue, P., Wooster, B., Tordoff, P. (2007) *The reliability of ProZone MatchViewer: a video-based technical performance analysis system*. International Journal of Performance Analysis in Sport **7.3**: 117-129.
- [6] Brillinger, D. R. (2007) *A potential function approach to the flow of play in soccer*. Journal of Quantitative Analysis in Sports **3** (1), 3.
- [7] Brillinger, D. R. (2010) *Modeling spatial trajectories*. Handbook of spatial statistics 463-476.
- [8] Carpita, M., Sandri, M., Simonetto, A., Zuccolotto, P. (2013) *Football mining with R*. Data Mining Applications with R.
- [9] Carpita, M., Sandri, M., Simonetto, A., Zuccolotto, P. (2015) *Discovering the Drivers of Football Match Outcomes with Data Mining*. Quality Technology & Quantitative Management **12.4** 561-577.

- [10] Catapult USA, Sports Ltd. (2015) *Wearable Technology for Elite Sports*. URL <http://www.catapultsports.com/>.
- [11] Cintia, P., Giannotti, F., Pappalardo, L., Pedreschi, D., Malvaldi, M. (2015) *The harsh rule of the goals: data-driven performance indicators for football teams*. Data Science and Advanced Analytics (DSAA) 36678. IEEE International Conference pp. 1-10.
- [12] Cintia, P., Rinzivillo, S., Pappalardo, L. (2015) *A network-based approach to evaluate the performance of football teams*. Machine Learning and Data Mining for Sports Analytics Workshop, Porto, Portugal.
- [13] Couceiro, M. S., Clemente, F. M., Martins, F. M., Machado, J. A. T. (2014) *Dynamical stability and predictability of football players: the study of one match*. Entropy **16** (2), 645-674.
- [14] Fonseca, S., Milho, J., Travassos, B., Araujo, D. (2012) *Spatial dynamics of team sports exposed by Voronoi diagrams*. Human movement science **31** (6), 1652-1659.
- [15] Gesmann, M., de Castillo, D. (2013) *Package 'googleVis'*. Interface between R and the Google Chart Tools.
- [16] Goldsberry, K. (2012) *Courtvision: New visual and spatial analytics for the nba*. 2012 MIT Sloan Sports Analytics Conference.
- [17] Goldsberry, K. (2013) *Pass Atlas: A Map of where NFL Quarterbacks throw the ball*. Grantland.
- [18] Heinz, S. (2014) *Practical application of motion charts in insurance*.
- [19] Hilpert, M. (2011) *Dynamic visualizations of language change*. International Journal of Corpus Linguistics **16** (4), 435-461.
- [20] Impire (2015) Impire AG. URL <http://www.bundesliga-datenbank.de/en/products/>
- [21] Losada, A. G., Theron, R., Benito, A. (2016) *BKViz: A Basketball Visual Analysis Tool*. IEEE Computer Graphics and Applications **36** (6), 58-68.
- [22] Metulini, R. (2017) *Spatio-Temporal Movements in Team Sports: A Visualization approach using Motion Charts*. Electronic Journal of Applied Statistical Analysis.
- [23] Metulini, R., Manisera, M., Zuccolotto, P. (2017) *Space-Time Analysis of Movements in Basketball using Sensor Data*. in Statistics and Data Science: new challenges, new generations. SIS2017 proceedings (Firenze University Press) ISBN: .
- [24] Moura, F. A., Martins, L. E. B., Anido, R. D. O., De Barros, R. M. L., Cunha, S. A. (2012) *Quantitative analysis of Brazilian football players' organisation on the pitch*. Sports Biomechanics **11** (1), 85-96.
- [25] Passos, P., Davids, K., Araujo, D., Paz, N., Minguens, J., Mendes, J. (2011) *Networks as a novel tool for studying team ball sports as complex social systems*. Journal of Science and Medicine in Sport **14.2** : 170-176.
- [26] Passos, P., Araujo, D., Volossovitch, A. (2016) *Performance Analysis in Team Sports*. Routledge.
- [27] Perin, C., Vuillemot, R., Fekete, J. D. (2013) *SoccerStories: A kick-off for visual soccer analysis*. IEEE transactions on visualization and computer graphics **19** (12), 2506-2515.
- [28] Pileggi, H., Stolper, C. D., Boyle, J. M., Stasko, J. T.: Snapshot (2012) *Visualization to propel ice hockey analytics*. IEEE Transactions on Visualization and Computer Graphics **18** (12), 2819-2828.
- [29] Polk, T., Yang, J., Hu, Y., Zhao, Y. (2014) *Tennis: Visualization for tennis match analysis*. IEEE transactions on visualization and computer graphics **20** (12), 2339-2348.
- [30] Sacha, D., Stein, M., Schreck, T., Keim, D. A., Deussen, O. (2014) *Feature-driven visual analytics of soccer data*. Visual Analytics Science and Technology (VAST), IEEE Conference on (pp. 13-22).
- [31] Saka, C., Jimichi, M. (2015) *Inequality evidence from accounting data visualisation*.
- [32] Santori, G. (2014) *Application of Interactive Motion Charts for Displaying Liver Transplantation Data in Public Websites*. In Transplantation proceedings **46** (7), 2283-2286. Elsevier.
- [33] Santos, J. L., Govaerts, S., Verbert, K., Duval, E. (2012) *Goal-oriented visualizations of activity tracking: a case study with engineering students*. In Proceedings of the 2nd international conference on learning analytics and knowledge, pp. 143-152. ACM.
- [34] Theron, R., Casares, L. (2010) *Visual analysis of time-motion in basketball games*. In International Symposium on Smart Graphics. Springer Berlin Heidelberg, pp. 196-207.

Sensor Analytics in Basketball

Metulini, R. and Manisera, M. and Zuccolotto, P.

- [35] Tracab (2015) *Tracab Player Tracking System*. URL <http://chyronhego.com/sports-data/player-tracking>.
- [36] Travassos, B., Araujo, D., Duarte, R., McGarry, T. (2012) *Spatiotemporal coordination behaviors in futsal (indoor football) are guided by informational game constraints*. *Human Movement Science* **31** (4), 932-945.
- [37] Travassos, B., Davids, K., Araujo, D., Esteves, P. T. (2013) *Performance analysis in team sports: Advances from an Ecological Dynamics approach*. *International Journal of Performance Analysis in Sport*, **13.1** : 83-95.
- [38] Turvey, M. T., Robert E. Shaw. (1995) *Toward an ecological physics and a physical psychology*. *The science of the mind: 2001 and beyond*, 144-169.
- [39] Wasserman, S., Katherine F. (1994) *Social network analysis: Methods and applications*, Vol. 8. Cambridge university press.

# Evaluation of a corporate physical activity program using mixed methods

D.Meyer<sup>1</sup>, S.Muir<sup>2</sup>, M. Weerasinghe Jayawardana<sup>3</sup>, D. Ho<sup>4</sup>, O. Sackett<sup>5</sup>

<sup>1</sup> Swinburne University of Technology dmeyer@swin.edu.au

<sup>2</sup> Swinburne University of Technology sdmuir@swin.edu.au

<sup>2</sup> Swinburne University of Technology mjawardana@swin.edu.au

<sup>4</sup> Virgin Pulse David.Ho@virginpulse.com

<sup>5</sup> Virgin Pulse Olivia.Sackett@virginpulse.com

## Abstract

Physical inactivity is associated with obesity and has been linked to cancer, diabetes, heart disease and depression. The resulting sick leave, absenteeism and productivity losses have severe economic consequences for businesses. Corporate physical activity programs are one way in which businesses can address this problem. In this study we evaluate the impact of such a program using mixed methods, focusing on performance measures relating to physical activity, nutrition, feelings of happiness and wellbeing. The large sample size (more than 18000 participants) allowed the use of text mining to provide context for when the program resulted in a recommendation to a friend. It also allowed the use of machine learning methods to identify what program modules and features influenced the success of the program and for whom the program was particularly successful. Particular strengths of the analysis include the use of a mixed methods design, the triangulation of results using baseline and post-program measures to supplement survey responses, the use of sophisticated text mining approaches, and the integration of the quantitative and qualitative modules of the study using machine learning approaches. The results will inform future developments of this and other programs designed to increase physical activity and improve employee health.

## 1 Introduction

Globally in 2014, more than 1.9 billion adults were overweight (39%), and of these over 600 million (13%) were obese (WHO, 2006). Obesity is particularly prevalent in Western countries such as Australia, where 63.4% of adults (11.2 million people) were classified as either overweight or obese in 2014-2015 (ABS, 2015). Such figures are particularly concerning considering that obesity is a proven risk factor for serious health problems, including cardiovascular disease and diabetes. Physical inactivity is one of the major contributors to obesity and poor health among adults, which has led to the World Health Organisation implementing physical activity guidelines which recommend at least 30 minutes of physical activity per day (WHO, 2010). However, many adults remain physically inactive with one in four adults not meeting these guidelines (WHO, 2017). Apart from regular physical activity, a healthy diet usually guards against obesity. A high intake of dietary non-starch polysaccharides/fibre as opposed



to a high intake of energy-dense, micronutrient-poor foods is recommended by Swinburn, Caterson, Seidell and James (2004).

Health initiatives, aimed at increasing physical activity and/or improving nutrition are therefore recognized as being vitally important. Workplace initiatives have been particularly recommended by Dishman, Oldenburg, O'Neal and James (1998) because people spend so much of their time at and traveling to and from work. Virgin Pulse is a global Software as a Service (SaaS) vendor addressing this need. One of the Virgin Pulse programs, called the Global Challenge (GC), consists of a 100-day virtual journey during which employees are placed in teams of seven, provided with an activity tracker and given access to an application through a web browser or mobile device. Teams compete with one another to accumulate steps, measured by the activity trackers. In addition to promoting physical activity, the program incorporates a number of modules which focus on encouraging improvement in sleep, nutrition and life-work balance. The program is gamified to encourage employees to develop healthy habits through education, goal setting and positive reinforcement using progress monitoring and (virtual) achievement awards.

This paper provides an assessment of the effectiveness of the GC program in terms of improving employee awareness of the importance of daily physical activity and a healthy diet, and improving happiness and wellbeing. In this paper we use both quantitative and qualitative data in our evaluation. We use text mining to analyze the results of an open-ended description of the GC program, and then, using machine learning tools we augment our quantitative evaluations with this qualitative data, thereby conducting a truly 'mixed methods' analysis as described by Creswell (2015).

In this paper we address four research questions. The first concerns the effectiveness of the GC program with improvements expected for the four outcome measures (daily physical activity, healthy nutrition, happy feelings and wellbeing). The second concerns the relationship between these four outcome measures with Happy Feelings expected to mediate the relationship between Physical Activity and Wellbeing and the relationship between Healthy Nutrition and Wellbeing. The third research question concerns the impact of the various program modules and features on the success of the program, with differences expected for the four outcome measures. Finally, the fourth research question concerns the importance of analyzing qualitative data in order to enhance the understanding acquired through a quantitative analysis.

## 2 Method

### 2.1 Data Collection

Qualitative and quantitative survey data were collected from more than 18000 participants in the Virgin Pulse GC program during the period May-September 2016. Self-reported awareness of the importance of daily physical activity and a healthy diet, as well as perceived happy feelings and wellbeing, were collected at the start and end of the 100-day program. Participant experience data was collected through an additional survey released approximately two weeks after program completion. In particular, this survey provided unstructured text responses to the question "In one or two sentences how would you describe the GC to a friend or colleague?" The online nature of the program meant that all data was collected through questionnaires provided by way of web links.

## 2.2 Global Challenge (GC) Measures

In this study the World Health Organisation measure of wellbeing (WHO5) was used as the primary outcome measure. Secondary outcome measures included awareness of the importance of daily physical activity and a healthy diet and perceptions of happy feelings, all measured on a 0-6 ordinal scale, with higher scores indicating better performance. An attempt was made to establish what were the important contributing program characteristics for the improvements seen in the outcome measures. In particular, perceptions of the best three modules of the program, chosen from GC Physical Activity, GC Nutrition, GC Sleep, GC Me and GC Balance, and perceptions of the best program features (e.g., virtual trophies) are investigated. All these variables were measured on a binary scale with one for a positive response and zero for a negative response. Finally, the open-ended responses to the request for a description of the GC program were analysed. Other data collected, such as stress, happiness, mental health and productivity were left for other studies.

## 2.3 Initial Statistical Analysis

The quantitative data were initially analysed using traditional statistical methods. In view of the very large sample size only results that were significant with  $\alpha < 0.1\%$  were reported, together with effect sizes where relevant. SPSS version 24 and AMOS version 24 were used for this initial analysis.

*Descriptive statistics* were accumulated for observed initial values and for improvements in awareness of the importance of daily physical activity and a healthy diet, perceptions of happy feelings and wellbeing. Addressing the first research question, paired tests were used to test for significant improvements, while Pearson correlations were used to test for significant linear associations between the outcome measures and their change scores. The paired t-tests were then repeated with a last observation carried forward (LOCF) substitution for missing post-program data, in order to provide protection against attrition bias.

A *3-stage hierarchical regression analysis for wellbeing (WHO5)* was used to address the second research question with baseline wellbeing scores included at stage 1 and change scores for awareness of the importance of daily physical activity and a healthy diet added at stage 2, and change scores for happy feelings added at stage 3. This allowed a test for mediation, and a comparison of standardised weights and total standardised effect sizes produced a ranking of the three secondary measures in terms of their impact on wellbeing.

*Additional Hierarchical Regression Analyses* were then conducted to determine which program modules and features were most important for predicting final levels of wellbeing (WHO5), happy feelings and awareness of the importance of daily activity and a healthy diet, while controlling for initial levels on these variables and demographic characteristics. The ordinal scales for happy feelings and awareness of the importance of daily physical activity and a healthy diet meant that an ordinal hierarchical logistic regression analysis was needed for these variables.

## 2.4 Text and Data Mining

The large data set containing both quantitative and qualitative data allowed a data mining approach for the evaluation of this program. Text mining was applied to the qualitative responses in the survey and regression trees were employed in order to predict improvements in awareness of the importance of

daily physical activity and healthy nutrition, happiness and wellbeing. Importantly the data were randomly split (40%, 30%, 30%) for training, validation and testing (Linoff and Berry. 2014). This meant that the regression trees could be fitted and validated before being tested on fresh data, ensuring that these models were not over-fitted and that reliable goodness of fit statistics could be obtained and compared across the various models. SAS Enterprise Miner version 14.1 was used for the Text Mining and Machine Learning analyses.

*Text parsing* was applied to the open-ended text responses of respondents providing a description of the GC program for a friend. This text mining involved creating a dictionary of terms using natural language processing (NLP). This includes identifying parts of speech and word stemming whereby words such as action, actions, activity and activities are all translated into a single "+act" noun. This dictionary of terms allowed the formation of a term by respondent matrix (sometimes called a term by document matrix) with the frequency for each term entered for all respondents.

*Text filtering* was then used to identify the terms with very high or very low frequencies, which were not useful for discriminating between respondents. The best discriminators tend to have total frequencies between  $n/10$  and  $n/100$  when there are  $n$  respondents (Chakraborty et al., 2013). Weights were assigned to terms in order to facilitate this discrimination process.

*Topic extraction* was used to determine how various client responses are associated with themes or ideas. These topics consisted of terms that commonly occurred in the same response, with each respondent given a score representing the strength of the association of their response for each topic. Each client response may relate to zero, one or more topics. The process of topic extraction is briefly described below. The term by respondent matrix described earlier tends to be very sparse with a very large number of terms, therefore requiring a reduction in dimensionality. This can be achieved through latent semantic indexing (LSI), sometimes called latent semantic analysis (LSA), using the weights described above. This procedure is a form of spectral value decomposition which is related to principal module analysis. Linearly independent dimensions were extracted from the weighted term by respondent matrix in such a way that terms that were important for any particular dimension tended to be related to each other, making interpretation easy. These dimensions were rotated and then used to define topics of interest. Each topic corresponds to a single dimension, with the term topic weights represented by term co-ordinates on the topic dimensions. Similarly, respondents were assigned a weight corresponding to each topic. Term cut-off values are threshold scores for term topic weights, to determine which terms belong to a topic. These thresholds are set one standard deviation above the mean term topic weight. Similarly, respondent cutoffs are set one standard deviation above the mean of the respondent topic weights. The Topic scores obtained for all respondents had very skewed distributions, making a regression tree approach more appropriate for the ensuing analyses of these scores.

*Regression trees* were then used in order to build models for the prediction of improvements in wellbeing and happiness and for the prediction of improvements in awareness of the importance of daily physical activity and healthy nutrition, using only the above topics as predictors. This was done in order to provide a "thick (contextual) description" of the GC program (Palinkas, 2014). Regression trees are constructed in a top-down recursive manner with all the data in the training data initially entered into a root node. An optimum choice of topics for splitting and an optimum choice of splitting points based on respondent scores is then employed in order to ensure that successive splits of the data into child nodes reduces the variance within each node, with the predicted improvement for any child node defined as the average change score for respondents within all such nodes. This recursive splitting process stops when the variance for the prediction errors is minimized using the validation data and the final model is

assessed using the test data. Advantages of this data mining approach over other methods, such as multiple linear regression or neural networks, include the following. There is no need to impute missing values, irrelevant predictor variables are automatically excluded and significant interaction and nonlinear terms are automatically included, there is no need for data transformation or assumptions of normality and, most importantly, the trees are easy to interpret, understand and check against domain knowledge. However, further refinements to prediction accuracy are possible using a technique called gradient boosting in which a series of regression trees are created which together form a single predictive model. A tree in the series is fit to the prediction errors from the earlier trees in the series. Importantly this method allows an accurate ranking of all the predictor variable in terms of their importance for explaining the variation in respondent change scores. This is done using the number of splits performed using each variable or by the reduction in error variance achieved by each variable when used for splitting.

*Mixed methods* analysis was utilized in the final analysis with a tree for improvements in wellbeing fitted in terms of all the quantitative variables and all the qualitative topic data. This approach allowed an evaluation of the importance of qualitative and quantitative data for determining why the GC program was so beneficial for the majority of people included in this study.

### 3 Results

#### 3.1 Initial Statistical Analysis

*Descriptive Statistics.* A total of 18674 people were included in this study, 56% female and 44% male. The average age was 43 years with a range of 18 to 76 and a standard deviation of 10.75 years. The average height was 171cms with a standard deviation of 10cms and the average weight was 78kg with a standard deviation of 19kg. Response rates for the 18674 people included in the sample were relatively high at baseline (e.g. 94% for WHO5) but falling by the end of the program (e.g. 72% for WHO5). Interestingly attrition rates for wellbeing and happy feelings were related to baseline levels for the awareness of the importance of good nutrition but they were not related to baseline levels for the awareness of the importance of daily physical activity. However, baseline awareness of the importance of healthy nutrition was only marginally lower for respondents who did not complete both WHO5 assessment (MN=3.75, SD=1.29 versus MN=3.82, SD=1.27) and marginally lower for those who did not complete both Happy Feelings assessments (MN=3.73, SD=1.29 vs MN=3.82, SD=1.27).

Table 1 provides summary statistics for the baseline and change scores for wellbeing (WHO5), happy feelings and awareness of the importance of daily physical activity (PhysAct) and healthy nutrition, with higher values indicating better performance. In response to the first research question significant improvements for awareness of the importance of daily physical activity and healthy nutrition as well as happy feelings and wellbeing were found ( $p < .001$ ), with large effect sizes in all cases except awareness of the importance of healthy nutrition for which the effect size was moderate. These conclusions were maintained even when conservatively biased tests were performed (LOCF for missing post-program data). There were negative correlations of moderate size between the baseline and change scores, suggesting that improvements were larger for respondents with relatively low baseline performance. Moderate strength correlations were also found between wellbeing and happy feelings and between awareness of the importance of daily physical activity and healthy nutrition.

Table 1. Descriptive Statistics for the Primary and Secondary Outcome Measures

	Initial Scores				Improvement Scores for GC Program			
	1.WHO5	2.Happy feelings	3.Phys. Act.	4.Nutri-tion	5.WHO5	6.Happy feelings	7.Phys. Act.	8.Nutri-tion
Mean	54.27	3.35	3.70	3.80	13.91	.65	.99	.53
Std. Dev.	19.07	1.28	1.35	1.27	18.07	1.30	1.43	1.33
Effect Size( $\eta^2$ )					.372*	.201*	.322*	.136*
LOCF Effect Size( $\eta^2$ )					.288*	.158*	.274*	.108*
Sample Size	17480	18190	18481	18277	13529	14292	15745	14467
Correlations								
1	1.00							
2	.65*	1.00						
3	.23*	.17*	1.00					
4	.23*	.19*	.54*	1.00				
5	-.51*	-.28*	-.14**	-.14*	1.00			
6	-.22*	-.53*	-.08*	-.09*	.52*	1.00		
7	-.11*	-.08*	-.70*	-.30*	.32*	.23*	1.00	
8	-.08*	-.07*	-.25*	-.56*	.33*	.26*	.42*	1.00

(\*  $p < .001$  for paired t-tests and Wilcoxon Rank Signed Rank test)

LOCF: Last Observation Carried Forward for missing data

A 3-stage hierarchical regression analysis for wellbeing was used to assess the relationship between wellbeing and the changes in happy feelings and changes in the awareness of the importance of daily physical activity and healthy nutrition while controlling for baseline wellbeing scores. Only 29% of the variation in the final wellbeing scores was explained by the initial wellbeing scores. This percentage increased to 39% when the improvement scores for awareness of the importance of daily physical activity and healthy nutrition were included in stage 2 and which then increased to 52% when improvements in feeling happy were added at stage 3. As illustrated in Figure 1 in addition to a direct effect of awareness of the importance of daily physical activity and healthy nutrition on final wellbeing, there is also an indirect effect through improvements in happy feelings. As shown by the standardised weights and total standardised effect sizes in Table 2, the most important variable for predicting final wellbeing levels was the baseline wellbeing scores followed by changes in happy feelings, then improvements in awareness of the importance of healthy nutrition and finally improvements in awareness of the importance of daily physical activity. These results suggest that improving wellbeing is a goal which needs a multi-faceted approach addressing all three of these criteria (happiness, nutrition and physical activity).

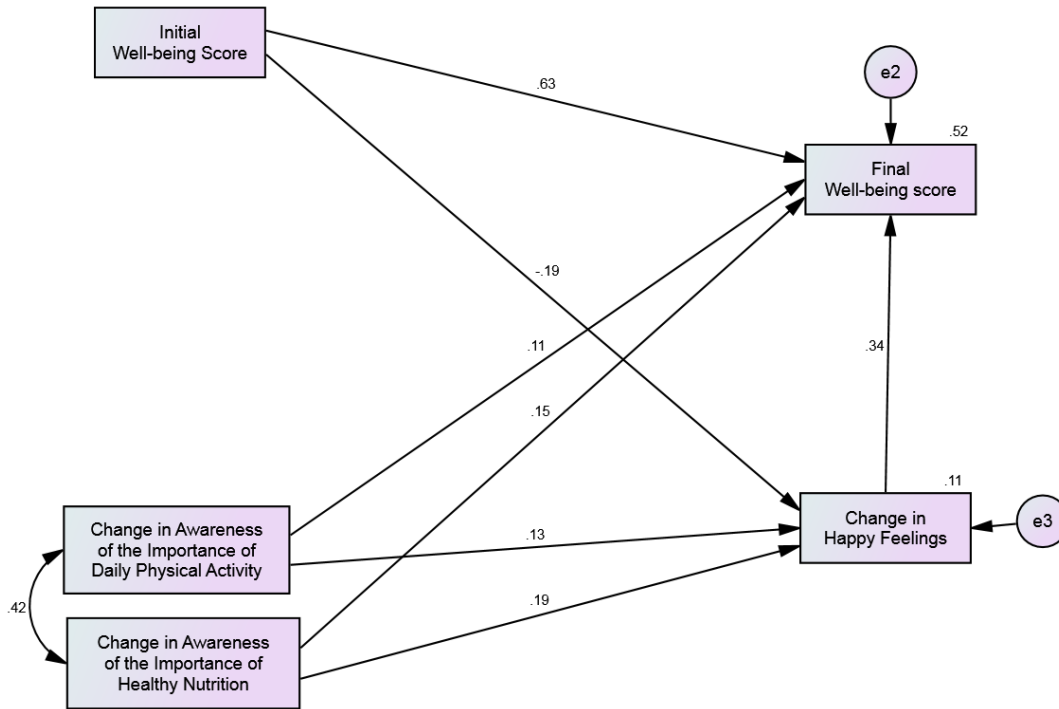


Figure 1: Drivers for Wellbeing in terms of the Change Scores for the Secondary Outcome Measures with Standardised Weights and R-Square Values shown

Table 2. Regression Analysis for Wellbeing (WHO5 Scores) after the GC Program

Predictors	Estimated Coefficient	Coefficient Standard error	Standardised Coefficient	Total Standardised Effect Size ( $\eta^2$ )
Baseline WHO5	.618*	.006	.641	.563
Change in Happy Feelings	4.994*	.094	.353	.344
Change Importance Healthy Nutrition	2.067*	.097	.150	.212
Change Importance Daily Physical Activity	1.549*	.090	.120	.160

*Hierarchical Regression Analyses* were then conducted to determine which program modules and features were most important for predicting final levels of wellbeing and happiness and awareness of the importance of daily physical activity and healthy nutrition, while controlling for initial levels on these variables and demographic characteristics. Ordinal logistic regressions were used for the happiness and awareness models because these variables were measured on a 0-6 ordinal scale. After entering baseline values at stage 1, demographic variables (gender, age, Body Mass Index) were entered at stage

2, with a binary evaluation of the program modules at stage 3 and a binary evaluation of the program features at stage 4. For all the binary measures a more positive evaluation was coded as one as opposed to zero.

Table 3 shows the percentage of variation explained for the final scores for wellbeing, happy feelings and the importance of daily physical activity and healthy nutrition, together with the significant factors at each stage. Age had a positive association with performance in all cases, and, in the case of the importance of daily physical activity, performance was worse for people with a higher body mass index (BMI). However, in the case of the importance of healthy nutrition and happy feelings, performance was better for people with higher BMI's. Females had better performance than males in terms of the importance of daily physical activity and healthy nutrition.

Table 3. Hierarchical Regression Analysis for Final Scores

	R-square for Final Scores (Nagelkerke for ordinal logistic regressions)			
Outcome measures	Importance daily physical activity	Importance healthy nutrition	Happy feelings	Wellbeing
Type regression	Ordinal logistic	Ordinal logistic	Ordinal logistic	Standard
Initial scores	11.4%	20.5%	24.5%	29.0%
Demographics	12.4% Age(+), BMI(-), Female(+)	21.0% Age(+), BMI(+), Female(+)	24.6% Age(+), BMI(+)	29.5% Age(+)
Best GC Modules	14.4% (All)	23.5% (All but physical Activity)	25.9% (All)	31.4% (All)
Best GC Features	16.8% (Mini_Challenge, Leaderboard, Ind_Mini_Leagues, Trophies, My_Stats)	24.3% (Mini_Challenge, Competitions, Trophies, GC_Community,)	26.9% (Mini_Challenge, Competitions, Trophies)	32.8% (Mini_Challenge, Leaderboard, Competitions, Ind_Mini_Leagues, Trophies, GC_Community, My_Stats)
Sample Size	15635	14375	14206	13449

These results suggest that in terms of these outcome measures the GC program is more successful for older people, with BMI and gender also impacting the success of the program. All modules of the program were beneficial, although the Physical Activity module had no significant association with healthy nutrition. Of the twelve program features, several were not seen as beneficial, however, the Mini-Challenges and Trophies were helpful for improving performance for all four of the above outcome measures.

A ranking of the importance of the program modules is provided in Table 4, based on t-scores (estimated coefficient/standard error). The GC Balance module was seen as the most important module in all cases except for awareness of the importance of healthy nutrition. The major benefit of the GC Balance module was a more positive outlook with this benefit reported by 58% of the people who completed this module of the program. Despite being a highly popular module of the program for 90%

of participants, GC Physical Activity was the least important module in all cases except for awareness of the importance of daily physical activity outcome measure.

Table 4. Ranking of the Best GC Program Modules at Stage 3 of Regression for each Outcome Measure using t-scores (1=most important, 5=least important)

Best GC Modules	Rankings for each Outcome Measure			
	Importance daily physical activity	Importance healthy nutrition	Happy feelings	Wellbeing
Physical Activity	2)10.59*	5)2.48	5)4.12*	5)6.29*
GC Me	4)8.49*	3)6.46*	4)5.58*	4)7.61*
GC Sleep	5)4.58*	4)6.38*	3)5.71*	3)8.53*
GC Nutrition	3)8.60*	1)17.26*	2)9.11*	2)12.14*
GC Balance	1)10.89*	2)10.43*	1)10.58*	1)12.17*

(\*  $p \leq 0.001$ )

These results provide an answer for the third research question, with the importance of the different program modules and features found to differ depending on the outcome measure. In particular, the low rankings for the Physical Activity Module of the program for happy feelings and wellbeing were a surprise. Even more of a surprise was the relatively low importance of many of the program features. It was only the Mini-Challenges and Trophies which appeared to have consistently beneficial impacts across the four outcome measures. However, what is most obvious in Table 3 is the low R-Square values, suggesting that it is not so much the individual modules and features of the program that drives its success as the composite whole.

The above quantitative results have done little to explain why the GC program is so successful in improving wellbeing, happiness and awareness of the importance of daily physical activity and healthy nutrition. We therefore now consider the words used by participants to describe the GC program to their friends and colleagues.

### 3.2 Text Mining

Attention now focuses on a qualitative module of the survey, namely responses to the question, “in one or two sentences, how would you describe the GC to a friend or colleague”. Parsing, filtering and topic selection produced the 25 topics displayed in Table 5. The number of terms associated with each topic and the number of respondents (# Docs) associated with each topic are also displayed in Table 5. The most common topic relates to good motivation and good fun (Topic 11) with 1002 respondents describing the GC in this way.



Table 5. Topics Extracted from the GC Description

Topic ID	Topic	Number of Terms	# Docs
1	fun,good fun,great fun,competitive,+great	20	689
2	+activity,+level,physical,+activity level,+physical activity	32	773
3	+healthy,+lifestyle,+healthy lifestyle,+help,+tool	39	889
4	+day,+step,+walk,aware,+little	33	675
5	â,f,ã,de,zu	32	130
6	+competition,friendly,+friendly competition,+colleague,gcc	32	436
7	health,+improve,+opportunity,+great,wellbeing	45	950
8	+great,+motivator,great motivator,moving,good motivator	31	878
9	+move,+encourage,aware,daily,good	20	501
10	+motivate,+exercise,moving,gcc,+walk	21	914
11	+good,+motivator,good fun,good motivator,+tool	29	1002
12	+challenge,+help,gcc,+team challenge,great challenge	41	693
13	team,building,team building,good,+great	46	469
14	+keep,+track,+help,+goal,fit	35	570
15	active,+encourage,aware,+great,+stay	25	573
16	fun,competitive,+exercise,+easy,+colleague	30	740
17	+challenge,fun,+help,+life,+goal	20	354
18	+motivation,great motivation,+great,+tool,good motivation	24	366
19	+team,+compete,+work,world,+colleague	45	677
20	+life,+healthy,style,+healthy life,life	44	539
21	+program,+encourage,physical,gcc,motivational	50	504
22	+exercise,daily,+help,gcc,aware	45	703
23	+step,+track,daily,+count,+easy	46	797
24	fitness,+improve,+level,+increase,+fitness level	40	750
25	+experience,+help,great experience,motivational,gcc	35	603

Figure 2 shows the trees that were constructed from these topics for improvements in awareness of the importance of physical activity and healthy nutrition as well as improvements in feelings of happiness and wellbeing. The greatest improvements in awareness of the importance of daily physical activity and healthy nutrition occurred for participants who scored relatively low on Topic 23, about “keeping track of the number of steps taken each day”. However, those who scored high on Topic 20 concerning “healthy lifestyle” tended to obtain more improvements regarding healthy nutrition and in terms of happy feelings. The topic dealing with “keeping track of goals” (Topic 14) was associated with improvements in wellbeing as was the topic referring to “motivation to get moving” (Topic 8). In all these trees the Count refers to the number of participants for the training and validation data in each box with the thick black line indicating the path for the majority of participants after each split.

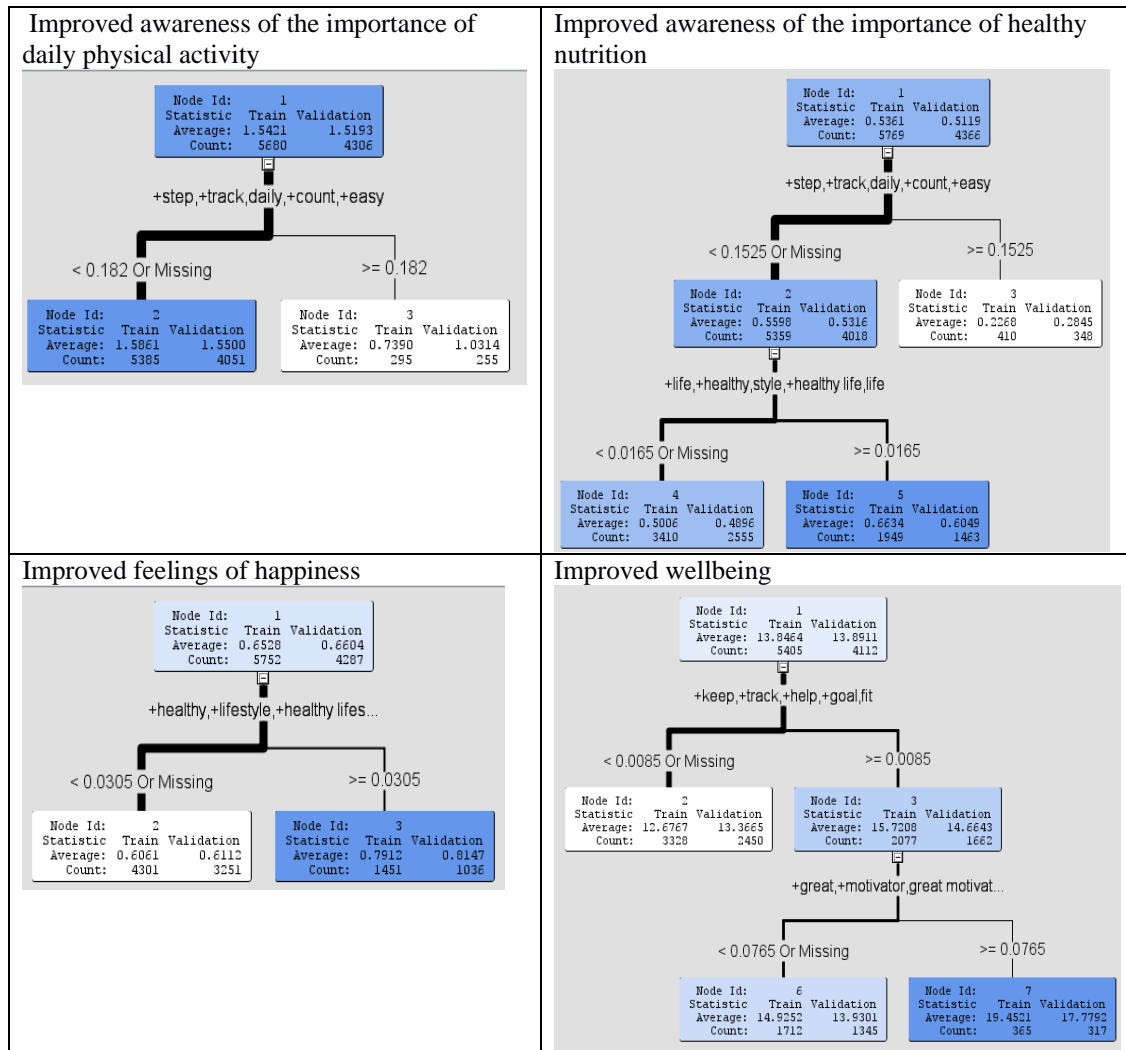


Figure 2: Trees for changes in awareness of the importance of daily physical activity and healthy nutrition, change in happy feelings and wellbeing using topics extracted from the GC descriptions.

