Proceedings of the 5th International Conference on Mathematics in Sport

Loughborough University, U.K.
29 June – 1 July 2015

Editors: Anthony Kay, Alun Owen, Ben Halkon, Mark King

Contents

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 1–10

L.C. Brandao, R.R. Del-Vecchio and J.C.C.B. Soares de Mello
Graph Centrality Analysis for the Evaluation of the 2014 Guanabara Cup 11–18

A. Corris, A. Bedford and I. Grundy
Predictive Modelling of Team Success in the t20 Big Bash League 19–24

S. Drikos and I. Ntzoufras
Skills Importance across Ages for Men's Volleyball 24–31

E. Dziedzic and G. Hunter
Predicting the Results of Tennis and Volleyball Matches Using Regression Models, and Applications to Gambling Strategies 32–40

S. Gleave
The World's Biggest Sports Database 41–45

S. Gleave and M. Taylor
The Ageing Game 46–51

I. Heazlewood and J. Walsh
Mathematical Models that Describe and Predict Performance Change in Mens and Womens Athletic Events: 1960-2014 52–59

I. Heazlewood and J. Walsh
Mathematical Models Predicting Performance in Track and Field at the 2011, 2013 and 2015 IAAF World Championships 60–65

N. Hirotsu and T. Ueda
Measuring Efficiency of a Set of Players of a Soccer Team and Differentiating Players' Performances by their Reference Frequency 66–71

J. Hunt, N. Sanders, A. Bedford and S. Morgan
Towards in-depth Performance Analysis for Wheelchair Rugby with the Australian Steelers 72–77

S. Kampakis, I. Kosmidis, E. Leng and W. Diesel
A Supervised PCA Logistic Regression Model for Predicting Fatigue-related Injuries using Training GPS Data 78–84

S. Knust
Construction methods for sports league schedules 85–90

S. Kovalchik
Comparative performance of models to forecast match wins in professional tennis: Is there a GOAT for tennis prediction? 91–96

G. Kyriakides, K. Talattinis and G. Stephanides
Raw Rating Systems and Strategy Approaches to Sports Betting 97–102

S. Lawrence
The Age Advantage in Association Football 103–109

S. Mancini
An Integer Programming Model to provide Optimal Line-up for Master Swimming Relay 110–114

P. Marek
Effects of Rule Changes in NHL 115–120

R. Minton
ShotLink and Consistency in Golf 121–126

K. Nurmi and J. Kyngäs
Some Interesting Sports League Models 127–132

B. O'Bree, J. Baglin and A. Bedford
Assessing the effect of player ability on scoring performance for holes of different par score in PGA golf 133–138

B. O'Bree and A. Bedford
Can betting in PGA golf be profitable? An analysis of common wagering techniques applied to outright winner market prices 139–143

A. Owen, D. Marshall, K. Sidorov, Y. Hicks and R. Brown
Assistive Sports Video Annotation: Modelling and Detecting Complex Events in Sports Video 144–148

A. Petrides, A.J. Purnell and G. Vinnicombe
A new algorithmic methodology for assessing, and potentially predicting, racing cyclists' fitness from large training data sets 149–156

P. Ritmeijer
Adapting the elo rating system for multi-competitor sports 157–162

J. Sargent
Batting at Three in Five-Day Cricket: a Descriptive Analysis 163–170

J. Schönberger
Parallel Scheduling of Several Non-Commercial Sport Leagues 171–176

B. Šedivá
Dynamic Forecasting Model Based on Ranking System for Hockey Match 177–182

J. Sellitti Rangel Junior
Statistical Modelling and Forecasting Association Football Results 183–188

J.C. Soares de Mello, C. Mourão, L. Angulo Meza and S. Gomes Júnior
Analysis of the 2014 Formula 1 Constructors' Championship using a modified Condorcet Method 189–194

R. Stefani
The Relative Competitive Balance of Male and Female Teams in World Championship Competition and the Relative Predictability of Official International Sports Rating Systems 195–201

J. Van Haaren and J. Davis
Predicting the Final League Tables of Domestic Football Leagues 202–207

S. Vinyes Mora, G. Barnett, C.O. ?a Costa-Luis, J. Garcia Maegli, M. Ingram, J. Neale and W.J. Knottenbelt
A Low-cost Spatiotemporal Data Collection System for Tennis 208–213

P. Volf
A competing risks model for the time to the first goal 214–219

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

Dilaksha Attanayake and Gordon Hunter

School of Mathematics, Kingston University, KT1 2EE, U.K.
dilaksha_r@yahoo.co.uk, G.Hunter@kingston.ac.uk

Abstract

Simulation games of cricket matches, using some random element, have been popular for many years dating back at least to the "Owzat" or "Howzat" dice-based game available at least since the 1970s, and more recently including many highly successful computer/video games. However, such games have tended to focus on entertainment value rather than serious application. Nevertheless, such probabilistic simulations could have real practical use. Data-driven probabilistic simulation models can be used to test the validity of various metrics of player performance by comparing the results of the simulations using these metrics with the statistics of real matches involving the same players. Furthermore, such simulations, if judged to be sufficiently realistic and reliable, could be used to estimate probabilities of outcomes of events within matches and of the possible results of actual matches. This could have applications to gambling, and to investigations of corrupt activities, such as match fixing and spot fixing. In this paper, we use data on real players and performance measures based on both batting averages and strike rates (for batsmen) or on both economy rates and strike rates (for bowlers) to develop probabilistic models of the evolution of an innings in Twenty-Twenty (T20) cricket. We perform computer simulations, using balanced teams of players for whom a reasonable amount of data is available, and investigate the effect of modifying the player performance metrics, and hence the probabilistic models, on the individual and team scores. We evaluate our models by comparing the distributions of scores from simulated innings with those of real innings involving the same players or teams. We also discuss potential applications of our models, including a suggestion for a method to highlight possible anomalous performances.

1 Introduction

Cricket is one of the world's most popular sports, particularly from the following it attracts both "live" at the matches and through TV and internet coverage. Twenty20 (T20) cricket – a fast short limited over format of the game, where each team is allowed to bat for a maximum of 20 overs – has become the latest popular form of cricket, attracting extensive and lucrative sponsorship in many countries. Predicting the outcomes of cricket matches is of interest to a wide variety of parties – the players, their coaches and managers, fans, sponsors and the gambling community, both gamblers and bookmakers/betting exchanges. However, extensive gambling on cricket has led to allegations of corrupt activity in relation to betting, and in some cases these have led to criminal convictions – for example, the case of the England v Pakistan Lords Test match of 2010, which resulted in three members of the Pakistani team being imprisoned, including their captain Salman Butt. Thus, 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 build on a recent pilot study [5] where we developed and applied a data driven probabilistic model of the progress of a T20 cricket innings, based on real data for actual international players. We discuss the inadequacy of "traditional" player performance indicators – batting and bowling averages, and batting run rates or bowling economy rates – for T20 cricket and propose some new measures of a player's quality in these contexts. We then develop a probabilistic model for a given batsman scoring runs off a specified bowler using these indicators and a modified form of a Poisson distribution. We then apply this model to simulate a set of England Sri Lanka "virtual" T20 international matches, and compare the distribution of simulated innings scores with those from real matches involving those teams. In the light of the results we obtained, we discuss the limitations of our model and the performance indicators we used and suggest ways in which they could be improved.

2 Performance Indicators for Individual Cricketers for Limited Overs Matches

The "traditional" method of comparing the performances of several batsmen (and bowlers) over several matches is via their batting (or, respectively, bowling) "averages". A batsman's "average" is calculated by dividing the total number of runs he has scored by the number of times he has been out. Similarly, a bowler's average is given by the number of runs he has conceded divided by the number of wickets he has taken. These – particularly the batting average, which gives too great a reward for innings where the batsman was not out – have received much criticism from cricket-loving professional statisticians [1, 6, 8]. A notable example of this is the case of W.A. Johnson, a specialist bowler in the Australian squad which toured England in 1953, who attained the very rarely-achieved feat of averaging over 100 as a batsman in an English first-class season, despite never exceeding 28 in any one innings – he totaled 102 runs over 17 innings, but had only been dismissed once! Damodaran [2] investigated the rate at which various batsmen from the Indian national team scored during the course of their innings, adopting an approach using "hazard functions" and conditional means. However, this was primarily to investigate and compare the different batsmen's confidence and potential at different points in their innings. The use of both batting and bowling averages is even less appropriate for limited overs cricket, where the aim for a batsman should be to score runs quickly rather than just accumulate runs without getting out. Similarly, the principal aim for a bowler in such matches should be to concede as few runs as possible, rather than necessarily taking a lot of wickets. However, for batsmen, run rate (runs scored divided by balls faced) alone is not necessarily sufficient either. Some tail-end "sloggers" may achieve a good run rate, but rarely stay in long enough to build a really valuable innings. Such a late-order batsman should surely not be rated higher than, say, a patient opening batsman who accumulates many runs at a steady, but lower, rate.

For test bowlers, a "Barnes score" has been recently proposed [3], comparing each bowler to the legendary S.F. Barnes, who managed to take 189 wickets in just 27 test matches between 1901 and 1914. This "Barnes score" is calculated by wickets taken divided by the product of bowling average (runs conceded per wicket taken) and strike rate (average balls bowled per wicket taken). This metric favours bowlers who take a lot of wickets, which is perhaps the primary concern in test or first class cricket, but less so in limited overs cricket. In a previous paper [5], we suggested that the rate at which runs are conceded (i.e. runs conceded divided by balls bowled) – namely the "economy rate" per ball – was an appropriate measure of a bowlers performance in limited overs cricket (with lower values indicating better bowlers), and

that any appropriate performance indicator for batsmen in this form of the game should take into account both the batsman's average number of runs per innings and the rate (in runs per ball faced) at which those runs were scored. In order to do this in a realistic way, in [5] we calculated a batsman's "performance indicator" (PI) as a product of his batting average, as a fraction relative to the best batting average found anywhere in our database of player statistics, and his average run rate, both for the same type of cricket match (in this case, T20) of current interest.

The "Barnes score", the batting PI of our previous paper, and the more conventional performance measures of batting and bowling averages, batting run rate and bowling economy and strike rate, all prove to be special cases of a more general performance measure:

$$\text{Performance} = (\text{Runs scored or conceded})^{\alpha}(\text{Balls faced or bowled})^{\beta}(\text{Wickets taken or lost})^{\gamma} \quad (1)$$

where α, β, γ are some constants for each model. The traditional batting and bowling averages both have $\alpha = 1, \beta = 0$ and $\gamma = -1$, whilst the batting run rate and bowling economy rate have $\alpha = 1, \beta = -1$ and $\gamma = 0$. The batting PI of our previous paper has $\alpha = 2, \beta = -1$ and $\gamma = -1$, whereas the "Barnes score" for bowlers has $\alpha = -1, \beta = -1$ and $\gamma = 3$. We believe that this more general model of performance is worthy of further investigation in the future.

We now propose some further batting and bowling performance indicators. Consider: $R =$ Runs scored (or conceded when bowling), $B =$ Balls faced (or delivered when bowling), $O =$ Number of times getting out (for batsmen) and $W =$ number of wickets taken (for bowlers). Our suggested batting performance indicators T_1, T_2 and T_3 can be defined as: $T_1 = R^2/(B \times O)$, $T_2 = R^2/[(B+1)\times(O+1)]$ and $T_3 = R^2/[O \times B+1]$ respectively. We normalize these indicators by dividing each value by the maximum of the considered data and multiplying by 100 (yielding T_1', T_2' and T_3' respectively). Therefore, the higher the indicator, the better the batsman is. In the cases of T_1', T_2' and T_3' , 100 is the highest possible value. T_1 was the batting PI used in [5]. Similarly, bowling performance indicators B_1 (the Barnes score), B_2 and B_3 can be defined as : $B_1 = W^3/(B \times R)$, $B_2 = R^2/[(W+1)\times(B+1)]$ and $B_3 = R^2/[W \times B+1]$ respectively. For T_2, T_3, B_2 and B_3, 1 is added to both the number of wickets taken (or lost) and the number of balls bowled (or faced) to avoid any difficulties due to division by zero, but these additions should make very little difference to regular bowlers. We normalize these indicators by dividing each value by the minimum of the considered data (yielding B_1', B_2' and B_3' respectively). Therefore, the lower the score, the better the bowler is, with the exception of the Barnes score B_1 and its normalized form B_1', for which high scores are better than low ones.

A selection of these batting and bowling performance indicators are then used in the development of probabilistic models, with parameters calculated using real player and match data, as described in section 3 below, and the results used in probabilistic simulations of innings of England v Sri Lanka T20 matches.

3 Data-Driven Probabilistic Model

"Markov-like" models have displayed considerable success for modelling two-person games, such as Tennis singles matches [7, 10, 11, 12]. However, such games are very controlled situations, whereas a ball-by-ball analysis of a cricket match, with many different combinations of bowlers, batsmen and fielders, presents a far bigger challenge.

Nevertheless, taking inspiration from the two player tennis model of Spanias & Knottenbelt [11], based on average statistics for each player over all the matches they have played (over a particular time window) on the professional circuit, we have developed a probabilistic,

data-driven model for ball-by-ball developments during limited overs cricket matches. For the purposes of this paper, we have decided to focus on T20 matches, although the model should be adaptable to any form of limited overs cricket, if the probabilities of the possible events are calculated from data appropriate to that genre of the game.

Pairs of teams of 11 players, for each of whom extensive summary statistics were available of their T20 performances (Number of matches played, innings batted, number of times out, runs scored, balls faced, batting average, batting run rate (runs per ball faced), number of fours and sixes scored, number of balls bowled, runs conceded, wickets taken, average runs conceded per ball), were selected. Batting and bowling "performance indicators" were calculated for each player, as described in section 2 above.

For each possible bowler-batsman combination from these two teams (where the bowler and batsman must come from opposing teams), a "modifier" was calculated by summing the deviations of the batsman's and bowler's performance indicators from their respective average values across the two teams of current interest. This modifier is then added to the mean of that batsman's average runs scored per ball and the bowler's average runs conceded per ball. This resulted in what we anticipate being a typical run rate for that particular bowler-batsman combination. This was then further adjusted, using the mean rates (per ball faced) which the batsman scored fours and sixes and the net average runs per ball (excluding fours and sixes) used as the "mean rate", μ, for a Poisson distribution:

$$P(X = x) = \frac{e^{-\mu}\mu^x}{x!} \tag{2}$$

where X is the number of runs scored off a given ball. The probabilities for $X = 4$ and $X = 6$ were then corrected based on the batsman's actual rates of scoring such boundaries, and the complete set of probabilities normalized to sum to 1. The probability of a batsman being dismissed on any given ball was calculated from the number of times the batsman has been out relative to the total number of balls he had faced in the T20 matches in our dataset.

Examples of the expected distributions of runs scored off a single ball for various bowler-batsmen combinations from our dataset can be found in [5], where the descriptions "weak", "good" and "strong" are used in terms of the appropriate players' batting and bowling performance indicators, based on the dataset we used, rather than being any comment on the players' talents. From these, it can be noted that, for all batsmen considered, the probability of n runs being scored off a single ball peaks at a relatively low value of n, this value which gives the maximum probability being higher for a "good" batsman than for a "weak" one facing the same bowler. Also, the probability of scoring a four is generally higher than that for scoring a three, and much higher than scoring a five (a very unusual event). Similarly, the probability of scoring a six is normally lower than that for a four (for the same bowler-batsman combination), but considerably higher than that of scoring a five. These observations, based on our "modified Poisson distributions", are consistent with everyday experience about the distributions of runs scored off a single ball in cricket.

4 Data Used and Design of Experiments

Based on our probabilistic model described in Section 3, we developed a computer program in Visual Basic to simulate limited overs cricket matches. The probabilities for a specified number of runs being scored, or the batsman being dismissed, off any given ball, when a specified batsman is facing a specified bowler, were given by our probabilistic model with parameters determined by the players data for the type of cricket matches of interest (in the

context of this paper, T20 matches) obtained from `http://www.espncricinfo.com/`. Teams of 11 international players, each of whom had played at least 30 completed T20 matches, were selected, including batsmen, bowlers and a wicket-keeper, and the appropriate probabilities for each batsman-bowler (where these were from different teams) combination computed. Pseudo-random numbers generated in the Visual Basic program were compared with cumulative probabilities of events, starting with the batsman being dismissed and followed by 0, 1, 2, 3, 4, 5, 6 or 7 runs being scored from each ball. (As would be expected, the probabilities of 5 or 7 runs being scored off a ball were extremely low.) For our computer simulation models, data from professional T20 matches involving 22 international players, each of whom had represented England or Sri Lanka (11 from each country) between February 2005 and the end of April 2014, were taken from the ESPN Cricinfo website `http://www.espncricinfo.com/`. The records of all these players contained batting data and many of them also had bowling data.

4.1 Traditional Statistics

A sample of the batsmen's records, using traditional batting summary statistics is given in Table 1 (a and b), and of the bowlers' records in table 2. The statistics are taken from all professional T20 matches involving those players over the specified time window, since relatively few T20 internationals have been played, so little data is available on these, and calculating model parameters based on very limited datasets tends to give unreliable results.

Country	*Player*	*Matches*	*Innings*	*Not Outs*	*Runs*	*Average*
England	M. Lumb	169	168	11	3865	24.61
England	A. Hales	102	101	7	2809	29.88
England	M.M. Ali	80	77	3	1596	21.56
Sri Lanka	K.J. Perera	55	53	4	1029	21.00
Sri Lanka	T.M. Dilshan	161	157	20	3637	26.54
Sri Lanka	M.D.S. Jayawardene	170	165	24	4051	28.73

Table 1a. Example player records for overall batting performance: traditional statistics. *Average* is the average runs scored per dismissal (i.e. *Innings* − *Not Outs*)

Player	*Balls Faced (BF)*	*Times out per BF*	*Average runs per ball*	*4s/BF*	*6s/BF*	*Performance Indicator (PI)*
M. Lumb	2746	0.05717	1.408	0.173	0.051	1.159
A. Hales	2009	0.04679	1.398	0.154	0.046	1.398
M. M. Ali	1284	0.05763	1.243	0.131	0.034	0.897
K.J. Perera	840	0.05833	1.225	0.102	0.054	0.861
T.M. Dilshan	3056	0.04483	1.190	0.141	0.025	1.057
M.D.S. Jayawardene	3129	0.04506	1.295	0.149	0.030	1.245

Table 1b. Example player records for overall batting performance: statistics explicitly used in our first model. The "*Performance Indicator" (PI)* used here was originally proposed in [5].

4.2 New Performance Indicator Statistics

Tables 3 and 4 show possible new performance indicators, as suggested in section 2 above, for some players from the two teams considered. Recall that, for the batting indicators, higher

Player	*Matches*	*Deliveries*	*Wickets*	*Runs*	*Average*	*Econ*	*Runs per ball*
James Anderson	44	933	41	1318	32.15	8.47	1.412
R.S.Bopara	179	2283	108	2873	26.60	7.55	1.258
Steve Finn	50	1058	53	1311	24.74	7.43	1.238
N. Kulasekara	74	1561	81	1924	23.75	7.39	1.232
Lasith Malinga	201	4355	265	4802	18.12	6.61	1.102
Thisara Perera	120	1917	104	2603	25.03	8.14	1.357

Table 2. Example player records for overall bowling performance. Econ (economy) is average runs conceded per over, and Average is average runs conceded per wicket taken.

values indicate "better" batsmen, whereas for bowlers, with the exception of the "Barnes' scores" B_1 and B_1', lower values indicate "better" bowlers.

Batsman	T_1	T_2	T_3	T_1'	T_2'	T_3'
Alexander Hales (Eng)	41.01	40.63	41.01	100.00	100.00	100.00
Mahela Jayawardene (SL)	37.2	36.92	37.2	90.70	90.86	90.70
Joseph Buttler (Eng)	36.43	35.96	36.43	88.84	88.49	88.84
Kumar Sangakkara (SL)	36.17	35.88	36.17	88.19	88.31	88.19
Christopher Woakes (Eng)	34.98	33.23	34.98	85.30	81.78	85.29
Eoin Morgan (Eng)	34.49	34.19	34.49	84.10	84.14	84.10
Tillakaratne Dilshan (SL)	32.41	32.17	32.41	79.02	79.18	79.02
Gary Ballance (Eng)	32.04	31.3	32.04	78.13	77.02	78.13
Thisara Perera (SL)	31.26	30.72	31.26	76.23	75.60	76.23
Angelo Mathews (SL)	30.41	29.97	30.41	74.15	73.76	74.15
Nuwan Kulasekara (SL)	14.65	14.09	14.64	35.71	34.66	35.71
Stuart Broad (Eng)	7.24	6.79	7.24	17.66	16.71	17.65
Lasith Malinga (SL)	6.80	6.62	6.80	16.59	16.29	16.59
James Anderson (Eng)	5.09	3.92	5.04	12.40	9.64	12.29

Table 3. New statistics for batsmen using three different indicators for both England and Sri Lanka teams.

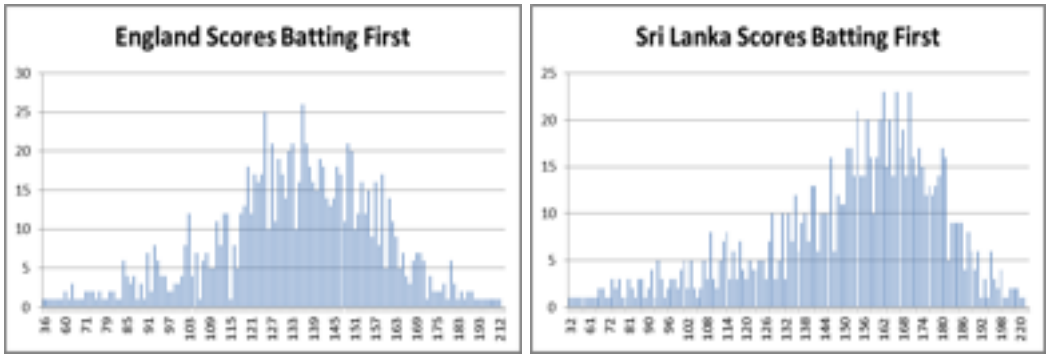
Bowler	*Avg*	*Econ*	B_1	B_2	B_3	B_1'	B_2'	B_3'
Lasith Malinga (SL)	18.1	6.61	0.89	19.90	19.98	16.06	1.62	1.00
Stuart Broad (Eng)	20.5	7.09	0.26	23.94	24.22	4.63	1.95	1.21
Seekkuge Prasanna (SL)	24.2	7.25	0.06	28.21	29.20	1.12	2.30	1.46
Nuwan Kulasekara (SL)	23.8	7.39	0.18	28.90	29.28	3.19	2.36	1.47
Steven Thomas Finn (Eng)	24.7	7.43	0.11	30.05	30.65	1.94	2.45	1.53
Moeen Munir Ali (Eng)	26.6	7.33	0.08	31.89	32.58	1.49	2.60	1.63
Ravi Bopara (Eng)	26.6	7.55	0.19	33.15	33.48	3.47	2.71	1.68
Thisara Perera (SL)	25.0	8.14	0.23	33.64	33.99	4.07	2.75	1.70
Angelo Mathews (SL)	29.3	7.25	0.09	34.89	35.44	1.68	2.85	1.77
Christopher Woakes (Eng)	26.0	8.17	0.16	34.97	35.44	2.80	2.85	1.77

Table 4. New statistics for bowlers using three different indicators for both England and Sri Lanka teams. Avg and Econ are the bowlers bowling averages and economy rates, as defined for Table 2.

5 Match Simulation Experiments

5.1 Simulations with Old Performance Indicators

Using same batting and bowling performance indicators described in [5], namely batting indicator T_1 and the bowler's "economy rate" per ball, an initial simulation experiment was run between England and Sri Lanka. Figure 1 and 2 show the distributions of scores for each country whilst batting first, over 1000 simulated innings. (Note that we focus on first innings since second innings scores will sometimes be restricted by the team batting second reaching their target score and winning the match!)



Figures 1 and 2. Distribution of total number of runs scored by England and Sri Lanka, as team batting first against the other, over 1000 simulated matches

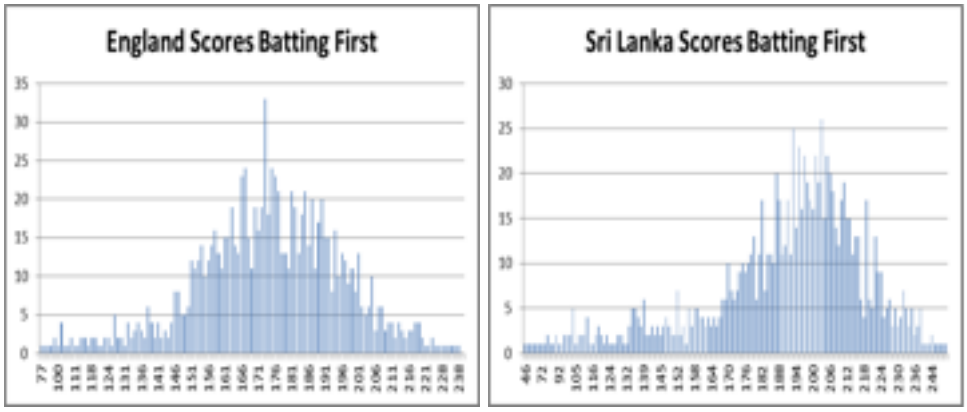
Based on the "raw" innings scores shown in Figures 1 and 2, the mean (with standard deviations in brackets) innings totals were 138.48 (24.82) for England and 151.06 (28.93) for Sri Lanka. For England batting, the mean number of wickets lost was 6.76 (2.21), and the team had been all out within the 20 overs 143 times out of 1000. The corresponding values for Sri Lanka batting were mean wickets lost 6.29 (2.85) and all out 270 out of 1000 innings. It can clearly be seen that the distributions of innings totals are considerably non-normal. However, it should be noted that the majority of the very low innings scores correspond to occasions where the team was dismissed within the 20 overs : for England, out of 92 team totals under 100 runs, 66 corresponded to innings where the team were all out, whereas for Sri Lanka batting, of 64 totals below 100, in 63 cases the team were bowled out. Removal of cases where the team was prematurely dismissed would remove part of the left-hand tail from the distribution of innings totals, yielding mean (and SD scores) of 163.32 (17.36) for Sri Lanka and 138.76 (20.31) for England.