### 3.3 Mixed Methods Analysis

The final tree for predicting improvements in wellbeing, incorporating all quantitative and qualitative data, was too large to display, but the regression model with stepwise selection of predictors shown in Table 6 had a similar R-Square value of 35%. As expected improvements in feelings of happiness was the most important predictor of improvements in wellbeing, followed by increased awareness of the importance of healthy nutrition and daily physical activity. After controlling for these predictors, the most beneficial module of the program was GC Sleep followed by GC Balance and the most beneficial program feature was the Trophies followed by the Leaderboard. Topic 1 (having fun and being

competitive) and 20 (healthy lifestyle) had a positive association with improvements in wellbeing, but there was a negative relationship with Topic 23 (keeping track of daily steps taken). The boosted regression tree provides some confirmation of these results while also suggesting that Topic 14 (“Helping to track fitness goals”) and Topic 8 (“Good motivation to get moving”) are also desirable aspects of the program.

Table 6: Regression Mixed Model Analysis with Boosted Relative Importance

Predictor	Regression with Stepwise Selection				Boosted Regression Tree	
	Estimated Coefficient	Std. Error	t-value	p-value	Number Splits	Importance Training /Validation Data
Improvements:						
Importance daily physical activity	1.85	.162	11.38	<.001	19	.15/0
Importance healthy nutrition	2.18	.174	12.50	<.001	22	.49/.41
Feeling happy	5.99	.165	36.25	<.001	29	1/1
Best Program Modules						
GCBalance	.78	.220	3.54	<.001	2	.09/.06
GCNutrition	.56	.228	2.46	.014		
GCSleep	1.03	.245	4.20	<.001	3	.08/.05
Best Program features						
Trophies	.49	.213	2.31	.021	5	.11/.07
Leaderboard	.51	.209	2.44	.015	4	.10/.03
Mini Challenge					3	.11/.09
Topics						
23	-8.74	2.99	-2.92	.004	3	.08/.05
20	7.63	2.95	2.59	.010	2	.07/0
1	5.17	2.51	2.06	.040		
14					7	.15/.07
8					3	.09/.03

## 4 Discussion

Too often have researchers and program developers been concerned with measuring merely the frequency of program and technology usage. It is time that program success be defined not by the amount but rather by the nature of its use. By identifying the specific features of the program that were most important to participants, the results of this study serve as a valuable contribution to the development of future web-based programs of this nature. Our results showed the Trophies and Mini-Challenges to consistently be the most important features of the GC. This suggests that future programs may benefit from incorporating unique challenges throughout their program as opposed to a single overall goal. In addition, the award of virtual trophies appears to be motivational for success in terms of physical activity,

nutrition, happiness and wellbeing goals. The GC Balance module also appears to tick several boxes. This module of the program was designed to promote life-work balance, but the results suggest that this module is also very helpful in regard to all four of the above outcome measures. However, it is important to recognize that, in isolation, individual modules and features of the program had little influence. Rather, it was the program in its entirety that was so effective. This is consistent with recent meta-analyses which have found that multi-module programs, utilising multiple behaviour change techniques, are superior to more solitary interventions (Webb et al., 2010). However, although the GC Physical Activity module was helpful for creating an awareness of the importance of daily physical activity (ranked second), it was ranked lowest for healthy nutrition, feelings of happiness and wellbeing.

Contrary to previous evaluations that have found greater improvements in outcome measures for those participants who are healthier at baseline (e.g., Ablah et al., 2015; Abraham, Crespin, & Rothman, 2015), the current evaluation found the greatest improvements for lower initial outcome levels. This is consistent with previous evaluations of the GC program which have found significant improvements among participants with poor wellbeing at baseline (Freak-Poli, Wolfe, Wong, & Peeters, 2014). This suggests that the GC program is effective in producing positive health behavior change among participants who are most at-risk and therefore, most in need of assistance.

Results from the current study also revealed a significant association between age, with older participants more likely to successfully engage in the GC. Previous studies investigating determinants of sustained participation have also reported older participants to be more likely to remain engaged in online workplace programs (Rongen, Robroek, Van Lenthe, & Burdorf, 2013). Effects of BMI were also observed in this study. In particular, it was found that improvements in awareness of the importance of daily physical activity were worse for participants with a higher BMI, while in the case of the importance of healthy nutrition and feelings of happiness, performance was better for participants with a higher BMI. These contradictory results are in line with what has been found in other studies. For instance, Abraham et al. (2015) found that participants with a higher BMI had more success in an incentive-based employer wellness program while Ablah et al. (2015) found that people with healthier BMI's performed better for minimal worksite interventions at small worksites. In this study females benefited more than males in terms of an awareness of the importance of daily physical activity and healthy nutrition. This supports the findings by Robroek, Brouwer, Lindeboom, Oenema and Budorg (2010) and Cook, Hersch, Schlossberg and Leaf (2015) in their evaluations of web-based health promotion programs but contradicts the study of Ross and Wing (2016) which showed greater success in the case of males for a similar programs.

Consistent with previous qualitative evaluations of the GC (Scherrer, Sheridan & Sibson, Ryan & Henley, 2010), the current study found employees to express a sense of enjoyment from participating in the program. Results from our qualitative analysis suggest that those who saw the GC program as enjoyable, competitive and promoting a healthy lifestyle were more likely to benefit in terms of feelings of happiness and awareness of the importance of a healthy diet. This is consistent with the results reported by Davey, Fitzpatrick, Garland and Kilgour (2009) who found enjoyment and challenge motives to be positive predictors of performance in a similar workplace health promotion program. The importance of enjoyment is noteworthy as such intrinsic motives have been found to be predictive of enhanced wellbeing (Ryan & Deci, 2000). By fostering a sense of enjoyment among participants, this may explain why the GC was so successful. Hence, programs targeting wellbeing should aim to incorporate features that participants are likely to enjoy engaging in. However, results for other features of the of GC were mixed as keeping track of steps served to reduce benefits relating to improved

awareness of the importance of daily physical activity and healthy nutrition. These descriptions also impacted on overall wellness which was also higher for those who found the GC helpful for tracking progress against fitness goals, a good motivator to get moving and great competitive fun.

However, it must be acknowledged that our sample only includes those that have completed the program and the final assessment. The assessment completion rate is good at over 70%, but we have not made any attempt to investigate the feelings of those who did not complete the program or the assessment. In addition, this paper has only considered outcomes relating to physical activity, nutrition, feelings of happiness and wellbeing. No attempt has been made to consider the effect of the program on mental health, stress, sleep or productivity in this paper. This will be addressed in other research.

### 4.3 Conclusions

The GC program is successful in increasing the awareness of the importance of daily physical activity and healthy nutrition. It is also successful in improving feelings of happiness and wellbeing. The qualitative analysis identified the motivation to adopt a healthy lifestyle, to meet goals and keep moving as important characteristics of the program, while the quantitative analysis identified the GC Balance module as having the greatest impact across these outcome measures. Trophies and Mini-Challenges are features of the GC which are also particularly beneficial. Most importantly the holistic nature of the GC is recognized as a key contributor to its success. However, the Physical Activity module of the program is perhaps worth more study. In particular, keeping track of steps appears to detract from the benefits of the program.

This paper has shown that a mixed methods analysis is well suited to an evaluation of programs such as the GC, with qualitative data used to add a richness and a context for the quantitative results. The use of text mining for the extraction of qualitative information is recommended because this information can then be easily combined with existing quantitative data. Future studies should consider employing similar techniques in order to provide further insights into the elements of programs that are responsible for success.

### References

- [1]Ablah, E., Dong, F., Konda, K., Konda, K., Armbruster, S. and Tuttle, B. (2015) *Early success is vital in minimal worksite wellness interventions at small worksites*. Health Education & Behavior, **242**(4):500–509.
- [2]Abraham, J. M., Feldman, R., Nyman, J. and Barleen, N. (2011). *What factors influence participation in an employer-based wellness program?* Inquiry, **48**:221-241.
- [3]Australian Bureau of Statistics (2015). *National Health Survey: First Results*, cat. no. 4364.0.55.001. 2015. <http://www.abs.gov.au/ausstats/abs@.nsf/mf/4364.0.55.001>. Accessed 10 Jan 2017.
- [4]Chakraborty. G., Pagolu, M. and Garla, S. (2013) *Text mining and analysis: practical methods, examples, and case studies using SAS*. SAS Institute Inc.
- [5]Cook, R. F., Hersch, R. K., Schlossberg, D. and Leaf, S. L. (2015) *A web-based health promotion program for older workers: Randomized controlled trial*. Journal of Medical Internet Research, **2015**;17(3):e82.
- [6]Creswell J.W. (2015) *A concise introduction to mixed methods research*, Thousand Oaks, California: Sage.
- [7]Davey, J., Fitzpatrick, M., Garland, R. and Kilgour, M. (2009) *Adult participation motives: Empirical evidence from a workplace exercise programme*. European Sport Management Quarterly, **9**(2):141–162.

- [8] Dishman, R. K., Oldenburg, B., O'Neal, H., & Shephard, R. J. (1998) *Worksite physical activity interventions*. American Journal of Preventive Medicine, **15**(4):344–361.
- [9] Freak-Poli, R. L. A., Wolfe, R., Walls, H., Backholer, K. and Peeters, A. (2011) *Participant characteristics associated with greater reductions in waist circumference during a four-month, pedometer-based, workplace health program*. BMC Public Health, **11**:824.
- [10] Linoff G.S. and Berry M.J. (2011) *Data mining techniques: for marketing, sales, and customer relationship management*. Wiley Computer Publishing.
- [11] Palinkas, L.A. (2014) *Qualitative and mixed methods in mental health services and implementation research*. Journal of Clinical Child and Adolescent Psychology, **43**(6); 851-861.
- [12] Robroek, S. J. W., Brouwer, W., Lindeboom, D., Oenema, A. and Burdorf, A. (2010) *Demographic, behavioral, and psychosocial correlates of using the website module of a worksite physical activity and healthy nutrition promotion program: A longitudinal study*. Journal of Medical Internet Research, **12**(3):e44.
- [13] Rongen, A., Robroek, S. J. W., van Lenthe, F. J. and Burdorf, A. (2013) *Workplace health promotion: a meta-analysis of effectiveness*. American Journal of Preventive Medicine, **44**(4):406-415.
- [14] Ross, K. M. and Wing, R. R. (2016) *Implementation of an internet weight loss program in a worksite setting*. Journal of Obesity, <http://doi.org/10.1155/2016/9372515>
- [15] Ryan, R., & Deci, E. (2000). Self-determination theory and the facilitation of intrinsic motivation. American Psychologist, **55**(1), 68–78. <https://doi.org/10.1037/0003-066X.55.1.68>
- [16] Scherrer, P., Sheridan, L., Sibson, R., Ryan, M., & Henley, N. (2010) *Employee engagement with a corporate physical activity program: the Global Corporate Challenge*. International Journal of Business Studies **18**(1), 125-139.
- [17] Swinburn, B.A., Caterson, I., Seidell, J.C. & James, W.P. (2004) *Diet, Nutrition and the prevention of excess weight gain and obesity*. Public Health Nutrition **7**(1A), 123-146.
- [18] Webb, T. L., Joseph, J., Yardley, L. and Michie, S. (2010) *Using the internet to promote health behavior change: A systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy*. Journal of Medical Internet Research, **12**(1):1-18.
- [19] World Health Organization (2006). *Obesity and overweight* (Fact sheet no. 311). <http://www.who.int/mediacentre/factsheets/fs311/en/>. Accessed 10 Jan 2017.
- [20] World Health Organization (2010). *Global recommendations on physical activity for health*. Geneva, Switzerland.
- [21] World Health Organization (2016). *Physical activity* (Fact sheet no. 385). <http://www.who.int/mediacentre/factsheets/fs385/en/>. Accessed 10 Jan 2017.

# Which one has more influence in female air pistol performance: experience or training?

Mon, Daniel\* and Díaz, Arturo\*\*

\* Universidad Politécnica de Madrid. Facultad de Ciencias de la Actividad Física y del Deporte. Spain  
danielmonl@gmail.com

\*\* Facultad de Ciencias del Deporte. Universidad de Murcia. ardi@um.es

## Abstract

**Introduction:** Many factors affect shooting performance. There seem to be a consensus in which the experience and the training time determine the performance in shooting sport. Unfortunately is not completely clear which one is more important, experience or training, especially in women. The objective of this study is to analyze the training and experience influence on the Olympic female air pistol performance. **Methods:** Twenty three female shooters who participated in a Spanish air pistol championship filled a questionnaire about her shooting experience in years and how much they trained every week in hours per week before the competition. The data obtained were correlated with the official competition performance. **Results:** Significant correlations were found between the participants' training and the performance  $R^2=0.17$ . **Discussion:** As previous studies have showed, both training time and shooting experience determine performance. Our data are in concordance with the previous studies. The training time seems to have a more direct influence in the performance than the experience. On the other hand, experience seems to affect performance when is combined with the body balance.

## 1 Introduction

Experience seems to be a factor affecting performance in shooting sports. However the level of influence in precision shooting sport is different depending on the authors. Values between  $r=0.19$  and  $r=0.72$  were previously obtained.

Last studies show that experience and training were related to the performance with values that oscillated between  $r^2=0.12$  and  $r^2=0.19$  respectively. Furthermore these results were equivalent to variation coefficients of 2.24% and 2.34% (Mon et al., 2015; Mon et al., 2014).

The experience plays an important role in the cognitive area of shooting. Many factors can determine the performance. The theoretical knowledge of the shooting aspects could have some kind of influence reducing the number of shoots with zero values (Chung et al., 2004).

Other studies explain that the experience is important to make the properly adjustments on the sights and on the eyes vision. The most experienced shooters are able to use the vestibular and kinesthetic information in a better way, taking an advantage to their less experienced sport mates (Aalto et al., 1990).

Regarding the focused time vision and quiet eye availability time, something similar happens in both modalities, clay target shooting and precision shooting. The experienced shooters are able to

Experience or training?

Mon, Daniel and Díaz, Arturo

maintain a focused vision for longer times than their less level mates, which seems to be an obviously advantage for shooting. Furthermore, this not only happens in standard situations, but also in situations with disturbing elements (Di Russo et al., 2003).

The performance can be modified in more ways by the experience. The physical components of the shot are affected by the experience too. One important factor is the cleanness pressure of the trigger. Experienced shooters could make a more uniform pressure on the trigger avoiding unnecessary movements which could be transmitted to the barrel of the gun and therefore decreasing the performance. Not only this factor is related with the experience, the body balance has some kind of relation too. There seem to be consensus in which the experienced shooters have less body sway movements than less experienced shooters (Goonetilleke et al., 2009).

Finally, the experience determined the anxiety level of the shooter during the competition. Moreover the self-confidence parameter can be modified as the shooter improves his competition experience (Chung et al., 2004).

On the other hand, the training positive effects on performance are widely accepted by the scientific community. For example, according to Gulbinskienė and Skarbalius (2009), the Lithuanians shooters have less performance than shooters from other countries due the less training volume.

The effectively training time is an important element in military people. There seem to be differences in the total score in the pre-test previous to the training and the post-test made after the training period (Chung et al., 2004).

Other professions which involve the use of weapons, like policemen, could improve the shooting performance with training too. Both, men and women improved their performance after a pistol training program time while they were on the police academy. Women improved up to a 136% and men up to a 45%, reducing the differences between genders at the end of the program (Vučković et al., 2008).

Some studies affirmed that visual training programs improves the skills of the eye, but this positive progress didn't have any repercussion on the pistol performance (Quevedo et al., 1999).

Other studies as well as confirming the effects of the training, explain how to train. Some of them explain that training should be individual, others talk about the benefits of combined training with relaxing exercise programs and some authors talk about the necessity of specific balance training exercises (Puglisi and La rocca, 2014).

As the feedback information is necessary in the training process, some authors explain how this feedback could be used. Feedback in real time, kinesthetic feedback and auditory feedback have been shown as positive on the performance (Mullineaux et al., 2012).

The main objective of this study is to find out what variable has more influence on the air pistol performance in female shooters: the training time or the shooting experience.

## 2 Method

The study was performed during an official competition. 23 female shooters who participated in a Spanish air pistol championship filled a questionnaire about their shooting experience in years and how much they trained every week in hours per week before the competition. The participation on the study was voluntary and open for all the participants.

The data obtained were correlated with the official competition performance. The performance was measured by electronic targets Sius Ascor. The competition was according to the ISSF rules and



Experience or training?

Mon, Daniel and Díaz, Arturo

regulations in every moment. The control of the competition, the pistols and the shooting clothes were done by the official Spanish shooting referees.

A force platform was used to measure the body balance during the official trainings too. The balance conditions were the same that in competition.

In accordance with the rules and regulations of the Spanish Royal Federation of Olympic Shooting eligibility to participate required a minimum of 320 points in sport pistol shooting in previous competitions (RFEDETO, 2012). None of the participants suffered injuries during the year previous to the competition.

## 2.1 Material and Procedure

A demographic data questionnaire was used, including age, hours of weekly training, suffered injuries, experience (in years), degree of importance of competition, perceived general status and perceived physical condition. The shooters completed the questionnaire the same day that they confirmed their inscription to the competition.

The experimental design of the present study was approved by the Spanish Royal Shooting Federation as well as the CSD (Spanish High Council for Sports). The study was performed following the guidelines of the Declaration of Helsinki, last modified in 2008. The authors certify that the present work was carried out in the absence of any financial, personal or other relationships with other people or organizations within three years of beginning the submitted work that could inappropriately influence, or be perceived to influence, the presented work and lead to a potential conflict of interest.

All analysis was implemented by use of SPSS version 17. The normal distribution of the variables was tested via implementation of the Kolmogorov-Smirnov test. For the analysis of the parameters that affected performance linear regressions and Pearson correlations were calculated. The level of significance was set to 0.05.

## 3 Results

The statistical analysis revealed for all the shooters that the overall performance was positively affected by the number of training hours ( $F_{1,21}=5,43$ ;  $p<0,05$ ), Significant correlations were found between the participants' training and the performance  $r^2=0.17$ . The linear regression equation predicting performance was the following:  $\text{Performance} = 9.37 + 0.04 * \text{Experience} - 0.006 * \text{Body sway}$ ,  $r^2=0.79$ .

## 4 Discussion

The scientific bibliography has shown that training is important in every sport to reach an optimal performance. Unfortunately there are big gaps between the different studies. In some studies the performance during the training period grew up a lot (136% on women) (Vučković et al., 2008), however our result are far away from this percentage. Our data give us a result of 17%. This percentage which explains the performance variance is obviously shorter than the results from Vučković et al. (2008). These differences could be due to the different participants. In our study the shooters were only athletes while in the other study they were cops. However our results are in accordance with Smith and Hagman (2000). Both studies showed similar results with a correlation between performance and training time of  $r^2=0.17$ .

Experience or training?

Mon, Daniel and Díaz, Arturo

Regarding to the influence of the experience on the performance, there are differences between the studies. Thereby, we were able to find values between  $r=0.72$  and  $r=0.19$ . Our data didn't show any correlation between experience and performance on women. Furthermore, the age plays a negative role on the performance; therefore, our results are not in accordance with the previous authors when we analyzed each variable independently. The differences in the results could be due the differences on the method and on the participants. While we have sport shooters other studies have military and police people. Another important thing that could explain these differences is that we made the study under actual competition conditions while the others made it under training conditions.

The experience didn't have influence on the performance when we analyzed the variables separately, but in contrast it was really important when we made the linear regression with all the variables together, as we can see on the performance equation in the results section. This can be interpreted as follow: the profile of an elite shooter could be someone who has big experience and lots of training but who is not very old, as the age has a negative influence on the performance.

Our results are in accordance with previous studies as the balance could affect the performance (Mon et al., 2014). Our data suggest that the combination of balance and experience in women could determine the performance equation. This could be explained by the fact that elite shooters in our data were relative young, with lots of experience and with many hours of training per week. Therefore this type of shooters could improve their balance during many years of training having an advantage with the other athlete's types.

## 5 Conclusion

The study concludes with the following: performance is related with both experience and training time, being the last one more important if the different variables are analyzed independently.

## 6 Acknowledgements

Many thanks for the support of the Sports Faculty (Murcia University).

## 7 References

- Aalto H, Pyykko I, Ilmarinen R, et al. (1990) Postural stability in shooters. *ORL J. Otorhinolaryngol. Relat. Spec.* 52: 232-238.
- Chung G, Cruz G, Vries L, et al. (2004) Determinants of Rifle Marksmanship Performance: Predicting Shooting Performance with Advanced Distributed Learning Assessments. DTIC Document.
- Di Russo F, Pitzalis S and Spinelli D. (2003) Fixation stability and saccadic latency in elite shooters. *Vision Res* 43: 1837-1845.
- Goonetilleke RS, Hoffmann ER and Lau WC. (2009) Pistol shooting accuracy as dependent on experience, eyes being opened and available viewing time. *Applied Ergonomics* 40: 500-508.
- Gulbinskienė V and Skarbalius A. (2009) Peculiarities of investigated characteristics of lithuanian pistol and rifle shooters' training and sport performance. *UGDYMAS KŪNO KULTŪRA*: 21.
- Mon D, Zakynthinaki MS, Cordente CA, et al. (2015) Finger Flexor Force Influences Performance in Senior Male Air Pistol Olympic Shooting. *PLoS ONE* 10: e0129862.
- Mon D, Zakynthinaki MS, Cordente CA, et al. (2014) Validation of a Dumbbell Body Sway Test in Olympic Air Pistol Shooting. *PLoS ONE* 9: e96106.
- Mullineaux DR, Underwood SM, Shapiro R, et al. (2012) Real-time biomechanical biofeedback effects on top-level rifle shooters. *Applied Ergonomics* 43: 109-114.

Experience or training?

Mon, Daniel and Díaz, Arturo

- Puglisi ML and La rocca R. (2014) Evaluation of postural balance in skeet shooting. *International Journal of Education and Research* 2.
- Quevedo L, Solé J, Palmi J, et al. (1999) Experimental study of visual training effects in shooting initiation. *Clinical and Experimental Optometry* 82: 23-28.
- RFEDETO. (2012) *Reglamento Técnico General para todas las Modalidades de Tiro*, Madrid: Real Federación Española de Tiro Olímpico.
- Smith MD and Hagman JD. (2000) Predicting rifle and pistol marksmanship performance with the Laser Marksmanship Training System. DTIC Document.
- Vučković G, Dopsaj M, Radovanović R, et al. (2008) Characteristics of shooting efficiency during a basic shooting training program involving police officers of both sexes. *Facta universitatis-series: Physical Education and Sport* 6: 147-157.

Experience or training?

Mon, Daniel and Díaz, Arturo

- Puglisi ML and La rocca R. (2014) Evaluation of postural balance in skeet shooting. *International Journal of Education and Research* 2.
- Quevedo L, Solé J, Palmi J, et al. (1999) Experimental study of visual training effects in shooting initiation. *Clinical and Experimental Optometry* 82: 23-28.
- RFEDETO. (2012) *Reglamento Técnico General para todas las Modalidades de Tiro*, Madrid: Real Federación Española de Tiro Olímpico.
- Smith MD and Hagman JD. (2000) Predicting rifle and pistol marksmanship performance with the Laser Marksmanship Training System. DTIC Document.
- Vučković G, Dopsaj M, Radovanović R, et al. (2008) Characteristics of shooting efficiency during a basic shooting training program involving police officers of both sexes. *Facta universitatis-series: Physical Education and Sport* 6: 147-157.

# The Application of Hurdle Models to Accurately Model 0-0 Draws in Predictive Models of Football Match Outcomes

A. Owen\*

\*De Montfort University, Leicester, UK. email address: alun.owen@dmu.ac.uk (until summer 2017)  
Coventry University, UK (from summer 2017)

## Abstract

A novel application of so-called hurdle models is presented in the context of association football match prediction models, which has the flexibility of being able to capture either the deflation or inflation of the probability of 0-0 scores. The deployment of the model in real time on week by week basis to develop one-round ahead predicted forecasts is discussed in the context of a Bayesian framework. The problem of forming appropriate informative prior distributions for parameters that need to satisfy sum to zero constraints is also developed. Using match results data from the Scottish Premier League, the use of the hurdle model approach is shown to improve the predictive accuracy of the probability of 0-0 scores and can lead to a potentially profitable betting strategy when used for betting on the 0-0 score market.

## 1 Introduction

Statistical models of association football match outcomes in the published research literature are often based on modelling the number of goals scored by the home and away teams, with much of this work focused on use of the Poisson or similar distributional assumptions. Examples include the much cited work of Maher (1982) and Dixon and Coles (1997), but an alternative bivariate poisson model is presented in Karlis and Ntzoufras (2003), and dynamic models are presented in Owen (2011) and Koopman and Lit (2015). More recently Boshnakov et. al. (2016) demonstrate improvements in model fit with the use of a Weibull based count model. Much of this work has reported the tendency for the models to under-estimate the probability of 0-0 scores, and solutions provided in the form of modifications to the model to inflate the model's probability for the 0-0 score.

However, we show in Section 2 that there are situations where these models actually over-estimate the probability of a 0-0 score (rather than under-estimating them) and so in fact deflation of the 0-0 score line is required (rather than inflation).

An alternative novel approach to modelling 0-0 scores is therefore proposed in Section 3, and the deployment of the model in real time on week by week basis to develop one-round ahead predicted forecasts is then presented in the context of a Bayesian framework in Section 4. The results of using this proposed model for betting on the 0-0 score markets are then summarised in Section 5 with an overall discussion in Section 6.

## 2 A typical Poisson model and problems with zero inflation

### 2.1 A Poisson model

The underlying model considered here is a simple Poisson goals model, which forms the basis of much of the goals based models in football within the existing literature. This can be described by denoting the number of goals scored by team  $i$  playing at home and team  $j$  playing away as  $X_{ij}$  and  $Y_{ij}$  respectively, and for a league with  $n$  teams, assuming that  $X_{ij}$  and  $Y_{ij}$  ( $i, j = 1, \dots, n; i \neq j$ ) can be modelled as (conditionally independent) Poisson random variables as follows:

$$X_{ij} \sim Po(\mu_{ij}), \quad Y_{ij} \sim Po(\lambda_{ij}). \quad (1)$$

The  $\mu_{ij}$  and  $\lambda_{ij}$  represent mean scoring rates for the home and away teams respectively, and are assumed to depend on (latent) attacking abilities of the two teams,  $a_i$  and  $a_j$ , and defensive abilities of the two teams,  $b_i$  and  $b_j$ . Typically these dependencies are specified via the log-linear predictors:

$$\log(\mu_{ij}) = a_i + b_j + g_H, \quad \log(\lambda_{ij}) = a_j + b_i + g_A, \quad (2)$$

where  $g_H$  and  $g_A$  reflect the underlying mean log-scoring rates at home and away respectively. Note that the log-linear predictors in (2) are also often stated equivalently as:

$$\mu_{ij} = \alpha_i \times \beta_j \times \gamma_H, \quad \lambda_{ij} = \alpha_j \times \beta_i \times \gamma_A. \quad (3)$$

In this parameterisation, attacking teams have larger attack parameter estimates, whilst better defending teams have lower defence parameter estimates. The parameters  $\gamma_H$  and  $\gamma_A$  in (3) reflect the underlying (geometric) mean scoring rates at home and away respectively, and  $\gamma_H/\gamma_A$  represents a home effect.

### 2.2 Problems with zero inflation

The main problem with inflating 0-0 scores that has hitherto been considered in the published literature, is that in some leagues the underlying model(s) used often actually over-estimate the probability of a 0-0 score (rather than under-estimating them), and so in fact deflation of the 0-0 score is required (rather than inflation). As an example, Figure 1 shows, for the Scottish Premier League (SPL) from 1994/95 to 2011/12, using the model defined in Section 2.1, the ratio of the mean 0-0 score probabilities (across all matches played each season) divided by the observed proportion of games that ended 0-0. This illustrates, in the SPL over that period, the model (defined in Section 2.1) produces 0-0 scores probabilities that often over-estimate the true proportion of games that end 0-0.

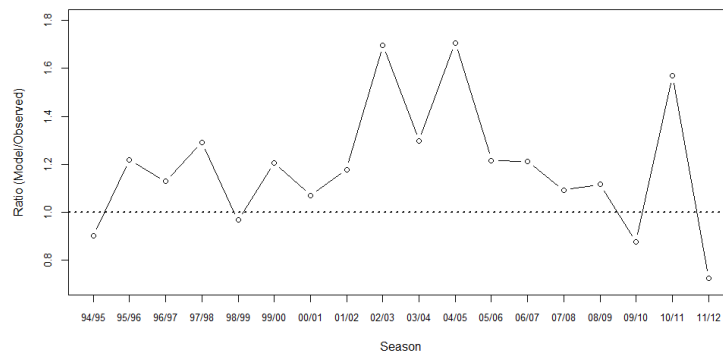


Figure 1. Ratio of mean 0-0 score probabilities divided by observed proportion of games ending 0-0.

None of the models in the existing literature facilitates deflation of the 0-0 score apart from the modified Poisson model of Dixon and Coles (1997). However, whilst their modification can actually facilitate deflation of the 0-0 score, in this case this would also require deflation of the 1-1 score at the same time, which is often not required. This paper therefore proposes an alternative novel approach to inflation of the 0-0 score probabilities which are based on the use of hurdle models which allows score inflation or deflation, or indeed no adjustments, whichever is best supported by the match data used in the model.

### 3 Hurdle models for 0-0 score probability deflation/inflation

#### 3.1 Overview

Hurdle models were first discussed in Mullahy (1986) but a good summary of these models can be found in Winkelmann (2008). These are in essence two-stage models, where in our case, the occurrence of 0-0 scores would be modelled as a Bernoulli variable as a first stage (the “hurdle”), whilst a truncated Poisson distribution is then used to model all other scores as a second stage. The first stage attempts to model the occurrence of 0-0 draws, whilst the second stage models all other scores, conditional on at least one goal is scored by one of the two teams involved in a match. This would seem to be an intuitive idea since association football matches tend to start with both teams more inclined to focus on not conceding a goal, and only focusing more resources in attack if they do indeed concede. Dixon and Robinson (1998) provides evidence to support this assertion by showing that scoring rates for teams increase once they concede a goal.

The basic (independent) Poisson probability model given in Section 2.1 can be stated as:

$$P(X_{ij} = x, Y_{ij} = y) = \frac{\mu_{ij}^x \lambda_{ij}^y \exp^{-(\mu_{ij} + \lambda_{ij})}}{x!y!}, \quad (4)$$

which can be adapted into a two-stage hurdle model, such that the first stage models the occurrence of 0-0 draws as a Bernoulli random variable with probability  $p$ . The second stage of the hurdle model then models all scores other than 0-0 as a truncated Poisson, conditional on the probability that at least one goal is scored in the match. The probability model for the hurdle model in this case is:

$$P(X_{ij} = x, Y_{ij} = y) = \begin{cases} p & x = y = 0 \\ (1 - p) \times \frac{\mu_{ij}^x \lambda_{ij}^y \exp^{-(\mu_{ij} + \lambda_{ij})}}{x!y!} & \text{otherwise.} \end{cases} \quad (5)$$

Considering the fact that in the underlying Poisson model, the probability of a 0-0 draw is given by  $\exp^{-(\mu_{ij} + \lambda_{ij})}$ , this suggests that a reasonable approach to modelling  $p$  might be to define this as:

$$p = c \times \exp^{-(\mu_{ij} + \lambda_{ij})}, \quad (6)$$

for some constant  $c$  ( $c > 0$ ) to be estimated from the model. If  $c=1$  the model is equivalent to the simple Poisson model of Section 2.1, whereas for  $c < 1$  the probability of a 0-0 draw is deflated, whilst for  $c > 1$  the probability of a 0-0 draw would be inflated.

The contribution to the likelihood (denoted as  $L_{ij}$ ) from a match involving teams  $i$  and  $j$ , where the observed goals are  $X_{ij} = x$  and  $Y_{ij} = y$ , would then be defined as follows:

$$L_{ij}(\mu_{ij}, \lambda_{ij}, c | X_{ij} = x, Y_{ij} = y) = \begin{cases} c \times \exp^{-(\mu_{ij} + \lambda_{ij})} & x = y = 0 \\ \left(1 - c \times \exp^{-(\mu_{ij} + \lambda_{ij})}\right) \times \frac{\mu_{ij}^x \lambda_{ij}^y \exp^{-(\mu_{ij} + \lambda_{ij})}}{x!y!} & \text{otherwise.} \end{cases} \quad (7)$$

Note that since  $0 \leq p \leq 1$ , then  $c$  must be constrained so that  $0 \leq c \leq \exp^{(\mu_{ij} + \lambda_{ij})} \forall i, j$ .

### 3.2 Match data and in-sample model fit

The model can be fitted by simply defining the full likelihood as the product of the terms shown in (7) across all matches of interest and minimising the full (log) likelihood using any suitable optimiser. However, we utilise a Bayesian framework with suitably non-informative priors for the model's parameters, making use of an MCMC approach via the OpenBUGS software, see OpenBUGS(2017). The key reason for this approach is that a Bayesian framework provides a very flexible framework when deploying the model for real to obtain one-match ahead predictions, since in that case information from past seasons is required to be incorporated into the estimation of parameters for the current season of interest. Note however, that in order to fit the full likelihood defined by (7) using OpenBUGS, the so called “zeros-trick” is required. Details of the zeros-trick” is given in Lunn et. al. (2013). We argue the case further for the Bayesian framework again in Section 4.

Data from the SPL is available electronically from [www.football-data.co.uk](http://www.football-data.co.uk) over the period from 1994/1995 to the present, although the only data we require are the names of the home and away teams and the goals scored by each team in each match. Data from 1994/95 through to 2011/12 was used for initial exploratory model fitting which is reported on in this section. The remaining five seasons from 2012/13 to 2016/17 are used later in in Section 5, to assess the model's out of sample performance when deploying the model for real to obtain one-match ahead predictions.

The model was fitted to each complete season's match results separately (using non-informative priors for each season) from 1994/95 through to 2011/12. This covered a total of 3,816 matches, each providing an in-sample model-fitted probability for the 0-0 score. As well as monitoring the estimate for  $c$  each season, it is also easy using OpenBUGS to monitor the lower 2.5% and upper 97.5% limits of the posterior distribution for  $c$ , which are shown in Figure 2. This illustrates (not surprisingly) that over this period, the estimate of  $c$  is below 1 for each season where Figure 1 suggested deflation of the 0-0 score was required. However, the lower 2.5% and upper 97.5% limits highlight the large standard error for this estimate when data from just a single season is used, although in some seasons (e.g. 2002/03, 2004/05 and 2010/11), the evidence for deflating the 0-0 score probability is quite clear.