5.2 Simulations with old batting indicator but new bowling performance indicator

The simulations were re-run, but this time basing our performance indicators on batting indicator T_1 (as before) and bowling indicator B_2, so that a bowler's quality is this time measured not just by his economy rate but also by the number of wickets he takes – conceding very few runs per ball is still considered good, but so is taking wickets frequently at low cost. The patterns for the innings totals were broadly similar to those for the simulations using the previous (original) performance indicators. Using the "raw" innings scores shown in Figure 3 and 4, the mean (with standard deviations in brackets) innings totals were 173.76 (22.91) for England

and 189.09 (31.42) for Sri Lanka. For England batting, the mean number of wickets lost was 6.78 (2.17), and the team had been all out within the 20 overs 129 times out of 1000. The corresponding values for Sri Lanka batting were mean wickets lost 6.26 (2.80) and all out 237 times out of 1000 innings.

Again, the distributions of innings totals are considerably non-normal, with the majority of the very low innings scores corresponding to the team being dismissed within the 20 overs : for England, there were just 6 team totals under 100 runs, all corresponding to innings where the team were all out, whereas for Sri Lanka batting, there were 20 innings totals below 100, and in all cases the team were bowled out. Removal of such cases where the team was prematurely dismissed again removed a large part of the left-hand tail from the distribution of innings totals, yielding mean (and SD scores) of 201.42 (17.34) for Sri Lanka and 178.37 (18.56) for England.



Figures 3 and 4. Distribution of total number of runs scored by England and Sri Lanka, as team batting first against the other, over 1000 simulated matches (using new bowling performance indicator B_2).

6 Discussion, Conclusions and Future Work

For actual T20 international matches involving either or both of these countries (between January 2005 and December 2014), the summary statistics are shown in Table 5 below. It would appear that, for the real innings and for both sets of simulations, on average Sri Lanka tend to have slightly higher innings totals than England when batting first, , but also have greater level of variation in their scores. The innings totals and numbers of wickets lost by the team batting first produced by both sets of simulations appear realistic. Although the totals produced using the second (new) set of player performance indicators are slightly higher than those from the first set, the differences are not that large. Further investigation, using more simulated games, on the distributions of scores, run rates and wickets lost could be useful. It would also be beneficial to widen this investigation, both performing more simulations with these same performance measures, and examining the distributions more carefully, and also with the others we have proposed above. We should also consider player performance indicators suggested by other authors, such as the "Richards Score" [4] which takes into account the fraction of the team total scored by the individual batsman and his rate of scoring relative to that of his team (thus, batsmen who are making large contributions to modest totals, and scoring faster than their team-mates, get rewarded), or the "Bowling Quality Index" proposed by Narayanan [9].

Team batting first	Number of matches	Min. innings score	Mean innings score	Max. innings score	Std dev	Min. wickets lost	Mean wickets lost	Max. wickets lost	Std dev
England (v all teams)	30	111	161.23	214	28.36	5	6.97	10	1.59
Sri Lanka (v all teams)	37	101	163.68	260	32.04	3	6.30	10	2.15
England (v Sri Lanka only)	6	132	157.17	190	22.40	5	6.17	9	2.56
Sri Lanka (v England only)	6	128	161.50	189	24.49	4	6.60	10	2.39

Table 5: Summary statistics of actual innings totals and wickets lost by England and Sri Lanka when batting first in actual T20 innings (abandoned games excluded).

A limitation of this model is that, subject to the quality of the batsmen currently "in" and the bowlers they are facing, the "pace" of the game – in terms of scoring rate and the rate at which wickets fall – would be expected to be uniform. Given a specific batsman facing a particular bowler, the scoring rate and probability of the batsman getting out would be independent of the stage of the game. This is not realistic – if a top order batsman is still "in" in the later overs, he would be expected to score more freely but take more risks (and hence have a higher probability of getting out) than when facing the same bowler much earlier in the innings – particularly if several wickets had fallen early on. Furthermore, we have not yet considered the effect of such factors, or of changing the performance indicators used, on what is probably the most important outcome – the overall result of the match for the team: did they win, lose or tie? Finally, if the models developed and tested are considered to be sufficiently reliable, simulations using them could be used to estimate probabilities of particular match outcomes or events within matches. These in turn to be used as a method to "flag up" potentially anomalous incidents, which might be associated with illegal gambling or attempts to fix matches or aspects of matches. There are, therefore, plenty of opportunities for extending this work.

References

[1] 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.

[2] U. Damodaran. ODI cricket: characterizing the performance of batsmen using 'tipping points'. In *Proceedings of 4th International Conference on Mathematics in Sport*, 2013.

[3] K. Date. The Barnes Standard. [online], 2013. `http://www.espncricinfo.com/blogs/content/story/671745.html`.

[4] K. Date. The Richards Standard for ODI batsmen. [online], 2013. `http://www.espncricinfo.com/blogs/content/story/692433.html`.

[5] G. Hunter and D. R. Attanayake. Can probabilistic computer simulations be used to detect irregularities in cricket matches? In *Proceedings of 6th International Conference on Semantic*

E-Business and Enterprise Computing (SEEC 2013), pages 7–13, 2013.

[6] G. Hunter and P. Thethi. Batting averages in English county cricket : What can we learn from a formal statistical analysis? *Mathematics Today*, 48(3):135–139, 2012.

[7] G. Hunter and K. Zienowicz. Can Markov models accurately simulate lawn tennis rallies ? In *Proceedings of 2nd IMA International Mathematics in Sport Conference 2009*, 2009.

[8] A.C. Kimber and A.R. Hansford. A statistical analysis of batting in cricket. *Journal of the Royal Statistical Society A*, 156(3):443–455, 1993.

[9] A. Narayanan. Bowling Quality Index re-visited: incorporating home/away and recent form. [online], 2011. `http://www.espncricinfo.com/blogs/content/story/620607.html`.

[10] P. Newton and K. Aslam. Monte Carlo Tennis : A Stochastic Markov Chain Model. *Journal of Quantitative Analysis in Sport*, 5(3), 2009.

[11] D. Spanias and W. J. Knottenbelt. Predicting the outcomes of tennis matches using a low-level point model. *IMA Journal of Management Mathematics*, 24:311–320, 2013.

[12] K. Zienowicz, A. Shihab, and G. Hunter. The Use of Mel Cepstral Coefficients and Markov Models for the Automatic Identification, Classification and Sequence Modelling of Salient Sound Events Occurring During Tennis Matches. *Journal of the Acoustical Society of America*, 123(5):3431, 2008.

Graph Centrality Analysis for the Evaluation of the 2014 Guanabara Cup

Luana C. Brandão, Renata R. Del-Vecchio, and João Carlos C. B. Soares de Mello

Federal Fluminense University, Niteroi, Rio de Janeiro, Brazil
luanabrandao@id.uff.br, renata@vm.uff.br, jcsmello@pesquisador.cnpq.br

Abstract

Round-robin tournaments rank participants according to their results in all games, considering that all victories have the same value. However, our understanding is that victories against strong opponents have more value than those against weak opponents. Hence, such tournaments should consider this interpretation in their final rankings, using centrality measurements that evaluate the influence of an element based on its neighbours centralities. In this context, we use eigenvector and layer centralities to evaluate team performance in the 2014 Guanabara Cup, considering this understanding.

1 Introduction

In contrast to knockout tournaments, where each game eliminates one contestant, in round-robin tournaments, all contestants compete against each other. Round-robin tournaments are considered fairer because the teams' final results rely on many games, whereas, in knockout tournaments, occasional performances have great influence in the final ranking.

Nevertheless, round robins require many more matches for completion, and therefore are usually designed for national tournaments. The strongest national football leagues, which are determined by the International Federation of Football History & Statistics [14], all have round robin formats, such as *Liga Nacional de Fútbol Profesional* (Spain), *Serie A* (Italy), Premier League (England), etc.

In such tournaments, the teams receive a pre-determined amount of points for each type of result, e.g., 3 points for each win, 1 point for each draw, and zero points for losses. Although the number of points is the same, winning different games presents different levels of difficulty for the team. Therefore, we understand that victories or draws against strong opponents should have more value for the final ranking than those against weak opponents.

For that, we represent a round-robin tournament using a directed graph, where edges denote non-losses, i.e., winnings and draws. Thence, we calculate centrality measurements that evaluate the influence of each element, according to its neighbours' centralities, namely eigenvector and layer [2] centralities. When applied to our graph, such measurements rank teams that beat opponents who, in turn, had many victories, better than teams that beat opponents who, in turn, had few victories.

The case study analysed herein is the Guanabara Cup, disputed by 16 clubs from the State of Rio de Janeiro. This football championship has been the object of previous studies, such as Alves et al [1], who analysed the championship's most incoherent results and their effects on the final ranking. In the present work, we use eigenvector and layer centralities, to relativize the importance of victories and draws, for the evaluation of team performance in the 2014 Guanabara Cup, which was a complete round robin for the first year.

2 Bibliographic Review

The basis of this study are directed graphs, identified by an ordered pair $G = (V, E)$, where V is the set of n vertices and E is the set of m edges. Each edge is formed by an ordered pair of vertices from V, i.e., $v_i, v_j \in V$, and is directed from v_i to v_j. Connected vertices may be called neighbours.

In this study, we use weakly connected graphs, i.e., when we replace the directed edge with undirected edges, we are able to find at least one path that connects any two vertices. Mathematically, we define a path as the graph $P = (V, E)$, where $V = \{v_1, \ldots, v_n\}$ and $[v_i, v_{i+1}] \in E$ for $1 \leq i < n$ [5]. In other words, a path is a sequence of vertices, with edges connecting each vertex to the following one.

A matrix of order n, $A = A(G) = [a_{ij}]$, where $a_{ij} = 1$, if $(v_i, v_j) \in E$ and $a_{ij} = 0$ if $(v_i, v_j) \notin E$ is called the adjacency matrix for G, which is asymmetric for directed graphs. The characteristic polynomial of the adjacency matrix is $p_G = \det(A(G) - \lambda I)$, where the roots $\lambda_1, \ldots, \lambda_n$ are eigenvalues for G. The eigenvalue λ_1, where $|\lambda_1| \geq |\lambda_i|$ for $1 \leq i < n$, is called the matrixs spectral radius.

Centrality measurements, in general terms, evaluate the importance of a vertex with regard to a network. However, there are different forms of measuring this importance, which is why there are different measures of centrality. In the present study, we use degree centrality, eigenvector centrality, and the methodology initially proposed by Bergiante et al [2], which, from now on, we will call layer centrality.

Degree centrality incorporates the original definition of centrality, proposed by Freeman [10], and corresponds to the degree of each vertex. Using a simple adjustment to the definition in Sinha and Mihalcea [20], we measure the vertices' degrees in the graphs herein as the difference between the sum of the outgoing arcs (out-degree), which represent a non-loss in our graph, and the sum of the incoming arcs (in-degree), which represent a non-win. In the adjacency matrix, it is the difference between the sum of the vertex's row and the sum of the vertex's column, as shown in (1).

$$C_D(v_i) = d(v_i) = \sum_{j=1}^{n} a_{ij} - \sum_{j=1}^{n} a_{ji}, \quad \text{where} \quad v_i \in V \tag{1}$$

Since, in the present study, we are interested in the opponents' centralities, we propose the use of the eigenvector centrality [4], where an element's centrality corresponds to the linear combination of the neighbours' centralities. Using the adjacency matrix, the eigenvector centrality x_i of a vertex v_i satisfies equation (2), or $Ax = \lambda x$, in matrix notation.

$$\lambda x_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n \tag{2}$$

In (2), the λ solutions correspond to the eigenvalues of the adjacency matrix, whereas $(x_1, \ldots, x_n)$ correspond to the matrix's eigenvectors. To calculate the vertices' centralities, we use the entries for the unitary positive eigenvector associated with λ_1 [3].

Since we only need λ_1 and the corresponding eigenvector, we could use the iterative power method to compute their approximations, if the matrix is diagonalizable, non-negative and corresponds to a connected graph [12]. In this method, we start by multiplying the matrix of order n, by an arbitrary non-null column vector with n elements y_0, e.g., the vector of all ones. The resultant vector, y_1, should be normalized and used to multiply the matrix in the following iteration, for repeated iterations as shown in (3).

$$y_k = Ay_{k-1} = A^2y_{k-2} = \cdots = A^k y_0 \tag{3}$$

From (3), it is possible to show equation (4) [12], in which r represents any given coordinate from the vectors.

$$\lim_{k\to\infty} \frac{(y_k)_r}{(y_{k-1})_r} = \lambda_1 \tag{4}$$

Therefore, the iterative process should be repeated until it converges with precision ϵ, i.e., when $\left| l_r^{(k)} - l_r^{(k-1)} \right| \leq \epsilon$, where $l_r^{(k)} = \frac{(y_k)_r}{(y_{k-1})_r}$.

However, the adjacency matrices for directed graphs are asymmetric, thus they may present complex eigenvalues, and eigenvectors with complex entries [22]. In other cases, particularly when there are elements with no status, i.e., teams that only lost in a given tournament, the eigenvector analysis produces meaningless results [4]. To avoid these mathematical obstacles, and also to simplify calculations, we propose a third ranking based on layer centrality, which considers, to some extent, neighbours' centralities.

This methodology begins by identifying the vertex (or vertices) with the smallest difference between the sum of the entries in the corresponding row(s) and the sum of the entries in the corresponding column(s), and ranking it in last place. Then we exclude from the adjacency matrix, the row(s) and column(s) associated with this vertex (or vertices). Thence, we repeat this process until all the vertices are ranked. This methodology does not provide a centrality measurement, only a centrality ranking, although this may be sufficient for the decision maker.

The layer centrality may be considered an intermediary between the other centrality measures presented herein. On one hand it has similarities with degree centralities, because the bottom of the ranking is highly dependent on the vertices degrees. More specifically, the last ranked element is the same in both rankings, because the first elimination in the layer methodology depends solely on the vertices degrees. Furthermore, it is relatively simple and allows a more natural interpretation of results.

On the other hand, the layer methodology considers, to a certain extent, the centralities of neighbours, because the elements connected to the eliminated vertices lose these connections in the following iteration. Hence, elements that present high degrees, due to connections with peripheral elements, lose these connections early in the iterative process and thus are ranked poorly. Therefore, the top of the ranking tends to dissimilate from the degree centrality ranking. In fact, for certain cases, still not fully studied, the layer centrality ranking provides results similar to eigenvector centrality.

3 Applications for Centrality Measurements

Centrality measurements have been used to evaluate the importance of elements in various types of networks. Del-Vecchio et al [8] analysed the Brazilian stock market; Gonçalves et al [11] applied centrality measurements to create a more harmonious urban environment; Hussain [13] used centrality measurements to destabilize terrorist network; Kirkland [17] explored a different notion of vertex centrality and applies it to food web data; Joyce et al [16] developed a centrality measurement for networks in the human brain; Page et al [19] proposed a centrality measurement for various applications on the World Wide Web; among many other studies that provided various applications of graph centrality, many of which proposed new measurements, more suitable for the case under consideration. In this scenario, Borgatti [6] classified network flows based on the kinds of trajectories that traffic may follow and the method of spread, to match each type of flow to the most suitable centrality measures.

Particularly, centrality measurements have also been applied to sports. Jiang et al [15] used centrality measurements to determine the best college coaches of the past centuries for

men's basketball, football, and baseball. The authors calculated the teams' skills using eigenvector centrality, to rank teams that beat strong opponents higher than those that beat weak opponents.

4 Case Study

From 1982 to 2013, except for the 1994 and 1995 seasons, the Guanabara Cup was the first stage for the Carioca Championship, whereas the Rio Cup was the second stage. The clubs were divided into two groups and played, for each stage, a qualifying phase, semi-finals and a final match. In the qualifying phase, the clubs from one group played against the clubs from the other group in the Guanabara Cup, whereas in the Rio Cup, the clubs from each group played against each other.

In 2014, the championship was restructured into a single stage with a qualifying phase, semi-finals, and a final match. The qualifying phase, called the Guanabara Cup, which is the object of this case study, consists of a complete round robin disputed by 16 clubs from the State of Rio de Janeiro. The final ranking for this phase is computed, attributing 3 points for each victory, 1 point for each draw, and zero for losses.

However, these clubs present very different strengths, inasmuch as they dispute different national championships – some dispute *Brasileirão Série A*, which is the top Brazilian football championship, while other clubs dispute other levelled championships, including *Brasileirão Série D*, the fourth and lowest tier of Brazilian football. In view of this fact, victories against different clubs in the Guanabara Cup, present particularly different levels of difficulty for the teams, and, thus, should be valued differently.

Therefore, we propose new performance evaluations for the 2014 tournament, which take the opponents strengths into consideration. For that, we use a directed graph to represent the tournament, in which the vertices symbolize the clubs, and the edges characterize a non-loss. In other words, if team x beat team y, there will be a single edge from x to y, whereas if x and y tied, there will be an edge from x to y and another from y to x.

In this graph, the degree centrality, calculated as in (1), measures the net balance of game winnings, i.e., the difference between victories and losses. This provides results equivalent to the Copeland method [7], which is commonly used in sports ranking [18, 21]. It is also equivalent to a tournament ranking that attributes 2 points for each victory, 1 for each draw, and zero for losses. Since the tournament under evaluation attributes 3 points for victories, the degree centrality ranking may be somewhat different than the official ranking.

Nevertheless, these methodologies evaluate all winnings equally, despite the differences between beating a strong and a weak team. Therefore, we also calculate the teams eigenvector centralities. Since the graph is connected, and the matrix is non-negative and diagonalizable [12], we may use the power method, described herein, to calculate λ_1, with $\epsilon \approx 10^{-3}$. This measurement, applied to our graph, evaluates teams that beat strong opponents better then those that beat weak opponents [15].

To overcome the aforementioned mathematical obstacles associated with eigenvector centralities, we also calculate the layer centrality ranking, which considers, to some extent, the differences between beating strong and weak opponents, when evaluating team performance. This third centrality is also more natural and intuitive, particularly for result interpretation.

We use the results for each match in the 2014 Guanabara Cup, available in FERJ [9], to build the adjacency matrix for the tournament, and subsequently calculate the aforementioned centrality measurements. Table 1 shows the rankings for the degree, the eigenvector, and the layer centralities, ordered by degree centrality. Table 1 does not present the official ranking

because it considers a different punctuation system, and thus it should not be compared to the rankings calculated herein. However, the official ranking is very similar to the degree centrality ranking, regardless of certain position inversions.

	Centrality Rankings		
Clubs	Degree	Eigenvector	Layer
Flamengo	1	1	2
Fluminense	2	2	1
Vasco da Gama	3	3	5
Boavista	4	4	6
Cabofriense	5	4	3
Friburguense	6	7	7
Nova Iguaçu	7	5	4
Macaé	8	8	9
Botafogo	9	11	12
Volta Redonda	10	13	11
Bangu	11	12	10
Resende	12	10	8
Bonsucesso	13	8	13
Madureira	14	15	14
Audax Rio	15	14	15
Duque de Caxias	16	16	16

Table 1 – Centrality Rankings for Clubs performance in the 2014 Guanabara Cup

We can observe from Table 1 that Cabofriense, Nova Iguaçu, and Resende were much better ranked in the second and third rankings than the first. Analysing the original data, we notice that Cabofriense tied with Fluminense; Nova Iguaçu tied with Cabofriense and with Vasco da Gama; and Resende tied with Boavista, Bonsucesso, and Nova Iguaçu, besides winning the game against Macaé. These are the games against relatively strong opponents that influenced the second and third rankings, but were given the same value as other games, against weak opponents, in the first ranking, and in the final tournament ranking. Furthermore, Botafogo lost positions in the second and third rankings because 4 out of the 9 earned points came from games against Bonsucesso, Madureira, Audax Rio and Duque de Caxias.

On the other hand, Bonsucesso is well ranked in terms of its eigenvector centrality, but the layer centrality ranks the team in the same position as the degree centrality. This happens because the bottom of the layer and the degree centrality rankings are alike, as previously explained herein. Similarly, Volta Redonda's position in the layer ranking is similar to the degree ranking, despite the team's bad evaluation in terms of eigenvector centrality, also because it is positioned toward the bottom of the degree ranking.

Interestingly, Flamengo is ranked first according to the first two rankings, but second according to layer centrality. Analysing the original data, we notice that Flamengo won almost every game in the tournament, except for their tie with Bangu and their loss against Fluminense, whereas Fluminense lost to Madureira and Botafogo, and tied with Bonsucesso, Cabofriense, Duque de Caxias, and Vasco da Gama. The position reversal happened because the layer methodology compares the teams that remain in the matrix at each step. Since Fluminense and Flamengo are the only teams left in the matrix for the last iteration, and since Fluminense beat Flamengo, it is ranked first. Therefore, in this case, the layer centrality ranking presents some similarities with finals in knockout tournaments, because when there are two teams left to rank, the winner of the match will occupy the first position.

Although Flamengo remains first in the eigenvector ranking, the relative difference between Flamengo's and Fluminense's eigenvector centralities (2.7% of the entire range) is much smaller then the relative difference between Flamengo's and Fluminense's degrees (21% of the entire range), as shown in Table 2. In other words, despite sustaining the first place in the second ranking, Flamengo partially loses its dominance in terms of eigenvector centrality.

Similarly, we notice, from Table 1, that Vasco da Gama is in third place in the first two rankings, but in fifth place in terms of layer centrality. From Table 2, we may observe that there is a small difference between Vasco da Gama and Fluminense in terms of degree centrality (5% of the entire range), but a greater difference between these two teams in terms of eigenvector centrality (11% of the entire range). In other words, although Vasco da Gama remains in third place in the eigenvector centrality ranking, it is further away from the second-place, than in the degree centrality.

Table 2 shows measurements for the degree and the eigenvector centralities, ordered by degree centrality.

	Centrality Measurements	
Clubs	Degree	Eigenvector
Flamengo	11	0.349
Fluminense	7	0.343
Vasco da Gama	6	0.320
Boavista	3	0.270
Cabofriense	3	0.286
Friburguense	1	0.236
Nova Iguaçu	0	0.270
Macaé	-1	0.222
Bangu	-2	0.214
Botafogo	-2	0.217
Volta Redonda	-2	0.214
Bonsucesso	-3	0.228
Resende	-3	0.218
Madureira	-4	0.182
Audax Rio	-6	0.184
Duque de Caxias	-8	0.144

Table 2 – Measurements for degree and eigenvector centralities

5 Conclusions

In the present study, we evaluated team performance in the 2014 Guanabara Cup tournament, disputed by teams with very different capacity levels, and whose format was a complete round robin for the first year. We used eigenvector and layer centralities to relativize the importance of each victory and draw, according to the opponents' strengths. We also calculated the degree centralities, to compare results.

The advantage of layer centrality over eigenvector centrality is that, besides avoiding mathematical obstacles related to asymmetric matrices, as well as complicated calculations, the first presents a more natural and intuitive interpretation than the second. Interestingly, the layer centrality ranking presented similarities with finals in knockout tournaments, because when ranking the first two teams, the match between them decides the first position. This is why

this methodology inverted the first and the second positions, when compared to the other rankings.

Future studies could consider the tournament's punctuation system, i.e., 3 points for victories, 1 point for draws, and zero points for losses, when building the adjacency matrix corresponding to the directed and weighted graph. In this case, the degree centrality, measured as the sum of the out-degrees, would be exactly the teams' total amount of points in the tournament. Moreover, other works may adapt the methodology proposed herein for double round-robin tournaments, such as *Brasileirão* and other football leagues.

Acknowledgements

We acknowledge the financial support of CNPq (Brazilian Ministry of Science and Technology) and Capes (Coordination for the Improvement of Higher Education Personnel).

References

[1] A.M. Alves, J.C.C.B. Soares de Mello, T.G. Ramos, and A.P. Sant'Anna. Weak rationality in football results: An analysis of the Guanabara Cup using the Bowman and Colantoni method. In *Proceedings of the 3rd IMA International Conference on Mathematics in Sport, Salford*, 2011.

[2] N.C.R. Bergiante, J.C.C.B. Soares de Mello, M.V.R. Nunes, and Paschoalino F.F. Aplicação de uma proposta de medida de centralidade para avaliação de malha aérea de uma empresa do setor de transporte aéreo brasileiro. *Journal of Transport Literature*, 5(4), 2011.

[3] P. Bonacich. Power and centrality: A family of measures. *The American Journal of Sociology*, 92(5):1170–1182, 1987.

[4] P. Bonacich and P. Lloyd. Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, 23:191–201, 2001.

[5] A. Bondy and U.S.R. Murty. *Graph Theory*. Springer, 2008. ISBN: 978-1-84628-969-9.

[6] S.P. Borgatti. Centrality and network flow. *Social Networks*, 27(1):55–71, 2005.

[7] A.H. Copeland. *Reasonable Social Welfare Function*. University of Michigan, Detroit, Michigan, 1951.

[8] R.R. Del-Vecchio, D.J.C. Galvão, L. Silva, and R.F.V.L. de Lima. Medidas de centralidade da teoria dos grafos aplicada a fundos de ações no brasil. In *XLI Simpósio Brasileiro de Pesquisa Operacional*, Porto Seguro, Brasil, 2009.