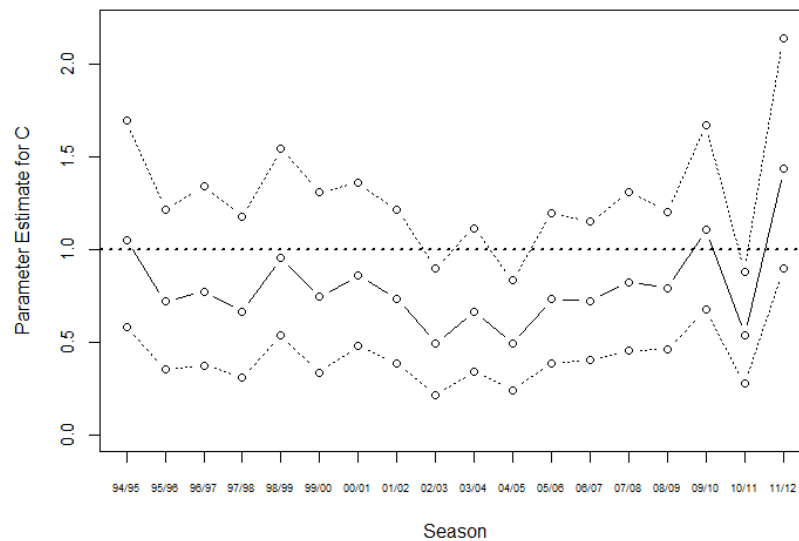


Figure 2. Estimates of  $c$  (—) with lower 2.5% and upper 97.5% posterior limits (-----).



With regard to assessments of model fit, since the underlying Poisson model has been much reported on previously, we focus here on the calibration of the Hurdle Model with respect to 0-0 scores. The in-sample 0-0 score probabilities, from the 3,816 matches using the Hurdle model, were divided into class-intervals based on the probability ranges (bins) of 0 to  $<0.015$ , 0.015 to  $<0.025$ , 0.025 to  $<0.035$ , etc. The observed proportion of matches in each interval that actually ended 0-0 were then examined, such that a better calibrated model should have observed proportions closer to the mid-point of these intervals. The class interval mid-points were taken to be 0.01, 0.02, 0.03 etc. Figure 3 plots the resulting observed proportions of matches ending 0-0 against these class interval midpoints for the hurdle model (plotted as +). For comparison, Figure 3 also shows the equivalent results for the model of Section 2.1 where  $c$  is constrained to be 1 (plotted as o), which is referred to here as the Non-Hurdle model. This illustrates the improved calibration with respect to modelling 0-0 score probabilities for the Hurdle model (+), compared to the Non-Hurdle model (o). This also suggests that the over-estimation of the 0-0 score probabilities with the Non-Hurdle model tends to occur across almost the full range of model-fitted probabilities. Note that the number of matches in each probability interval up to and including 0.105 to  $<0.115$  was almost always well above 100, but for probability intervals above that range the number of matches was always well below 100, which explains the greater variability in the calibration plot for the higher probability intervals.

The overall mean 0-0 score probability with the Hurdle model is 0.065 which compares better with the observed proportion of 0.069 (Table 1) than the overall mean 0-0 score probability of 0.078 with the Non-Hurdle model.

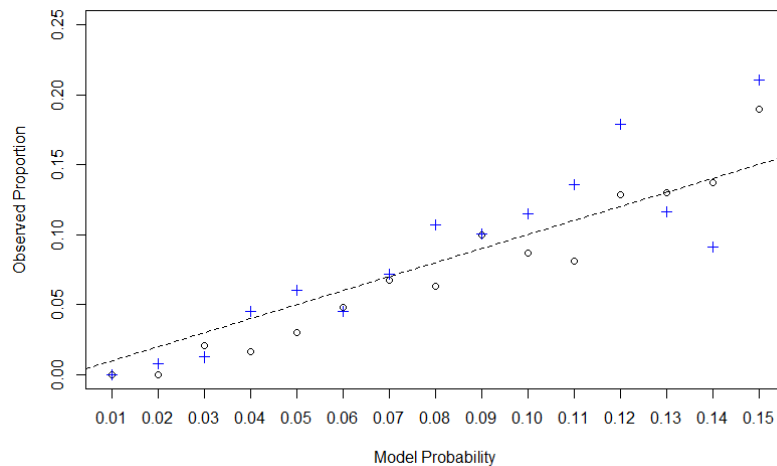


Figure 3. Calibration plot for the Hurdle model (+) and Non-Hurdle model (o).

Model comparison in the Bayesian framework also often includes assessments of the Deviance Information Criterion (DIC). However, this is less useful in this case since the Hurdle model differs from the Non-Hurdle model mainly only in terms of modelling the probability of both teams scoring zero goals. Hence the difference in DIC between the two models is as expected very small and was always between -4 and +3 (Hurdle DIC minus Non-Hurdle DIC). There was a positive difference in the DIC in favour of the Hurdle model in 11 seasons, a close to zero difference in 3 seasons and a negative difference in favour of the Non-Hurdle model in just 4 seasons. This does provide further support for the Hurdle model, but differences in DIC of less than 5 are often reported as negligible.

An alternative and more appropriate tool for comparing the two models in this case is to make use of Bayes Factors, which are the posterior odds (given match data  $\{X,Y\}$ ) that the Hurdle model ( $M_1$ ) is a better model than the Non- Hurdle model ( $M_0$ ). The Bayes Factor in this case can be expressed as:

$$\frac{p(M_1|\{X,Y\})}{p(M_0|\{X,Y\})} = \frac{p(M_1)}{p(M_0)} \times \frac{p(\{X,Y\}|M_1)}{p(\{X,Y\}|M_0)}, \quad (8)$$

which equates to the ratio of the prior odds in favour of  $M_1$  multiplied by the likelihood ratio.

An assessment of the Bayes Factors for each season from 1994/95 to 2011/12 was undertaken using prior odds of 1, so that the results indicate which of the two models the likelihood provides more support for. The resulting Bayes Factors are plotted in Figure 4 which shows that although there is sometimes little to choose between the two models, there are a significant number of seasons when the Hurdle model is much preferred over the Non-Hurdle model. This is most notable in the same three seasons identified earlier where there was clear evidence to support deflation of the 0-0 score probability.

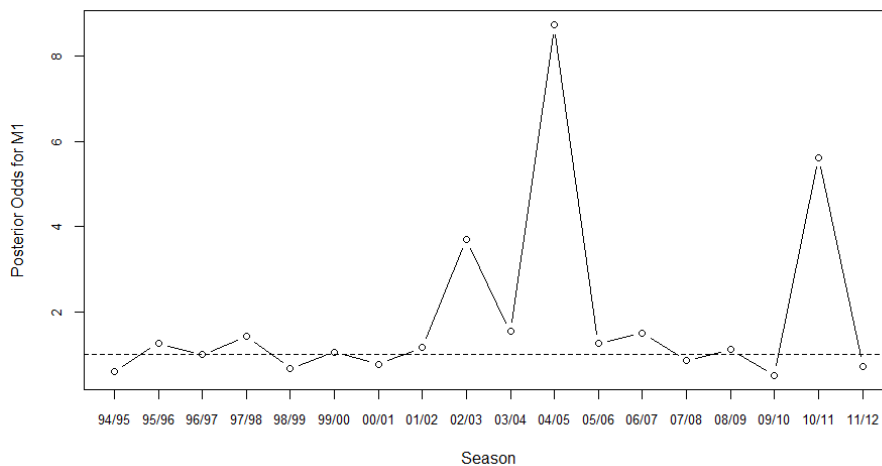


Figure 4. Bayes Factors for  $\frac{p(M_1|\{X,Y\})}{p(M_0|\{X,Y\})}$  in support of the Hurdle Model

#### 4 A Bayesian framework for the deployment of the model for future match prediction

The occurrence of 0-0 scores only occurred in 7% of the matches on average in the SPL during the period 1994/95 to 2011/12. Modelling and ultimately estimating  $c$  can therefore be subject to a relatively large standard error as indicated earlier (Figure 2). When deploying the model for real we would therefore want to incorporate data from a large number of previous seasons, i.e. all seasons for which we have data we believe is relevant. However, when modelling and estimating the team-based (log) attack and defence abilities  $a_i$  and  $b_i$ , typically we would want to include only the last 2 or 3 seasons of match results data to estimate these, since data from much further in the past is often considered to be much less relevant to teams' future performance. This difference in the amount of past data utilized in the model is easily facilitated in a Bayesian framework via the use of suitable informative prior distributions for the various parameters to be estimated.

For the parameter  $c$  we specify a gamma prior as:

$$c \sim Ga(r_c, m_c), \tag{9}$$

where  $r_c$  and  $m_c$  are initially estimated by examining the match results data over the full period from 1994/95 to 2011/12. Values for  $r_c$  and  $m_c$  in the priors for 2012/2013 onwards can then be based on the properties of the posterior distribution for  $c$  at the end of the relevant previous season on a rolling basis.

To complete the specification of the hurdle model in the Bayesian framework we define normal priors for the log-attack and log-defence parameters,  $a_i$  and  $b_i$ , and gamma priors for  $\gamma_H$  and  $\gamma_A$ , as follows:

$$a_i \sim N(m_{a_i}, \sigma^2), \quad b_i \sim N(m_{b_i}, \sigma^2), \tag{10}$$

$$\gamma_H \sim Ga(r_{\gamma_H}, m_{\gamma_H}), \quad \gamma_A \sim Ga(r_{\gamma_A}, m_{\gamma_A}). \tag{11}$$

Note that  $\sigma^2$  represents the prior variance, which for simplicity is also assumed to be common to all teams and to both the log-attack and log-defence parameters. Here the value for  $\sigma^2$  is taken to be 0.02 (precision = 50) since this was reported in Owen (2011) as being of a suitable magnitude for optimising one-match ahead predictive performance in the SPL. The values for  $m_{a_i}$ ,  $m_{b_i}$ ,  $r_{\gamma_H}$ ,  $m_{\gamma_H}$ ,  $r_{\gamma_A}$  and  $m_{\gamma_A}$  in the priors specified in (10) and (11) can be derived by fitting the model to a suitable number of previous seasons (2 or 3?), and using the final parameter estimates as the basis for forming suitable priors for the next season. Priors for the team-based log-attacking abilities  $a_i$  and log-defensive abilities  $b_i$  for the promoted teams are an obvious issue to consider, and here we based these on the mean of the end of season estimates for the teams that had been promoted in the past since 1994/95.

Finally, in order to ensure unique identifiability of the parameters, two constraints are required, and although other alternatives are also possible, here we use the following constraints:

$$\sum_{i=1}^n a_i = 0, \quad \sum_{i=1}^n b_i = 0. \tag{12}$$

The use of the constraints given by (12) imply that the attack and defence parameters for each team are given relative to an average team (one which has mean log-attack and mean log-defence parameters of 0). One problem this presents however, is that informative prior distributions for these ‘‘constrained’’ team-based attacking and defensive abilities can only be specified for  $n-1$  of the  $n$  teams, with the priors for the  $n^{\text{th}}$  team being derived via equation (12). Specifying appropriate informative priors in this situation is therefore problematic, but a solution to this problem was presented in Owen (2011) as follows:

For the attack parameters we define:

$$\mathbf{m}_\alpha = [m_{\alpha_1}, m_{\alpha_2}, \dots, m_{\alpha_n}]^T. \tag{13}$$

The priors for the attack parameters in (10) can then be expressed as  $\boldsymbol{\alpha} \sim N(\mathbf{m}_\alpha, \mathbf{W})$ , where  $\mathbf{W}$  is a diagonal prior variance matrix with entries  $\sigma^2$  and zeros on the off-diagonal. It can then be shown that the identifiability constraint (12) will hold if  $\mathbf{1}_n^T \mathbf{m}_\alpha = 0$  where  $\mathbf{1}_n$  is the  $n \times 1$  matrix such that  $\mathbf{1}_n^T = [1, 1, \dots, 1]$ , and if the prior variance matrix  $\mathbf{W}$  is modified to the variance-covariance matrix  $\mathbf{R}$  as follows:

$$\mathbf{R} = \frac{n\sigma^2}{(n-1)} \left( \mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \right), \tag{14}$$

where  $\mathbf{I}_n$  is the  $n \times n$  identity matrix. Note that the multiplying factor  $n/(n-1)$  is incorporated so that the variances on the diagonals reflect the variances that are originally specified.

However, the variance-covariance structure given by (14) then presents a new problem, since  $\mathbf{R}$  is not of full rank and hence has no inverse. OpenBUGS cannot be used directly to sample from a multivariate normal prior with this variance-covariance structure due to the lack of a suitable inverse. This problem can be overcome by sampling from suitable univariate conditional distributions, but this may result in a loss of efficiency. A more efficient approach again presented in Owen (2011), is to sample values for a  $(n-1) \times 1$  vector of unconstrained parameters  $\mathbf{c} = [\theta_1, \theta_2, \dots, \theta_{n-1}]^T$ , from a normal distribution with zero mean and variance-covariance matrix  $\mathbf{S}$  given by:

$$\mathbf{S} = \frac{n\sigma^2}{(n-1)} (\mathbf{I}_{n-1} + \mathbf{1}_{n-1}\mathbf{1}_{n-1}^T) \quad (15)$$

If a new vector of parameters  $\mathbf{u}$  is then calculated as  $\mathbf{u} = \mathbf{J}\mathbf{c}^*$ , where  $\mathbf{c}^* = [\theta_1, \theta_2, \dots, \theta_{n-1}, 0]^T$  and

$$\mathbf{J} = \begin{pmatrix} \mathbf{I}_{n-1} - \frac{1}{n}\mathbf{1}_{n-1}\mathbf{1}_{n-1}^T & \frac{1}{n}\mathbf{1}_{n-1} \\ -\frac{1}{n}\mathbf{1}_{n-1}^T & \frac{1}{n} \end{pmatrix}, \quad (16)$$

it can be shown that the resulting values of  $\boldsymbol{\alpha} = \mathbf{m}_\alpha + \mathbf{u}$  represent the required sampled values of the attack parameters with the required evolution structure and variance-covariance structure given by (14), and with the identifiability constraint (12) holding. A similar approach to that described above was also applied to the defence parameters, but is not described here for brevity.

## 5 Results of future match predictions used for betting

The model described in Section 3, using the framework developed in Section 4, was retrospectively deployed on a round by round basis (where a round represents the set of weekend or mid-week fixtures) to the SPL during the five seasons from 2012/13 to 2016/17. Note that the 2016/17 only included data on matches up to and including 16<sup>th</sup> April 2017 that was available at the time of writing. This was implemented using OpenBUGS with code written in R (R Core Team, 2015) and calling OpenBUGS from within R when required using the R2OpenBUGS package. The sampled values typically displayed very good mixing behaviour, and running 2 chains of the sampler for 10,000 iterations, with the first 3,000 iterations being discarded as a burn-in, were assessed as more than adequate for estimation purposes. Out of sample predictive estimates of the one-round ahead 0-0 score probabilities for future matches, were derived by monitoring the proportion of samples where the home and away goals totalled zero, with missing values recorded as the goals data for those matches.

Note that the model can only be deployed to predict the outcomes of matches from Round 2 onwards in each season due to the need for data from at least the first round of matches to form the likelihood component of the Bayesian model. Clearly model fit performance will increase quite rapidly as each new round of match data becomes available and so we examine the one round-ahead predictive performance of 0-0 scores during the second half of each season. Examining these match results over the second half of five seasons gave a total of 523 matches of which 37 resulted in a 0-0 score. Our interest in this section relates to the performance of the model over these matches with respect to its performance in the betting markets related to the 0-0 score. There are two main markets where the model presented is of potential relevance; betting on the 0-0 score directly and betting that

the total number of goals will be under 0.5. Here we concentrate on the actual 0-0 score market and make use of historical odds for this market derived from [www.oddsportal.com](http://www.oddsportal.com). Since the odds reported on that site are UK odds which do not include the stake, an estimate of the market's implied probability of a 0-0 score is given by  $1/(1+\text{average odds})$ . Note however that this will be a slight over-estimate of the implied probability due to the over-round present in bookmakers' odds.

We examine the potential return if a betting rule is used so that a unit stake is bet on the 0-0 score if  $p_M > m \times p_B$ , where  $p_M$  is the one-round ahead model predicted probability,  $m$  is some margin (typically  $>1$ ) reflecting a potential advantage the model has over the market odds and  $p_B$  is the market implied probability. Figure 5 shows the Return On Investment (ROI) from betting unit stakes using this betting rule against a range of margins from 0.95 to 1.15. The ROI is simply the total profit/total stakes placed and allows comparisons with other staking strategies (not considered here). The fact that a positive return is realised for margins a little under 1.00 is symptomatic of the over-estimate of the market's implied probabilities discussed earlier. As an illustration of the potential for a positive return, betting whenever  $p_M > 1.1 \times p_B$  would have yielded a ROI of around 44% across a total of 86 bets. This is an unusually large return in these types of markets, but this needs to be tempered with the fact that the example is based on just 86 bets. Indeed, Figure 5 also shows bootstrapped (10% and 90%) percentiles for the estimated sampling distribution of the ROI, which indicates that whilst there is strong evidence of positive returns, there is still some uncertainty that this was not just by chance. However this does highlight the potential advantage the hurdle model provides over the information currently reflected in the market odds.

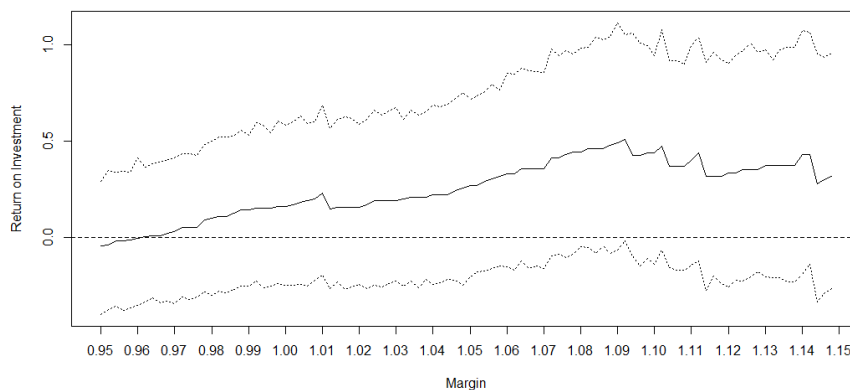


Figure 5. Return on Investment for a range of margins ( $m$ ) when betting where  $p_M > m \times p_B$

## 6 Discussion

A novel approach to improving the accuracy of predictive probabilities of 0-0 scores in association football has been presented. This is a simple modification to a basic Poisson goals model, much reported on in the published literature and which forms the basis for models in common use by many sports modelling practitioners, bookmakers and professional bettors. Since hurdle models can be used as a means of either increasing or decreasing the number of observed zeros in a Poisson model, they are suitable for cases where inflation or deflation of the 0-0 score, or indeed no adjustment, is required. No published work is evident in relation to the application of hurdle models in the context of

football outcomes and so the work presented here would represent a novel application of this type of model. In fact, most of the published work in relation to the application of hurdle models appears to concentrate on modelling zero-inflation, as opposed to zero-deflation. It has been demonstrated that the approach using the example of the SPL leads to improved calibration of the model fitted probabilities for 0-0 scores and ultimately to a potentially profitable betting strategy when used for betting on the 0-0 score market. The deployment of the model in real time on week by week basis to develop one-round ahead predicted forecasts was discussed in the context of a Bayesian framework, and a solution to handling the problem of forming appropriate informative prior distributions for parameters that need to satisfy sum to zero constraints was also presented. The approach presented can easily be adapted to allow a hurdle component for 0-0 scores to be incorporated into other underlying models that are different to the simple Poisson considered here. The work presented is very much work in progress and so further planned work includes examining this approach in other models such as the dynamic Poisson model of Owen (2011), the bivariate Poisson model of Karlis and Ntzoufras (2003) and the Weibull count model of Boshnakov et. al. (2016).

## References

- BOSHNAKOV G., KHARAT T. & MCHALE I.G. (2017) A bivariate Weibull count model for forecasting association football scores. *International Journal of Forecasting*, **33**(2), 458-466. <https://doi.org/10.1016/j.ijforecast.2016.11.006>
- DIXON M.J. & COLES S.G. (1997) Modelling association football scores and inefficiencies in the football betting market. *Applied Statistics* **46**, 265–280.
- DIXON M. & ROBINSON M.E. (1998) A birth process model for association football matches. *Statistician*, **47**, 523–538.
- KARLIS D. & NTZOUFRAS I. (2003) Analysis of sports data by using bivariate Poisson models. *Statistician*, **52**, 381-393.
- KOOPMAN S.J. & LIT R. (2015) A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. *Journal of the Royal Statistical Society Series A*, **178**, 167-186. <https://doi.org/10.1111/rssa.12042>.
- LUNN D., JACKSON C., BEST N., THOMAS A. & SPIEGELHALTER D. (2013) The BUGS Book: A Practical Introduction to Bayesian Analysis. Chapman and Hall, London.
- MAHER M.J. (1982) Modelling association football scores. *Statistica Neerlandica* **36**, 109–118.
- MULLAHY J. (1986) Specification and testing of some modified count data models. *Journal of Econometrics*, **33**(3), 341–365.
- OPENBUGS. (2017) OpenBUGS Home Page. [ONLINE] Available at: <http://www.openbugs.net/w/FrontPage>. [Accessed 7 May 2017].
- OWEN A. (2011) Dynamic Bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. *IMA Journal of Management Mathematics*, **22** (2), 99-113. <https://doi.org/10.1093/imaman/dpq018>.
- R CORE TEAM (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: <https://www.R-project.org/>.
- WINKELMANN R. (2008) *Econometric Analysis of Count Data* 5th ed. Springer, Berlin.

# Flow Network Motifs Applied to Soccer Passing Data

David A. Perdomo Meza

Analytics FC + email address: dperdomomeza@gmail.com

## Abstract

Network Motifs are important local properties of networks, and have lately drawn increasing attention as promising concepts to unearth structural design characteristics of complex networks. In this document, we push the boundaries of the existing body of literature which has used this theory to study soccer passing networks by attempting to uncover unique team passing network structure, and make a rigorous attempt to formalise a theoretical framework in which to carry out and evaluate these analyses. We contribute to the existing body of knowledge by proposing a framework based on repeatability in which to establish the ideal length of flow motifs with which to study soccer passing networks, and also by considering spatial classifications of flow motifs to achieve greater precision in our claim to discover unique team passing network style.

## 1 Introduction

The collaborative, dynamic and fluid nature of soccer has lent itself to be studied under the lens of network theory. Authors such as Peña and Touchette (2012) have used passing networks to try and extract insight and information about individual team and player performance. In the context of network theory research in soccer, the concept of *flow motifs* has lately taken the spotlight. Originally introduced by Milo et al. (2002), network motifs attempt to quantify local structural properties that are essentially inherent to each individual network's nature. In the authors' words, the Theory of Network Motifs is an attempt to "uncover the basic building blocks of most networks" (Milo et al., 2002).

The study of flow motifs in soccer passing networks is off to a promising start. Previous authors have used the concept to analyse unique local structure in individual teams' passing networks which can be likened to discovering individual teams' *playing style* or what we will refer to as team *passing network profiles*, acknowledging the limitations of this claim.

In this document, we will contribute to the theory by reviewing the current state of research in this area found in the literature and establish rigorous definitions in a more generalised version of the theory than that previously presented, as well as broadening the conceptual arsenal by proposing a theoretical framework based on the concept of "*repeatability*" to evaluate the objective of discerning unique team passing network style.

We also contribute to the body of knowledge by considering and tentatively answering the question about the ideal length of motifs, and by introducing a classification of flow motifs using spatial variables that furthers the cause of obtaining unique team *passing profiles*.

The document is organised as follows: in Section 2 we make the formal mathematical definitions of the theory and discuss their interpretation in the context of soccer passing networks, as well as discussing the current state of research found in the literature. In Section 3 we replicate the 'state of the art' results using

the data set we had available, before introducing the framework based on repeatability and using it to answer the question about ideal motif length and to explore the spatial classification of motifs. Section 4 concludes on our results and Section 5 closes out by proposing some directions for future work.

## 2 Flow Network Motifs

Lets begin by defining the central concepts of the Flow Network Motifs theory:

**Definition 1.** A *flow network* is a connected directed graph  $G(V, E, \gamma)$  with  $\gamma$  a fixed path such that  $\forall (v, u) \in E$ , we have that  $\gamma(t) = (v, u)$  for some  $t \in \mathbb{N}$ .

**Definition 2.** A *flow network isomorphism* is a mapping  $\psi : G(V, E, \gamma) \rightarrow G'(V', E', \gamma')$  such that the induced mapping  $\psi : G(V, E) \rightarrow G'(V', E')$  is a graph isomorphism and  $\psi(\gamma(t)) = \gamma'(t)$ . We denote the fact that two flow networks are isomorphic as  $G \sim G'$ .

**Definition 3.** For a number  $k \in \mathbb{N}$ , the *class set of k-motifs* is defined as  $FN(k) / \sim$  where  $FN(k) := \{G(V, E, \gamma) \mid G \text{ is a flow network and } \text{length}(\gamma) = k\}$  and  $\sim$  is the equivalence relationship induced by flow network isomorphisms. Additionally, for a class  $[\omega] \in FN(k) / \sim$  and a directed graph  $G(V, E)$ , an *occurrence* of  $[\omega]$  in  $G$  is a subgraph  $G'(V', E') \subset G(V, E)$  such that  $\omega \sim G'$  (which is equivalent to saying that there is a copy of  $\omega$ 's path  $\gamma$  in  $G$ ). These occurrences are also referred to as *k-motifs* of  $G$ .

The literature has adopted the convention of denominating motif classes by a representative set of acronyms that uniquely represent each class. For example, for the focal case of  $k = 3$ ,  $FN(3) / \sim$  has five elements illustrated below.

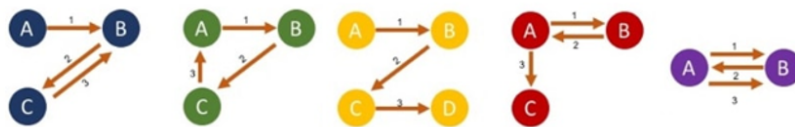


Figure 1: For  $k = 3$  there are 5 different classes of flow motifs:  $ABCB$ ,  $ABCA$ ,  $ABCD$ ,  $ABAC$  and  $ABAB$ .

The uniqueness of the representation by the acronyms is ensured by the equivalence classes under ‘ $\sim$ ’: even accepting labeled nodes,  $ABAC \sim BABC$ .

### 2.1 Network Motifs in Soccer Passing Networks

The connected components of soccer passing networks (i.e. uninterrupted passing sequences) evidently fit the definition of *flow networks* with players as nodes, passes between them as directed edges and passing sequences as *paths*. Figure 2 illustrates the occurrence of network motifs for  $k = 3$  in a passing sequence of the England national team.

Gyarmati, Kwak and Rodriguez (2014) used passing information from Spanish La Liga to determine the relative frequency with which each team used each motif type for  $k = 3$ . Considering each team as a vector



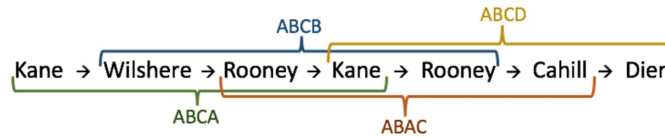


Figure 2

in  $\mathbb{R}^5$ , they used principal component analysis to establish that FC Barcelona had a relative use of the motif types that distinguished them from other teams; and in addition test the frequency of 3-motifs of teams' passing network versus their expected occurrence in randomly generated networks (in random networks  $ABAB$  happens less frequently than  $ABCD$  for example) to establish that the relative frequencies of motifs correspond to unique characteristics of passing networks and should be an object of study. Building off their work, Peña and Navarro (2015) try to extrapolate the method to a player level by considering that players have 'roles' within team motifs: for example, in an  $ABAC$  sequence, a player could have been the  $A$ ,  $B$  or  $C$  node. For  $k = 3$ , counting for all 5 motif types a player could have performed 15 different 'roles'. Using clustering algorithms on this way of seeing players in  $\mathbb{R}^{15}$ , the authors attempted to answer the question "Who can replace Xavi?". Bekkers and Dabadghao (2017) use the theory of passing motifs in relation with expected goals to investigate further into the applications of these concepts. In general, the narratives of these papers has been around attributing a *style* to teams and players according to the distribution of the vectors operationalised from the motifs found in their passing networks through the use of various data mining techniques.

In the following sections we will build on these aforementioned approaches and provide a theoretical framework in which to evaluate the objective of assigning *styles* to teams' passing networks using the theory of Flow Network Motifs, expand on the existing methodologies and ideas, and provide some empirical results using the available data.

### 3 Data, Methods and Results

The data that was used here was Opta's F24 event files, which log, classify, locate on the pitch and timestamp every 'on the ball event' in a football match of a covered league. For the research presented here we had access to the full data set for the 2014-15 and 2015-16 seasons of the English Premier League as well as the 2015-16 season of the Bundesliga, which we used to extract all passing sequences occurring in these seasons and crucially who the players involved in each sequence were and in which order, which allow us to identify the flow motifs present in the sequences.

*Remark 1.* We log **all** occurrences of flow motifs in passing sequences independent of the actual length of the sequence, so using the example from Figure 2, a 6-pass long sequence will log 4 different occurrences of 3-motifs, 3 occurrences of 4-motifs, 2 occurrences of 5 motifs and a single occurrence of a 6-motif.

The results from Hierarchical Clustering and Principal Component Analysis of a methodology replicated from Gyarmati, Kwak and Rodriguez (2014) applied to the Premier League data are shown in Figure 3.

Of special interest is the signaling out of Leicester as the team with the most unique style in this period in addition to its "style" being closer to traditionally powerful teams like Manchester City and Arsenal,

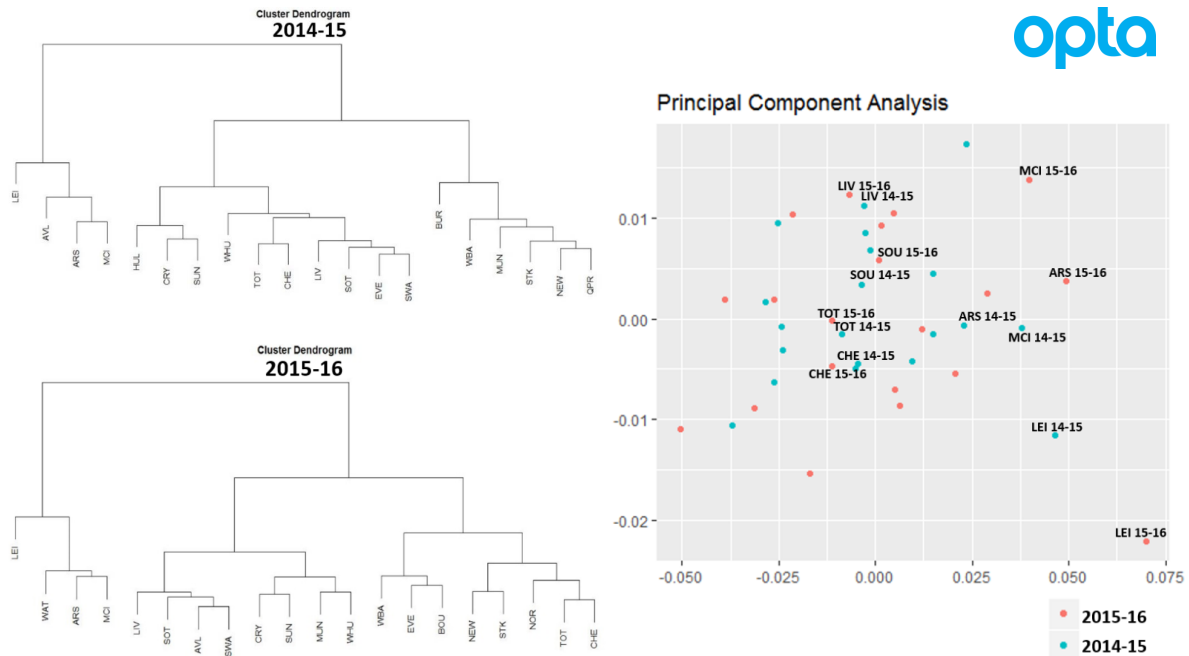


Figure 3: Using data provided by Opta, each team is seen as an observation in  $\mathbb{R}^5$  where each entry corresponds to the relative frequency of the occurrences of each of the 5 motif classes for  $k = 3$

especially in light of its surprising title achievement in the 2015-16 season. It is also noteworthy that even though in the 2014-15 season Leicester finished 14th, the most similar style to Leicester 15-16 is precisely Leicester 14-15 (see the Appendix for boxplot graphs of each team’s distribution of the frequency of each class for each match in the 2015-16 season).

### 3.1 Repeatability

A key objective of any methodology with a ‘style’ narrative in soccer data analytics is *repeatability*: the variability of results intended to represent stable qualities like *style*, *potential* or *quality* for specific teams or players should be small over short periods of time. In the context of underlying *style* of passing networks, there wouldn’t be much trust in a method that said a team had a certain style at some point in a season and a diametrically different one a few fixtures later.

In approaches such as this one in which “*stylistic*” characteristics are operationalised and vectorised in a euclidean observation space, the distances between observations conceptually serve as a proxy for *similarity* of style. The preliminary results shown above for two consecutive seasons of soccer passing network data provide some rough evidence that the study of flow motif frequencies can help discern some unique and *repeatably* characteristics of teams’ passing networks. However, the existing literature has failed to formalise the study of the *repeatability* in this context. To verify repeatability and establish a theoretical framework in which to ‘measure’ repeatability that can compare different methodologies, we propose to randomly divide

a team's motifs into two different a priori unlabeled sets and operationalising a measure of “success” in assigning more similar *styles* to teams' two sets of motifs in relation to sets of motifs of other teams. In the next section, we will introduce a rigorous framework in which to operationalise and implement this notion of repeatability in tandem with analysing the results for different choices of  $k$ .