[9] FERJ. Campeonato Estadual da Série A de Profissionais de 2014 – Taça Guanabara. [online], 2015. `http://www.fferj.com.br/Campeonatos`, Accessed on 5 March 2015.

[10] L.C. Freeman. Centrality in networks: I conceptual clarification. *Social Networks*, 1:215–239, 1979.

[11] J.A.M. Goncalves, L.S. Portugal, and C.D. Nassi. Centrality indicators as an instrument to evaluate the integration of urban equipment in the area of influence of a rail corridor. *Transportation Research Part A*, 43(1):13–25, 2009.

[12] R.A. Horn and C.R. Johnson. *Matrix Analysis (2nd ed.)*. Cambridge University Press, New York, 2013.

[13] D.M.A. Hussain. Destabilization of terrorist networks through argument driven hypothesis model. *Journal of Software*, 2(6):22–29, 2007.

[14] IFFHS. The worlds strongest national league 2014. [online], 2015. `http://www.iffhs.de/the-worlds-strongest-national-league-2014/`, Accessed on 6 March 2015.

[15] T.S. Jiang, Z.T. Polizzi, and C.Q. Yuan. A networks and machine learning approach to determine the best college coaches of the 20th-21st centuries, 2014. *CoRR* abs/1404.2885.

[16] K.E. Joyce, P.J. Laurienti, J.H. Burdette, and S.S. Hayasaka. Algebraic connectivity for vertex-deleted subgraphs, and a notion of vertex centrality. *PLoS ONE*, 5(8):e12200, 2010.

[17] S. Kirkland. Algebraic connectivity for vertex-deleted subgraphs, and a notion of vertex centrality. *Discrete Mathematics*, 310(4):911–921, 2007.

[18] V.R. Merlin and D.G. Saari. Copeland method ii: Manipulation, monotonicity, and paradoxes. *Journal of Economic Theory*, 72(1):148–172, 1997.

[19] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[20] R. Sinha and R. Mihalcea. Using centrality algorithms on directed graphs for synonym expansion. In *Proceedings of the 24th International Florida Artificial Intelligence Research Society Conference. Palm Beach, Florida*, 2011.

[21] J.C.C.B. Soares de Mello, L.F.A.M. Gomes, E.G. Gomes, and M.H.C. Soares de Mello. Use of ordinal multi-criteria methods in the analysis of the formula 1 world championship. *Cadernos Ebape.BR*, 3(2):1–8, 2005.

[22] W. Wei, J. Pfeffer, J. Reminga, and K.M. Carley. Handling weighted, asymmetric, self-looped, and disconnected networks in ORA. Technical report, Carnegie Mellon University, Pittsburgh, 2011.

Predictive Modelling of Team Success in the t20 Big Bash League

Aaron Corris, Anthony Bedford and Ian Grundy

RMIT University
s3238499@student.rmit.edu.au

Abstract

Due to the shortened nature of a T20 cricket match compared to Test or 50 over cricket, many commentators believe the matches to be little more than luck. This paper looks at the Australian domestic t20 competition, known as the Big Bash League, to look at whether its possible to predict matches with a better than 50-50 chance at a team versus team level. An Elo model was built to predict the results of matches and compared against the results of predicting the market favourite to win each match. The Elo model predicted more matches correctly than picking the favourite according to the market, but picking the market favourite had the best result in any one season. Home Ground Advantage, winning the toss and which team batted first were factors investigated, however none of those factors had a significant impact. Ultimately, the Elo model appears to be a reasonably good indicator for team strength in the Big Bash League.

1 Introduction

This paper investigates whether the performance of teams in the Australian Twenty-20 (t20) cricket league known as the Big Bash League (BBL) can be modelled using an Elo model. Initially developed for chess, the Elo model has been adapted for use in a wide variety of sports and competitions such as association football (soccer), tennis, volleyball, Australian Rules football, Scrabble, baseball and video games (eSports). However, there does not appear to be any attempts to apply such a model to domestic t20 cricket leagues such as the BBL or the Indian Premier League (IPL) even though there seems to be sound logic for doing so.

The Elo ranking model rates a team based off their past results. In any given match, the two teams have their pre-game rating which provides a probability of winning the match for each team. This rating then changes based off the result of the match. The winning team's rating will increase, while the losing team's rating will decrease, with the magnitude of the change being dependent on the pre-game ratings. The winning team effectively takes the ranking points from the losing team as their rankings will change by the same amount, but in opposite directions. A team will gain fewer ranking points if they defeat a lower rated team than if they defeat a higher rated team. Similarly, a team that loses to a higher rated team will lose fewer points than a team that loses to a lower rated team. Thus, the model corrects for inaccuracies relatively quickly as an overrated team will lose a lot more points when defeated by an underrated team.

2 Data and Methodology

Each BBL match since its inception in December 2011 was collated into a spreadsheet. The data collected were the wickets and runs for each team, how many balls were faced in each innings, who won the match, who won the toss, who batted first, and the pre-match odds. An

Elo model was constructed on the data, using the equations

$$Exp_{Home} = \frac{1}{1 + 10^{(R_{Away} - R_{Home})/\theta}}$$

and

$$Exp_{Away} = \frac{1}{1 + 10^{(R_{Home} - R_{Away})/\theta}}$$

for the expected win probabilities of home and away teams respectively, and

$$R'_{Home} = R_{Home} + K(Obs_{Home} - Exp_{Home})$$

and

$$R'_{Away} = R_{Away} + K(Obs_{Away} - Exp_{Away})$$

to adjust the ratings of the respective teams after each match. The model uses

$$\theta = 400$$

and

$$K = 39$$

as its parameters.

Analysis of the data and the model was done using the statistical software SPSS, while the optimisation of the model was done using @Risk Optimizer.

3 Results

The initial tests done were to investigate whether winning the toss, batting first or playing a match at your home ground had any noticeable impact on the likelihood of winning. It was hypothesised that playing at home and winning the toss would both be desirable for a team and therefore improve the chances of winning, while choosing to bat first would have no effect. Rain affected matches were excluded from this analysis since it was unknown how these factors might differ when playing conditions change and of the 136 matches only 8 were rain affected so an adequate sample of matches to compare any differences could not be found.

The following SPSS output table shows the proportion of matches won by the home team (home), the team batting first (bat1st) and the team that won the toss (toss), compared to the assumption that these factors would have no effect.

	Test value = .5					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
home	.352	127	.725	.01563	-.0721	.1034
bat1st	.000	127	1.000	.00000	-.0878	.0878
toss	-.176	127	.860	-.00781	-.0956	.0800

Table 1: SPSS output of T-test for advantage of home ground, batting first and winning the toss

Contrary to expectation, it appears as though none of the factors had any significant effect on the likelihood of a team winning a match. This was quite a strange result as logically it would be expected that a team would have familiarity with their home ground and the associated conditions and therefore have some level of advantage compared to playing at a somewhat foreign ground. Similarly, the lack of any advantage for winning the toss was quite confusing as in theory winning the toss should provide a team with the opportunity to take advantage of the best conditions for batting or bowling as a result. This was investigated further on a season by season basis.

Season		Test value = .5					
		t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
						Lower	Upper
2012	home	1.316	28	.199	.12069	-.0671	.3085
	bat1st	.926	28	.362	.08621	-.1045	.2769
	toss	.183	28	.856	.01724	-.1762	.2107
2013	home	-.701	31	.488	-.06250	-.2442	.1192
	bat1st	-1.063	31	.296	-.09375	-.2737	.0862
	toss	-.701	31	.488	-.06250	-.2442	.1192
2014	home	.171	32	.865	.01515	-.1648	.1951
	bat1st	.516	32	.609	.04545	-.1338	.2248
	toss	-.171	32	.865	-.01515	-.1951	.1648
2015	home	.000	33	1.000	.00000	-.1771	.1771
	bat1st	-.339	33	.737	-.02941	-.2062	.1474
	toss	.339	33	.737	.02941	-.1474	.2062

Table 2: SPSS output of T-test for advantage of home ground, batting first and winning the toss split by season

Splitting the matches by season showed that there was no trend in the changing of any of these factors. This suggests that the lack of advantage is not due to the players adapting to the factors, but rather the factors having little to no appreciable effect on the outcome of the match. It was considered that perhaps a team based analysis may show some sort of advantage to some teams, however as the teams had only played an average of 33 matches each there was not sufficient evidence to conclude anything on a team by team basis.

As home ground advantage, batting first and winning the toss were all deemed to lack any significance for who won the game, the sole important factor for the optimisation of the Elo model was the K weighting for the change in ratings post-game. The seasonal breakdown of the model predictions can be seen in the following table.

Season	**Correct**	**Matches**	**% Correct**
11-12	20	31	64.52%
12-13	20	35	57.14%
13-14	20	34	58.82%
14-15	22	34	64.71%
Total	82	134	61.19%

Table 3: Correctly predicted BBL matches by season according to optimised model

The mean squared error (MSE) of various combinations of previous seasons was minimised to find the best result for the 2014-15 season. The best fit was found to be a K-value of 39,

resulting in an overall prediction rate of 61.2% and a 64.7% prediction rate in the 2014-15 season. This was a rather consistent result from season to season, with the predictions always exceeding 57% within a season, but never exceeding 65%.The significance of this becomes clear when comparing this result to the market.

Predicting matches based on the pre-game odds of the market was expected to be somewhat successful as bookmakers do not want to lose money off an event, and if they are inaccurate they are more likely to do so. The seasonal breakdown of the market predictions can be seen in the following table.

Season	**Market %**
11-12	48.39%
12-13	51.43%
13-14	70.59%
14-15	44.12%
Total	53.73%

Table 4: Proportion of correctly predicted matches by season according to market odds

Interestingly, the 2013-14 season was extremely predictable according to the market, with a whopping 70.6% of matches being predicted correctly. Despite that, the overall predictive rate of the market was barely better than 50%, and two of the four seasons were worse than 50%. It's very strange that the market is performing so poorly in this case, as they're generally considered to be the benchmark. This suggests that the methodology being used by the bookmakers for modelling t20 cricket has some major flaws in it.

Ultimately, it appears that the Elo model is a useful tool for predicting the match results in the BBL.

4 Discussion

As mentioned above in the results, the methodology behind the market odds appears to be somewhat flawed. While some of this may be explained by poor wagering choices from customers of the bookmakers, there's too much difference between the Elo model and the market performance for it to simply be biased by the customers. My hypothesis is that the bookmakers evaluate the ability of a team by adding up the theoretical contributions of each of the likely members of the playing XI. This method may not accurately account for factors such as team balance at selection, tactics from the captains, harmony of the team, etc and thus there exists a discrepancy between the theoretical sum total of the player contributions, and the overall team performance.

One such example could be the performance of the Melbourne Stars, who have widely been considered by experts to have the strongest squad in multiple BBL seasons, but have consistently performed worse than expected. If one simply adds up the expected individual contributions of the Stars players, it would exceed that of any other team. However, in a match scenario the intra and inter team interactions end up with the Stars losing to (on paper) inferior opposition. As the Elo model only looks at the team performances, these factors are intrinsically taken care of by the model while the market may not do so.

One factor that deserves further consideration in the future is altering the K factor based on margin of victory. This is simple for matches where the team batting second fails to chase the target, as the margin of victory is then simply the difference between the runs scored at the close of each innings. However, when the team batting second successfully chases down

the target score, they win by x number of wickets with y number of balls to spare. While it is obvious that a team which wins by 6 wickets with 10 balls to spare has won by a greater margin than a team which wins by 3 wickets with 5 balls to spare, it is much trickier to decide whether a 3 wicket win with 10 balls to spare is a greater margin of victory than a 6 wicket win with 5 balls to spare. Furthermore, there's no easy way to compare a win by runs to a win by wickets with balls to spare.

Possible solutions to the margin of victory dilemma include looking at the net run rates or the Duckworth-Lewis projection. Net run rate is used as the tiebreaker for teams on equal wins and is quite simple to calculate, however it fails to acknowledge the wickets in hand constraint for winners who bat 2nd. Using the Duckworth-Lewis projection would likely give a much better margin of victory in terms of runs, however the Duckworth-Lewis calculations are significantly more complex. While this has been looked at in 50 over cricket, the shortened t20 game may yield different results since the reduction in total deliveries for the match means it becomes a much stricter constraint.

Further data is required for results on whether certain teams are favoured by playing at home, winning the toss and batting first. While the limited data available suggests that Perth have some level of advantage both when they play at home and when they bat first, there is simply not enough data to make any reliable conclusion on this. This simply requires more matches to be played so the data exists.

Finally, it is worthwhile determining whether this model can also be applied to the IPL and even international t20 cricket. I suspect that it will fall apart for international t20 cricket as those matches tend to be few and far between, potentially resulting in drastic changes in the selected teams between matches. However, a very similar model should be applicable to the IPL as it is a very similar format to the BBL being franchise based round robin tournaments played over 1-2 months.

5 Conclusion

The Elo model appears to work quite well at predicting matches in the BBL. It is quite consistent from season to season unlike the market predictions, and has outperformed the market in three of the four seasons and overall across all four seasons. Playing at your home ground, winning the toss and batting first had no appreciable effect on the probability of winning a match despite expert knowledge suggesting they would have some level of impact. So while a considerable level of luck is involved in a t20 match, it does appear to be somewhat predictable with a reasonable level of confidence.

Skills Importance across Ages for Men's Volleyball

S. Drikos and I. Ntzoufras

Faculty of Physical Education and Sport Sciences, National and Kapodistrian University of Athens
sdrikos@gmail.com

Abstract

Volleyball is a competitive team sport whose main objective is to score the most points by grounding the ball to the opponents side of the court. The numbers of points a team scores is primarily based on the execution of the skills of the game. Due to the hierarchical structure of the game events follow stable patterns: serve outcome, pass–set–attack (complex 1) outcome, Serve–block–dig–set– counter attack (complex 2) outcome. The outcome consists of four possibilities: To win or to lose a point, continuation of the action with the ball on the team's side or with the ball on the opponent's side. In this paper we provide a method to calculate the importance of every skill on the outcome of the Volleyball game for youth, juniors and men teams. We base our analysis on Fellingham's model for importance scores. We use data from the performance analysis of the winning team from the most recent world championships in all ages for male volleyball. In order to arrive at posterior distributions we use a Markov chain transition matrix. To estimate the Markovian transition matrix, we assume a multinomial likelihood with a Dirichlet prior on the transition probabilities. The prior distribution of skills is formed by expert coaches' opinion. Our main purpose is to examine if key skills are consistent across ages in male volleyball.

1 Introduction

Volleyball is a competitive team sport in which the aim of each team to score first a required number of points by grounding the ball to the opponent's side of the court. Volleyball belongs to the same team sports category with Tennis, Badminton and Table Tennis. All of them are described as *Net and Wall games* [6].

In international tournaments, a volleyball match consists of at least three and a maximum of five sets with the winning team reaching first three earned sets. Each set is played until one team reaches a score of 25 points with at least two points difference from its opponent. The number of points a team scores is primarily based on the efficiency of the team skills of the game. Volleyball has three scoring skills (serve, attack, block) and three non scoring skills (pass, set, dig). For each team a maximum of three contacts of the ball is allowed before the ball crosses the net to the opponent's court. Each team in its effort to pass the ball to the other court exhausts all the contacts most of the times. A rally begins when a member of one team serves the ball across the net to the opposite team. Thus there are two complexes of the game. Complex 1 consists of pass, set and attack (or attack 1) and complex 2 consists of serve, block, defense and counterattack (or attack 2).

Quantitative analysis is not something new in the area of Volleyball but most of the times it is restricted to characterizing players' or teams' performance per season, per match or in more detail per set. Thus there are elements for the importance of each skill for seasons, tournaments, matches, sets, even for ambivalent sets [2], but there are very little data for the importance of skills and the level of their execution on the ending of a single rally.

In this work we implement a method to calculate the importance of every skill on the outcome of the Volleyball game for youth, juniors and men teams. We base our analysis on

Fellingham's model [4] for estimating the importance scores. Our main purpose is to examine whether the key skills are consistent across tournaments for different ages in male volleyball.

Due to the hierarchical structure of the game, events follow consistent patterns: serve outcome, pass–set–attack 1 (complex 1) outcome, serve–block–dig–set–counter attack 2 (complex 2) outcome. All the attacks, except for those following a serve's pass, are attack 2. The outcome consists of three possibilities: to win a point, to lose a point or to the continuation of the game on the opponent's side. Due to the data collection method, the continuation of the game in the opponents side cannot be considered as a final try since the ball does not cross the net.

2 The data

All recorded data refer to the performance of the winning team from the most recent world championships of National teams in all age categories for male volleyball. National World Championships for three age categories (youth, juniors and men) currently take place every four years.

All data record the performance on selected matches of the latest World national team champions (Poland in Men, 2014; Russia in juniors and in youth, both for 2013).

We record and analyze only the actions of the selected team when the ball is in its court separately for both complexes. All skills were categorized in two complexes, except for the serve and serve's pass. Thus we have set 1 and 2, block 1 and 2, defense 1 and defense 2, free ball 1 and free ball 2. For each skill we use a different ordinal grading system on tactical scale with one being a poorly performed skill and the maximum to be a skill performed in the optimal way. For serve, jump and float and for pass, against jump or against float serve we use a six-level ordinal tactical scale [1]. Sets, either after serve's pass or after defense, are divided according to the position on the court and the tempo of the ball. We include five possible positions of attack of two tempos for each set. We have also added a variety of attack moves: attack as second contact from the setter, attack out of system and direct attack when the ball is driven directly from the opponent's court. For block, either after serve or after attack, we use a three-level ordinal scale [5].

For defense, we use two states: free ball and dig, either after serve or after attack. Finally, the skill of *set error* records the case where the recorded team makes an unforced error during the second touch.

All recorded skills and their properties are summarized in Table 1. The data are recorded in a 60×62 matrix denoted by $\mathbf{y}$ and each element y_{ij} recording the number of occurrences of each subsequent moves and skills (how many times skill S_j follows skill S_i). Column and row 60 refer to the continuation at the field of the opponent and the last two columns are terminal moves resulting to a point in favor or against the team under study.

Table 1. Performance ratings for all skills

3 Method of analysis

In this paper we use the methodology of Fellingham's and Reese [3] to evaluate the importance of each skill. According to their approach, the coefficient of the skill importance for the i-th recorded skill is defined as the ratio of the posterior mean of P_i over its corresponding standard deviation, that is

$$I_i = \frac{E(P_i|y)}{\sqrt{V(P_i|y)}}$$

Skill (main)	Skill (sub)	Levels	Complexes	Tempos
Serve	Serve ace	6		
(Jump & float)	The ball drives directly to the serving team's court	5		
	The ball is on receiving team's court but with just one option for attack	4		
	The ball is on receiving team's court with two options for attack	3		
	The receiving team has all the options for attack	2		
	Serve error	1		
Pass	The receiving team has all the options for attack in optimum tempo	6		
(vs Jump & vs Float serve)	The receiving team has all the options for attack	5		
	The ball is on receiving team's court with two options for attack	4		
	The ball is on receiving team's court but with just one option for attack	3		
	The ball drives directly to the serving team's court	2		
	Pass error	1		
Block	Block kill	3	2	
	The ball remain to the blocking team's court	2	2	
	The ball remain to the attacking team's court	1	2	
Setting location	Left Front Side		2	2
	Right Front side		2	2
	Right Back side		2	2
	Middle Front side		2	2
	Middle Back side		2	2
	Out of system		2	
	Setter's Tip or attack in 2nd touch		2	
	Setting error		2	
	Direct attack		2	
Defense	Dig		2	
	Free ball		2	

2 complexes indicate that the skills are recorded for both complex 1 and 2 separately; 2 tempos indicate that the skills are recorded for quick and high tempos

where P_i is the probability that this skill will eventually end up in a point in favor of the team under study after two subsequent game moves and is calculated by

$$P_i = P(Y_{t+1} = \text{point}^+ | Y_t = S_i) + \sum_{k=1, k\neq i}^{n} P(Y_{t+2} = \text{point}^+ | Y_{t+1} = S_k) P(Y_{t+1} = S_k | Y_t = S_i) \tag{1}$$

with n being the number of skills, $P(Y_{t+1} = \text{point}^+ | Y_t = S_i)$ being the probability of scoring a point in favor of the team under study after a skill S_i and $P(Y_{t+1} = S_k | Y_t = S_i)$ is the transition probability from skill $S_i$to skill S_k. We further assume that the scoring for each skill is not influenced by the time point, hence

$$P_i = P(Y_{t+1} = \text{point}^+ | Y_t = S_i) = P(Y_{t+2} = \text{point}^+ | Y_{t+1} = S_k).$$

This ratio gives the opportunity to attach an importance score to every to every skill and its corresponding levels or categories. Every time the ball is on the side of the observed team a sequence of events takes place that follow specific schemes: serve–outcome, pass–set1–attack1–outcome and defense–set2–attack2–outcome. We assume that these schemes are of first order Markov chains. We record these sequences in a transition probabilities matrix where data of the matrix represent the probability to move from one state to another. Because we observed three teams, we created three transition matrices with dimensions 60×62 (two states related to point earned or lost by the team are terminal), containing the transitions for jump and float serves, passes against jump and float serves, sets after serve's pass by location and tempo, sets after defense by location and tempo, block after serve, blocks after attack, free balls and digs and possible outcomes. Because of the hierarchical structure of the game, there are sequences that are not feasible, such as moving from an excellent pass to an ace serve. These cases are restricted to zero probability.

We use a simple Bayesian model to estimate the transition probabilities and through them the success probabilities P_i as defined in (1). For simplicity, we denote by $\pi_{ik} = P(Y_{t+1} = S_k | Y_t = S_i)$. For each row (i.e. skill), we assume multinomial likelihood

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

and $\sum_{k=1}^{n+2} \pi_{ik} = 1$, for each i. We use a conjugate Dirichlet prior distribution of the type

$$f(\pi_{i1}, \dots \pi_{i,n}, \pi_{i,n+1}, \pi_{i,n+2} | a_{i1}, \dots a_{i,n}, a_{i,n+1}, a_{i,n+2}) \propto \prod_{k=1}^{n+2} \pi_{ik}^{a_{ik}-1}$$

where in each row the prior parameters a_{ij} can be elicited by using the coaches as an expert's opinion. Prior estimations were conducted by expert coaches in the three specific ages of men's Volleyball. Due to the difficulty of passing the notions of transition probabilities prior elicitation was a tedious and difficult task. For this reason, we have decided to attach only a low weight to the experts/coaches opinion. Hence, the prior parameters were set equal to the elicited transition probabilities of experts multiplied by $0.1 \times N_i$ for each skill S_i and hence account for additional 10% of data points.

All skill scores were calculated using a simple Monte Carlo scheme of 10,000 iterations.

4 Results

With the use of importance scores we estimate the impact of performance in a specific skill and the uncertainty of it. Uncertainty connects directly to the number of executions of each skill. The skills which are performed more often receive higher importance scores. Comparing importance scores across teams or age categories, the ranking of importance scores should be compared. We summarize results by focusing on importance scores for each level of every skill for the three age categories (see Table 2). An analysis of the importance score for each skill (based on Table 2) follows.

Serve: Aces either from jump or from float serves are skills with high importance score in the three age categories. In Men's category all levels of float serves are more important compared to the similar level of jump serves. Also serves level 2 have higher importance score than serves levels 3 and 4 either if there are float or jump serves. This is due to the nature of importance scores as was explained above. An easy serve level 2 is executed more times than serves levels 3 and 4 during a match but the serving team maintains the right to fight for the point even if circumstances are against it because the ball was in its opponent with all attacking abilities.

Pass: A difference between youth or juniors and mens teams is the importance of pass accuracy. According to our scale with pass level 6 a team has all the possible choices to set the ball in optimum tempo and with pass level 5 has also all the possible choices but not in the fastest way. In men and junior teams passes level 6 (either against jump or float serves) and passes of level 5 have higher importance scores than all the other levels of pass. In youth teams this did not happen. Accurate pass is more important against float serve than against jump serve. Pass levels 5 and 6 against float serve are the most important scores for youth team.

	Men		**U21**		**U19**	
Block kill AS	948.1	(1)	315.5	(1)	165	(1)
SrvJump 6	281.2	(2)	139.1	(3)	57.2	(3)
Block kill	225.2	(3)	168.5	(2)	57.6	(2)
SrvFloat 6	114.4	(4)	57.6	(4)	15.7	(5)
Passin Float 6	27.9	(5)	18.3	(6)	14.3	(6)
Passin Jump 6	27.8	(6)	15.5	(9)	6.1	(22)
Passin Float 5	27.8	(7)	18.3	(5)	13.2	(7)
Passin Jump 4	24.9	(9)	14.1	(10)	8.2	(14)
Passin Float 4	22.8	(10)	14	(11)	11	(8)
Attack 1 MF quick	21.9	(11)	16.5	(8)	9.4	(10)
Passin Jump 5	27	(8)	16.9	(7)	3.3	(38)
SrvFloat 2	17.9	(12)	13.7	(12)	9.3	(11)

Table 2. Importance Scores for volleyball skills (the rankings of the skills are provided in brackets); the table is sorted according to the men's skill scores.
[Table continued on next page.]