### 3.2 Choice of $k$ and Repeatability

Previous authors have focused on a choice of  $k = 3$ , but to our knowledge no theoretical backing has been given to support that this is indeed the best choice for  $k$  with which to study soccer passing networks. We can intuitively predict that very low choices of  $k$  like 1 or 2 will fail to differentiate or uncover unique traits of teams' passing networks, while extremely high choice of  $k$  will suffer from several damaging defects, like for example smaller data sets (less 20-pass sequences are completed per season than 3-pass sequences) or in general “overshooting” the underlying style or intent of teams: the flow motifs of 20-pass sequences are so numerous, specific and subject to ‘randomness’ over longer periods of time, that the frequency of their occurrences most likely fails to correspond to underlying team style or intent of play but are rather a random outcome unrelated to the underlying team structure. The question remains, however, whether there is a more appropriate choice of  $k$  that strikes the right balance between capturing enough unique team structure in their passing network without overshooting it. For context, Figure 4 shows the number of occurrences of  $k$ -passes long passing sequences for  $k \in [3, 4, 5, 6, 7]$  in the 2015-16 season of the Premier League, as well as the number of *classes* encountered for each value of  $k$ :

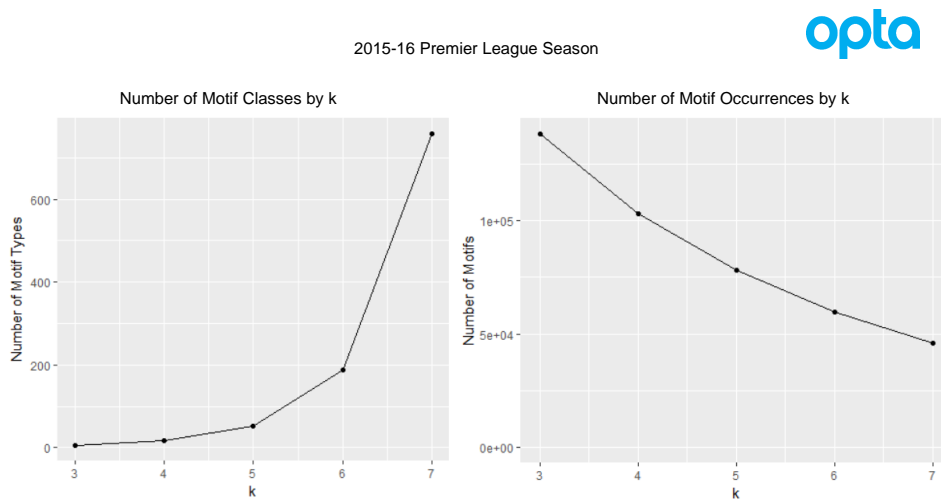


Figure 4: Data provided by Opta. The graph on the left shows that for  $k = 3, 4$  and  $5$ , all the 5, 15 and 52 theoretically possible classes of motifs occur in the data; while for  $k = 6$  and  $7$  there are 187 and 759 classes present in the data, which don't match the total amount of theoretically possible classes. The graph on the right plots the number of occurrences of motifs in the data for each  $k$ , going from 138,432 3-motifs to 45,820 7-motifs.

To answer this question, we propose the following methodology: Using data from the 2015-16 Premier League season, for each choice of  $k$ , we take each team's observations of  $k$ -motifs and randomly divide them into two different sets as if artificially considering them as two sets corresponding to two different teams. Defining  $\mathfrak{R} := \#(FN(k) / \sim)$  (for example when  $k = 3$ ,  $\mathfrak{R} = 5$ ), in the same way as the methodology outlined above for  $k = 3$ , each team can be seen as a vector in  $\mathbb{R}^{\mathfrak{R}}$  where each entry corresponds to the relative frequency of a type or *class* of  $k$ -motif. Given our random division of teams' motifs, for the 2015-16 season data we have 40 (2 sets for all 20 teams) vectors in  $\mathbb{R}^{\mathfrak{R}}$ . If our belief that unique team information is properly and repeatably represented in this vectorisation by teams' position in the  $\mathbb{R}^{\mathfrak{R}}$  observation space is correct, then the euclidean distance between a team's two vectors corresponding to their two sets of motifs should be comparatively smaller to the distances to other teams' vectors. In addition, this effect should be clearer for choices of  $k$  that better capture the unique underlying structure of teams' passing networks; so the 'strength' of this effect serves as a comparable indicator for different choices of  $k$ . Specifically, we define a *repeatability index* or *percentage* as:

$$\left( \sum_{i \text{ in teams}} 2 - \frac{\text{order}(v_i^1, v_i^2) + \text{order}(v_i^2, v_i^1)}{N - 1} \right) / N, \quad (1)$$

where  $N$  is the total of teams in the sample and for each team  $i$ ,  $v_i^1, v_i^2$  denote its two sets of vectors and  $\text{order}(v, u)$  the position of vector  $u$  in the list of all the other  $N - 1$  vectors ordered from closest to farthest away from  $v$  using the scaled euclidean distance of  $\mathbb{R}^{\mathfrak{R}}$  (meaning its on a scale from 1 to  $N - 1$ ). Therefore, the closer the *repeatability percentage* is to 1, the 'closer' the styles attributed to teams' two sets of vectors on average; which is precisely our intuitive notion of *repeatability*.

*Remark 2.* The reader might wonder why instead of the strict ordered position we didn't consider the z-score distance between teams' two sets of motifs, which seems like a more robust and orthodox approach. However, the reason this wasn't done is that in order to compare different choices of  $k$  (and therefore different values of  $\mathfrak{R}$ ), the distances for each choice are measured in spaces of different dimensions and we have no way of knowing whether the 'dimensional' structure of the data was affecting the comparability of the distances for different choices of  $k$ .

Figure 5 shows the average repeatability percentage for each choice of  $k$  from 3 to 7 for 100 trials of randomly dividing all 20 teams' motifs into two sets using data from the 2015-16 season of the Premier League.

We immediately find evidence of the "overshooting" of underlying structure for higher choices of  $k$ . For  $k = 6, 7$  the repeatability percentage is 57.8% and 52.3% respectively. Considering that the repeatability percentage of a method that randomly assigned distances between vectors is expected to be 50%, this means that by  $k = 7$  we've lost almost all comparable structure. The choice with the highest repeatability is satisfactorily  $k = 3$  with 82.7%.

### 3.3 Spatial Classification of Passing Motifs

As mentioned previously, the used data files also provide an  $x, y$ -coordinates location on the pitch for each logged event. An additional objective which we explored was the possibility of deepening our unique representation of a team's passing network style by including a classification by motifs' spatial characteristics.

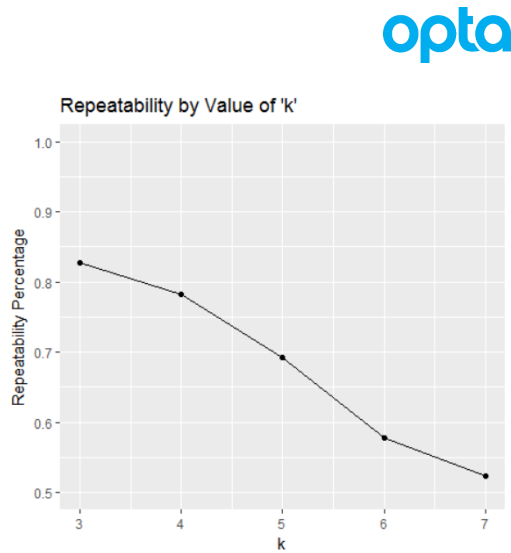


Figure 5: Data provided by Opta.

There are two different operationalisations of the available spatial information which we wanted to test against the data:

1. Directly use the start and end  $x,y$ -coordinates of each pass in a motif so that a  $k$ -motif will have an associated vector of spatial variables in  $\mathbb{R}^{4k}$ .
2. Use the ‘*polar*’ coordinates approach of  $r$  and  $\theta$  so that a  $k$ -motif will have an associated vector of spatial variables in  $\mathbb{R}^{2k}$  as is shown in Figure 6.

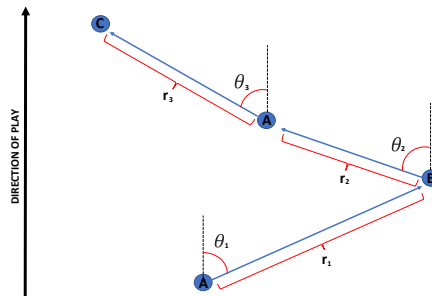


Figure 6

Under our assumption that differentiating motifs by their spatial attributes can help identify unique team structure, we want to operationalise these characteristics in a manageable way which we can use with confidence as a representation of a team style that differentiates it from others. The way we propose of doing this is using k-means clustering on the motif observations in the spatial variables observation space to obtain different categories of motifs and then compare the vectors of relative frequencies of teams' motif types in terms of what category they fall under.

In other words, using  $k$ -motifs and  $n$  clusters determined by spatial variables for  $k, n \in \mathbb{N}$ , teams will be observations in  $\mathbb{R}^{k*n}$ : for each of the  $k$  motif types, we have the relative frequency with which that motif type/class was performed in the form of each of the  $n$  spatial categories.

*Remark 3.* In this approach there are two forms of categories: the motif classes (i.e. *ABAC*) and the category corresponding to clusters from spatial data. Therefore a motif will be of a certain class **and** also of a certain spatial category. Knowing that considering relative frequencies rather than raw amounts worked well in the original approach, we similarly consider the relative frequencies of these  $k*n$  pair of categories.

We can use the same 'repeatability methodology' as above to determine whether these vectorised representations exhibit more unique structure for teams, and crucially can also use the comparability of the repeatability percentages to decide what the correct number of categories  $n$  is to introduce as a parameter to the k-means algorithm, such that more unique structure is captured without *overshooting* it by using excessive detail. Figure 7 shows the average results for 100 random trials for both the 'x,y-coordinates' and polar coordinates ideas; where we divide teams' motifs into two sets and consider the distances between observations in  $\mathbb{R}^{5*n}$  for  $k = 3$  (therefore  $k = 5$ ) and  $n \in [1, 2, 3, \dots, 50]$ .

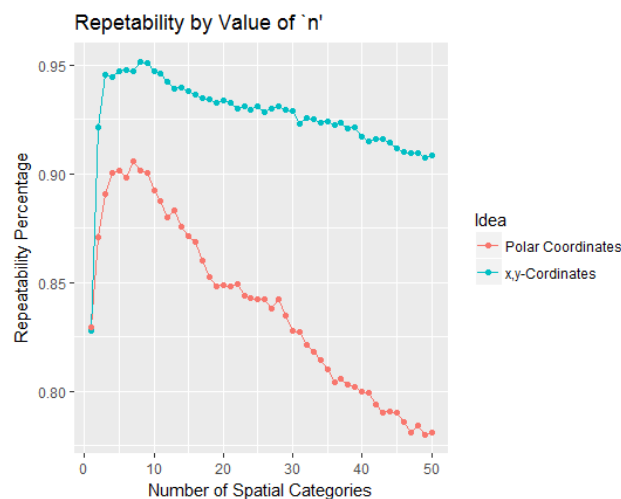


Figure 7: Both vectorisations begin with the same repeatability at  $n = 1$  (which is equivalent to the original methodology of Gyarmati, Kwak and Rodriguez (2014)), then start achieving greater repeatability than this original approach by encoding more underlying structure before reaching an 'optimum' and starting to 'overshoot' this structure and decrease in repeatability.

The maximum repeatability is achieved at  $n = 8$  spatial categories of motifs using  $x, y$ -coordinates, with a repeatability index up to 95.2%! This is an extremely high repeatability, with teams' two sets of motifs being

the closest two observations more often than not. The potential applications for this unique identification of team passing network profile are as endless as they are exciting. We can look to discover profiles that are more hurtful to certain defensive set-ups, predict match outcomes, match transfer targets to teams according to how well their profiles fit, etc. Knowing that the underlying structure of passes moving across the pitch accommodates 8 categories can be important theoretical background for exercises of an entirely different nature like modeling of games as Markov Processes for example.

*Remark 4.* Even though the  $x, y$ -coordinates classification worked better in general than the polar coordinates classification, there's value in the fact that the latter also achieved impressive improvements in repeatability (90.6% at  $n = 7$ ). This is because the repeatability of the  $x, y$ -coordinates classification is confounded with how deep into the opposition's half a team plays (therefore their motifs will have "high" values of  $x$  coordinates). It's not that this information isn't valuable, but it's information which is widely known and available in basic soccer knowledge about teams that dominate opposition. In contrast, the repeatable structure for the Polar Coordinates classification is purely geometrical and not confounded with 'obvious' things we already know, which makes it interesting even if its not as highly repeatable as the  $x, y$ -classification.

Figure 8 below shows the results of hierarchical clustering and principal component analysis applied to the observation space of passing network profiles for 2015-16 Bundesliga and Premier League teams.

The results have several interesting talking points. First of all, there is a clear distinction between the passing profiles of teams from the Bundesliga and teams from the Premier League. Only 4 teams are in general closer to teams from the opposite league, all from the Bundesliga: Bayern Munich, Bayer Leverkusen, FC Ingolstadt and Darmstadt. Relegation headed Ingolstadt and Darmstadt are both heavy outliers of the data set so this fact might not be meaningful in their case.

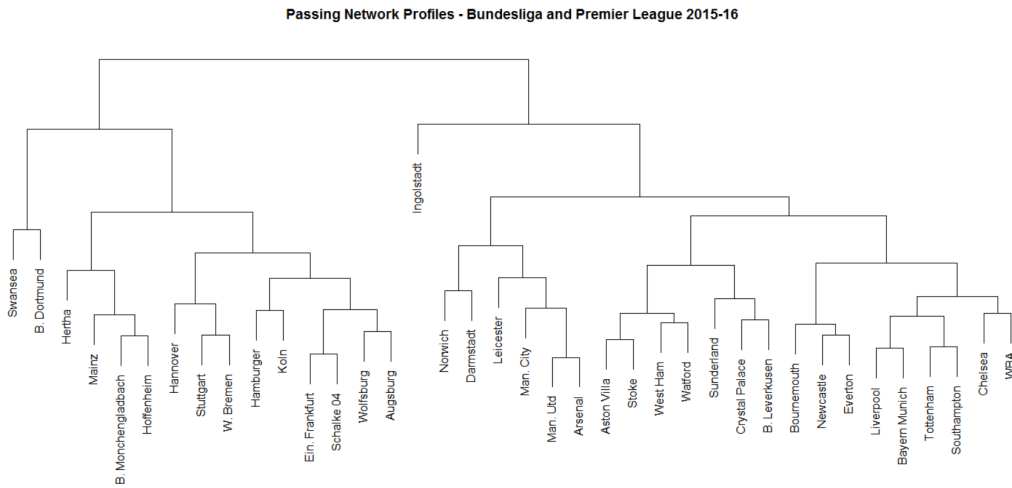
Overall, the high-performing teams of both leagues are separated into two clusters: Manchester United, Arsenal, Manchester City and Leicester in one; and Bayern Munich, Liverpool, Tottenham and Chelsea in another (along with West Bromwich Albion and Southampton). Borussia Dortmund have a more distinctive profile.

In the appendix you can find some examples of '*passing network profiles*' for Premier League and Bundesliga teams for  $k = 3$  and  $n = 8$  for  $x, y$ -coordinates classification.

## 4 Conclusions

The theoretical framework and results presented here are important since to our knowledge, no previous work had attempted to provide this level of rigour and formality to an evidently exciting and promising area of soccer data analytics. We have established a conceptual dossier with which to evaluate the results of the theory and establish best practice in some of the parametric choices, and in addition we have pushed the boundaries of stylistic identification of teams' passing networks by analysing the inclusion of spatial classifications of flow motifs. We are now armed with a high degree of theoretical confidence based on repeatability underpinning our claims of underlying passing network *profiles*, from which we can take firm and assured steps towards using these *stylistic findings* to answer some more applied questions in soccer data analytics in areas such as tactics, pre-match preparation and recruitment.

Some of the results also have far-reaching consequences beyond the application of flow motifs. Soccer is a highly complex sport: what happens on a soccer pitch during a match is very dynamic and interdependent; and as a result, in contrast to other traditional sports like baseball and basketball, data analytics have failed



Team Passing Network Profiles

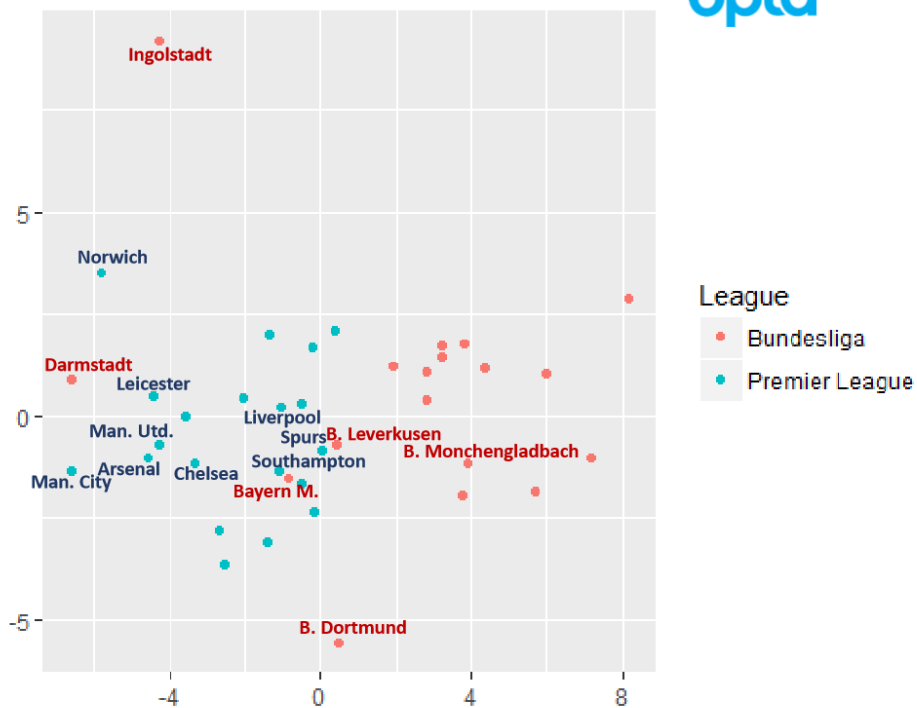


Figure 8



to really penetrate deep into the inner workings of soccer. This difficulty is mainly owed to dealing with the dynamic sequentiality of soccer events. Breaking down the sequentiality into meaningful constituent blocks is an important horizon for soccer analytics research. Considering that the grand majority of soccer events are passes (over 80% according to research by Cintia, Rinzivillo and Pappalardo (2015)), the fact that  $k = 3$  is the optimum length in which to study passing motifs sheds a light on the optimum length of constituent time blocks of soccer matches in which we may aim to break down our analyses to operationalise the complex sequentiality of the information and process the raw data into meaningful representations for teams and players.

## 5 Future Work

We presented above a methodology to deepen our extraction of unique passing network *profiles* of different teams, and briefly discussed some of its potential applications. Future work should focus on exploring some of these more applied possibilities, underpinned by the theoretical confidence provided by the framework presented here that we are indeed uncovering underlying structure inherent to teams' passing networks.

Another interesting area to explore would be to take on board the objective of having repeatable style operationalised in an observation space through the use of flow motifs, but approaching it from a 'Metric Learning' point of view. Also, having provided evidence that using spatial variables of  $k$ -motifs (particularly the  $4k$  start and end  $x, y$ -coordinates of each pass) can aid the objective of identifying repeatable team styles, the possibility of operationalising this unique team information by attributing to each team a probability distribution in  $\mathbb{R}^{4k}$  instead of going through the discretisation used here through  $k$ -means clustering should be explored. This approach would provide greater robustness that perhaps would allow us to delve deeper into unique team styles, but would do so at the expense of simplicity of the manipulation and operationalisation of these 'style findings'. However, if a creative approach managed to deploy these results in the Distribution Space of  $\mathbb{R}^{4k}$  and turn them into applications such as predicting outcomes or matching transfer targets, the implications could be massive.

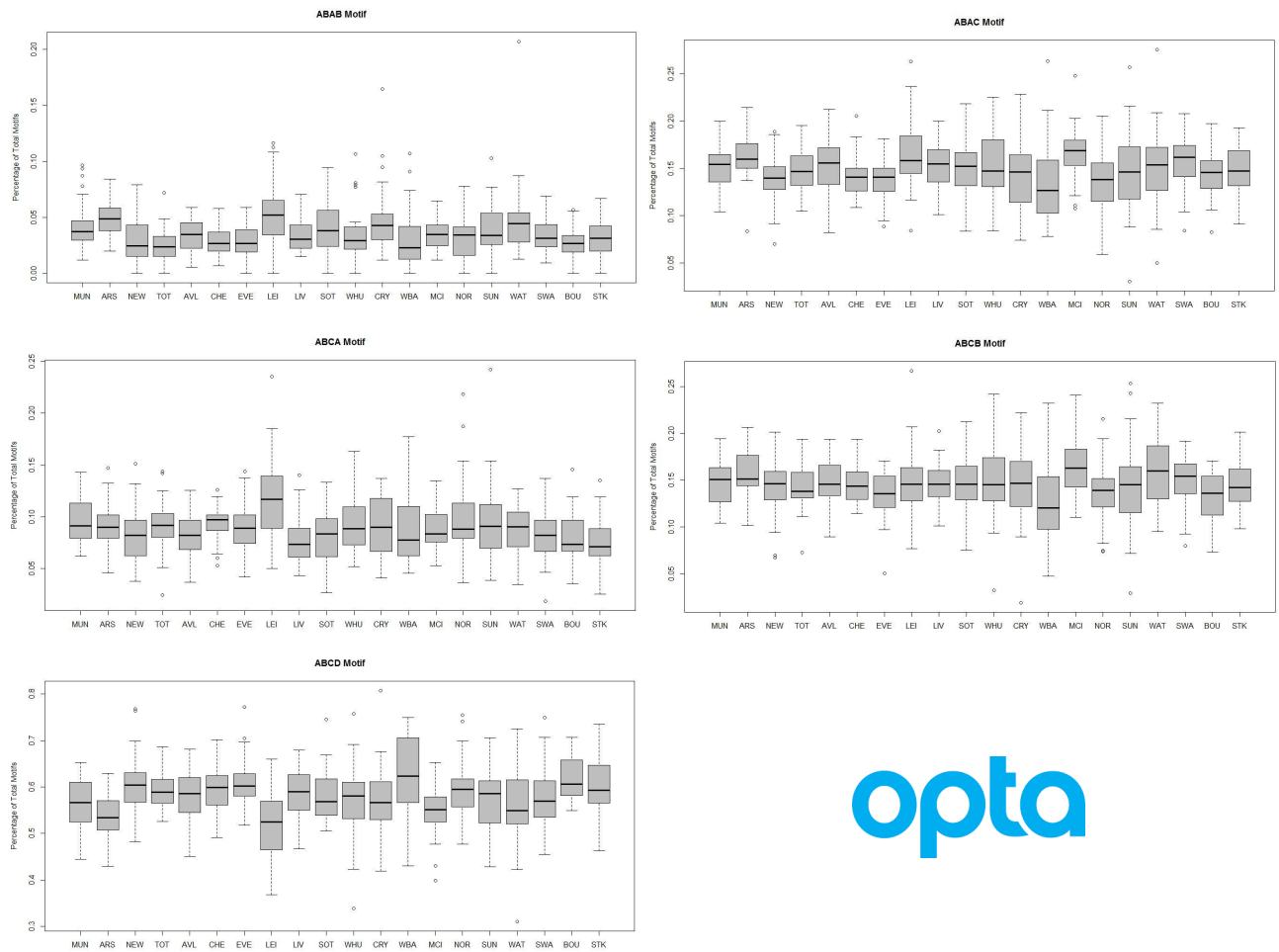
Another important dimension which we haven't explored here is unique passing network structure at a player level. Authors such as Peña and Navarro (2015) and Bekkers and Dabadghao (2017) have made headway on this matter, and the implications for recruitment and talent identification of having robust passing network *profiles* at player level could be revolutionary. This topic merits research documents in its own right, but the theoretical framework presented here can be a starting point to adapt to a player dimension which will allow linking the results of teams and players. On another note, it is also worth exploring the results of flow motifs applied to soccer networks of a different nature. As an example, we can envisage a network where the pitch is divided into a grid of zones and the network represents the ball going from zone to zone (zones being the nodes). Flow motifs may constitute an adequate language in which to operationalise the sequential character of these other networks.

Finally, the spatial category boxplots in the appendix reveal that the computation of the spatial clusters isn't converging efficiently. This computational aspect should be addressed in future work.

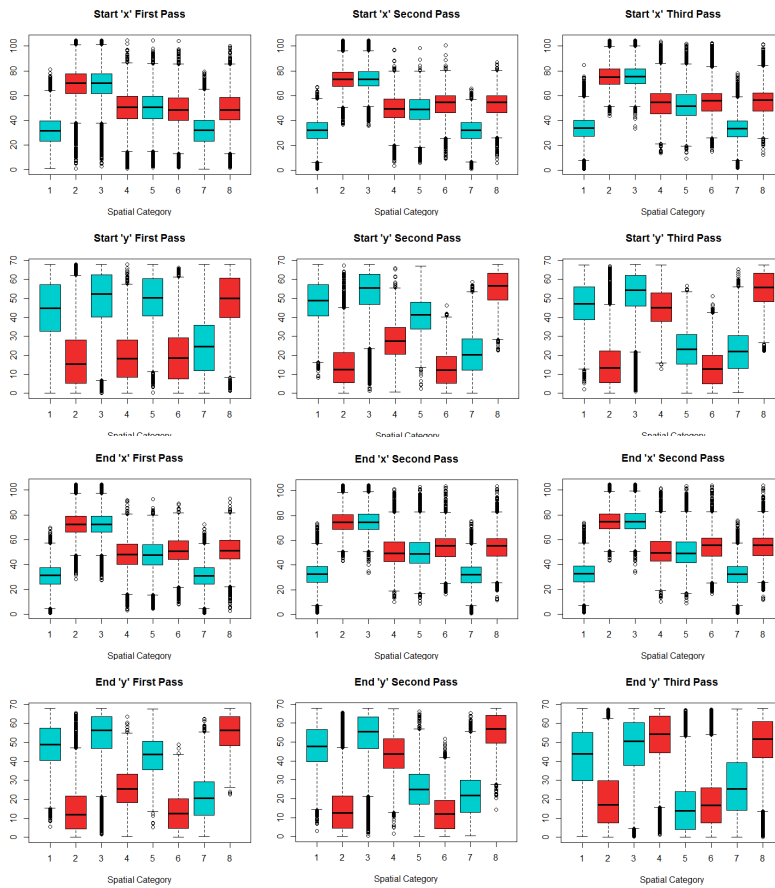
## Appendix

### Boxplots of Motif Class Relative Frequency By Team By Match ( $k = 3$ )

Based on Opta's 2015-16 Premier League Data



**Boxplots of Spatial Category Variables ( $k = 3, n = 8$ )**



## Examples of Team Passing Network Profiles

Based on 2015-16 season Opta data

### Bayern Munich

	Category 1	Category 2	Category 3	Category 4	Category 5	Category 6	Category 7	Category 8
ABAB	0.35%	0.49%	0.39%	0.08%	0.04%	0.47%	0.36%	0.46%
ABAC	1.92%	2.67%	2.21%	0.87%	0.87%	2.51%	1.77%	2.44%
ABCA	1.07%	1.58%	1.10%	0.46%	0.51%	1.26%	1.04%	1.01%
ABCB	1.63%	2.22%	2.01%	1.12%	1.18%	2.24%	1.81%	2.08%
ABCD	7.08%	7.88%	7.31%	9.92%	9.32%	5.48%	7.13%	5.67%

### B. Dortmund

	Category 1	Category 2	Category 3	Category 4	Category 5	Category 6	Category 7	Category 8
ABAB	0.45%	0.59%	0.38%	0.13%	0.17%	0.92%	0.69%	0.59%
ABAC	1.71%	2.38%	2.32%	1.28%	1.37%	3.52%	2.38%	3.00%
ABCA	0.83%	1.14%	1.26%	0.59%	0.51%	1.73%	1.18%	1.40%
ABCB	1.69%	2.02%	2.18%	1.54%	1.68%	3.07%	2.40%	2.53%
ABCD	5.03%	5.66%	6.09%	9.09%	9.05%	5.86%	5.81%	5.77%

### Manchester City

	Category 1	Category 2	Category 3	Category 4	Category 5	Category 6	Category 7	Category 8
ABAB	0.20%	1.01%	1.00%	0.13%	0.01%	0.34%	0.19%	0.50%
ABAC	1.29%	3.44%	3.48%	0.65%	0.69%	2.81%	1.27%	2.85%
ABCA	0.65%	2.33%	2.11%	0.40%	0.30%	1.47%	0.39%	1.31%
ABCB	1.17%	3.24%	3.62%	1.07%	0.88%	2.55%	1.23%	2.56%
ABCD	4.36%	9.05%	7.67%	8.75%	8.83%	6.39%	4.20%	5.62%

### Leicester City

	Category 1	Category 2	Category 3	Category 4	Category 5	Category 6	Category 7	Category 8
ABAB	0.42%	1.41%	1.38%	0.03%	0.08%	0.58%	0.29%	0.85%
ABAC	1.67%	3.05%	3.50%	0.61%	0.48%	2.71%	1.75%	2.52%
ABCA	1.38%	2.10%	2.73%	0.45%	0.24%	1.75%	1.30%	1.43%
ABCB	1.65%	2.39%	3.03%	0.72%	0.96%	2.18%	1.54%	2.28%
ABCD	6.29%	6.74%	7.17%	6.37%	7.17%	6.34%	6.13%	6.34%

### Tottenham

	Category 1	Category 2	Category 3	Category 4	Category 5	Category 6	Category 7	Category 8
ABAB	0.43%	0.47%	0.61%	0.01%	0.04%	0.19%	0.42%	0.35%
ABAC	2.52%	2.03%	2.56%	0.71%	0.95%	1.82%	1.92%	2.42%
ABCA	1.38%	1.61%	1.52%	0.54%	0.41%	1.28%	0.97%	1.35%
ABCB	2.70%	1.91%	2.38%	1.15%	0.95%	1.47%	1.80%	2.06%
ABCD	7.65%	6.85%	6.79%	9.00%	9.28%	6.02%	6.88%	6.63%



## References

- [1] Bekkers, J. and Dabadghao S. *Flow Motifs In Soccer: What Can Passing Behavior Tell Us?*. 1st ed. MIT Sloan, 2017. Available at: <http://www.sloansportsconference.com/content/flow-motifs-soccer-can-passing-behavior-tell-us/> (Accessed: 19 Apr. 2017).
- [2] Cintia, P., Rinzivillo, S. and Pappalardo, L., 2015, September. A network-based approach to evaluate the performance of football teams. In *Machine Learning and Data Mining for Sports Analytics Workshop, Porto, Portugal*.
- [3] Gyarmati, L., Kwak, H. and Rodriguez, P., 2014. Searching for a unique style in soccer. *arXiv preprint arXiv:1409.0308*.
- [4] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. and Alon, U., 2002. Network motifs: simple building blocks of complex networks. *Science*, 298(5594), pp.824-827
- [5] Peña, J.L. and Navarro, R.S., 2015. Who can replace Xavi? A passing motif analysis of football players. *arXiv preprint arXiv:1506.07768*.
- [6] Peña, J.L. and Touchette, H., 2012. A network theory analysis of football strategies. *arXiv preprint arXiv:1206.6904*.

# The Social Doubles Tournament Problem

F. Salassa<sup>a</sup> and F. Della Croce<sup>b,c</sup>

a) Politecnico di Torino, DIGEP, Corso Duca degli Abruzzi, 24, 10129 Torino, Italy. fabio.salassa@polito.it

b) Politecnico di Torino, DAUIN, Corso Duca degli Abruzzi, 24, 10129 Torino, Italy. federico.dellacroce@polito.it

c) CNR, IEIIT, Torino, Italy

## Abstract

A novel recreational sport scheduling problem, denoted as the Social Doubles Tournament Problem, is investigated. The problem asks for finding a schedule for doubles matches, as in tennis, where teams are not fixed a priori but have to be selected in order to generate a round robin tournament where every player must partner every other player once and must be opposed to every other player twice. Also, whenever a quadruplet  $i, j, k, l$  is selected for a match such that the couple  $i, j$  is opposed to the couple  $k, l$ , then the same quadruplet will have to be selected two other times, once with the couple  $i, k$  opposed to the couple  $j, l$  and the other time with the couple  $i, l$  opposed to the couple  $j, k$ . The problem can be formulated as an integer programming model with binary variables and schedules can be obtained for limited size problem instances. We discuss the strong connections of the social doubles tournament problem with combinatorial design and in particular with Balanced Incomplete Block Designs and with another tournament scheduling problem, that is the Social Golfer Problem. By means of these connections, we are able to provide schedules of the social doubles tournament problem for instances with up to 64 players.

## 1 Introduction

Sports related problems, and particularly sport scheduling problems, attract the attention of researchers and practitioners in many research domains such as mathematics, statistics, economics.

Applications and methods for solving scheduling problems in sports constitute a flourishing research field. The number of surveys (see Dinitz et al. (2006), Drexler and Knust (2007), Easton et al. (2004), Kendall et al. (2010), Rasmussen and Trick (2008), Ribeiro (2012), Wright (2009)) witnesses the great interest in these topics. In [20], hundreds of references are listed pertaining the field of sport scheduling. These references are classified according to the problem they model, e.g. round robin tournament, traveling tournament, referee assignment to cite a few. A classification of solution methods is also proposed. This classification varies from combinatorial design to exact solution algorithms such as branch & bound or heuristic approaches from a practical point of view. Besides, from a theoretical perspective, also complexity and approximation issues are investigated. Finally a list of different sports is depicted which includes both single player and team based sports.

Standard sport scheduling problems typically involve  $2n$  teams that have to play against each other once (resp. twice) in single (resp. double) round robin tournaments. This basic problem has a large number of variants as highlighted before. If no further constraints are present, the construction of a single round robin

tournament can be easily obtained in linear time in the number of matches (see e.g. Blest and Fitzgerald (1988)) and has a nice graph-theoretic interpretation: the construction of a round robin tournament with an even (odd) number of players corresponds to the problem of finding a 1-factorization (near-1-factorization) of a complete undirected graph with an even (odd) number of vertices, Mendelsohn and Rosa (1985).

Most of the literature (see, e.g. Della Croce and Oliveri (2006) and Nemhauser and Trick (1998)) concerning double round robin tournaments deals with sport leagues where there are further constraints related to the so-called home-away pattern: each match is played by two teams at the home of one of these teams and hence each team must alternate home matches and away matches as much as possible; also, some couples of teams may need a complementary home-away schedule as, for instance, they share the same stadium.

In this work we introduce a novel recreational sport scheduling problem denoted as the *Social Doubles Tournament Problem* (SDTP). SDTP relates to teams composed by two players as in tennis doubles matches or whist or bridge players. The problem consists in creating a round robin team tournament where teams (composed by pairs of players) are not fixed apriori but have to be generated while creating rounds and every player must have every other player both as a partner and as an opponent. In addition, the problem asks, whenever 4 players meet, to do so for three consecutive matches in a mini-tournament where players exchange their partners. The problem can be encountered when designing recreational competitions for sports like tennis and badminton or for card games like Bridge and, more in general, when the schedules are based on non-predetermined couples of players.

The paper is organized as follows. In Section 2, SDTP is introduced and an integer programming formulation with binary variables is proposed. Section 3 connects SDTP to combinatorial design with emphasis on Balanced Incomplete Block Designs (BIBDs) and to the social golfer problem. Correspondingly, schedules of SDTP for instances with up to 64 players are presented. Section 4 concludes the paper with final remarks.

## 2 Problem Description and Integer Programming Formulation

SDTP can be defined as follows. A round robin tournament for doubles must be scheduled and each player must partner once with all players and must be opponent twice to all players. In addition, each selected quadruplet  $i, j, k, l$  of players must appear three times in the schedule: a first time when team  $i, j$  is opposed to team  $k, l$ , a second time when team  $i, k$  is opposed to team  $j, l$  and a third time when team  $i, l$  is opposed to team  $j, k$ . In such a way the quadruplet involves a mini-tournament among the considered four players.