	Men		**U21**		**U19**	
Dig 2	16.8	(14)	11.4	(15)	7.4	(18)
Dig	15.8	(15)	10.6	(17)	8.2	(13)
Free ball	15.6	(16)	10.7	(16)	7.5	(17)
Attack 1 LS quick	17.2	(13)	12.7	(13)	16	(4)
Attack 2 out of system	10.9	(24)	9.3	(20)	6.8	(21)
Attack 2 BRS quick	9.9	(27)	3.5	(45)	7.6	(16)
Free ball2	15.3	(17)	10	(18)	7.1	(19)
SrvJump 2	14.3	(18)	6.6	(26)	4.9	(27)
Attack 1BRS quick	11.5	(22)	5.8	(30)	3.3	(37)
Attack 1 FRS quick	14.1	(19)	12.7	(14)	10.7	(9)
Attack 1 MB quick	12.3	(20)	5.6	(34)	2.3	(45)
Attack 2 FRS quick	10.5	(25)	8.3	(23)	8.6	(12)
Attack 2 MB quick	7.7	(35)	6	(28)	3.5	(35)
Continue Block 3	7.7	(34)	4.6	(40)	3.4	(36)
Passin Jump 3	8.5	(30)	5.8	(31)	2.3	(44)
Passin Float 3	8.5	(29)	5.6	(33)	4.5	(30)
Attack 1 out of system	6.7	(37)	4.9	(39)	4.9	(28)
SrvJump 3	9.3	(28)	7.1	(25)	4.4	(32)
SrvJump 4	7.5	(36)	4.9	(38)	4.8	(29)
SrvFloat 3	12	(21)	8.7	(21)	7.9	(15)
Passin Float 2	5.7	(40)	2.5	(48)	4	(33)
Attack 1 LS high	8.4	(31)	5.8	(32)	5	(26)
SrvFloat 4	8.3	(32)	6.4	(27)	6	(23)
Attack 2 LS quick	11	(23)	9.4	(19)	5.9	(24)
Attack 2 MF quick	10.1	(26)	8.6	(22)	6.9	(20)
Attack 2 STR TIP	2.9	(49)	4.4	(41)	2.9	(40)
SrvJump 5	3.7	(47)	3.5	(44)	2.8	(41)
Attack 1 MB high	1.9	(51)	0.3	(53)	0.3	(51)
Attack 1 STR TIP	4.5	(44)	3.7	(43)	3.1	(39)
Attack 2 LS high	6.3	(38)	5.8	(29)	2.5	(43)
Attack 2 FRS high	5.4	(41)	1.4	(51)	2.7	(42)
Continue Block AS3	6.2	(39)	5.1	(37)	4.5	(31)
Attack 2 BRS high	3.5	(48)	3.3	(47)	1.3	(50)
Direct attack	8	(33)	8.2	(24)	5.9	(25)
Attack 1 FRS high	4.3	(45)	3.3	(46)	1.5	(48)
SrvFloat/	5.1	(42)	5.5	(35)	3.8	(34)
Continue Block AS2	0.6	(53)	0.3	(52)	NaN	(53)
Passin Jump/	5	(43)	5.2	(36)	2.2	(46)
Attack 1BRS high	4.1	(46)	3.8	(42)	2	(47)
Attack 2 MB high	2.1	(50)	1.6	(49)	1.4	(49)
Continue Block 2	1.2	(52)	1.5	(50)	0.3	(52)

Table 2 (continued). Importance Scores for volleyball skills (the rankings of the skills are provided in brackets); the table is sorted according to the men's skill scores.

Attack 1: For all ages quick tempo attack after serves pass is more important than attack with high tempo. A differentiation is that for youth team, attacks from back row are not as important as front row attacks. The importance of back row attack (middle and right side) is getting higher as the age category increased. As attack out of system we have described the

attack independently of location and tempo when setting comes from any member of the team except the setter. In this type of attack the youth team has higher position in the importance scores ranking than in Men and Juniors teams respectively. According to these and in connection to the limited importance of passing accuracy against jump serve we can think that a difference between men or juniors teams is the rhythm of the offensive game. Youth teams must prepare better for a slower offensive tempo than older and more experienced players. Middle front side attack and left side's attack both in quick tempo are the most efficient attack types for men and juniors. All levels of well organized attack 1 have higher importance scores than attacks 2. This is a clear message about the importance of complex 1 in male volleyball.

Attack 2: Attack out of system is for all ages in the top two positions for attack 2. It is one of the most important attack skills during complex 2. Volleyball coaches have to work more on details concerning unpredictable situations such as when the setter is not able to set the ball according to the teams offensive plan and another player is getting to set it in a safe manner. For all ages setting quick tempo is more efficient than high tempo attacks.

Direct attack: Importance of direct attack when the ball comes directly from the other court is getting lower as age category increases. Direct attack is more important for youth and junior teams than for men.

Block: From all blocking situations the most important is Kill Block after serve. Kill block and serve ace are the two direct methods to break the point of the opponent. In Male Volleyball contrary to Male tennis the serving team has a disadvantage and if they win the point we call it a break point. Thus kill block when serve is the previous action is the most important option of defending abilities.

As for the defending skills Free ball and Dig, it is important to define that dig or free ball can take place in two occasions: either when responding to opponents action or from a technical spike to opponents block when circumstances are not ideal for a powerful attack especially because attack is out of system or because previous actions (pass and set) are not so successful. Free ball and dig in both complexes are in the second ten in ranking of importance scores for male teams across ages. Either after serve or after attack defending abilities gives the teams the opportunity to claim a point.

5 Conclusions

It is important to recognize that this analysis applies only for these specific teams. But these teams being the winners of the respective age category in world championships reflect the highest level of volleyball game for each category. Because of the class of the teams it is reasonable to extend our conclusions to teams in world level for these ages. According to our results importance of Volleyball skills across ages did not vary too much. High class teams as world champions use almost same ways to win points during a game. It would be a great challenge to use data of performance analysis as prior information for a specific team. The more detailed the data, the more standardized a team's profile is, and the more weighty the prior information.

Our target was to propose a performance analysis system more extended and more applicable to men's teams, with the ability to rank the importance not only for every volleyball skill but for every level of the scale we use to describe skills. This method could help volleyball coaches to highlight important skills per age category and to allocate training time more efficiently.

References

[1] Rocha. C. and Barbanti. V. An analysis of the confrontations in the first sequence of game action in brazilian volleyball. *Journal of Human and movement studies*, 50(4):259–272, 2006.

[2] S. Drikos and G. Vagenas. Multivariate assessment of selected performance indicators in relation to the type and result of a typical set in mens elite volleyball. *International Journal of Performance Analysis of Sports*, 11:85–95, 2011.

[3] G. Fellingham and C. Reese. Rating Skills in International Men's Volleyball. Unpublished report to the USA National Men's Volleyball Team, 2004.

[4] M. Miskin, G.W. Fellingham, and L.W. Florence. Skill importance in womens volleyball. *Journal of Quantitative Analysis in Sports*, 6(2), 2010.

[5] J.M. Palao, J.A. Santos, and A. Urena. Effect of team level on skill performance in volleyball. *International Journal of Performance Analysis of Sports*, 4(2):50–60, 2004.

[6] B. Reed and P. Edwards. *Teaching Children to Play Games*. White Line Publishing, Leeds, 1992.

Predicting the Results of Tennis and Volleyball Matches Using Regression Models, and Applications to Gambling Strategies

Edyta Dziedzic and Gordon Hunter

School of Mathematics, Kingston University, London, KT1 2EE, U.K.
edyta.dziedzic3@gmail.com , G.Hunter@kingston.ac.uk

Abstract

Predicting the outcomes of sports matches, based on information available in advance of the match, is of interest to a variety of parties – players, coaches, fans, bookmakers and gamblers. Some previous approaches have used computer simulations and/or stochastic models based on player statistics to predict such outcomes. In this paper, we instead investigate the use of regression models, using player or team rankings, previous performance in the same tournament, continent of origin and of the tournament, in predicting such outcomes of matches in the major "Grand Slam" tennis tournaments and international volleyball tournaments. We compare the use of "official" rankings of teams and players with ranking produced by a variant of a "page ranking" algorithm for internet web pages in these models, and study the success rate of each in predicting the results of matches not used for training the parameters of the models. We compare our results with those based on predicting that the higher ranked player/team, or the bookmaker's favourite, would always win. Our models show a modest advantage over both simpler schemes. We then apply our models to both simple (fixed stake) and Kelly "odds overlay" betting strategies, placing virtual bets on each match. Our models were trained on "Grand Slam" results up to and including Wimbledon 2014 and the Volleyball 2014, and tested on the 2014 US Tennis Open and the 2014 Volleyball World Championships, neither of which was used in training the models. We find that in most cases, a Kelly betting strategy based on our model performs better than simpler approaches, and on average gives a positive return (unlike some of the other strategies).

1 Introduction

Predicting the outcomes of sporting encounters is or interest to many parties – not only players or teams, their coaches and managers, but also gamblers and bookmakers, who want to obtain or maintain a competitive advantage over their rivals.

Over recent years, there has been a rise in quantitative predictive modelling of results for many sports, but over the same period there have been increasing allegations and evidence of corrupt activities, particularly in relation to gambling – for example in cricket [10, 11, 16] and Association Football (Soccer) [2, 14]. Such scandals have tarnished the reputations of several sports, in which traditionally "fair play" has been considered paramount, and could lead to great reductions in revenue due to fans becoming disillusioned and/or loss of income from sponsorships and advertising. Thus, both the sporting and legal authorities would welcome any tools becoming available which would facilitate the detection, quantification and prevention of such irregular and illegal activities. Reade & Akie [13] have attempted to do this for Association Football (Soccer) by using econometric forecasting methods to predict the probability of a given game ending in a draw and comparing this to what would be expected on the basis of the odds offered by bookmakers. In cases of corruption – either in terms of match-fixing by players or

of illegal gambling – there would be expected to be a significant discrepancy between the odds offered (once an allowance has been made for bookmakers' profit margins) and the probability of the match result. The well-known Duckworth-Lewis method [4], trained on the statistics from many real matches, has become an established method for determining fair targets in situations where a limited overs cricket match is interrupted by bad weather or other disturbances. In principle, this method could also be used to detect possible cases of corruption. However, as noted by Reade & Akie [13], such models need careful development before they can hope to provide evidence of sufficient reliability to be admissible in court.

In this paper, we describe the development of logistic regression models, using player or team rankings, previous performance in the same tournament, continent of origin and of the tournament, in predicting such outcomes of matches in the major "Grand Slam" tennis tournaments and international volleyball tournaments. We compare the use of "official" rankings of teams and players with ranking produced by a variant of a "page ranking" algorithm for internet web pages in these models, and study the success rate of each in predicting the results of matches not used for training the parameters of the models. We also compare a variety of simple and "odds weighted" (e.g. Kelly) betting strategies used in conjunction with these models to investigate which (if any) could yield a positive net return.

The remainder of this paper is organized as follows. In section 2, we give an overview of some other work relating to the modelling of tennis and volleyball matches. In section 3, we give details of our sources of data. In section 4, we briefly describe the two methods used for the ranking of teams and players, then we give details the development of our models. The results of applying our models to predict the results of subsequent tournaments – i.e. applying them to previously "unseen" data are given in section 5, and we state our conclusions and propose some further word in section 6.

2 Related Work

There has been considerable previous work using Markov-like models and simulations to model tennis, both at the level of the individual point or rally [6] and at the levels above the individual point : points to games, games to sets and sets to matches [1, 8, 9, 15]. Rather less work has been carried out on modelling volleyball, although [5] have also applied Markov-like models to this.

Regression modelling provides an alternative approach to predicting match outcomes – instead of using players' or teams' probabilities of winning points whilst serving or receiving, the logistic regression modelling approach takes a more "zoomed out" view, modelling players' overall performance at the level of winning or losing matches (in both tennis and volleyball, each player or team either "wins" or "loses" each match there are no draws or ties), based on evidence which would have been known prior to the actual match – the players' or teams' "rankings", the venue of the match, how the players or teams have done previously in the same tournament, etc. This approach creates a model for the relative probabilities of winning or losing, based on the outcomes of real matches for which the results and player/team "features" (or independent variables) were known. Such a model can then be used in reverse – given the "features" as "predictors" for given players or teams, the probability of a specified one of them winning a match between them can be estimated. Although applied to Association Football (Soccer), for which there is the additional result of a draw possible, this is essentially the approach taken by Read & Akie [13], but they apply an "ordered probit" regression model to allow for the three possible outcomes.

3 Datasets

The data used was set to be the best representation of the top tennis players and volleyball teams. We have only included the most prestigious events in the calendar for which most points towards the official rankings can be gained.

In the tennis events calendar there are four "Grand Slam" Tournaments organized each year. Results for all such tournaments organized between 2009 and 2013 were collected from the Tennis-Data website www.tennis-data.co.uk and formatted using MS Excel and SAS version 5.1 using a recorded macro technique. Data for the volleyball national teams was much harder to acquire, as no single source of data was available. The major three tournaments for the volleyball national teams are the Olympic Games, World Championships and World Cup, each organized every 4 years. In addition, the World League (for Men) and Grand Prix (for Women) are organized every June/July after the regular international teams' season is closed. Results for all national teams tournaments held between 2008-2013 were copied into MS Excel from over a 100 webpages on the FIVB (International Volleyball Federation) website www.fivb.org and formatted using templates and macros written in SAS version 5.1. Analysis of the raw data obtained showed that the official ranking methods give good estimates for the current teams' or players' strengths, in terms of predicting match outcomes. In the 20 "Grand Slam" tennis tournaments organised between 2009-2013, 71.5% of men's and 68.6% of womens games were won by the higher ranked player. Further analysis also showed that almost half of the matches lost by a higher ranked player occurred between players with less than 30 places difference in their rankings. For volleyball, 71% and 74% of all men's and women's matches respectively were won by the higher ranked team on tournaments organized between 2008-2013. We also observed that half of the matches where a higher ranked team suffered a loss was against a team within five places of them in the official ranking. These values provide a "baseline benchmark" for comparison with the predictive performance of our models.