The proposed problem strongly relates to the *Whist Tournament Problem* where the aim is to compute a round robin tournament for doubles like in SDTP but the additional mini-tournament requirement is absent. The Whist Tournament Problem definition dates back to the 19th century in a work of Moore (1896). The existence of Whist Tournaments for  $4n$  players ( $n$  integer) was proved by Baker (1975) using a variety of construction techniques. Main construction techniques are based on given formats where, starting from a first round with specific properties, the remaining rounds can be constructed cyclically. Research on Whist Tournaments (and variants) is still active, see the survey paper of Anderson (1995). Whist Tournaments with  $4n + 1$  players can still be produced by introducing a a bye for each player (Baker 1975). Correspondingly, the number of rounds is  $4n + 1$ . For  $4n + 2$  and  $4n + 3$  players, the situation changes dramatically as a single bye does not allow a feasible schedule and a balanced schedule is required where each player must have the same number of byes. This last requirement induces unpractical schedules due to the minimum number of rounds needed to balance byes among players, e.g. with 6 players where the number of rounds is 30.

SDTP is essentially a decision problem on the existence of a *whist*-like round robin schedule subject to the mini-tournament constraints where every player has every other player once as a partner and twice as an opponent. This mini-tournament requirement for each quadruplet has the goal of concentrating the matches of a single player. Indeed, by assigning a couple of hours to quadruplet  $i, j, k, l$ , three long sets can be played in the considered time period where in each set the players exchange teammates. Further, the above requirement can be seen as a *fairness* requirement as, whenever a quadruplet of players is selected, all possible combinations of teammates/opponents are considered. Notice that with  $4n$  players a round robin tournament foresees  $4n - 1$  rounds. As every quadruplet determines 3 rounds for each player, the total number of disjoint rounds is equal to  $\frac{4n-1}{3}$  and must be integer. Hence, SDTP allows feasible solutions with  $4n$  players where  $4n \equiv 4 \pmod{12}$ , that is 4, 16, 28, 40, 52, 64 etc. >From now on, we will denote by  $SDTP_\alpha$  an SDTP with  $\alpha$  players.

An interesting aspect of such balanced tournaments is that a single player ranking can be obtained even though doubles matches are performed. Since each player must have every other player as a partner and as an opponent the same number of times, a simple ranking methods based on match (or set) wins can be designed to end up with a single player ranking at the end of the league.

### 2.1 ILP Formulation

For any suitable  $SDTP_{4n}$ , we consider all possible quadruplets  $i, j, k, l$  inducing a double match among the four players of the quadruplet. Notice that every quadruplet  $i, j, k, l$  induces three matches corresponding to the three different combinations of partners and teammates of the quadruplet. Let denote by  $M = \binom{4n}{4}$  the total number of quadruplets. Correspondingly, the total number of possible matches is  $3M$ .

All players have to play in every round, that is,  $n$  quadruplets must be selected in every round. Notice that, once quadruplet  $i, j, k, l$  is defined, the three possible double matches  $(i, j \text{ vs } k, l)$ ,  $(i, k \text{ vs } j, l)$  and  $(i, l \text{ vs } j, k)$  can be represented by the three ordered quadruplets  $[ijkl]$ ,  $[ikjl]$  and  $[iljk]$  where teammates and opponents are univocally determined by their position in the quadruplet assuming as teammates players in positions 1, 2 and 3, 4 respectively. This is shown in Figure 1 where teammate and opponents of player  $i$  are considered:  $j$  is the teammate of  $i$  (left side of the figure) while  $k$  and  $l$  are both opponents of  $i$  (right side of the figure).

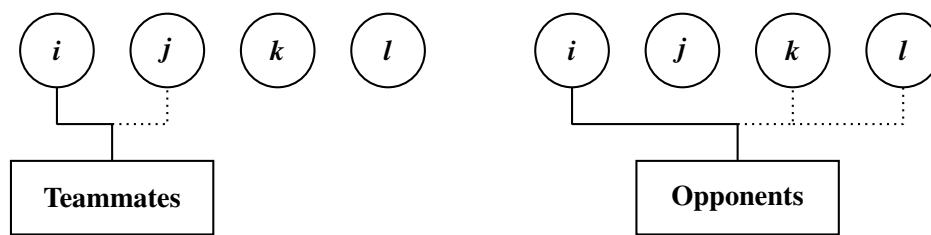


Figure 1: Example of teammate and opponents definition for Player  $i$  based on the positioning in the tuple.

We search then for a schedule with  $\frac{4n-1}{3}$  disjoint rounds where each quadruplet appears only once. The quadruplet is also ordered and induces therefore a specific double match as indicated in Figure 1 so that we can derive various coefficients for that match defining players as teammates or opponents. Then, each round



is triplicated in order to consider the other two possible matches of each selected quadruplet so as to get  $4n - 1$  rounds.

We present here an integer linear programming (ILP) formulation of the problem by making use of the notation and parameters introduced in Table 1.

$4n$	Number of players
$M$	Number of quadruplets
$R$	Number of rounds ( $=\frac{4n-1}{3}$ )
$i$	Index used for players
$j$	Index used for players
$m$	Index used for quadruplets
$r$	Index used for rounds
<i>Coefficients:</i>	
$C1_{mi}$	$= \begin{cases} 1 & \text{if player } i \text{ is present in quadruplet } m \\ 0 & \text{otherwise} \end{cases}$
$C2_{mij}$	$= \begin{cases} 1 & \text{if both players } i \text{ and } j \text{ are present in quadruplet } m \\ 0 & \text{otherwise} \end{cases}$
<i>Decision variables:</i>	
$x_{mr}$	$= \begin{cases} 1 & \text{if quadruplet } m \text{ is assigned to round } r \\ 0 & \text{otherwise} \end{cases}$

Table 1: Parameters and notations.

Since  $SDTP_{4n}$  is a decision problem, the ILP formulation presents no (that is fictitious) cost function subject to the following constraints.

$$\sum_{m=1}^M x_{mr} = n \quad \forall r = 1, \dots, R. \quad (1)$$

$$\sum_{m=1}^M C1_{mi} * x_{mr} = 1 \quad \forall i = 1, \dots, 4n; r = 1, \dots, R. \quad (2)$$

$$\sum_{m=1}^M \sum_{r=1}^R C2_{mij} * x_{mr} = 1 \quad \forall i = 1, \dots, 4n - 1; j = i + 1, \dots, 4n. \quad (3)$$

$$x_{mr} \in \{0, 1\} \quad \forall m = 1, \dots, M; r = 1, \dots, R. \quad (4)$$

Here, constraints (1) sets the number  $n$  of quadruplets to be selected in the matches for each disjoint round. Constraints (2) state that each player must be present exactly once in each disjoint round, namely, in each round all players must be different. Constraints (3) force each couple of players to be present in a quadruplet exactly once. Constraints (4) set variables to be binary.

By running CPLEX 12.6 on the above ILP model, a feasible solution for 16 players is obtained in few seconds on a standard laptop. Table 2 depicts the related schedule once the disjoint rounds determined by the ILP solver have been triplicated in order to fulfill the mini-tournament requirement. Larger size problems could not be handled by means of this approach.

	Match 1		Match 2		Match 3		Match 4	
	Team 1	Team 2	Team 3	Team 4	Team 5	Team 6	Team 7	Team 8
Round 1	1 2	3 4	5 6	7 8	9 10	11 12	13 14	15 16
Round 2	1 3	2 4	5 7	6 8	9 11	10 12	13 15	14 16
Round 3	1 4	2 3	5 8	6 7	9 12	10 11	13 16	14 15
Round 4	1 5	11 15	2 7	9 13	3 6	10 14	4 8	12 16
Round 5	1 11	5 15	2 9	7 13	3 10	6 14	4 12	8 16
Round 6	1 15	5 11	2 13	7 9	3 14	6 10	4 16	8 12
Round 7	1 6	9 16	2 8	11 14	3 5	12 13	4 7	10 15
Round 8	1 9	6 16	2 11	8 14	3 12	5 13	4 10	7 15
Round 9	1 16	6 9	2 14	8 11	3 13	5 12	4 15	7 10
Round 10	1 7	12 14	2 5	10 16	3 8	9 15	4 6	11 13
Round 11	1 12	7 14	2 10	5 16	3 9	8 15	4 11	6 13
Round 12	1 14	7 12	2 16	5 10	3 15	8 9	4 13	6 11
Round 13	1 8	10 13	2 6	12 15	3 7	11 16	4 5	9 14
Round 14	1 10	8 13	2 12	6 15	3 11	7 16	4 9	5 14
Round 15	1 13	8 10	2 15	6 12	3 16	7 11	4 14	5 9

Table 2: A solution for SDTP<sub>16</sub>

### 3 SDTP, combinatorial design and the social golfer problem

SDTP<sub>4n</sub> shares strong connections with combinatorial design and in particular with *Balanced Incomplete Block Designs* (BIBDs) Colbourn and Dinitz (2006). We recall that a BIBD is a collection  $\mathcal{B}$  of  $b$  subsets (called blocks) of a finite set  $\mathbb{X}$  of  $v$  elements, such that every block has the same number  $k$  of elements, and each pair of distinct elements appear together in the same number  $\lambda$  of blocks. The standard way to abbreviate these structures is  $(v, k, \lambda)$ -BIBDs.

The following basic properties (reported in Colbourn and Dinitz (2006)) hold.

**Property 1.** Any element of  $\mathbb{X}$  occurs in exactly  $r$  blocks where  $r = \frac{\lambda(v-1)}{(k-1)}$ .

**Property 2.** A BIBD has exactly  $b = \frac{vr}{k}$  blocks.

An important class of BIBDs are the so-called *resolvable BIBDs* (RBIBDs) where the blocks can be partitioned into sets called *parallel classes*, each of which forms a partition of the elements of the BIBD. The set of parallel classes is called a resolution of the design. An important family of resolvable designs are the *affine planes*. Affine planes of order  $n$  are resolvable  $(n^2, n, 1)$ -BIBDs. While not all  $(n^2, n, 1)$ -BIBDs are resolvable, for example  $(36, 6, 1)$ -BIBD is not resolvable, it is known that given  $q$  a prime power, a  $(q^2, q, 1)$ -BIBD is always resolvable.

As an example, consider the following  $(9, 3, 1)$ -BIBD where  $\mathcal{A}$  represents a set of parallel classes of  $\mathbb{X}$ .

$$\begin{aligned} \mathbb{X} &= \{1, 2, 3, 4, 5, 6, 7, 8, 9\} \text{ and,} \\ \mathcal{A} &= \{(123, 456, 789), \\ &\quad (147, 258, 369), \\ &\quad (159, 267, 348), \\ &\quad (168, 249, 357)\}. \end{aligned}$$

*Remark 1.* We remark that (also from the definition of BIBDs and Property 1) each element is present in the same number of blocks and every couple of elements occurs once.

The following property (reported in Ray-Chaudhuri and Wilson (1973) ) holds.

**Property 3.** *Let  $k \geq 2$  be an integer. For  $v \geq k$ , there exists a resolvable  $(v, k, 1)$ -BIBD if and only if  $v \equiv k \pmod{k(k-1)}$ .*

The following proposition holds.

**Proposition 1.** *SDTP $_{4n}$  with  $4n \equiv 4 \pmod{12}$  constitutes a resolvable BIBD.*

*Proof.* We show that SDTP $_{4n}$  with  $4n \equiv 4 \pmod{12}$  is equivalent to a resolvable  $(4n, k, 1)$ -BIBD for  $k = 4$ . Indeed, for  $k = 4$ , the  $(4n, k, 1)$ -BIBD corresponds to a  $(4n, 4, 1)$ -BIBD where  $4n \equiv 4 \pmod{12}$  corresponds to  $4n \equiv 4 \pmod{4 \cdot 3}$  that is  $4n \equiv k \pmod{4k(k-1)}$ . Hence, the  $(4n, k, 1)$ -BIBD is resolvable. Also, by considering set  $\mathbb{X} = \{1, 2, \dots, 4n\}$  of  $4n$  elements, the requirement induced by Remark 1 on the division of the corresponding set of quadruplets  $\mathcal{A}$  into parallel classes, coincides to the set of constraints (1) – (3) of the ILP formulation of SDTP $_{4n}$  if each class is considered as a disjoint round of the tournament.  $\square$

As an example, SDTP $_{16}$  corresponds to a  $(16, 4, 1)$ -BIBD which is resolvable. SDTP $_{4n}$  is also strictly correlated to the *Social Golfer Problem* (SGP) (Triska and Musliu (2012)). In SGP, the aim is to schedule  $g \times p$  golfers in  $g$  groups of  $p$  players for  $w$  weeks so that no two golfers play in the same group more than once. We denote the problem as  $(g, p, w)$  SGP. The original problem asks for the maximal value  $w$  such that the instance can be solved.

The following proposition holds.

**Proposition 2.** *SDTP $_{4n}$  is equivalent to  $(n, 4, \frac{4n-1}{3})$  SGP. Correspondingly, the resolvable  $(4n, 4, 1)$ -BIBD is equivalent to  $(n, 4, \frac{4n-1}{3})$  SGP.*

*Proof.* In SDTP $_{4n}$ , there are  $4n$  players and  $\frac{4n-1}{3}$  disjoint rounds where each couple of players appears in a quadruplet exactly once and in each round the set of  $4n$  players is partitioned into  $n$  disjoint quadruplets. But then, if we substitute the players with the golfers, we consider  $\frac{4n-1}{3}$  weeks and  $g = n$  groups composed by  $p = 4$  players, we get a  $(n, 4, \frac{4n-1}{3})$  SGP. The second part of the proposition directly derives from the proof of Proposition 1.  $\square$

In [21], the solution of  $(7, 4, 9)$  SGP (equivalent to SDTP $_{28}$ ) is reported. Hence, by triplicating the corresponding disjoint rounds of SDTP $_{28}$  we get the related schedule shown in Table 3.

By applying GECCODE [22], a constraint programming solver, to the equivalent  $(16, 4, 21)$  SGP, we managed to generate a solution for SDTP $_{64}$  that is shown in Table 4 (due to space limitation, only disjoint rounds are depicted). The solver required few seconds only on a standard laptop machine. However, it was not able to solve neither SDTP $_{40}$  nor SDTP $_{52}$  within 3600 of CPU time.

	Match 1		Match 2		Match 3		Match 4		Match 5		Match 6		Match 7	
	Team 1	Team 2	Team 3	Team 4	Team 5	Team 6	Team 7	Team 8	Team 9	Team 10	Team 11	Team 12	Team 13	Team 14
Round 1	1 2	3 4	5 6	7 8	9 10	11 12	13 14	15 16	17 18	19 20	21 22	23 24	25 26	27 28
Round 2	1 3	2 4	5 7	6 8	9 11	10 12	13 15	14 16	17 19	18 20	21 23	22 24	25 27	26 28
Round 3	1 4	2 3	5 8	6 7	9 12	10 11	13 16	14 15	17 20	18 19	21 24	22 23	25 28	26 27
Round 4	1 5	2 1 2 5	2 6	13 17	14 18	22 26	7 9	19 27	8 10	15 23	3 11	16 28	4 12	20 24
Round 5	1 21	5 2 5	2 13	6 17	14 22	18 26	7 19	9 27	8 15	10 23	3 16	11 28	4 20	12 24
Round 6	1 25	5 2 1	2 17	6 13	14 26	18 22	7 27	9 19	8 23	10 15	3 28	11 16	4 24	12 20
Round 7	1 6	24 2 8	2 5	15 19	16 20	23 2 7	8 11	17 26	7 12	13 22	3 9	14 25	4 10	18 21
Round 8	1 24	6 2 8	2 15	5 19	16 23	20 2 7	8 17	11 26	7 13	12 22	3 14	9 25	4 18	10 21
Round 9	1 28	6 2 4	2 19	5 15	16 27	20 2 3	8 26	11 17	7 22	12 13	3 25	9 14	4 21	10 18
Round 10	1 9	17 2 3	2 10	14 28	5 1 1	13 2 4	6 12	18 2 7	16 19	21 26	3 7	15 20	4 8	22 2 5
Round 11	1 17	9 2 3	2 14	10 28	5 1 3	11 2 4	6 18	12 2 7	16 2 1	19 2 6	3 1 5	7 20	4 2 2	8 2 5
Round 12	1 23	9 1 7	2 28	10 1 4	5 2 4	11 1 3	6 2 7	12 1 8	16 2 6	19 2 1	3 20	7 1 5	4 2 5	8 2 2
Round 13	1 7	16 1 8	2 8	21 2 7	5 1 2	14 2 3	15 1 7	22 2 8	6 1 1	20 2 5	3 10	19 2 4	4 9	13 2 6
Round 14	1 16	7 1 8	2 21	8 2 7	5 1 4	12 2 3	15 2 2	17 2 8	6 20	11 2 5	3 19	10 2 4	4 13	9 2 6
Round 15	1 18	7 1 6	2 27	8 2 1	5 2 3	12 1 4	15 2 8	17 2 2	6 2 5	11 20	3 2 4	10 1 9	4 2 6	9 1 3
Round 16	1 11	19 2 2	2 12	16 2 5	6 9	15 2 1	5 10	20 2 6	14 1 7	24 2 7	3 8	13 1 8	4 7	23 2 8
Round 17	1 19	11 2 2	2 16	12 2 5	6 1 5	9 2 1	5 20	10 2 6	14 2 4	17 2 7	3 13	8 1 8	4 2 3	7 2 8
Round 18	1 22	11 1 9	2 25	12 1 6	6 2 1	9 1 5	5 26	10 20	14 2 7	17 2 4	3 18	8 1 3	4 28	7 2 3
Round 19	1 8	14 20	2 7	24 2 6	6 10	16 2 2	13 1 9	23 2 5	5 9	18 2 8	3 12	17 2 1	4 11	15 2 7
Round 20	1 14	8 20	2 24	7 2 6	6 16	10 2 2	13 2 3	19 2 5	5 18	9 2 8	3 17	12 2 1	4 15	11 2 7
Round 21	1 20	8 1 4	2 26	7 2 4	6 2 2	10 1 6	13 2 5	19 2 3	5 28	9 1 8	3 21	12 1 7	4 27	11 1 5
Round 22	1 12	15 2 6	2 11	18 2 3	7 10	17 2 5	13 20	21 2 8	8 9	16 2 4	3 5	22 2 7	4 6	14 1 9
Round 23	1 15	12 2 6	2 18	11 2 3	7 1 7	10 2 5	13 2 1	20 2 8	8 16	9 2 4	3 22	5 2 7	4 14	6 1 9
Round 24	1 26	12 1 5	2 23	11 1 8	7 2 5	10 1 7	13 2 8	20 2 1	8 2 4	9 1 6	3 27	5 2 2	4 19	6 1 4
Round 25	1 10	13 2 7	2 9	20 2 2	3 6	23 2 6	4 5	16 1 7	7 1 1	14 2 1	8 1 2	19 2 8	15 1 8	24 2 5
Round 26	1 13	10 2 7	2 20	9 2 2	3 2 3	6 2 6	4 1 6	5 1 7	7 1 4	11 2 1	8 1 9	12 2 8	15 2 4	18 2 5
Round 27	1 27	10 1 3	2 22	9 20	3 2 6	6 2 3	4 1 7	5 1 6	7 2 1	11 1 4	8 2 8	12 1 9	15 2 5	18 2 4

Table 3: A solution for SDTP<sub>28</sub>.

	Match 1 Match 9	Match 2 Match 10	Match 3 Match 11	Match 4 Match 12	Match 5 Match 13	Match 6 Match 14	Match 7 Match 15	Match 8 Match 16
<b>Round 1</b>	1 2 3 4 33 34 35 36	5 6 7 8 37 38 39 40	9 10 11 12 41 42 43 44	13 14 15 16 45 46 47 48	17 18 19 20 49 50 51 52	21 22 23 24 53 54 55 56	25 26 27 28 57 58 59 60	29 30 31 32 61 62 63 64
<b>Round 4</b>	1 5 9 13 33 37 41 45	2 6 10 14 34 38 42 46	3 7 11 15 35 39 43 47	4 8 12 16 36 40 44 48	17 21 25 29 49 53 57 61	18 22 26 30 50 54 58 62	19 23 27 31 51 55 59 63	20 24 28 32 52 56 60 64
<b>Round 7</b>	1 6 11 16 33 38 43 48	2 5 12 15 34 37 44 47	3 8 9 14 35 40 41 46	4 7 10 13 36 39 42 45	17 22 27 32 49 54 59 64	18 21 28 31 50 53 60 63	19 24 25 30 51 56 57 62	20 23 26 29 52 55 58 61
<b>Round 10</b>	1 7 12 14 33 39 44 46	2 8 11 13 34 40 43 45	3 5 10 16 35 37 42 48	4 6 9 15 36 38 41 47	17 23 28 30 49 55 60 62	18 24 27 29 50 56 59 61	19 21 26 32 51 53 58 64	20 22 25 31 52 54 57 63
<b>Round 13</b>	1 8 10 15 33 40 42 47	2 7 9 16 34 39 41 48	3 6 12 13 35 38 44 45	4 5 11 14 36 37 43 46	17 24 26 31 49 56 58 63	18 23 25 32 50 55 57 64	19 22 28 29 51 54 60 61	20 21 27 30 52 53 59 62
<b>Round 16</b>	1 17 33 49 9 25 41 57	2 18 34 50 10 26 42 58	3 19 35 51 11 27 43 59	4 20 36 52 12 28 44 60	5 21 37 53 13 29 45 61	6 22 38 54 14 30 46 62	7 23 39 55 15 31 47 63	8 24 40 56 16 32 48 64
<b>Round 19</b>	1 18 35 52 9 26 43 60	2 17 36 51 10 25 44 59	3 20 33 50 11 28 41 58	4 19 34 49 12 27 42 57	5 22 39 56 13 30 47 64	6 21 40 55 14 29 48 63	7 24 37 54 15 32 45 62	8 23 38 53 16 31 46 61
<b>Round 22</b>	1 19 36 50 9 27 44 58	2 20 35 49 10 28 43 57	3 17 34 52 11 25 42 60	4 18 33 51 12 26 41 59	5 23 40 54 13 31 48 62	6 24 39 53 14 32 47 61	7 21 38 56 15 29 46 64	8 22 37 55 16 30 45 63
<b>Round 25</b>	1 20 34 51 9 28 42 59	2 19 33 52 10 27 41 60	3 18 36 49 11 26 44 57	4 17 35 50 12 25 43 58	5 24 38 55 13 32 46 63	6 23 37 56 14 31 45 64	7 22 40 53 15 30 48 61	8 21 39 54 16 29 47 62
<b>Round 28</b>	1 21 41 61 9 29 33 53	2 22 42 62 10 30 34 54	3 23 43 63 11 31 35 55	4 24 44 64 12 32 36 56	5 17 45 57 13 25 37 49	6 18 46 58 14 26 38 50	7 19 47 59 15 27 39 51	8 20 48 60 16 28 40 52
<b>Round 31</b>	1 22 43 64 9 30 35 56	2 21 44 63 10 29 36 55	3 24 41 62 11 32 33 54	4 23 42 61 12 31 34 53	5 18 47 60 13 26 39 52	6 17 48 59 14 25 40 51	7 20 45 58 15 28 37 50	8 19 46 57 16 27 38 49
<b>Round 34</b>	1 23 44 62 9 31 36 54	2 24 43 61 10 32 35 53	3 21 42 64 11 29 34 56	4 22 41 63 12 30 33 55	5 19 48 58 13 27 40 50	6 20 47 57 14 28 39 49	7 17 46 60 15 25 38 52	8 18 45 59 16 26 37 51
<b>Round 37</b>	1 24 42 63 9 32 34 55	2 23 41 64 10 31 33 56	3 22 44 61 11 30 36 53	4 21 43 62 12 29 35 54	5 20 46 59 13 28 38 51	6 19 45 60 14 27 37 52	7 18 48 57 15 26 40 49	8 17 47 58 16 25 39 50
<b>Round 40</b>	1 25 45 53 9 17 37 61	2 26 46 54 10 18 38 62	3 27 47 55 11 19 39 63	4 28 48 56 12 20 40 64	5 29 41 49 13 21 33 57	6 30 42 50 14 22 34 58	7 31 43 51 15 23 35 59	8 32 44 52 16 24 36 60
<b>Round 43</b>	1 26 47 56 9 18 39 64	2 25 48 55 10 17 40 63	3 28 45 54 11 20 37 62	4 27 46 53 12 19 38 61	5 30 43 52 13 22 35 60	6 29 44 51 14 21 36 59	7 32 41 50 15 24 33 58	8 31 42 49 16 23 34 57
<b>Round 46</b>	1 27 48 54 9 19 40 62	2 28 47 53 10 20 39 61	3 25 46 56 11 17 38 64	4 26 45 55 12 18 37 63	5 31 44 50 13 23 36 58	6 32 43 49 14 24 35 57	7 29 42 52 15 21 34 60	8 30 41 51 16 22 33 59
<b>Round 49</b>	1 28 46 55 9 20 38 63	2 27 45 56 10 19 37 64	3 26 48 53 11 18 40 61	4 25 47 54 12 17 39 62	5 32 42 51 13 24 34 59	6 31 41 52 14 23 33 60	7 30 44 49 15 22 36 57	8 29 43 50 16 21 35 58
<b>Round 52</b>	1 29 37 57 9 21 45 49	2 30 38 58 10 22 46 50	3 31 39 59 11 23 47 51	4 32 40 60 12 24 48 52	5 25 33 61 13 17 41 53	6 26 34 62 14 18 42 54	7 27 35 63 15 19 43 55	8 28 36 64 16 20 44 56
<b>Round 55</b>	1 30 39 60 9 22 47 52	2 29 40 59 10 21 48 51	3 32 37 58 11 24 45 50	4 31 38 57 12 23 46 49	5 26 35 64 13 18 43 56	6 25 36 63 14 17 44 55	7 28 33 62 15 20 41 54	8 27 34 61 16 19 42 53
<b>Round 58</b>	1 31 40 58 9 23 48 50	2 32 39 57 10 24 47 49	3 29 38 60 11 21 46 52	4 30 37 59 12 22 45 51	5 27 36 62 13 19 44 54	6 28 35 61 14 20 43 53	7 25 34 64 15 17 42 56	8 26 33 63 16 18 41 55
<b>Round 61</b>	1 32 38 59 9 24 46 51	2 31 37 60 10 23 45 52	3 30 40 57 11 22 48 49	4 29 39 58 12 21 47 50	5 28 34 63 13 20 42 55	6 27 33 64 14 19 41 56	7 26 36 61 15 18 44 53	8 25 35 62 16 17 43 54

Table 4: A solution for SDTP<sub>64</sub> (only disjoint rounds).

## 4 Conclusions

The problem of scheduling a round robin tennis tournament of doubles matches where teams have to be generated so that every player must have every other player as a partner once and as an opponent twice has been considered in this paper. The problem requires also the quadruplet of every match to appear three times considering all possible combinations of partners and opponents. An integer programming formulation of the problem has been provided that can handle instances of limited size. The problem is also connected to combinatorial design and to the Social Golfer Problem, allowing to determine feasible schedules with up to 64 players where, however, instances with 40 and 52 players remain currently unresolved to the best of our knowledge. This peculiarity suggests as future research, that it is worthy to investigate whether instances of the problem with  $4^k$  players ( $k$  integer) may show up specific properties that induce an ad-hoc approach for that family of instances of SDTP.

## Acknowledgements

We would like to thank Massimiliano Lembo for having introduced us to SDTP and for the on court testing of our solution to a 16-player instance.

## References

- [1] Anderson I. (1995) *A hundred years of whist tournaments*, Journal of Combinatorial Mathematics and Combinatorial Computing **19**, 129-150.
- [2] Baker R. (1975) Whist tournaments. *Proceedings of Southeastern Conference on Combinatorics, Graph Theory, and Computing*, Florida Atlantic University, Boca Raton, FL, pp. 89-100
- [3] Blest, D.C., Fitzgerald D.G. (1988) *Scheduling sports competitions with a given distribution of times*, Discrete Applied Mathematics **22**, 9-19.
- [4] Colbourn, C.J., Dinitz, J.H. (2006) *Handbook of Combinatorial Designs*, 2nd edition, CRC Press.
- [5] Della Croce, F., Oliveri, D. (2006) *Scheduling the Italian Football League: an ILP-based approach*, Computers and Operations Research **33**, 1963-1974.
- [6] Dinitz, J.H., Froncek, D., Lamken, E.R., Wallis, W.D. (2006) Scheduling a tournament, in: C.J. Colbourn, J.H. Dinitz (eds.), *Handbook of Combinatorial Designs*, CRC Press.
- [7] Drexler, A., Knust, S. (2007) *Sports league scheduling: graph- and resource-based models*, Omega **35**, 465-471.
- [8] Duran, G., Guajardo, M., Miranda, J., Saure, D., Souyris, S., Weintraub, A., Wolf, R. (2007) *Scheduling the Chilean soccer league by integer programming*, Interfaces **37**, 539-552
- [9] Easton, K., Nemhauser, G., Trick, M. (2004) Sports scheduling, in: J.T. Leung (ed.): *Handbook of Scheduling*, CRC Press, 52.1-52.19.
- [10] Kendall, G., Knust, S., Ribeiro, C.C., Urrutia, S. (2010) *Scheduling in sports: An annotated bibliography*, Computers and Operations Research **37**, 1-19.
- [11] Mendelsohn E., Rosa A., (1985) *One-factorizations of the complete graph – a survey*, Journal of Graph Theory **9**, 43-65.

- [12] Moore E.H. (1896) *Tactical Memoranda I-III*, American Journal of Mathematics **18**, 264-303.
- [13] Nemhauser G.L., Trick M.A.(1998) *Scheduling a major college basketball conference*, Operations Research **46**, 1-8.
- [14] Rasmussen, R.V., Trick, M.A. (2008) *Round robin scheduling - a survey*, European Journal of Operational Research **188**, 617-636.
- [15] Ray-Chaudhuri, D.K., Wilson R.M. (1973) The existence of resolvable block designs, in: *Survey of combinatorial theory* (Proc. Internat. Sympos., Colorado State Univ., Fort Collins), 361-376.
- [16] Recalde, D., Torres, R., Vaca, P. (2013) *Scheduling the professional Ecuadorian football league by integer programming*, Computers and Operations Research **40**, 2478-2484.
- [17] Ribeiro, C.C. (2012) *Sports scheduling: problems and applications*, International Transactions in Operational Research **19**, 201-226.
- [18] Triska, M., Musliu, N. (2012) *An improved SAT formulation for the social golfer problem*, Annals of Operations Research **194**, 427-438.
- [19] Wright, M. (2009) *50 years of OR in sport*, Journal of the Operational Research Society **60**, 161-168.
- [20] [http://www2.inf.uos.de/knust/sportssched/sportlit\\_class/node1.html](http://www2.inf.uos.de/knust/sportssched/sportlit_class/node1.html)
- [21] [http://www.mathpuzzle.com/MAA/54-Golf%20Tournaments/mathgames\\_08\\_14\\_07.html](http://www.mathpuzzle.com/MAA/54-Golf%20Tournaments/mathgames_08_14_07.html)
- [22] <http://www.gecode.org/>

# The Championship Timetabling Problem - Construction and Justification of Test Cases

Jörn Schönberger

Technische Universität Dresden, Germany, joern.schoenberger@tu-dresden.de

## Abstract

We consider a sport league timetabling task from table tennis. The timetables for several leagues (forming the championship) have to be setup simultaneously. So-called inter-league constraints require a coordination of the slot assignment for the meetings belonging to different leagues. In particular, meetings with different teams of a club must be assigned into different slots. These inter-league timetabling constraints represent limited venue capacities and the need to preserve table tennis player substitution opportunities between the different teams of a club. We describe the definition and parameterization of test scenarios with different levels of difficulty. Within a Monte-Carlo-experiment we demonstrate the adequateness of these test instances to serve as test field for the evaluation of championship timetabling algorithms.

## 1 Introduction

Timetabling refers to the selection of one slot out of a finite set of given time slots. The selected slot determines the starting time of an event or of an operation. In the context of sports, such an event is a meeting between two opponent teams. Therefore, sports league timetabling refers to the determination of time slots for all meetings that take place according to the applied competition program. For example, the competition program of a round robin system of play requires that each pair of teams in a league meet exactly once.

In Germany, non-commercial but organized sport in non-profit-oriented clubs plays an important role in the sector of leisure activity. The organization of non-commercial sport leagues becomes more and more challenging. This is mainly caused by demographic changes, reduced availability of venues (in most of the times gymnasiums at public schools) as well as sportsmen and less volunteers who are willing to takeover the responsibility to organize club-based sport events.

One general approach to cope with the aforementioned challenges is to centralize planning competencies. There are three reasons why centralization of planning has come into the focus of public discussion in non-commercial sports. First, a reduced number of volunteers taking care about the organization of sport leagues is needed. Second, scarce venues can be used in a more efficient way and, third, the quality of event planning can be increased significantly since the remaining volunteers are motivated and often highly skilled. However, in order to be able to centralize organizational tasks it is necessary to collect all information relevant for planning. Here, internet-based information exchange platforms can be used today.

Table tennis is one of the most favorite sport in Germany. Here, according to the website of the German Tabletennis Association ([www.tischtennis.de](http://www.tischtennis.de)) more than 590,000 players regularly participate in organized individual and team competitions. More than 9,300 clubs delegate more than 45,000 teams. In the last years,



two additional challenges related with the determination of timetables for table tennis leagues have been revealed. First, the access to the venues becomes more and more restricted. The number of days that can be used for meetings of two opponents reduces more and more since the clubs have less money available or the venue is allocated for other activities. Consequently, all teams of a club compete for the remaining slots and conflicts occur if the managers of different leagues determine the timetables for their leagues independently. Second, the availability of sportsmen decreases more and more due to increased workload and/or numerous other private obligations. As a countermeasure, the player substitution opportunities have been extended in the last years. Members of lower ranked teams can be used to replace absent members of higher ranked teams if necessary to re-complete teams. However, in order to be able to exploit these substitution opportunities it is necessary to coordinate the meetings of different teams of a club participating in the championship in different leagues.

During the last decade, the authorities responsible for the organization and planning of the table tennis competitions in Germany have developed and installed a common information exchange platform for the table tennis sport in Germany in the internet `dttb.click-tt.de` and another coupled information website `www.mytischtennis.de`. More than 90% of all information regarding the organization of leagues as well as the collection of results are submitted via this platform. In addition, this platform is used by some regional associations to collect the data needed by the league manager(s) to setup a league timetable. Hence, a major prerequisite for the centralized and simultaneous timetable generation for several leagues in order to meet the two aforementioned challenges is fulfilled. However, there are no algorithms that can be used to solve this so-called championship timetabling problem.

Within this paper we report first research results dedicated to the simultaneous timetable generation for two or even more leagues. The limited capacity of venues as well as the need to preserve the player substitution opportunities create constraints that couple the timetable generation tasks in different leagues (inter-league constraints). Particular attention is paid to the construction and test of adequate test cases that can be used to evaluate adequate timetable generation algorithms. Here, we will propose a test case generation scheme that enables the specification of scalable test instances with different challenges. We identify those problem parameters that make an instance more or less challenging w.r.t. to the inter-league constraints.

The decision problem is discussed in Section 2. Section 3 reports the approach to define the aforementioned test problem instances. Section 4 presents a very simple randomized timetable generator. Section 5 reports about the setup of computational experiments and the corresponding numerical results.

## 2 The Championship Timetabling Problem

### 2.1 Literature

Most of the reported research addressing the determination of timetables for sport events focuses on professional sports (Kendall et al., 2010). The major driver behind these approaches is the economic need to maximize the attractiveness of the sport events. The events have to attract the maximal possible number of spectators into the venue. The generation of a maximal amount of revenues from sold tickets and catering is also in the focus of the organizers. Furthermore, the realization of a quite high TV audience rating (Andreff and Szymanski, 2006) is addressed. Another important scheduling goal is the need to cope

with limited resources needed to realize the sport event. Here, limited access to venues, the availability of players (Della Croce et al., 1999) or the availability of referees (Trick et al., 2012; Farmer et al., 2007) are prominent examples of scarce resources in the organization of sport events.

The organization of non-commercial sport leagues has received only minor attention so far compared to the professional sports. As an example we refer to the articles Knust (2010) as well as Della Croce et al. (1999).

Moody et al. (2010) investigate a situation in which timetables for several divisions in youth sports are determined for shared venues. Larson et al. (2014) report about approaches to solve a problem that comprises the schedule determination of several divisions. Schönberger (2015) proposes a mixed-integer linear program for a championship timetabling problem. In this contribution it is demonstrated within computational experiments that even the identification of a feasible problem solution of a very small instance with only two leagues comprising six teams each is a quite challenging and time consuming task. Hence, heuristic algorithms are more promising candidates to tackle larger instances of the championship timetabling problem.

## 2.2 Problem Statement

A set of leagues each with a competition program is given. These leagues form the championship. The competition program of a league contains all meetings  $(i, j)$  between two teams  $i$  as well as  $j$  of a league. We call  $i$  the home team in this meeting (or the host team) and team  $j$  the away team (or the guest team). All teams are collected in the set  $T$ .

It is necessary to select a time slot for each meeting in each league from a given set of slots called the season  $S$ . The season  $S$  is partitioned into the fall season  $S^{fall}$  as well as the spring season  $S^{spring}$ . All slots in the fall season precede the slots in the spring season. The home team  $i$  of a meeting has to provide the venue in which the meeting takes place. It can restrict the slot selection for its home meetings by specifying a subset  $H_i$  of the season. Only those slots contained in  $H_i$  can be used in the timetable for home meetings of team  $i$ . In order to consider the unavailability of sportsmen each team  $j$  specifies slots that cannot be used for a meeting that involves this team as the guest team. These slots form the set  $B_j$ . A feasible time slot for the meeting  $(i, j)$  has to be drawn from the set  $H_i \cap B_j^C$ . The set  $B_j^C$  denotes the complementary set of  $B_j$  and contains all slots that are not blocked by team  $j$ .

A feasible schedule of a single league fulfills the following four constraints:

- $R_0$ : For each meeting  $(i, j)$  with host team  $i$  and guest team  $j$  the selected slot must be an element of the set  $H_i \cap B_j^C$  (*feasible slot*).
- $R_1$ : For each team at most one meeting can be selected in a slot (*single slot usage*).
- $R_2$ : A slot has to be selected for each meeting in the competition programs of the leagues in the championship (*program covering*).
- $R_3$ : The meeting  $(i, j)$  has been assigned into a slot of the fall season if and only if the meeting  $(j, i)$  has been assigned into a slot of the spring season (*season alternation*).

A club is allowed to delegate one or more teams to the championship. These team can be members of different leagues. If two or even more teams of a club  $c$  participate in the leagues of the considered championship then only  $N_c^{max}$  home meetings of teams of club  $c$  can be assigned into a slot. If more than

this number of meetings be in the venue at the same time then the regular conduction of a meeting cannot be guaranteed. This is the first inter-league restriction to be considered.

All teams of a club are consecutively numbered with a team number. The best performing team is labeled by 1 and the best players of the club form this team. The second best team is numbered by 2, and so on. Players of a team labeled by  $i + 1$  are allowed to substitute regular players of the team numbered by  $i$  in case of sportsmen unavailability. In order to preserve this substitution opportunity it is not desired to select the same slot for two meetings in which the  $i$ -th as well as the  $i + 1$ -th team of the same club are involved. This is the second inter-league restriction to be considered.

In order to achieve the required coordination among the timetables of the leagues in the championship, the following two inter-league constraints must be fulfilled.

- $R_4$ : The number of meetings of a club  $c$  in its home venue in one slot must not exceed  $N^{max}$  (*venue capacity*)
- $R_5$ : If a meeting with the  $i$ -th team of a club is assigned into slot  $s$  then no meeting of the  $i + 1$ -th team of this club must be assigned into the same slot (*substitution opportunity*).

It is intended to distribute the meetings of a team regularly over the two seasons. Therefore, a least number of unused slots should be considered in a timetable. We have decided to formulate this requirement as the objective function that evaluates a timetable. Each team specifies the sets of allowed home meeting slots and blocked away meeting slots individually but, doing so, it cannot be guaranteed that the least required number of unused slots between two consecutive meetings of a team can be realized. Therefore, we count the number of times in which the desired number of unused slots between two consecutive meetings of a teams is not reached. This number is used to evaluate the set of timetables in the championship.

### 3 Generation of Test Cases for the Championship Timetabling Problem

The championship timetabling problem is a rather new problem. No adequate test cases are available to evaluate algorithms developed to simultaneously generate timetables for several leagues while the two aforementioned inter-league constraint types  $R_4$  as well as  $R_5$  are taken into account. Therefore, we decided to define suitable test problems. Particular attention is paid to the parameterization of the scenarios so that different levels of difficulty can be realized.

#### 3.1 Composition of the Leagues

We consider a championship that comprises three leagues. Their competition programs are double round-robin programs in all cases. Each league is formed by 10 teams so that 30 teams are delegated by 20 clubs. Each competition program now contains 90 meetings leading to 270 meetings to be assigned into slots of the season. For two of the three leagues we have defined a minimal number of one unused slot as necessary between two consecutive meetings of a team ( $N^{max} = 1$ ). In the remaining league no unused slot is expected to be inserted between two consecutive meetings of a team.

Two of the clubs delegate three teams each. These six teams are randomly distributed over the three leagues so that each of these two clubs delegates exactly one team into a league. Three clubs delegate two

teams each. Again, these six teams are randomly assigned to a league so that not more than one team per club becomes a member of a league. The remaining 18 clubs participate each with one team in the championship. These 18 teams are randomly distributed over the three leagues so that each league is filled up with ten teams.

In reality a club is allowed to delegate two or even more teams into the same league. In this case, a violation of the inter-league constraint  $R_5$  become often unavoidable. Therefore, the check of  $R_5$ -feasibility between a pair of two teams of a league must be extended by some exception rules. However, in order to keep the presentation as clear as possible, we have decided to avoid situations in which two or even more teams of a club are found in the same league.

### 3.2 General Setup of the Season

The complete championship comprises 90 non-overlapping slots into which meetings can be positioned. Each meeting in the competition programs of the leagues in the considered championship has to be assigned into one of these slots. We number the slots by 0 up to 89. The first 45 slots, numbered by 0 to 44 form the first season  $S^{fall}$  (fall season) while the remaining 45 slots, numbered by 45 to 89 form the second season  $S^{spring}$  (spring season).

Both seasons  $w \in \{1, 2\}$  are split up into three equidistant league-periods  $P_l^w$ . Within the league period  $P_l^w$  only meetings of league  $l$  are allowed to be positioned. We specify the league periods  $P_1^1 := \{0, 1, \dots, 14\}$ ,  $P_2^1 := \{15, 16, \dots, 29\}$ ,  $P_3^1 := \{30, 31, \dots, 44\}$ ,  $P_1^2 := \{45, 46, \dots, 59\}$ ,  $P_2^2 := \{60, 61, \dots, 74\}$  as well as  $P_3^2 := \{75, 76, \dots, 89\}$ . For example, the meetings of league 1 in the fall season have to be assigned into the slots  $P_1^1 := \{0, 1, \dots, 14\}$  but the meetings of this leagues in the spring season must all be put into the slots from  $P_2^2 := \{60, 61, \dots, 74\}$ . We will exploit this strict separation of slots for meetings of different leagues to define a parameter that represents the the level of difficulty of an instance w.r.t. the two inter-league constraints.

### 3.3 Determination of Tentative League-Timetables in the Unconstrained Case

We setup a tentative timetable for each league in order to define the sets  $H_t$  as well as  $B_t$  for each team  $t$ . The leagues are treated one after another. For each league, we first decide randomly about the distribution of the meetings  $(i, j)$  and  $(j, i)$  among the fall and the spring season but ensure that the number of meetings in both seasons of this league is equal. Therefore, constraint  $R_3$  is fulfilled. After that, we know the two competition programs of each league for both the fall as well as the spring season. The inter-league constraints  $R_4$  as well as  $R_5$  are fulfilled since each league uses slots from its own league period only. All slots are interpreted as available for all meetings in the competition program so that constraint  $R_0$  is fulfilled. All meetings are assigned to a slot so that also  $R_2$  is fulfilled. In order to ensure that  $R_1$  is also fulfilled, we first group all meetings in rounds. Each round contains the maximal number of meetings of a league so that each team of this league is involved in at most one meeting of round. For example, the meetings  $(1, 2)$ ,  $(3, 4)$ ,  $(5, 6)$ ,  $(7, 8)$  and  $(9, 10)$  form a round. We therefore need 9 slots (having 15 available) since we have 9 rounds. Next, for each round, we randomly select a slot and assign all meetings of this round to the selected slot. Each slot is selected only once. After we have defined a slot for each round we randomly shift the matches to other slots taking into account that no constraint violations is induced.

### 3.4 Selection of Home Meeting Slots

We select  $\alpha$  slots from each of the two corresponding league periods into which meetings as the host team can be assigned. These slots form the set  $H_t$  of possible home meeting slots specified by team  $t \in T$ . In a first step, we put all those slots into  $H_t$  in which a meeting with team  $t$  as host takes place. The remaining slots are randomly selected from the so far unselected slots of team  $t$  independently whether these slots are already used by team  $t$  in an away meeting but these slots are taken from the corresponding league period only. For large  $\alpha$ -values the complete league period is available to schedule meetings in the venue of a team.

### 3.5 Selection of Blocked Slots for Away Meetings

After we have selected the possible home meeting slots for team  $t$ , we proceed with the selection of those  $\beta$  slots that are blocked for away meetings of team  $t$ . We draw these slots from the remaining slots that have not yet been used as away meeting slots in the tentative timetables generated in Subsection 3.3. If the number of these slots is less than  $\beta$ , then we select the remaining slots randomly from the corresponding league period.

### 3.6 Mixing the Slots of Several Leagues

No club delegates more than one team into a league (see the remark at the end of 3.1). All meetings of different leagues are assigned into slots of different league periods. Therefore, we can be sure that no inter-league constraints will be violated in the tentative timetables.

With respect to the evaluation of multi-league scheduling algorithms, it is necessary to provoke such constraint violations in a controllable fashion. Therefore, we enrich the set of allowed home meeting slots  $H_t$  for each team  $t$  of a club  $c$  by those slots that are used in other leagues by the other teams of club  $c$ . Let  $t$  be a team and  $c_t$  be the club that delegates  $t$ . We collect the slots in which the other teams of club  $c_t$  are scheduled in the set  $S_{c_t}(t)$ .  $\gamma$  slots from  $S_{c_t}(t)$  are arbitrarily selected and added to the set  $H(t)$  of allowed home meeting slots. If  $\gamma$  is less than  $|S_{c_t}(t)|$  we select the remaining  $|S_{c_t}(t)| - \gamma$  possible home meeting slots randomly from the remaining slots that are already used.

### 3.7 Parameterization

After we have determined the procedure to generate test instances with different numbers and collections of possible home meeting slots and blocked slots for away games, we can use it to generate test instances with different parameter values. We generate instances with  $\alpha \in \{5, 7, \dots, 15\}$  possible home meeting slots. Furthermore, we select the values of the number of blocked home games  $\beta$  from  $\{0, 2, \dots, 10\}$ .

In order to setup test scenarios with different levels of difficulty w.r.t. to inter-league constraints we vary the number  $\gamma$  of slots taken from foreign league periods. We select the  $\gamma$ -values from the set  $\{0, 2, 4, \dots, 20\}$ . All random drawings of slots are repeated with ten different seeding values  $\omega$ . We define  $|\{5, 7, \dots, 15\}| \cdot |\{0, 2, \dots, 10\}| \cdot |\{0, 2, 4, \dots, 20\}| \cdot |\{0, 1, \dots, 9\}| = 6 \cdot 6 \cdot 11 \cdot 10 = 3960$  different test championships each represented by the quadruple  $(\alpha, \beta, \gamma, \omega)$ .

## 4 Monte Carlo Timetable Generation Approach

In order to prove that these test championships represent real challenges of different levels of difficulty by varying the parameter  $\gamma$ , we have setup initial computational experiments with a quite simple timetabling algorithm.

We use a simple random timetable generating procedure to draw slots for all meetings in the championship. The main purpose of the application of this procedure is not to minimize the objective function value. Instead, we aim at demonstrating that the insertion of additional slots already used by other teams of a club increases the chance to generate violations of inter-league constraints.

The randomized timetabling procedure works as follows: At first, we randomly assign the two meetings  $(i, j)$  as well as  $(j, i)$  to the two seasons. Next, we draw a slot at random from the right season for the meeting  $(i, j)$ . We give preference to those slots that are declared as potential home meeting slots by the host team  $i$ , which are not blocked by the guest team  $j$ , and which is not yet used neither by team  $i$  nor by team  $j$ . If such a slot does not exist anymore, we select a blocked slot in order to avoid that two meetings of a team are assigned into the same slot. We proceed similarly for the meeting  $(j, i)$  in the other season. Consequently, we cannot prevent the appearance of violations of constraints corresponding to  $R_0$  if  $\gamma > 0$ .

No additional efforts are made to avoid violations of the inter-league constraints associated with  $R_4$  as well as  $R_5$  since we want to inspect the frequency of their appearance after a variation of  $\gamma$ .

## 5 Experiments

### 5.1 Setup of Computational Experiments

We apply the random timetable generator with five different seeding values to all 3960 test championships so that overall 19800 individual experiments are conducted. For each experiment, we save the following performance indicator values. The number of violations of constraint  $R_i$  ( $i = 0, 4, 5$ ) are saved in  $r_i$  and the objective function value of the generated set of timetables is stored in  $obj$ .

We first calculate the average values  $r_i(\alpha, \beta, \gamma)$  ( $i = 0, 4, 5$ ) for each combination of the three parameters  $\alpha, \beta, \gamma$  that determine the type of the considered test championship. In the same way, the average objective function value  $obj(\alpha, \beta, \gamma)$  is calculated from the observed objective function values.

Next, we scale the values of all indicators into the interval  $[0, 1]$  in order to prepare a comparative presentation of the result. We use the formula (1) to map the  $obj(\alpha, \beta, \gamma)$ -values into the interval  $[0, 1]$  getting the values  $obj^*(\alpha, \beta, \gamma)$ . In the same way we calculate  $r_0^*, r_4^*, r_5^*$ .

$$obj^*(\alpha, \beta, \gamma) = \frac{obj(\alpha, \beta, \gamma) - \min_{\alpha, \beta, \gamma} obj(\alpha, \beta, \gamma)}{\max_{\alpha, \beta, \gamma} obj(\alpha, \beta, \gamma) - \min_{\alpha, \beta, \gamma} obj(\alpha, \beta, \gamma)} \quad (1)$$

### 5.2 Results

We have conducted the aforementioned experiments with the goal to show that a variation of the parameter  $\gamma$  leads to a variation of the level of difficulty of a test case. Therefore, we study now the achieved results for the three different  $\gamma$ -values 0 (expected low chance for inter-league constraint violations), 10 (expected

moderate chance for inter-league constraints violations ) as well as 20 (expected high chance for inter-league constraints violations).

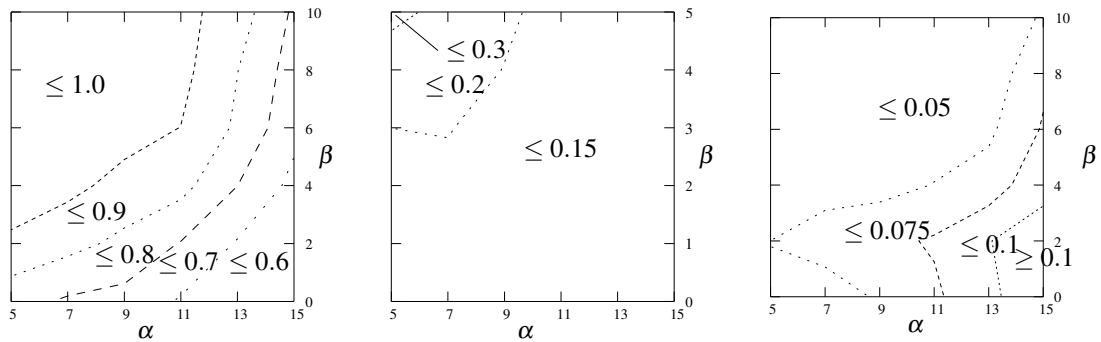


Figure 1: Scaled objective function values  $obj^*(\alpha, \beta, \gamma)$  for  $\gamma = 0$  (left plot),  $\gamma = 10$  (middle plot) as well as  $\gamma = 20$  (right plot)

At first, we analyze the objective function values. The scaled average objective function value decreases if we increase the sets of allowed home meeting slots (increase of  $\gamma$ ) as shown by the plots in Fig. 1. Without adding any slot from another league-period ( $\gamma = 0$ , left plot in Fig. 1) the scaled objective function values varies between 30% and 100%. After the insertion of  $\gamma = 10$  additional slots allowed to host home meetings (middle plot in Fig. 1), the objective function values are significantly lower and vary between 5% and 30%. If the set of allowed home meeting slots has been enriched by  $\gamma = 20$  slots then the scaled objective function values remain below 15%. We see that the additional home game slots support the generation of timetables with sufficient slots between two consecutive meetings of a team.

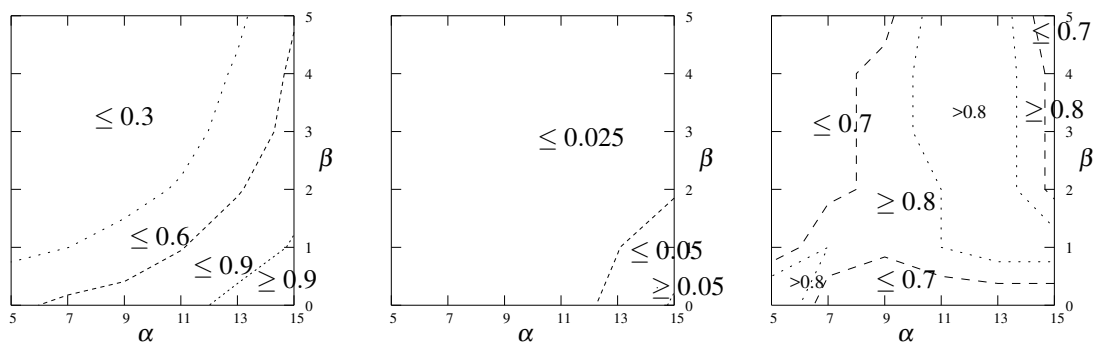


Figure 2: Scaled number  $r_0^*$  of violations of constraint  $R_0$  for  $\gamma = 0$  (left plot),  $\gamma = 10$  (middle plot) as well as  $\gamma = 20$  (right plot)

The number of violations of constraint  $R_0$  (use of a blocked slot) is quite high if  $\gamma = 0$  (left plot in Fig. 2). We observe scaled constraints violation numbers up to 100%. Also for  $\gamma = 20$ , we observe quite

high  $r_0^*$ -values of more than 80%. For  $\gamma = 10$ , only few violations of  $R_0$  are detected. For all three  $\gamma$ -values, we observe the most violations of  $R_0$  if the number of home meeting slots in the original league-period is high ( $\alpha \geq 13$ ) and few slots are blocked for away meetings ( $\beta < 1$ ). Here, the random timetable generation procedure selects slots from the wrong league-period with high frequency.

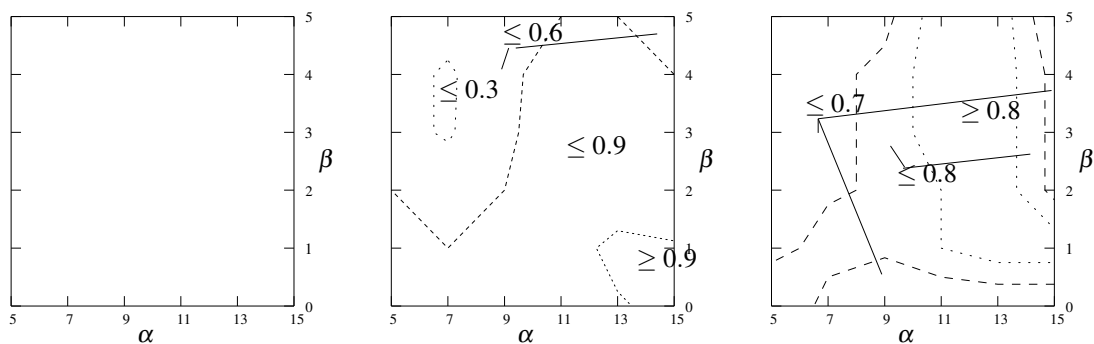


Figure 3: Scaled number  $r_4^*$  of violations of constraint  $R_4$  for  $\gamma = 0$  (left plot),  $\gamma = 10$  (middle plot) as well as  $\gamma = 20$  (right plot)

Obviously, in case that we do not add home meeting slots to  $H_i$  that could cause a capacity exceeding in a venue, no violations of restriction  $R_4$  appear (left plot in Fig. 3). However, The insertion of a moderate number of potential conflicting slots makes the championship timetable generation much more complicated. A quite high number of venue capacity exceeding is detected for  $\gamma = 10$  (middle plot in Fig. 3). The capacity exceeding is also observed in the scenarios with  $\gamma = 20$ . The maximal number of these violations is smaller compared to the  $\gamma = 10$ -scenarios. However, if  $\gamma = 20$  then, independently of  $\alpha$  and  $\beta$ , the least scaled number of violations of  $R_4$  is 60%. In the scenarios with  $\gamma = 10$  the scaled number of violations of  $R_4$  falls below 30% in some scenarios. This observation supports the assumption that an increase of  $\gamma$  lifts the level of difficulty of the timetable generation for the complete championship.

The analysis of the number of violations of  $R_5$  (preservation of substitution opportunities) presented in Fig. 4 reveals similar impacts of the increase of  $\gamma$  with respect to the second inter-league constraint.

In summary, the observed experiments indicate that the enrichment of the number of allowed home meeting slots with slots already used by other teams of a club makes the championship timetabling task more challenging. Therefore, The proposed suite of artificial test instances seems to be adequate to serve for the evaluation of algorithms designed to solve the championship timetabling problem.

## 6 Conclusions and Outlook

We have discussed the championship timetabling problem in which timetables for several leagues must be determined simultaneously. Inter-league constraints prevent the decoupling of the timetable generation tasks for the leagues. We have given a semi-formalized problem description in which a planning objective function as well as the the intra-league as well as the inter-league planning restrictions are emphasized. With respect to the newly introduced inter-league constraints we have derived a comprehensive set of test problem instances.



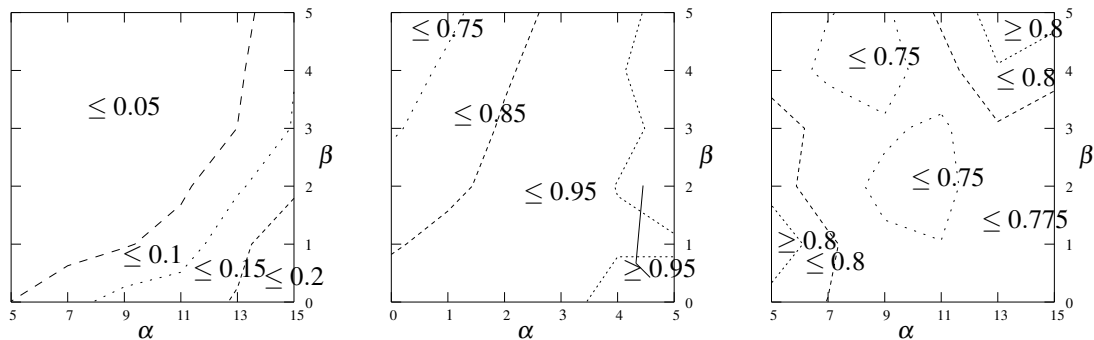


Figure 4: Scaled number  $r_5^*$  of violations of constraint  $R_5$  for  $\gamma = 0$  (left plot),  $\gamma = 10$  (middle plot) as well as  $\gamma = 20$  (right plot)

It could be demonstrated for a Monte-Carlo-approach-like random timetabling algorithm that the proposed parameterization of the test problem instances leads to instances with different difficulty levels. From the observed results we conclude that the proposed test cases are adequate to serve as base for the evaluation of more sophisticated algorithms developed to solve the championship timetabling problem.

## References

- Andreff, W., Szymanski, S., 2006. Handbook on the Economics of Sport. Edward Elgar Publishing.
- Della Croce, F., Tadei, R., Ascoli, P., 1999. Scheduling a round robin tennis tournament under courts and players availability constraints. *Annals of Operations Research* 92 (0), 349–361.
- Farmer, A., Smith, J. S., Miller, L. T., 2007. Scheduling umpire crews for professional tennis tournaments. *Interfaces* 37 (2), 187–196.
- Kendall, G., Knust, S., Ribeiro, C. C., Urrutia, S., 2010. Scheduling in sports: An annotated bibliography. *Computers & Operations Research* 37 (1), 1 – 19.
- Knust, S., 2010. Scheduling non-professional table-tennis leagues. *European Journal of Operational Research* 200 (2), 358 – 367.
- Larson, J., Johansson, M., Carlsson, M., 2014. An integrated constraint programming approach to scheduling sports leagues with divisional and round-robin tournaments. In: Simonis, H. (Ed.), *Integration of AI and OR Techniques in Constraint Programming: 11th International Conference, CPAIOR 2014, Cork, Ireland, May 19-23, 2014. Proceedings*. Springer International Publishing, pp. 144–158.
- Moody, D., Noy, A., Kendall, G., 2010. Youth sports leagues scheduling. In: MCCollum, B., Burke, E., White, G. (Eds.), *Proc. PATAT 2010*. pp. 283–293.
- Schönberger, J., 2015. Scheduling of sport league systems with inter-league constraints. In: Kay, A., Owen, A., Halkon, B., King, M. (Eds.), *Proc. MathSport International 2015*. pp. 171–176.
- Trick, M. A., Yildiz, H., Yunes, T., 2012. Scheduling major league baseball umpires and the traveling umpire problem. *Interfaces* 42 (3), 232–244.

# What Performance Data Tells Us about PEDs in Olympic Athletics and Swimming

Ray Stefani

California State university, Long Beach, USA  
Raystefani@aol.com

## Abstract

A statistical analog to the biological passport is used to identify possible uses of performance enhancing drugs among Olympic champions in athletics and swimming and to identify the effect of anti-doping efforts. During the 28 years from 1960-1988, athletics Olympic champions improved 21% in throwing events, 12% in jumping events and 5% in running events. After 28 years under increased anti-doping introduced in 1988 and with the biological passport introduced in 2009, Olympic throwing champions were 2% worse while jumping and running performances were about the same. Female champions improved 14% from 1960-1988, compared to 7% for men. From 1988-2016, women were 1.4% worse while men were about the same. The winners were so juiced with PEDs in 1988 that after an entire generation under anti-doping, champions in athletics are slight worse than in 1988. In swimming, the 1972 and 1976 winners showed unusually large percent improvements, coincident with the introduction of goggles (dramatically increasing training time) and new events. Complaints of PED use by East German women appears justified in that male and female champions had improved the same from 1960-1972; but female champions, mostly from East Germany, improved 1.5% more than men in 1976. Post 1976, male and female Olympic champions improved at the same rate. In 1996 after anti-drug efforts began in swimming, for the first time in Olympic history, swimming champions were worse than four years before. Janet Evans complained that Ireland's Michelle Smith must have been taking PEDs, after Smith improved and won three gold medals with mediocre times. The fact that an elite swimmer assumed that performances ought to be worse in 1996 indicates that elite swimmers probably had ceased taking some form of PED, as Mark Spitz hinted. The fact that only one medal in swimming has been stripped compared to 47 in athletics may belie subtle PED use in swimming.

## 1 Introduction

Efforts to catch cheaters who use performance enhancing drugs (PEDS) in the Olympics and to take away their medals only spans about the last 50 years of the nearly 2,800 years of Olympic history from 776 BC. At the Ancient Games, athletes took hallucinogens, opium juice and strychnine, [5], [7]. Strychnine was used as late as 1908 when Thomas Hicks won the marathon, yet nearly died for the effects of that drug, [10]. Later-to-be-US-General George

Patton took an injection of opium prior to the running portion of the 1912 Olympic Modern Pentathlon (in which he had his best placement), [10].

The International Olympic Committee (IOC) formed a Medical Commission in 1967, charged with the goal of catching PED cheaters, [5]. That duty moved to the World Anti-Doping Agency (WADA) in 1999. From 1967 until 2009, athletes were tested directly for any PED on a list of prohibited substances. Beginning with 2009, the “biological passport” was instituted in which regular samples are taken. Technicians look for changes in markers that indicate that a PED had been taken, even though the dosage is no longer out-of-bounds. Agents in the field report any comments or actions that suggest an athlete ought to be investigated, [9].

As of 10 Feb. 2017, 129 medals had been stripped, [3]: athletics (47), weightlifting (45), cross country skiing (9), wrestling (9), cycling (5), equestrian (3), gymnastics (2), shooting (2), and one each in alpine events, biathlon, boxing, judo, modern pentathlon, rowing and swimming (Greg DeMont in 1972, who could now obtain a therapeutic exemption for an asthma medication).

This paper will employ a “statistical passport” as an analog to the biological passport. Given that about 30% of the 300 Summer Olympic events are in the very popular sports of athletics and swimming, those two sports will be analyzed. The statistical marker will be the historical progression of the percent improvement per Olympiad ( $%I/O$ ) of Olympic champions. A search will be made for oddities not explainable by known historical events such as wars and boycotting as well as sport-specific (non-drug) changes causing improvement. We will also look for statements made by athletes that arouse suspicion and then examine  $%I/O$  in detail.

We define  $%I/O$  as the percent by which measured winning distance is increased over an Olympiad or elapsed winning time is decreased. If  $x_n$  and  $x_{n+1}$  represent the winning performances in Olympics  $n$  and  $n+1$  respectively, then  $%I/O$  is given by  $100 (x_{n+1}/x_n - 1)$  for an event like the shot put while for a timed event we take the negative of that term,  $100 (1 - x_{n+1}/x_n)$ .

## 2 Athletics

$%I/O$  is displayed in Figure 1 for the span of the modern Games (women first competed in 1928). The variable-terrain marathon and walking events are not included. A large value is understandable for 1900 when the Olympics expanded from an invitational competition to a world phenomenon. There were improvements in  $%I/O$  building to WW1, a drop in 1920 caused by the interruption of training and a rebound in 1924. That pattern repeated through WW2 with a rebound in 1952. For 1956-1976, the ensuing Cold War and national pride led to significant improvements with slowly diminishing values as we would expect of extended competition under the same conditions, leading to decreased returns. The boycotting of 1980 and 1982 caused fluctuation. The rebound in 1988 is expected; but, the size of the improvement is suspiciously large. It was a watershed moment in 1988 when Ben Johnson of Canada lost his 100 m world record and gold medal due to PED use, causing the IOC to ramp up anti-PED efforts. The result was the negative improvement in 1992 followed by small positive improvements through 2008, after which the biological passport was introduced. The ensuing negative improvement in 2012 was followed by a small recovery in 2016. Obviously,

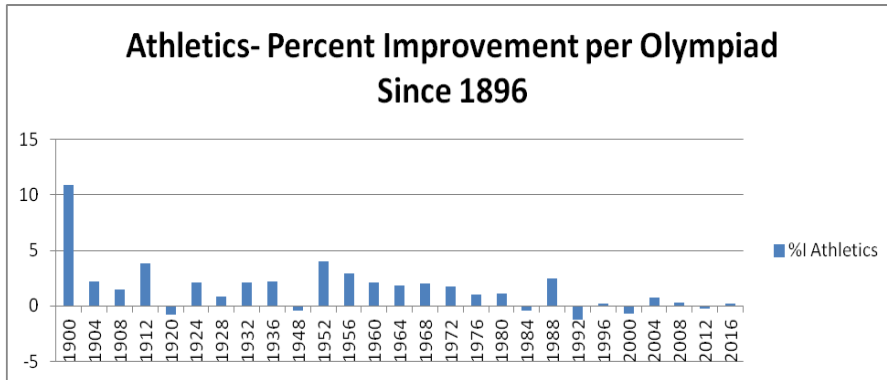


Figure 1

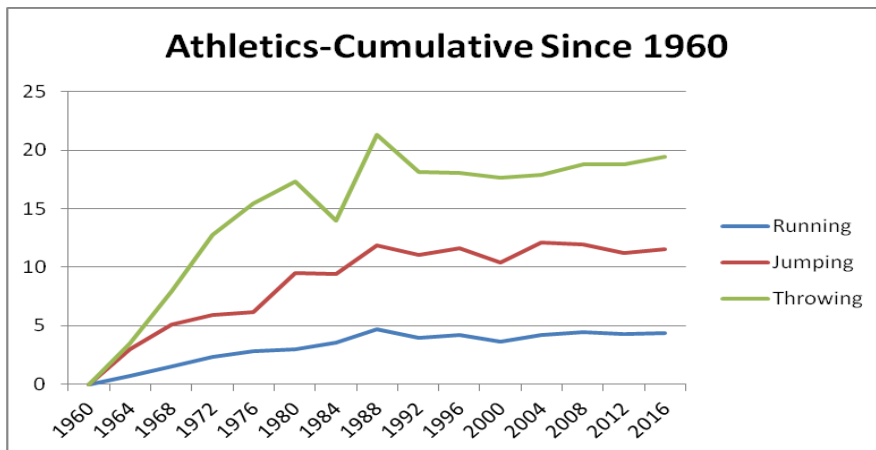


Figure 2

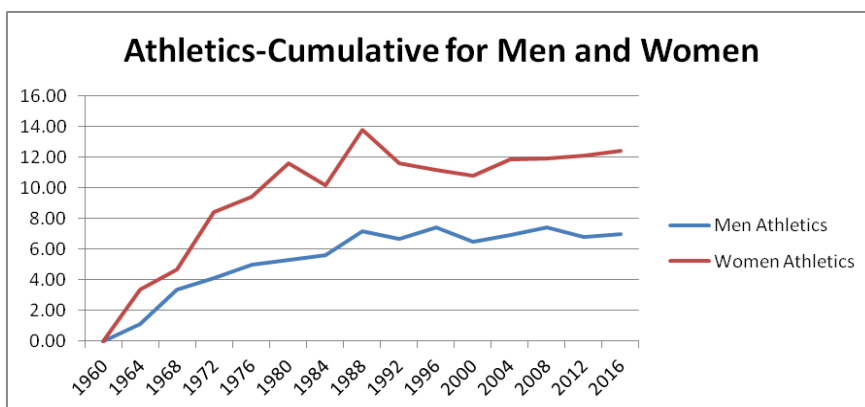


Figure 3

there was a large negative effect of anti-PED efforts just after 1988 with a smaller but visible negative effect after 2008.

Figures 2 and 3 are intended to evaluate the effects of anti-PED activity in 1988 by plotting cumulative improvement, thereby contrasting the 28 years prior to 1988 with the 28 years after 1988. In Figure 2, during the 28 years from 1960-1988, athletics Olympic champions improved 21% in throwing events, 12% in jumping events and 5% in running events. After 28 years under increased anti-doping introduced in 1988 and with the biological passport introduced in 2008, Olympic throwing champions became 2% worse while jumping and running performances were about the same. The strength-related throwing events were more negatively affected under more-stringent anti-PED efforts.

In Figure 3, female athletics champions improved 14% from 1960-1988, compared to 7% for men. From 1988-2016, women were 1.4% worse while men were about the same. Thus, female athletics champions were more negatively affected under more-stringent anti-PED efforts than their male counterparts.

Figures 1-3 tell us that the winners were so juiced with PEDs in 1988 that after an entire generation under anti-doping, Olympic champions in athletics were slightly worse in 2016 than in 1988.

### 3 Swimming

The first three Olympic swimming competitions are not useful for an historical study as the first competition in 1896 was in a harbor with treacherous waves, the second was held in the Seine, swimming with the current, while the third, in 1904, was in a pool, using yards and not meters. Thus, Figure 4 shows the history of  $%I/O$  in swimming starting in 1908, when performances were first reasonably comparable to today. Women first competed in 1912. Unlike athletics, WW1 and WW2 did not cause negative improvement in swimming. As with athletics, national pride urged on by the cold war lead to understandably consistent improvement from 1952 to 1968, followed by suspiciously large improvements in 1972 and 1976. Those two years warrant a closer look. Along with understandably lower  $%I/O$  during the 1980 and 1984 boycott years, it is reasonable to have smaller  $%I/O$  values in recent years in a mature sport, but it is indeed surprising to have the only incidence of a negative improvement in swimming in 1996 at Atlanta, during an Olympic year not beset by the influences of a war or by boycotting. Attention now focuses on 1972, 1976 and 1996.

Are there sport-specific conditions that might explain the large improvements in 1972 and 1976? Two possible factors come to mind. First, from 1960 to 1968, the number of events about doubled from 15 to 29. Larger-than-average improvements can be expected as new events are contested for the second and third times in 1972 and 1976. Second, swimming goggles were developed in 1968, were first used in competition in 1970 and were first used in the Olympics in 1976, [1]. Goggles diminished the detrimental effects of chlorinated water on the eyes, dramatically increasing practice time which would improve performances by enhancing training and therefore power output, as was the case with the rowing machine, [6]. We then expect short-term larger-than-normal improvements due to new events and goggles, as seen in 1972 and 1976.

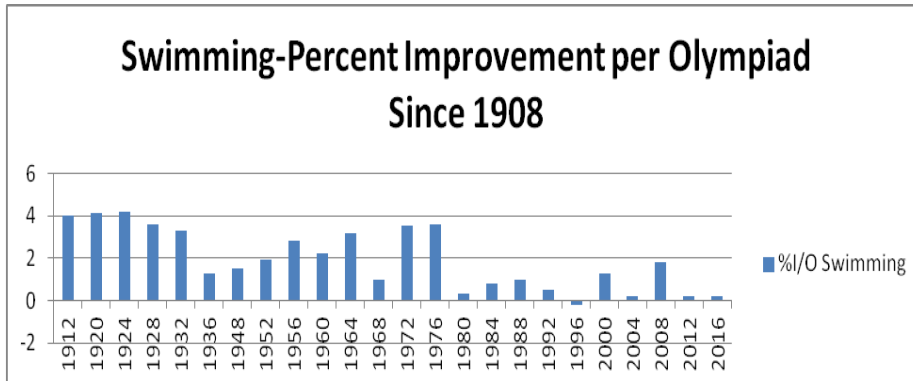


Figure 4

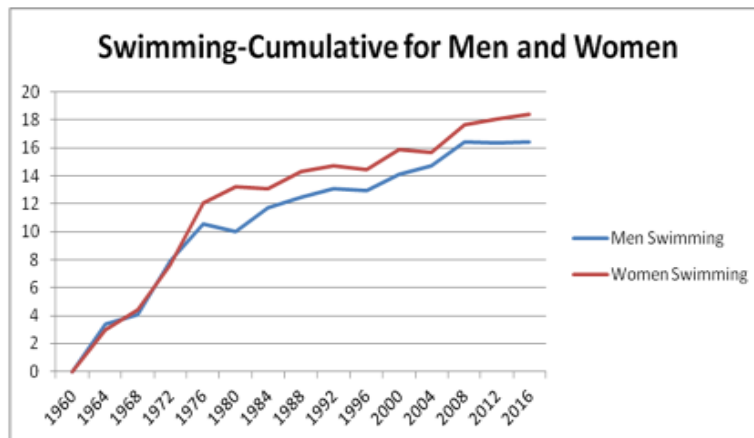


Figure 5

Event	Michelle Smith's Gold in 1996	Gold in 1992	Gold in 1988	Gold in 1984
400 m freestyle	4:07.25	4:07.18	4:03.85	4:07.10
20 m IM	2:13.99	2:11.65	2:12.59	2:12.64
400 m IM	4:39.18	4:36.54	4:37.76	4:39.24

Table 1 Michelle Smith's Winning Times in 1996 Compared to Winning Times in Previous 3 Games

There is a darker side to 1976. East German female swimmers dominated the medal count in 1976, after which American swimmer Shirley Babashoff complained about their training methods, [4]. Rutemiller [4] went on to report a retrospective “One hundred and sixty-seven former East German (DDR) athletes will be financially compensated through Germany’s Olympic Committee for the systematic doping of DDR athletes from 1973 through 1989. When told of this fact, Shirley Babashoff’s first comment was, ‘Only 167 Athletes!’.”

Can we expose the East German women’s unusual success statistically? Figure 5 plots the cumulative improvement of male and female Olympic champions over 56 years, as was done in Figure 3 for athletics. Male and female Olympic swimming champions improved at virtually identical rates from 1960 through 1972, but female champions improved 1.5% more than men in 1976, with East German women taking home most gold medals. With East German women competing but Western men not competing, women improved more than men in 1980, followed by the opposite situation in 1984. As all nations competed at Seoul in 1988, women and men were about 1.5% apart as in 1976. After 1988, men and women followed parallel paths of improvement. It does indeed appear that the East German women achieved an historically usual improvement in 1976, as Babashoff implied.

We move to the 1996 Games at Atlanta where swimmers performed worse than in 1992, coincident with stepped-up anti-PED activity. Lupica [2] wrote “Janet Evans, who handed the Olympic torch to Muhammad Ali at the end of the opening ceremonies, hinted that a faster swimmer, Ireland’s Michelle Smith, got faster because she uses performance-enhancing drugs. ‘It’s questionable . . . suspicious,’ Evans said. Evans finished ninth in the morning heats of the 400-meter freestyle. Last night, Smith ended up with Evans’ gold medal in the 400 freestyle and laughed when asked about drugs afterward. ‘I think it’s funny, actually,’ she said. ‘I’m the most tested Irish athlete . . . if you want to say [I’m using] drugs, look at all my drug tests’.”

Table 1 contains the suspicious (to Evans and other elite female swimmers) three gold-medal performances of Smith in 1996. Smith’s 400 m time was worse than Evan’s winning time in 1992. Evans had only to match her 1992 time to win but she was worse. None of Smith’s three winning time would have won in 1992 and 1988. Only one time would have won in 1984, in the 400 m IM, and that would only have been by 0.06 seconds. Smith performed at a personal best with historically mediocre (retro) times. The evidence suggests, albeit circumstantially, that Evans spoke for other elite athletes who had changed some training regimen, giving reason to suspect someone who performed at a personal best, rather than perform at a reduced level of proficiency.

Sixteen years later, in [8], Mark Spitz is quoted as saying “is the IOC testing for every drug that is performance enhancing drug out there? And the answer is, of course they are not and we know as a fact that a lot of the performance enhancing drugs that they test for on that list are old school. That they haven’t been taken by the elite athletes for years.” Spitz was talking about the then performances of Chinese swimmers. He seems to be indicating that swimmers had taken “old-school PEDs” but discontinued the practice. The negative performances of 1996 are suspiciously stationed right after the “old-school PEDs” became tested for.

## 4 Conclusions

Of the 129 Olympic medals stripped for PED use as of February 2017, 47 were in athletics. Following the crackdown on PED use in 1988 and the introduction of the biological passport in 2009, winning performances in athletics declined remarkably. During the 28 years before 1988, athletics Olympic champions improved 21% in throwing events, 12% in jumping events and 5% in running events. In the 28 years, post-1988, throwing champions became 2% worse while jumping and running performances were about the same. Female athletics champions improved 14% in the 28 years pre-1988, compared to 7% for men. Post-1988, female champions women were 1.4% worse while men were about the same. As of the 2016 Olympics, an entire generation had gone by without improving over 1988, indicating how prevalent the use of PEDs was in 1988.

The fact that only one medal in swimming has been stripped compared to 47 in athletics may belie subtle PED use in swimming. Led by the East German women, female swimming champions in 1976 showed a unique gain of 1.5% in improvement compared to men, a value never repeated. Later reporting indicated rampant use of PEDs in East Germany, although none were disqualified in swimming. In 1996 after anti-drug efforts began in swimming, for the first time in Olympic history, swimming champions were worse than four years before. An elite US swimmer, among others, complained that Ireland's Michelle Smith must have been taking PEDs, after Smith improved and won three gold medals with mediocre times. The fact that elite swimmers assumed that performances ought to be worse in 1996 suggests that elite swimmers probably had ceased taking some form of PED, as hinted by Mark Spitz in a 2012 interview.

## References

- [1] [https://en.wikipedia.org/wiki/History\\_of\\_competitive\\_swimwear](https://en.wikipedia.org/wiki/History_of_competitive_swimwear). Downloaded 5 October 2016.
- [2] Lupica, M. (1996). Comments Put Swimmer in Over Her Head. *NY Daily News*, 23 July 1996.
- [3] [https://en.wikipedia.org/wiki/List\\_of\\_stripped\\_Olympic\\_medals](https://en.wikipedia.org/wiki/List_of_stripped_Olympic_medals). Downloaded 25 February 2017.
- [4] Rutemiller, B. (2013). Doping's Darkest Hour; the East Germans and the 1976 Montreal Games. *Swimming World*, 28 November 2013.
- [5] <http://sportsanddrugs.procon.org/view.timeline.php?timelineID=000017>. Downloaded 12 January 2016.
- [6] Stefani, R.T (2008). The Physics and Evolution of Olympic Winning Performances. *Statistical Thinking in Sports*. Taylor and Francis: Boca Raton, London, NY, 2008.
- [7] Stefani, R.T (2016). Ancient Olympics: Events, Superstars, Cheating, Technology and Women's Role. *Proceedings of the 13<sup>th</sup> Australasian Conference on Mathematics and Computers in Sport*. Melbourne Australia, 11-13 July 2016.
- [8] Swimmer's Daily (2012). <http://www.swimmersdaily.com/2012/12/19/mark-spitz-questions-doping-in-swimming/>. Downloaded 15 January 2017.
- [9] WADA (2017). <https://www.wada-ama.org/en/questions-answers/athlete-biological-passport>. Downloaded 15 January 2017.
- [10] Walleinsky, D. and Loucky, J. (2012) *The Complete Book of the Olympics 2012 Edition*. Aurum: London, 2012.



# The Nappy Factor in Golf: The Effect of Children on the Sporting Performance of Professional Golfers

R.A. Syme

University of Salford, Salford, England  
r.a.syme@salford.ac.uk

## Abstract

This study investigates the effects of children upon the sporting performance of fathers who are professional golfers. Biographical and sporting data for 225 professional golfers are used to estimate fixed-effects regressions. In line with other studies, it is found that performance and earnings improve significantly after the birth of the first child and that this declines after each subsequent child. The fatherhood premium, or ‘nappy factor’, is estimated to be an increase in earnings of 10% for any first child, but this rises to 16% if the first-born child is a son and remains the only child. This study suggests that the rank-order nature of tournaments and the non-linear distribution of prize money within professional golf creates positive incentives to increase work effort, but that tournaments also increase pressure, particularly towards the end of tournament, and that ability to perform under pressure is increased if the player has a son as a first-born child.

## 1 Introduction

On 10<sup>th</sup> April, 2016, Danny Willett won the Masters Tournament and, in doing so, became the first Briton to do so in 20 years. While some may have pointed to his success as being unremarkable given that he ranked no.12 in the World Rankings and had already won a Tour event in the previous two months, it was the fact that he was only playing because his son had been born 12 days early that created significant media attention around a minor golf betting concept: the nappy factor. That term, created by Keith Elliott in his annual golf betting book (Elliott 1996), suggested that golfers received a psychological boost from becoming a new father and this increased their chances of winning tournaments.

While this may seem like ex post fitting, there has been ample evidence that parenthood does have a significant financial impact. For mothers, the ‘motherhood penalty’ has been estimated to be typically in the range 5-10% (see Waldfogel (1997), Lundberg and Rose (2000), Budig and England (2001)), though Wilde, Batchelder and Elwood (2010) find that the reduction in earnings is much higher (20-24%) for higher-skill women, and motherhood itself can explain over half of the gender earnings wage gap (Juhn and McCue, 2017). For men, the ‘fatherhood premium’ has received much less study and the results are rather more varied: Lundberg and Rose (2000), Simonsen and Skipper (2008), and Budig (2014) find parenthood raises hourly male earnings in the range 4-9%, while Cools and Strom (2016) find a small but negative effect of fatherhood on earnings and Wilde, Batchelder and Elwood (2010) also find a negative effect on earnings five or more years after first birth.

One explanation for the mixed results on the ‘fatherhood premium’ is that, even if there was a positive psychological effect from parenthood, it is rather difficult for a father to increase his hourly earnings around the same time as the birth of his child. Even if work motives and productivity were increased, any increase in earnings is likely, on average, to take a significant period of time to materialise. That is not the case with individual sports in which earnings are primarily derived from the prize money associated with weekly tournaments. In such sports, the impact of parenthood can lead to immediate increases in productivity and earnings if the ‘nappy factor’ has any validity, as has been suggested with the Danny Willett example above.

## 2 Data and Empirical Specification

### 2.1 Data

The dataset consists of biographical information relating to 284 players contained within the ‘Player Profiles’ of the PGA Tour’s ‘2016-17 Media Guide’ (PGA Tour 2016), while scoring and prize money data was obtained via the PGA Tour’s Shotlink Intelligence programme.

The biographical data contained the names and date of birth for each player’s children. In the 32 instances in which no dates of birth were listed, searches were made using previous Media Guides, news sites and player Twitter accounts. In most cases, the dates of birth were found. The names were used to identify the gender of the child and corroborated via news sites or Twitter if necessary. Players were excluded from the final sample if they were not members of the PGA Tour, were first-time members (‘rookies’) of the PGA Tour in the 2016-17 season, had children by marriage rather than birth, or had children by adoption rather than birth. This left a final sample of 225 players with biographical information relating to children’s gender and date of birth. As no dates are given for marriage, it was not possible to also include data relating to marital status.

For these 224 players, scoring and prize money data was collected via ShotLink for the period January 1996 to September 2016 (the conclusion of the 2015-16 season). One issue with using scoring data is that a player’s score in any round may be affected by the weather or the difficulty of the course, so the data was normalised via the following equation:

$$\text{‘Adjusted’ score} = \text{actual score} - \text{average score that day} + \text{Sagarin field quality} \quad (1)$$

Normalising on the average score for all players on a given day takes into account the variance in scores according to different weather and course conditions, but the average score per day may also vary according to the quality of players competing on any given day, so the measure of field quality was computed as the average Sagarin score for each player competing that day. Current values of the Sagarin scores per player can be found via the golfweek.com site and past ‘Adjusted’ scores can be found via the tour-tips.com site, which were then cross-referenced with the player scoring and prize money obtained via ShotLink. For ease of expression, these ‘Adjusted’ scores will be termed simply as ‘scores’ in the remainder of this paper.

Rather than use yearly dummies to correct for inflation within the prize money over the 20-year sample period, the prize money for five PGA Tour tournaments which were held in all 20 years was used to create a prize money index where 2016 = 100 (in 1996, the value is 29.1). This is important as it not only corrects for inflation in prize money, but accounts for the double-digit growth in prize funds between 1997-2004, which then slowed to no more than 2% per annum thereafter until 2015. This index is then used to control for (prize money) inflation when using the prize money data from ShotLink. The five tournaments were the Phoenix Open, Arnold Palmer Invitational, Players Championship, Memorial Tournament and Tour Championship.

## 2.2 Theoretical Model

While the associated reduction in human capital from a period out of the labour force has long been an explanation for the ‘motherhood penalty’ (see Mincer and Polacheck (1974) for an early exposition), Becker (1985) has suggested that parents may divert time and efforts from market work activities to home activities as children increase the relative return of the latter. In this approach, individuals have the following indirect utility function

$$U = U(t_H, e_H, Y) \quad (2)$$

where  $t_H$  = time spent at home,  $e_H$  = effort expended at home, and  $Y$  = income. Any time not spent at home is assumed to be spent at work ( $t_m$ ) and, similarly, any effort not expended at home is assumed to be expended at work ( $e_m$ ). Individuals maximise utility subject to the following income constraint

$$Y = w_m(e_m)t_m + v \quad (3)$$

where  $w_m$  = wage rate, which is a positive function of human capital, and  $v$  = non-wage income. The first-order conditions are as follows

$$\frac{\partial U}{\partial t_H} = \frac{\partial U}{\partial Y} \cdot w_m \quad (4)$$

$$\frac{\partial U}{\partial e_H} = \frac{\partial U}{\partial Y} \cdot \frac{\partial w_m}{\partial e_m} \cdot t_m \quad (5)$$

Which means that the marginal utility of time and effort in market work and the home will be equalised. Becker uses this model to explain a division of labour in the allocation of effort between husbands and wives: there is a specialisation effect within the household with wives generally increasing their focus on home production and husbands increasing their focus on market work, while the degree of specialisation will depend on their relative market wages and productivities within the home. Further, the greater the impact of work effort on wages, the greater the degree of specialisation on market work.

In this study, it is not the household division of labour that is of interest, but the extent to which fathers may amend their allocation of effort towards work following the birth of a child and the degree to which that reflects the responsiveness of earnings to effort expended at work. It is also of interest to study how the effort and earnings of fathers is affected by the both the number of children and the gender of children as it is likely that fathers’ marginal utilities from time and effort spent at home will be different following the first-born child to those following the birth of subsequent children and be different according to the gender of the child.

## 2.3 Prize Money and Incentives in Professional Golf

In the previous section, one important consideration in the Becker model is the responsiveness of wages to work effort. Within the human capital framework such increases in effort can have positive impacts upon wages, though largely after a significant delays because of the nature of labour contracts. Such delays are predicted to be much smaller within professional golf because of the tournament nature of output. Lazear and Rosen (1981) were the first to suggest that an individual’s earnings based only on output or rank relative to that of other competitors may have positive incentive efforts and, in terms of

the PGA Tour, Ehrenberg and Bognanno (1990) found strong support for the proposition that the level and distribution of prize money does influence players' performance.

In a standard PGA Tour event involving 156 players, there is a 'cut' after two rounds with only the top-70 players and ties progressing to the next round. Any player who does not progress beyond the cut will not earn any prize money. For those that play the full four rounds, the prize money distribution is highly non-linear with the winner receiving 18% of the total prize fund, the runner-up receives 10.8%, the third-placed player receives 6.8% and so on. Beyond this, unless a player has won a tournament on the PGA Tour, his playing rights for the following season will be determined by his rank in the FedEx Cup. Points are awarded for the FedEx Cup with a very similar distribution to the prize money for each tournament.

This means that any reduction in effort or performance will reduce both the player's earnings that week and the likelihood of remaining on the PGA Tour the following season. Conversely, any increase in effort that leads to a high finish in any tournament will lead to a disproportionately higher increase in earnings and in the chances of remaining on the PGA Tour the following season.

## 2.4 Empirical Specification

The empirical specification is based on the Becker model outlined above but with two parallel analyses. Given the highly non-linear distribution of prize money in tournaments on the PGA Tour, two separate equations are estimated for productivity (the player's score) and earnings. The general specification is

$$Y_{it} = \alpha + \beta \mathbf{X}_{it} + \gamma \mathbf{Z}_{it} + (\delta_i + \varepsilon_{it}) \quad (5)$$

where  $Y_{it}$  is outcome of interest, either score or  $\ln(\text{earnings})$ ;  $\mathbf{X}_{it}$  is a vector of human capital explanatory variables: age, plus its quadratic term, number of years as a professional golfer prior to joining the PGA Tour, and a dummy variable indicating the player's first season on the PGA Tour;  $\mathbf{Z}_{it}$  is a vector of child variables: number of children, ages of children, and gender of children. The subscripts  $i$  and  $t$  indicate the individual and time, respectively;  $\delta_i$  is the unobserved heterogeneity term, assumed to be individual-specific and time-invariant, so each equation will be estimated using individual fixed effects; while  $\varepsilon_{it}$  is an error component,  $\text{NID}(0, \sigma_\varepsilon^2)$ . Given the significantly higher prize money offered in Major Championships and WGC events, dummy variables for these events are also included in the earnings specification.

## 3 Results

### 3.1 The Effect of Children on Sporting Performance and Earnings

In the tables of regression results that follow, only the coefficients relating to the vector of child variables are reported. In all instances, the coefficients on the vector of human capital variables are of the expected sign and, for all but the 'numbers of years as a professional before joining the PGA Tour' variable on the earnings equations, they are all statistically significant at the 5% level. The coefficients on each variable are expected to have the opposite sign on the score and earnings regressions.

Table 1 reports the effects of children on the scores and earnings of players on the PGA Tour. It is estimated that players with children earn approximately 10% more per tournament than single men and that this effect is not a one-time increase but the effect on lowering score (i.e. increasing productivity) and raising earnings increases the longer that the player has been a father.

The results in columns (3) and (6) confirm the significant positive impact of the first- and second-born children upon the sporting performance of fathers. That effect on their productivity and earnings is

<i>Dependent variable</i>	Score (1)	Score (2)	Score (3)	Ln(Money) (4)	Ln(Money) (5)	Ln(Money) (6)
<i>Independent variable</i>						
Child	-0.093*** (0.032)	-0.084*** (0.032)		0.104*** (0.026)	0.099*** (0.026)	
Fatherhood Years		-0.013*** (0.004)			0.016*** (0.003)	
One Child			-0.098*** (0.033)			0.103*** (0.026)
Two Children			-0.078* (0.042)			0.102*** (0.033)
Three or more Children			0.231*** (0.058)			-0.041 (0.046)
$\bar{R}^2$	0.153	0.155	0.153	0.157	0.157	0.157
Number of observations	43,228	43,228	43,228	26,641	26,641	26,641

Table 1 – The Effect of Children upon Scores and Earnings

Notes: \*\*\*, \*\*, and \* indicates significance at 1%, 5% and 10% level, respectively. Additional regressors include age (and square terms), numbers of years as a professional before joining the PGA Tour, dummy variables for rookie season (all equations), plus dummy variables for Major Championships and WGC events (earnings equations only). Standard errors in parentheses.

strongest for the first-born child, and while there is a decline in the productivity boost from having two children, productivity is still significantly higher than for players with no children. In terms of earnings, the 10% increase in earnings per tournament is maintained if the player has two rather than one child.

However, such positive effects of having children end after the birth of the second child. While there had been a decline in productivity following the birth of the second child, but it is strongly negative from the third child onwards. Similarly, the strong positive effect on earnings from the birth of a first child that are maintained with the birth of the second child no longer exist after the birth of the third child. The increased home effort from three or more children appears to have had a significant impact on work effort for professional golfers in this sample.

### 3.2 The Effect of Child Gender on Sporting Performance and Earnings

This section looks at the impact of child gender upon the market outcomes of fathers. In this respect, the equations suggest different influences of gender for a single child. For the productivity equation, there is only a minor difference in the impact of the gender of the first-born child, but that is not the case for the earnings equation. Whereas a first-born daughter has no significant effect on earnings, father earnings increase significantly if the first-born child is a son. The impact of a son upon a new father's earnings is now estimated to be 16% rather than the 10% found in Table 1 which did not consider the gender of the child.

In the Ehrenberg and Bognanno (1990) study, they found that the incentives arising from the prize distribution were not apparent in the first two rounds before the cut, but rather they were concentrated in the final round. In this respect, Table 2 shows that the gender of a single child may have no differential effect on player performance other than when the player is near the top of the leaderboard and small gains in productivity can have large impacts on earnings. Where the single child is a son, that increased

<i>Dependent variable</i>	Score (1)	Score (2)	Ln(Money) (3)	Ln(Money) (4)
<i>Independent variable</i>				
One Child (Boy)	-0.093** (0.043)		0.161*** (0.034)	
One Child (Girl)	-0.127*** (0.046)		0.057 (0.035)	
Two Children (Boy, then Girl)	-0.040 (0.062)		0.119** (0.048)	
Two Children (Girl, then Boy)	0.072 (0.061)		-0.037 (0.047)	
Two Children (Both Boys)	-0.029 (0.059)		0.096** (0.047)	
Two Children (Both Girls)	-0.475*** (0.072)		0.219*** (0.054)	
At Least Three Children	0.210*** (0.058)		-0.053 (0.046)	
Number of Boys		0.058 (0.045)		0.096*** (0.036)
(Number of Boys)^2		0.011 (0.017)		-0.042*** (0.014)
Number of Girls		-0.263*** (0.052)		0.126*** (0.041)
(Number of Girls)^2		0.134*** (0.023)		-0.076*** (0.019)
$\bar{R}^2$	0.154	0.152	0.156	0.150
Number of observations	43,228	43,228	26,641	26,641

Table 2 – The Effect of Child Gender upon Scores and Earnings

Notes: as outlined in Table 1.

effort away from the home appears to lead to much more successful performance under the pressure of competing for large sums of money during the final round of a tournament.

For players with two children, the productivity improvements following fatherhood have largely disappeared once a second child has been born. The exception being the case of two daughters, which is an unusual result not found in other studies. The earnings equation is rather more standard. As outlined above, there is a significant increase in earnings for new fathers with a son and even though this effect declines following the birth of a second child, earnings are still 10-12% higher than for players with no children. The only situation in which the impact of having two children does not positively impact upon the earnings of father is where the first-born is a daughter and the second-born is a son, which again emphasises the positive earnings effect of a first-born son.

This result is largely supported by the results in columns (2) and (4). Given the coefficients on quadratic terms, the largest improvement in productivity over players with no children occurs when players have one child and it is a daughter. However, the largest improvement in earnings over players with no children occurs when the player has one son and one daughter. The results from Table 2 confirm

<i>Dependent variable</i>	Score (1)	Ln(Money) (2)
<i>Independent variable</i>		
First Child, aged 0-2	-0.090*** (0.033)	0.092*** (0.026)
First Child, aged 3-5	-0.067 (0.049)	0.158*** (0.038)
First Child, aged 6-8	-0.130** (0.061)	0.225*** (0.047)
First Child, aged 9-11	-0.432*** (0.072)	0.384*** (0.056)
First Child, aged 12-14	-0.536*** (0.086)	0.543*** (0.067)
First Child, aged 15-17	-0.607*** (0.106)	0.509*** (0.082)
Second Child, aged 0-2	-0.010 (0.043)	-0.053 (0.033)
Second Child, aged 3-5	0.099* (0.059)	-0.078* (0.045)
Second Child, aged 6-8	0.348*** (0.075)	-0.271*** (0.058)
Second Child, aged 9-11	0.562*** (0.095)	-0.429*** (0.073)
Second Child, aged 12-14	0.887*** (0.125)	-0.587*** (0.097)
Second Child, aged 15-17	0.918*** (0.177)	-0.313** (0.132)
At Least Three Children	0.258*** (0.050)	-0.143*** (0.039)
$\bar{R}^2$	0.155	0.160
Number of observations	43,228	26,641

Table 3 – The Effect of Child Age upon Scores and Earnings  
Notes: as outlined in Table 1.

those of Table 1 that having up to two children raises earnings, but that also earnings are far higher if the first-born child is a son.

### 3.3 The Effect of Child Age on Sporting Performance and Earnings

The results so far consistently suggest that having children raises the sporting performance and earnings of golfers. However, it is now worth considering whether this is a temporary or a permanent effect. Tables 3 reports the results around this issue with three-year bands for each of the first two children. Different groupings of age bands were trialled within the specification of the estimation equations and the results were found to be consistent across these different groupings. The previous results have

suggested that any improvement in productivity or earnings is reversed if a father has three or more children, so the age bands are only reported for the first two children.

In terms of earnings, there is a permanent, positive increase in earnings following the birth of the first child and this effect increases in size as the age of the child increases. In the early years of fatherhood, there is an initial positive impact on productivity, but this becomes statistically insignificant while the child is 3-5 years old, while thereafter the effect becomes significant and increases with the age of the child. One possible explanation may be that the marginal utility of home time is at its highest level when the first-born child is youngest and while this leads to incentives to increase work effort while the father is away from home, the former declines with the increasing age of the child but the latter doesn't.

For the second-born child, it is very different story. For the early years, there is little impact upon the father's productivity levels and earnings, but from the age of six onwards, there is a significant and increasingly negative impact upon the father's productivity and earnings that will, at least, partially offset the positive impact from the first-born child.

### **3.4 The Effect of First Child Age and Gender on Sporting Performance and Earnings**

The results of the previous section suggest that fatherhood significantly increases sporting performance and earnings, but that this effect starts to diminish once the second child reaches school age, or maybe reaches the age at which (s)he will be taught to play golf. This section looks at whether the first-child effect is gender-specific and permanent. In this respect, the results are unambiguous and confirm those found in Section 3.2. Having a son as the first-born child significantly raises earnings and this effect is both permanent and increases in value during the son's childhood years. By contrast, if the first-born child was a daughter, the positive effect on earnings only arises once the child has reached the age of nine, but remains permanent thereafter. Across all six age groupings, the coefficient in the earnings equation is larger for the first-born son than the first-born daughter. This is further evidence of the gender differential in the impact of children upon the earnings of fathers, though these results suggest that the differential is largest at younger ages.

In terms of the productivity equation, there is some similarity across the two genders for the first-born child. For both, there is a positive effect on father productivity during the early childhood years, this effect then diminishes considerably before becoming significant and positive from the ages of 6-8 (sons) or 9-11 (daughters). Comparing the coefficient for the first-born child across the productivity and earnings equations, there is near-symmetry for the daughter age groups, but not for the son age groups. Up to the age of 5, a first-born son is estimated to have a negligible effect on the father's productivity, but a statistically significant effect on the father's earnings. Such a difference reinforces the general finding from the previous sections that having a son is suggested to increase the ability of fathers to compete under the pressure of contention within tournaments and thus increase earnings more so than any general increase in productivity.

There were no such gender differences found for the age bands for the second child. In line with the results of the previous Section, for both sons and daughters, there was no significant effect upon either productivity or earnings while the second-born child was aged 5 or less. Similarly, once the second-born child reaches the age of six, there is a significant negative and permanent effect on the father's productivity and earnings.

Such consistency of findings across the productivity and earnings equations relating to the second child suggest that the differential effect of sons over daughters to increase fathers' wages occurs only for the first-born child.



<i>Dependent variable</i>	<b>Score (1)</b>	<b>Ln(Money) (2)</b>
<i>Independent variable</i>		
First Child, Boy aged 0-2	-0.076* (0.042)	0.129*** (0.034)
First Child, Boy aged 3-5	-0.009 (0.056)	0.198*** (0.044)
First Child, Boy aged 6-8	-0.231*** (0.066)	0.290*** (0.052)
First Child, Boy aged 9-11	-0.457*** (0.078)	0.388*** (0.061)
First Child, Boy aged 12-14	-0.498*** (0.094)	0.595*** (0.072)
First Child, Boy aged 15-17	-0.523*** (0.113)	0.537*** (0.084)
First Child, Girl aged 0-2	-0.094** (0.046)	0.056 (0.036)
First Child, Girl aged 3-5	-0.118* (0.064)	0.073 (0.050)
First Child, Girl aged 6-8	0.044 (0.080)	0.073 (0.063)
First Child, Girl aged 9-11	-0.353*** (0.095)	0.332*** (0.074)
First Child, Girl aged 12-14	-0.564*** (0.117)	0.453*** (0.090)
First Child, Girl aged 15-17	-0.857*** (0.168)	0.448*** (0.130)
Second Child, aged 0-2	-0.013 (0.043)	-0.044 (0.033)
Second Child, aged 3-5	0.093 (0.059)	-0.061 (0.046)
Second Child, aged 6-8	0.322*** (0.076)	-0.277*** (0.059)
Second Child, aged 9-11	0.552*** (0.097)	-0.445*** (0.074)
Second Child, aged 12-14	0.940*** (0.129)	-0.599*** (0.100)
Second Child, aged 15-17	0.936*** (0.177)	-0.299** (0.141)
At Least Three Children	0.260*** (0.050)	-0.135*** (0.039)
$\bar{R}^2$	0.156	0.155
Number of observations	43,228	26,641

Table 4 – The Effect of First Child Age and Gender upon Scores and Earnings

Notes: as outlined in Table 1.

## 4 Conclusions

Given that the sporting performance of professional golfers is at a peak during their 20s and 30s, it should be no surprise that periods of sporting success will coincide with periods of fatherhood. This paper has investigated whether there is any causal link between fatherhood and sporting success.

The primary result is that, even after accounting for human capital factors, there is a 10% increase in earnings associated with fatherhood and this rises to 16% if the first-born child is a son and remains the only child. Further children reduce and eventually reverse the fatherhood premium, but this process does not start until the second child has reached school age.

The secondary result is that having a son as a first-born child has differential effects on the productivity and earnings of fathers. While the gender of the first-born child has only minor effects on the productivity of fathers, the effect on earnings is significantly higher if the first-born child is a son. This study has suggested that the rank-order nature of tournaments and the non-linear distribution of prize money within professional golf has created positive incentives to effort as well as significant increases in pressure for players near the top of the leaderboard towards the end of tournament. However, the ability to perform under that tournament pressure is increased if the player has a son as a first-born child.

## References

- [1] Becker, G.S. (1985) Human Capital, Effort, and the Sexual Division of Labour, *Journal of Labour Economics*, 3(1), S33-S58
- [2] Budig, M.J. (2014) The Fatherhood Bonus and the Motherhood Penalty: Parenthood and the Gender Gap in Pay. Report prepared for the Third Way Next Research Report series
- [3] Budig, M.J. and England, P. (2001) The Wage Penalty for Motherhood, *American Sociological Review*, 66(2), 204-225
- [4] Cools, S. and Strom, M. (2016) Parenthood Wage Penalties in a Double Income Society, *Review of Economics of the Household*, 14, 391-416
- [5] Ehrenberg, R.G. and Bognanno, M.L. (1990) Do Tournaments Have Incentive Effects, *Journal of Political Economy*, 98(6), 1307-24
- [6] Elliott, K. (1996) *Elliott's Golf Form 1997*. Barnsley: Rowton Press
- [7] Juhn, C. and McCue, K. (2017) Specialization Then and Now: Marriage, Children, and the Gender Earnings Gap across Cohorts, *Journal of Economic Perspectives*, 31(1), 183-204
- [8] Lazear, E.P. and Rosen, S. (1981) Rank-Order Tournament as Optimum Labour Contracts, *Journal of Political Economy*, 89(5), 841-864
- [9] Lundberg, S. and Rose E. (2000) The Effects of Sons and Daughters on Men's Labour Supply and Wages, *The Review of Economics and Statistics*, 84(2), 251-268
- [10] Mincer, J. and Polacheck, S. (1974) Family Investment in Human Capital: Earnings of Women, *Journal of Political Economy*, 82(2), S76-S108
- [11] PGA Tour (2016) *Official Guide 2016-17*. Ponte Vedra Beach: PGA Tour
- [12] Simonsen, M. and Skipper, L. (2008) An Empirical Assessment of Effects of Parenthood on Wages, *Advances in Econometrics*, 21, 359-380
- [13] Waldfogel, J. (1997) The Effect of Children on Women's Wages, *American Sociological Review*, 62(2), 209-217
- [14] Wilde, E.T., Batchelder, L. and Elwood, D.T. (2010) The Mommy Track Divides: The Impact of Childbearing on Wages of Women of Differing Skill Levels, Working Paper 16582, National Bureau of Economic Research

# Compare the superiority of Japanese Collegiate Baseball Leagues

T. Toriumi\* (Keio University Institute of Physical Education, Japan)

\* bird@keio.jp

## Abstract

The Tokyo Big 6 Baseball League (BIG6) and the Tohto University Baseball League (TOHTO) are the top two collegiate baseball leagues in Japan. Both leagues consists of six teams and are highly competitive; however, there is no opportunity for the teams in the two leagues to compete directly against each other except for the national championships. In this report, I calculate the strengths of six teams across both leagues and compare the superiority of them. Using the Bradley-Terry model, I estimate the strengths of each team in both leagues. The results indicate that the first team in TOHTO had the highest strength of 0.1882, followed by the first team in BIG6 at 0.1166. The results of the corresponding strengths of the first-ranked to sixth-ranked teams are as follows: 0.1166, 0.0630, 0.0430, 0.0310, 0.0172, and 0.0023 in BIG6 and 0.1882, 0.1115, 0.0848, 0.0574, 0.0406, and 0.0256 in TOHTO. On comparing the strengths of the same rankings in both leagues, the strengths of every team in TOHTO are higher than those of the corresponding teams in BIG6. Thus, it can be concluded that TOHTO is superior to BIG6.

## 1 Introduction

There are 26 university baseball federations in various parts of Japan, each federation having its own league in which each first division comprises six universities, and all leagues are held twice a year in spring and autumn. The first 26 teams from each spring league participate in the Japan National Collegiate Baseball Championship (JNCBC). Additionally, those from autumn league participate in district preliminary rounds, and the selected 11 teams can participate in another national championship, the Meiji Shrine Baseball Tournament (MSBT).

There are two widely known leagues within 26 leagues: the Tokyo Big Six Baseball League (BIG6) and the Tohto University Baseball League (TOHTO). BIG6 started in 1925, with its games being held at the same six universities every spring and autumn. Hence, the league structure is unchanged since its inception, and there are no team changes from season to season. TOHTO was established in 1931 and, currently, it includes four divisions with 21 participating universities. The first to third divisions consist of six teams and the fourth division of three teams.

The competition levels of BIG6 and the first division in TOHTO are one of the highest among 26 leagues. For example, the representatives of BIG6 and TOHTO can automatically participate in MSBT without going through the district preliminaries. Moreover, the opportunity for both teams to play against each other in JNCBC and MSBT every year is only at the finals because of the lottery method used. Nevertheless, the number of wins in JNCBC of both TOHTO and BIG6 is 24 times, while the Kansai Big6 Baseball League recorded only 6 wins, which is ranked third. Similarly, the number of wins

Compare the superiority of Jpn Baseball Leagues

Toriumi, T.

in MSBT of TOHTO is 15, that of BIG6 is 13, and those of Kansai Big6 Baseball Leagues and Tokyo Metropolitan Area University Baseball League are 5 each, which ranks them both third, at quite a distance from the previous two. Many players belonging to BIG6 and TOHTO become professional baseball players after graduating from their universities every year.

Therefore, a controversy exists as to which league is stronger between BIG6 and TOHTO, as they typically do not compete against each other. For further information, Toriumi and Watada (2017) summarise the league operation methodology and other details.

## 2 Objective

Here, I use the match results of BIG6 and TOHTO to estimate the strengths of each rank in both leagues. By comparing the first teams within the 26 university leagues participating in JNCBC and MSBT, I outline the features of BIG6 and TOHTO.

## 3 Method

I use data on matches conducted within BIG6 (1,088 matches), TOHTO (982 matches), and the national championships of JNCBC and MSBT (519 matches) from 2001 to 2015 (30 seasons). Data on TOHTO from the spring of 2005 are excluded because the season was played with only five teams. Using the Bradley–Terry model, I estimate and compare the strengths of each league rank as follows: 1) estimate the strengths of 26 teams from their results in JNCBC and MSBT; 2) calculate the strengths of each rank within BIG6 and TOHTO; 3) match the strengths of the first teams in both leagues obtained in step 2 and the strengths of both leagues obtained in step 1; and 4) normalise the strengths of each rank in both leagues.

The strengths used are the averages of the 30 seasons for the 15 analysed years. Toriumi and Watada (2017) describe the derivation and calculation method of the Bradley–Terry model in detail. In this study, I perform the iterative calculation 8,188 times.

## 4 Results

Fig. 1 shows the strengths of each league representative at JNCBC and MSBT from 2001 to 2015. The results of step 1 indicate that the first team in TOHTO had the highest strength of 0.1882, followed by that of the first team in BIG6 with 0.1166. The next is the Tokyo Metropolitan Area University Baseball League representative with 0.0903.

Fig. 2 shows the results of step 4, indicating the strengths of the first to the sixth teams in BIG6 and TOHTO. Comparing the strengths of the same ranking positions in both leagues, the strengths in TOHTO are higher than the corresponding teams in BIG6.

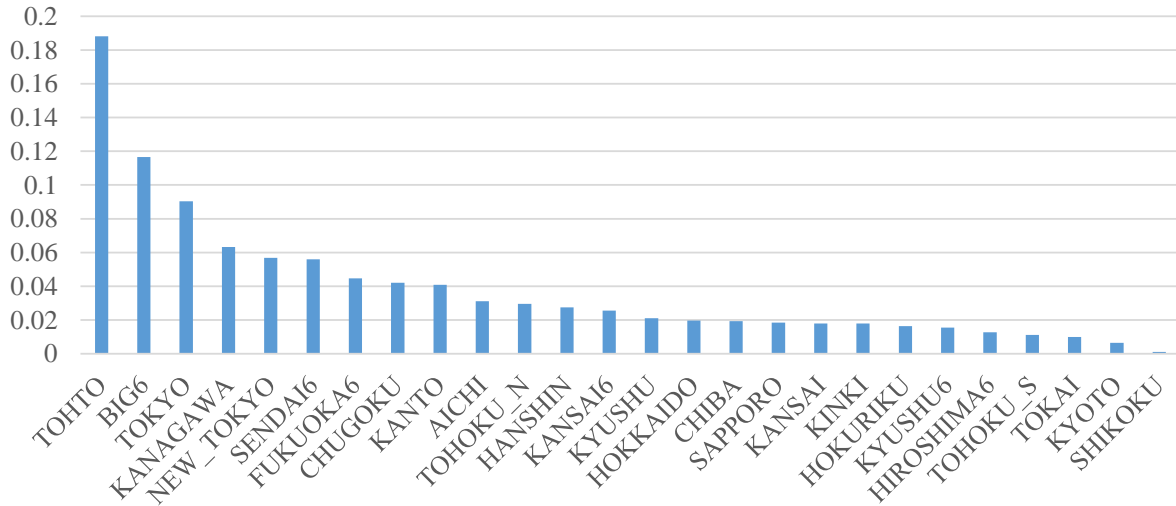


Figure 1. The strengths of each league representative at JNCBC and MSBT.

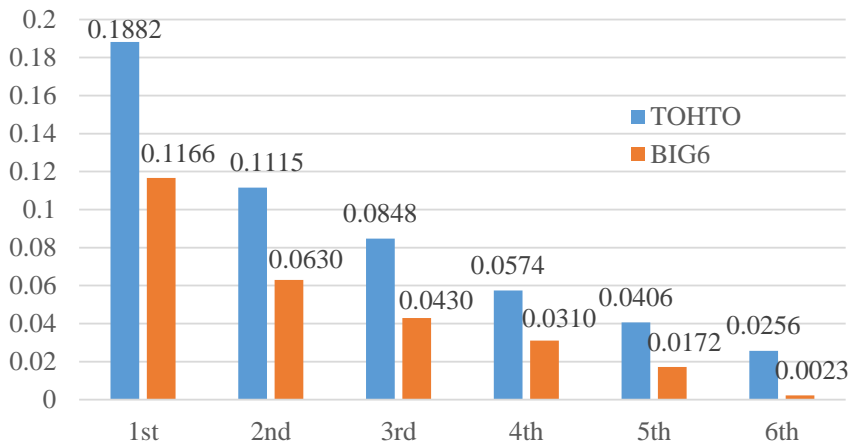


Figure 2. The strengths of the first to the sixth teams in BIG6 and TOHTO.

## 5 Discussion

### 5.1 Strength comparison of first teams in BIG6 and TOHTO

Comparing the strengths of the first teams in BIG6 and TOHTO, the strength of the first team in TOHTO is 0.1882 and that of BIG6 is 0.1166. Therefore, the strength of the first team in TOHTO is 61.4% higher than that in BIG6 because TOHTO won the championship and held the second position relatively more

in JNCBC and MSBT than BIG6 did. That is, the number of times that TOHTO held the champion title was 39 and that of BIG6 was 37, and the number of times that TOHTO held the second place was 32 and that of BIG6 was 23. Additionally, by comparing the strengths obtained through the Bradley–Terry model, it is possible to calculate the winning rate when the two teams play against each other. Hence, the winning rate of the first team in TOHTO is 61.7% ( $0.1882 / (0.1166 + 0.1882) = 0.617$ ). The actual results between the two teams are six wins and two losses for TOHTO, the winning rate being 75.0%. The total number of games between the two teams is not enough but the winning rate obtained by the Bradley–Terry model corresponds with the actual results.

## 5.2 Team comparison below second ranks in BIG6 and TOHTO

When teams of the same rank (i.e. second or lower) in BIG6 and TOHTO play against each other, the winning rates for the teams belonging to TOHTO are shown in Table 1. The winning rate of the sixth team is as high as 91.9%, while those of the other rankings are from 63.9% to 70.2%. This is because BIG6 consists of only the same six teams, while TOHTO consists of 21 teams, with the sixth-ranked team playing a replacement game against the first team of the second division every season. It is thus considered that the strength of the sixth team in TOHTO does not decrease below a certain threshold. As the winning rates are around 65% from the second to fifth ranks, it can be considered that the strengths of the team belonging to TOHTO are superior to those of BIG6.

Table 1. The winning rate for the teams belonging to TOHTO against the same rank of BIG6.

Ranking	Win rate[%]
1st	61.7
2nd	63.9
3rd	66.4
4th	64.9
5th	70.2
6th	91.9

## 5.3 Strength comparison between each ranking position in BIG6 and TOHTO and teams participating in JNCBC and MSBT

Only the first teams in the 26 university leagues are not allowed to participate in the national championships, but if all teams in BIG6 and TOHTO participate, given their strengths, it would be interesting to identify what rankings they can reach. Table 2 summarises the strengths and ranking positions of the first teams in each league and every other team in BIG6 and TOHTO. As such, the second and third teams in BIG6 and from the second to fourth teams in TOHTO have the same strengths as the best eight in the national championships, and the fourth teams in BIG6 and the fifth team in TOHTO have the minimum strengths of the teams participating in MSBT. However, it would be difficult for the fifth team or below in BIG6 and sixth team in TOHTO to pass the first rounds of the national championships with their current strengths.

Compare the superiority of Jpn Baseball Leagues

Toriumi, T.

Table 2. The strengths and ranking positions of the first team in each league and every other team in BIG6 and TOHTO.

Ranking	Strength	League	Ranking	Strength	League
1	0.1882	TOHTO	12	0.0274	HANSHIN
2	0.1166	BIG6	13	0.0256	KANSAI6
	<u>0.1115</u>	<u>TOHTO 2<sup>nd</sup></u>		<u>0.0256</u>	<u>TOHTO 6<sup>th</sup></u>
3	0.0903	TOKYO	15	0.0211	KYUSHU
	<u>0.0848</u>	<u>TOHTO 3<sup>rd</sup></u>	16	0.0197	HOKKAIDO
4	0.0633	KANAGAWA	17	0.0193	CHIBA
	<u>0.0630</u>	<u>BIG6 2<sup>nd</sup></u>	18	0.0185	SAPPORO
	<u>0.0574</u>	<u>TOHTO 4<sup>th</sup></u>	19	0.0179	KANSAI
5	0.0568	NEW_TOKYO		0.0179	KINKI
6	0.0559	SENDAI6	20	<u>0.0172</u>	<u>BIG6 5<sup>th</sup></u>
7	0.0446	FUKUOKA6	21	0.0164	HOKURIKU
	<u>0.0430</u>	<u>BIG6 3<sup>rd</sup></u>		0.0155	KYUSHU6
8	0.0420	CHUGOKU	22	0.0127	HIROSHIMA6
9	0.0408	KANTO	23	0.0112	TOHOKU_S
	<u>0.0406</u>	<u>TOHTO 5<sup>th</sup></u>	24	0.0099	TOKAI
10	0.0311	AICHI	25	0.0065	KYOTO
	<u>0.0310</u>	<u>BIG6 4<sup>th</sup></u>		<u>0.0023</u>	<u>BIG6 6<sup>th</sup></u>
11	0.0296	TOHOKU_N	26	0.0011	SHIKOKU

#### 5.4 Winning rate comparison between BIG6 and TOHTO

Regarding the strengths of the same ranking positions in both leagues, TOHTO holds higher strengths for every ranking. I calculate the winning rates of each rank against every other rank in both leagues and summarise them in Tables 3 and 4. Comparing these tables, in BIG6, the strength of the sixth team with 0.0023 is 86.7% smaller than that of the fifth team with 0.0172 as shown in Fig.2, so the winning rates of sixth team against any other team can be as low as 11.6% or even less. This means that the sixth team rarely wins even when playing opposite the fourth or fifth teams. However, in TOHTO, the winning rates of the sixth team against the fourth and fifth teams are 30.8% and 38.6%, respectively, which are higher than in BIG6. Additionally, in TOHTO, there are replacement games held after the league games between the sixth and first teams of the second division, making it interesting to follow the sixth team's classification, as opposed to BIG6. Regarding the championship winner, the winning rates of the first team against the second and third teams in BIG6 are 64.9% and 73.1%, respectively, which are little higher than in TOHTO at 62.8% and 68.9%, respectively. This means that it would be more difficult to predict the championship winner team in TOHTO than in BIG6.

Compare the superiority of Jpn Baseball Leagues

Toriumi, T.

Table 3. The winning rates [%] of each rank against every other rank in BIG6.

	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>	4 <sup>th</sup>	5 <sup>th</sup>	6 <sup>th</sup>
1 <sup>st</sup>		64.9	73.1	79.0	87.1	98.1
2 <sup>nd</sup>	35.1		59.5	67.0	78.5	96.5
3 <sup>rd</sup>	26.9	40.5		58.1	71.4	95.0
4 <sup>th</sup>	21.0	33.0	41.9		64.3	93.2
5 <sup>th</sup>	12.9	21.5	28.6	35.7		88.4
6 <sup>th</sup>	1.9	3.5	5.0	6.8	11.6	

Table 4. The winning rates [%] of each rank against every other rank in TOHTO.

	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>	4 <sup>th</sup>	5 <sup>th</sup>	6 <sup>th</sup>
1 <sup>st</sup>		62.8	68.9	76.6	82.2	88.0
2 <sup>nd</sup>	37.2		56.8	66.0	73.3	81.3
3 <sup>rd</sup>	31.1	43.2		59.6	67.6	76.8
4 <sup>th</sup>	23.4	34.0	40.4		58.6	69.2
5 <sup>th</sup>	17.8	26.7	32.4	41.4		61.4
6 <sup>th</sup>	12.0	18.7	23.2	30.8	38.6	

## 6 Conclusion

Using the Bradley–Terry model, the strengths of each ranking position in BIG6 and TOHTO are estimated. Comparing the strengths of the first teams in both leagues, I find that the team in TOHTO is 61.4% higher than in BIG6. In addition, the strengths of the teams from the second to fifth ranking positions in TOHTO are higher than in BIG6. Thus, it can be concluded that TOHTO is superior to BIG6.

## Appendix

I am grateful to Emeritus Professor H. Watada of Keio University for collaboration on the early stages of this work. I would like to thank the members of Downing College of the University of Cambridge for their hospitality during my visit, when the main results of this paper were obtained.

## References

- [1] Toriumi, T. and Watada, H. (2017) *A comparative study of the team strengths calculated by mathematical and statistical methods and points and winning rate of the Tokyo Big6 Baseball League*. Bulletin of the institute of physical education, Keio University **56**, 45-53.



# Applying Occam's Razor to the Prediction of the Final NCAA Men's Basketball Poll

John A. Trono

Saint Michael's College  
One Winooski Park  
Colchester, VT 05439 (USA)  
jtrono@smcvt.edu

## Abstract

Several approaches have recently been described that attempt to match how the coaches vote in the final poll that is taken after the men's NCAA basketball championship concludes (Trono and Yates 2014, 2015). The new strategy presented here: is more straightforward than the prior approaches; has demonstrated reasonably high correspondence with the actual, final polls; and its results can be generated much more easily than the most accurate of the aforementioned approaches.

## 1 Introduction

Once the National Collegiate Athletic Association (NCAA) selection committee has completed its deliberations, and that year's men's basketball tournament bracket has been made public, the following three weeks of tournament games generate more interest (at least within the United States) than any other sporting event that spans this length of time. During the first few days after the bracket is announced, millions of individuals will each spend a varying amount of time agonizing over how to complete their bracket – in an attempt to be the winner of various types of pools.

Strategies for making successful predictions of all 63 tournament games, which are played over six 'rounds', are as unique as the teams competing therein. However, this paper will *not* focus on the prediction of games that occur during the tournament, but instead, it will focus on the prediction of what the final rankings will be – as generated by the coaches who vote in the final poll which is taken *after* the championship game has been played.

## 2 The first prediction strategy

In a previous study, a weighted, least squares regression model was used to predict the number of votes each team would achieve in the aforementioned final poll (Trono and Yates 2014). The chosen equation used three objective quantities to generate a team's predicted vote total: its winning percentage (PCT), multiplied by 100; the number of NCAA tournament wins ( $W$ ) + 1 earned by that team; and the team's power rating (PR), as determined by the power rating system (Carroll, Palmer and Thorn 1988). The plus one, in  $W + 1$ , positively distinguishes those teams who received an NCAA tournament bid, and lost their only game, over teams who were not invited. National Invitation Tournament (NIT) wins were also deemed to be worth one quarter of an NCAA tournament win in that study.

Even though this model was reasonably accurate, in both a team's rank – and its predicted vote count, the weights in this model were improved, using Monte Carlo techniques, to try and match each team's final rank more accurately by eliminating the goal to match the predicted final vote total as closely as

possible. Using the final polls from 1993-2007 as the training set, the most accurate set of weights that were found appear in the following, 'best' Monte Carlo equation (MCB):  $6.68605 * PCT + 17.64763 * PR + 88.24644 * (W + 1)$ . The constant term in the original regression equation was dropped because an accurate final vote total is no longer relevant since the accurate order of the teams in the predicted final poll is now the primary goal.

The accuracy of this model was determined by the Spearman Correlation Coefficients (SCCs) for the top 15, 25 and 35 teams in the final polls. The goal was to maximize the sum of these three values, and when applied to the training set, the MCB model produced an average, yearly sum of 2.57721, with the three individual, average SCC values ranging between 0.85 and 0.865, respectively. For the nine subsequent years, the average sum is 2.51756, and the small drop in performance comes from a 0.04 decrease in the average SCC-25 value, and another 0.02 decrease in the average SCC-35 observed result.

### 3 Subsequent prediction strategy

While the article describing how the MCB model came to be was still in production, during April of 2013, several revelations occurred which eventually led to a more accurate approach (Trono and Yates 2015). The primary idea behind the first improved model is that teams with the same number of NCAA tournament wins typically remain in the same relative order as they appeared in the penultimate coaches' poll, which occurs right after all of the conference tournaments – that determine the teams who will receive each conference's automatic bid to the NCAA tournament – have finished. The number of teams ranked in between such teams will usually shrink, depending upon where these teams reside in the penultimate ranking, but most teams don't usually leap frog over other teams – with the same number of NCAA tournament wins – from the penultimate to the final vote.

While the vote totals which appear in the penultimate coaches' poll are certainly not truly '100% objective', once these totals are known, a prediction which incorporates them is deterministic, i.e. reproducible. And because past behavior is typically a reliably accurate predictor of future behavior, those coaches who thought that team A should be ranked higher than team B will probably continue to think so unless significant, new evidence is produced, i.e. team B shows that it actually *is* a better than team A by winning one or more NCAA tournament games than team A wins. Therefore, it seems acceptable to incorporate this penultimate coaches' poll information into a prediction model – in an effort to improve the overall accuracy of such a model.

The secondary idea was to simply add to the penultimate coaches' poll vote total a bonus reward quantity which is associated with the number of NCAA tournament wins each team accrues. This reward should probably increase more as the number of such wins increases, i.e. the difference between the bonus that is associated with four tournament wins, instead of three, should be larger than the bonus difference between three and two wins, respectively.

To test this approach, the first reward strategy was to simply use the sequence inherent in Zipf's Law, starting with  $1/7$  as the bonus for those teams who were invited to the NCAA tournament, but lost their only game. (The recently added category of 'play-in' games are excluded from the total representing a team's number of NCAA tournament wins.) The bonus for a team with one win would be  $1/7 + 1/6$ , and so on, up to the bonus for the champion:  $1/7 + 1/6 + 1/5 + \dots + 1/2 + 1/1$ . The initial vote count used with this approach would also first be normalized into the zero to one range by dividing the number of votes a team receives in the penultimate coaches' poll by the maximum vote total that a team could receive. The final, bonus increment, i.e.  $+ 1$ , for the champion then guarantees that they will be the top team in the final poll, using this particular reward approach – and that specific outcome has occurred in all previous final polls since 1993, when this poll was reintroduced after its previous manifestations in 1953, 1954, 1974 and 1975.

A basic investigation of this idea produced an SCC sum for this approach (ZPF) above 2.7 (where the SCC sum for the MCB model was almost 2.58), and every one of three SCC values were around the 0.9 level, or higher – on average, over the fifteen years in the training set. This initial, encouraging result fueled a deeper study including other possible bonus reward quantities that might predict the final poll more effectively than how well these rewards, that mimicked Zipf's Law, performed.

The individual SCC sums were all well above 2.7, for the ZPF model, in the following years (2008-12, which were the years available when the first study – describing the MCB model – was submitted), except for the SCC sum in 2010. The primary cause of this lower SCC sum (in 2010) was because Xavier, a #6 seed who won two NCAA tournament wins, only garnered four votes in the penultimate coaches' poll (and were ranked as the #33 team), and was therefore predicted to be the #26 team in the final poll, according to the ZPF model, but was actually ranked #14. (The MCB model predicted them to be #15.)

This large disagreement dramatically impacted the SCC-15 value. A similar situation occurred in 2013, where the 9<sup>th</sup> seeded (and unranked) Wichita State team played their way into the Final Four, and eventually was ranked as the #4 team in that year's final poll. Even with four NCAA tournament wins to their credit, being unranked meant that the only quantity contributing to their team's total, that would determine their predicted rank – as far as the ZPF model was concerned, would be the specific tournament reward for their four NCAA tournament wins. So, the ZPF model predicted them to be the #13 team (whereas the MCB model predicted #8).

On average, there were 46.8 teams that appeared in the penultimate coaches' poll during the 15 year timespan in the training set, and so over 300 division one teams received zero votes in that poll, and are therefore all equal in some sense, at least with regards to models like ZPF – from the strong teams like Wichita State (26-8) down to a team with zero wins (e.g. Grambling, which was 0-28 in 2013). It seemed like the final poll predictions would be more accurate if each team could be assigned some quantity that would help to distinguish, i.e. separate somewhat, all the teams that reside in the unranked majority.

Thankfully, the tournament selection ratio (TSR) seemed to be an appropriate metric for the task (Trono 2013). The TSR value for any team is between zero and one, and half of this quantity is based on the two penultimate polls (populated by coaches and sportswriters, i.e. the AP poll for the latter, and, both polls are weighted equally), and so the highest TSR value a team can attain, that is not ranked in both polls, is 0.5. Eight computer based rating/ranking systems contribute the other 0.5. A trimmed Borda count is used amongst the eight, where four rating systems include each game's margin of victory (MOV), and the other four ranking systems do not. Wichita State's TSR value in 2013 was 0.44952, and their rank was therefore improved to #6 once this component was incorporated into a modified ZPF model. (Xavier also moved up to #16 – at the end of the 2010 season – when including this component.)

A total of five prediction models were deemed the most accurate in the follow-up study, where other patterned, bonus reward sequences were evaluated against the training set – and were added to either the normalized, penultimate coaches' poll value, or the TSR (Trono and Yates 2015). The first five rows in Table 1 contains the results for these models (where OCC will be described shortly). ZP2 follows the same bonus reward pattern as ZPF, except that the first denominator is 8, and the last is 2. PR2 uses prime numbers as denominators (17, 13, 11, 7, 3, 2), but uses two as the numerator instead. LN2 begins with 0.1, and then adds 0.2, then 0.3 and so on until finally adding 0.7 to the championship runner-up's bonus reward value in the sequence (2.1), so that the champion's bonus value becomes 2.8. Finally, the 50T approach attempts to use 50% as the amount of increase from one term to the next. The initial value was chosen by trying to maintain a two digit, decimal value for each term as well as have the final bonus reward value be roughly one larger than the runner-up's value. With these constraints in place, the first value was chosen to be 0.24, which produced the following bonus sequence: 0.24, 0.36, 0.54, 0.81, 1.21, 1.81 and 2.71. The baseline model was created for comparison purposes, and it essentially orders teams

using the number of NCAA tournament wins that they earned – utilizing the sum of a team's PCT and PR values to break the ties that would occur when teams have the same number of tournament wins in this simple, benchmarking model. Therefore, the bonus reward sequence for the baseline model was simply the values from one through seven.

To be able to produce the MCB model predictions, or any of the three other models in Table 1 that rely on the TSR, one would need to have access to at least the PR values, because the power rating system contributes two of the eight values in the TSR's computer-based component: one with, and one without, MOV. The rating system devised by Jeff Sagarin, whose ratings can be found in *USA Today*, is also one of the eight TSR computer based systems, but needless to say, without access to the TSR values, these three models would not be beneficial to someone attempting to make predictions about the final coaches' poll.

Table 1 – Spearman correlation coefficients (for a variety of models)

Model Acronym	Model-type, or, uses Pen/TSR	Training Set 93-2007	Predictions 2008-16
ZPF	Pen.	2.77810	2.77184
ZP2	TSR	2.80695	2.84642
LN2	Pen.	2.79890	2.84439
50T	TSR	2.78121	2.86102
PR2	TSR	2.77118	2.84387
MCB	Lin. Regress.	2.57721	2.51756
OCC	Multiplicative	2.75766	2.73884
Baseline	Win-based	1.80075	1.53183

At least for the ZPF and LN2 models, all that is required, to make predictions, is the penultimate coaches' poll, and knowledge of the bonus reward sequences used therein. The OCC system also essentially only needs the penultimate coaches' poll's actual ranks for each team, as the normalization step, which is required for ZPF (and LN2), is unnecessary.

## 4 One basic operation

While studying the distribution of how NCAA tournament wins match up with the penultimate and final polls – both with regards to lists of final ranks, and which integer pairs (the penultimate rank, and NCAA win total) earned that specific rank, as well as lists of which final ranks were derived from the same penultimate poll rank (and accompanying NCAA tournament win count) – a very basic (and simple) strategy seemed worth investigation.

Historically, most teams with one NCAA tournament win, who were ranked between roughly #25 and 'the teens', in the penultimate coaches' poll, have tended to end up very close to the same final rank, whereas those same teams moved up in the final poll with more wins, and fell in the ranking, with zero tournament wins.

It would be nice if one could say that a team ranked #P in the penultimate coaches' poll, and who earned W NCAA tournament wins, would end up being ranked as the #R team in the final coaches' poll. However, a team's final rank is also related to how other teams fared in the NCAA tournament. For instance, three wins will typically move a team higher in the rankings if the teams above said team only won two or fewer tournament games.

In an attempt to model this particular behavior, the strategy was to simply divide the team's penultimate rank by  $2^{(\text{wins}-1)}$ , and then order the teams by these new rank values to determine the final ranking. It is hard to see how this prediction can be determined in a simpler fashion, so, as long as this approach is fairly accurate, it appears that the Gordian knot of the aforementioned, computationally sophisticated strategies, will have been severed by this idea – and explains why it is called OCC, which is short for Occam.

## 5 Implementation decisions for the OCC approach

Before this particular approach can be effectively evaluated, several parameters must be established. Since a team's penultimate rank is required to produce the rank value, which will determine said team's predicted final rank, some value must be used for all teams that are unranked, i.e. those teams who receive zero votes in the penultimate coaches' poll. Of all the years in the training set, 1993 had the most teams (57) that received at least one vote. So, for initial investigation purposes, 60 seemed to be a reasonable value for the penultimate rank of all the unranked teams. The initial results after making this temporary decision were somewhat promising; however, several minor adjustments seemed necessary before the optimal value for the penultimate rank could be determined.

First off, this particular approach produced some ties since there are instances where a team, with a penultimate rank of  $P$ , that also earned  $W$  NCAA tournament wins, will have the same rank value as any other team with  $W+1$  wins that also had a penultimate rank of  $2*P$ . To alleviate such situations, a small value should be added to  $P$  before the rank altering division operation is performed. Using the training set, adding a small positive value produced somewhat more accurate results than adding a small negative value, so 0.05 was chosen though most any value under one half should yield the same results as the positive quantity selected here.

Secondly, it seemed appropriate to add a small integer constant to the penultimate rank, before the rank altering operation is applied, to lessen the numerical advantage that the #1 team (in the penultimate coaches' poll) might have – given the specifics of this particular strategy. Simply adding one to the penultimate rank maximized the SCC sum results for the years in the training set, so now the best rank constant, for the unranked teams, can be searched for.

So, if a team's penultimate rank is  $P$ , then  $P + 1.05$  will be divided by  $2^{(\text{wins}-1)}$ , to produce that team's ranking value, which determines its final rank, after all the rank values are placed into ascending order. For unranked teams, the most accurate value for  $P$ , using the training set, was found to be 67, though variations in the average SCC sums were typically quite small (0.002 to 0.011) when changing this unranked value by plus or minus one. For instance, when examining the values of  $P$  in the range from 57 to 76, the average SCC sums varied from 2.7204 up to 2.75866 – the latter, when using 67.

Once 67 was chosen, the two previous parameter evaluations were revisited, and it was confirmed that it was still best to add one to the penultimate rank as well as adding a small constant (for tiebreaking purposes, regarding the teams' rank values) before performing the appropriate division operation. The NIT divisor of four, as chosen for the aforementioned MCB model, also worked well for the OCC model as that divisor gave the NIT champion the equivalent of 1.5 NCAA tournament wins. The OCC model tended to place a previously unranked NIT champion somewhere between #30 and #40 – which is where the NIT champion has usually ended up in the final coaches' poll except on five occasions (all before 2002); three of those occasions, the NIT champion was #28 or #29, and in 1993 they were #25 as well as #24 in 1997. Only 1.9 and 2.1 were evaluated instead of 2, in the OCC's rank-altering division operation, and neither produced more accurate results. Teams that receive votes in the penultimate coaches' poll, and who are invited to the NIT, are therefore sometimes predicted to be ranked above the NIT champion, as occurred in 2016, and is illustrated in Table 2, where NIT wins are designated by “#” in the wins column.

Table 2 – Various ranks and rank values (for 2016)

Pen. Rank	NCAA Wins	Rank Value	Pred. Rank	Final Rank
1	3	0.513	3	3
2	0	6.100	8	7
3	5	0.253	2	2
4	3	1.263	5	6
5	3	1.513	6	5
6	6	0.220	1	1
7	4	1.006	4	4
8	0	18.100	21	14
9	1	10.050	15	11
10	0	22.100	22	19
11	2	6.025	7	8
12	2	6.525	9	9
13	1	14.050	17	16
14	1	15.050	18	20
15	2	8.025	11	12T
16	0	34.100	29	22
17	2	9.025	13	12T
18	0	38.100	33	24
19	2	10.025	14	15
20	2	10.525	16	18
21	0	44.100	37	29
22	0	46.100	38	31
23	0	48.100	39	28
24	1	25.050	23	25
25	1	26.050	24	27
26	3	6.763	10	17
27	1	28.05	25	33
28	1	29.05	26	30
29	1	30.05	27	26
30.5	2	15.775	19	21
30.5	"2"	36.271	31	40T
32	2	16.525	20	23
33	1	34.050	28	34T
34.5	1	35.550	30	32
36	1	37.050	43	37
37.5	1	38.550	34T	34T
37.5	1	38.550	34T	40T
40	0	82.100	50T	38
40	1	41.050	36	40T
67	4	8.506	12	10
67	"5"	48.119	40	34T
67	"4"	57.223	41	39

Table 2 illustrates the predictions that were made by the OCC model after the 2016 NCAA tournament completed. Within this table, a penultimate rank of 67 indicates that that team was unranked, while teams listed at 37.5 implies that they were both tied for the #37 rank. (Dayton, which was also ranked #34, did not receive any votes in the final coaches' poll, and therefore does not appear in Table 2.) Three teams received the same number of votes, and since all three were tied for 39<sup>th</sup> in the penultimate coaches' poll, the rank of 40 represents their average rank:  $(39 + 40 + 41) / 3$  (though only two of them appeared in the final poll). The number of NIT tournament wins, for a team, are enclosed in double quotes in Table 2, and a 'T' within the final rank column indicates teams that were tied for that rank – in that poll. (The SCC values for OCC in 2016 were: 0.86339, 0.88827, and 0.90640, for the top 15, top 25 and top 35 teams respectively.)

## 6 Conclusion

This study has illustrated that a very straightforward, exponential update of a team's integral rank position, in the penultimate poll – which is taken just before the NCAA men's basketball tournament begins – models quite accurately the position where teams will be ranked in the final poll, which is taken once that tournament is completed. The magnitude of these updates relates directly to a team's performance during the NCAA tournament (and even less so for the NIT) where a team's regular season accomplishments are purportedly captured in the rank they are assigned in the penultimate poll. This new model (OCC) compares quite favorably with the slightly more accurate, but much more time consuming to generate, previous models that were briefly described here, and are described in more detail in the last two references below.

## References

- [1] Carroll, B., Palmer, P., and Thorn, J. (1988) *The Hidden Game of Football*, Warner Books.
- [2] Trono, J. (2013) *Evaluating Regional Balance in the NCAA Men's Basketball Tournament using the Tournament Selection Ratio*. In Proc. of the 4<sup>th</sup> International Conference on Mathematics in Sport.
- [3] Trono, J., and Yates, P. (2014) *How Predictable is the Overall Voting Pattern in the NCAA Men's Basketball Post Tournament Poll?* *Chance*, 27(2):4-12.
- [4] Trono, J., and Yates, P. (2015) *Predicting the NCAA Men's Postseason Basketball Poll More Accurately*. 27<sup>th</sup> European Conference on Operational Research.

ISBN 978-88-6938-058-7



9 788869 380587