3.1 Ranking Schemes and Regression Models

For many supporters, the official rankings are a source of identifying the current top performers. The prestige that comes with being listed as one of the top players or teams is often associated with extensive media coverage of every game, playing at the biggest arenas, and can lead to lucrative sponsorship deals. However, rankings can also sometimes be misleading. For example, in 2013, Caroline Wozniacki was ranked as the top tennis player for 23 weeks despite failing to win a single "Grand Slam" Tournament in her career to date. The Official Ranking only takes the most recent results into account; therefore, if a player or a team is weakened by injuries it can often underestimate their performance after they come back from the recovery period. Additionally, official rank calculations are used for seeding the top 32 tennis players in most of the major tournaments. This makes it very hard for lower ranked players to climb up the rankings, as they often play their first matches in each tournament against a much stronger player. Therefore, the seeding systems often result in lower ranked players not finishing at positions highly rewarded within the official ranking schemes. The official rankings are also used as a qualification requirement for some tournaments, preventing some of the lower ranked players gaining valuable points and experience. In Volleyball, the top 12 teams, according to the current World Rankings, are seeded for the first two positions in a set number of "pools" at tournaments. This seeding system guarantees that no two teams from the top 4 will be in the same pool, as teams are divided into at least four pools at each of the major tournaments. Volleyball tournaments are usually composed of Preliminaries ("Pool" Stage) and "Finals" (Single Elimination Stage). This means that, at the preliminary stage of the competition,

losing a match does not mean immediate elimination from the tournament, but only the teams who won the most "pool" matches qualify to the "Finals" stage, where the loser of each match is then eliminated.

In [3] an alternative method of ranking is considered. The "Page Ranking" method described by them is derived from an algorithm used by internet search engines such as Google to rank web pages by relevance, and was first applied to tennis by Radicchi [12]. The Page rank calculates players' strength only using players' performance against other players which means that, contrary to the official ranking method, all matches played thorough the duration of each tournament are considered, whereas the official ranking only takes the final result for each player. However the page rank algorithm averages players' performances and, as a result, underestimates the current strength for players and teams who have substantially improved their skills recently, or overestimates their current strength if they have not performed at their best recently or have suffered an injury. The page rank could also give misleading results if a player only recently started competing, as not many past results would then be available.

Official rankings of players or teams are computed by the appropriate authorities – the ATP (Association of Tennis Professionals) for men's tennis, the WTA (Women's Tennis Association) for womens tennis, and FIVB (Federation International de Volley-Ball) for both men's and women's volleyball. In the case of tennis, the World Rankings are based on each player's results over a period of the previous 52 weeks. Points are awarded to a winner or player knocked-out in a particular round at different tournaments. The most are awarded according to how far the player progressed in each tournament, with more points being available for the more prestigious tournaments, which tend to attract stronger players, than for minor ones. For example, the winner of a "Grand Slam" tournament – Wimbledon and the Australian, French and U.S. Opens – would gain 2000 points, whereas the winner of an ATP 500 tournament would be awarded just 500 points. Each player must compete in a certain number of tournaments of specified caliber in order to gain a ranking, and a player cannot climb up the ranking ladder quickly by just playing in a large number of the less prestigious tournaments.

The situation for Volleyball is somewhat more complicated, since the official FIVB rankings are based on the results achieved at the following major tournaments: Olympic Games and their qualification tournaments, World Cup, World Championships and their qualification tournaments, Continental Championships, Continental and World qualification tournaments, World League (Men), and World Grand Prix (Women). The first four of these are organised every four years, whereas the World League and Grand Prix take place every year. For the first three, and all their qualification tournaments, points awarded are retained in the rankings for 4 years. For Continental Championships, points are retained for 2 years, and some points are also awarded to the best non-qualifying teams. The World League (Men) and World Grand Prix (Women) are organised every year, so points from those are awarded after each tournament's completion but are retained for just one year. More ranking points are available to teams which do well in the most prestigious tournaments, namely the Olympic Games, World Cup and World Championships, than for the others. As these top competitions are organised only every four years, sometimes a team's current strength does not reflect the position obtained the last time those tournaments were held, which is one of the weaknesses of the current FIVB ranking system, retaining points from those results for a period of 4 years.

In our approach, we obtain probabilities of match outcomes using models based on the page ranking method [3, 12] as well the current official rankings, and compare the predictive power of both approaches. The probability modelled was for a given player or team to lose a match. To allow true comparison in terms of improving the goodness of fit, only those predictor variables for which direct equivalents for both methods could be found were considered. It is

well known that sports vary in popularity and level of funding and sponsorship between different countries. For example, in tennis, 68% of all male and 72% of female active professional tennis players respectively considered in this project are European. Furthermore, 81% of the top 32 male players are also European. Although the players' continent of origin was considered as a possible variable for inclusion into the logistic regression models, it was found to be highly correlated to the players' ranking, which was a more powerful predictor, so the continent of origin was excluded from all the models.

As outlined in the previous sections, just using the players' or teams' rankings to predict the outcome of each match is quite successful. However, we noted that in our approach including the previous results in the most recent equivalent tournament provides a slight uplift on the prediction rate. This is especially true for tennis where all of the Grand Slam tournaments are organized on a different surface which affects performance of some of the players. All tennis and volleyball models were built using different combinations of variables based on their official/page rank and previous performance in a tournament of the same type, and the optimal coefficients for the logistic regression models computed using maximum likelihood estimation in SAS, with respect to the data for 2009-13 for tennis, and 2008-13 for volleyball.

Each model developed to predict the tennis results in a combination of the following variables: page or official ranking difference between a higher and a lower ranked player, best round for the player of primary interest the previous instance of that tournament, ranking of the higher ranked player, ranking of the player for who the probability is measured. Only statistically significant parameters were retained in the final models.

The final two logistic regression models for men's tennis, using page rankings and official rankings respectively, are as follows:
Model MT1 :

$$\ln[p_{ij}/(1-p_{ij})] = 5.34+0.04*\text{pagerankdiff}-0.07*\text{Hpagerank}-1.202*\text{bestround}+0.04*\text{playerpagerank}$$

Model MT2 :

$$\ln[p_{ij}/(1-p_{ij})] = 4.99+0.03*\text{offrankdiff}-0.05*\text{Hoffrank}-1.17*\text{bestround}+0.02*\text{playeroffrank}$$

where p_{ij} represents the probability that player i loses to player j, pagerankdiff and offrankdiff represent the difference in the rankings ("official" or "page", as appropriate), Hpagerank or Hoffrank the higher ranking of the players, playerpagerank or playeroffrank the rankings of player i, and bestround represents the latest round in the same tournament reached by player i the last time he/she played in it, set to zero if no such instance was recorded. This dependence on the previous version of the same tournament can be regarded as a "surrogate" variable for the surface (grass, clay, hard court) on which that tournament is played.

The optimal models for women's tennis are :
Model WT1 :

$$\ln[p_{ij}/(1-p_{ij})] = 0.138+0.0029*\text{pagerankdiff}+0.604*\text{bestrounddiff}+0.0116*\text{playerpagerank}$$

Model WT2 :

$$\ln[p_{ij}/(1-p_{ij})] = 0.042+0.0076*\text{offrankdiff}+0.566*\text{bestrounddiff}+0.016*\text{playeroffrank}$$

where bestrounddiff is the difference between the rounds attained by the higher and the lower ranked players respectively the last time that tournament was played.

For mens volleyball, the best-fitting models are :

Model MV1 :

$\ln[p_{ij}/(1-p_{ij})] = 0.0134 - 0.176 * \text{pagerankdiff} - 0.359 * \text{prevLast4} + 0.312 * \text{teampagerank} 0.315 * \text{Lpagerank}$

Model MV2 :

$\ln[p_{ij}/(1-p_{ij})] = 0.383 - 0.136 * \text{offrankdiff} - 0.805 * \text{prevLast4} + 0.201 * \text{teamoffrank} 0.231 * \text{Loffrank}$

and for womens volleyball : Model WV1 :

$\ln[p_{ij}/(1-p_{ij})] = 0.158 - 0.237 * \text{pagerankdiff} - 0.264 * \text{bestround} + 0.379 * \text{teampagerank} 0.399 * \text{Lpagerank}$

Model MV2 :

$\ln[p_{ij}/(1-p_{ij})] = 0.226 - 0.153 * \text{offrankdiff} - 0.581 * \text{bestround} + 0.256 * \text{teamoffrank} 0.271 * \text{Loffrank}$

where prevLast4 takes thevalue of 1 if the team reached semi-finals or finals stage the last time that tournament was played, or 0 otherwise.

The validation and testing of these models in application is described in section 5 below.

4 Results and Applications to Betting

4.1 Predictive power of the logistic models and validation

We now investigate how successful our models are at predicting the results of matches. In Table 1 below, we show the fractions of matches of each type which our models correctly predict the winner, including the "baseline" model which always just chooses the higher ranked player. In some cases, the tournaments considered were in the dataset used to train the models, but others (indicated with an asterisk) were not, so these cases are genuine tests of the models' ability to predict results in a general context. It can be seen that, in most cases, our regression models outperform the baseline model, and the better model usually does so by a considerable amount.

4.2 Applications to Gambling

Five betting strategies were employed using the probabilities obtained from the logistic regression models: (i) A simple fixed stake bet on the bookmakers favorite – a virtual bet was placed on the favorite indicated by the bookmakers – this was done both using the bookmakers' quoted odds, and with the odds scaled in such a way that the corresponding win or lose probabilities would sum to one; (ii) A simple fixed stake bet was placed on the favorite as indicated by the probabilities calculated from the better of the logistic regression models; (iii) A bet of constant stake was placed on the player or team for which the more "generous" odds were offered. This was decided by subtracting the odds predicted by our model from the bookmakers' odds, and placing a bet on the player or team for which that difference was greatest; (iv) A simple fixed stake bet was placed on the bookmakers' favorite, but only if the odds were more generous than for the opponent; (v) Kelly "Odds Overlay" betting [7], which is a strategy that adjusts the stake of the bet according to the probabilities computed from our model and the bookmakers' odds offered. A fraction f of the gambler's current capital is placed on a bet, calculated as:

$$f = (p.d - 1)/(d1)$$

where p is the gambler's estimate of his probability of winning, and d is the 'decimal odds' offered by the bookmaker for a particular player. A "Kelly" gambler will place a bet based on

Gender	**Sport**	**Tournament(s)**	**Matches**	**Won by Higher Ranked Player/Team**	**Page Rank Model**	**Official Rank Model**
Men	Tennis	"Grand Slams" 2009-2013	1830	71.53%	83.28%	81.15%
Men	Tennis	US Open 2014*	90	64.44%	64.44%	76.67%
Women	Tennis	"Grand Slams" 2009-2013	876	67.92%	73.76%	70.50%
Women	Tennis	US Open 2014*	74	77.00%	89.20%	90.01%
Men	Volleyball	World Cup, World Championships, Olympic Games, 2008-2013	750	71.10%	70.40%	71.60%
Men	Volleyball	World Championships 2014*	89	68.50%	68.50%	71.90%
Women	Volleyball	World Cup, World Championships, Olympic Games, 2008-2013	643	74.50%	77.90%	72.60%
Women	Volleyball	World Championships 2014*	70	78.50%	80.00%	75.70%

* No data from these tournaments were used in training the models, but data from all the other tournaments (of the appropriate sport) listed were included when computing the optimal parameters in each model.

Table 1: Correct match prediction proportions for our models, compared with the "baseline" of always selecting the higher ranked player or team.

the 'edge' value, defined as: $(d1) - (1 - p)/p$, such that the gambler: (a) places a bet if the 'edge' is positive, or (b) does not place a bet if the 'edge' is zero, or (c) places a bet on the opponent if the 'edge' is negative (and the corresponding 'edge' for the opponent is positive). All bookmakers' odds used were found on the betexplorer.com and tennis-data.co.uk websites.

Player A decimal odds (our model)	Player B decimal odds (our model)	Player A bookmakers' decimal odds	Player B bookmakers' decimal odds
1.37	3.72	1.66	2.10
1.46	3.15	1.66	2.10
1.57	2.76	1.90	1.80
1.24	5.22	3.25	1.33

Table 2: Sample of decimal odds obtained from our model compared with the bookmakers' odds for the same men's tennis matches. Note that, due to the bookmakers incorporating a profit margin into their odds, we do not expect the bookmakers' odds to correspond to probabilities which sum to one.

The betting strategies were applied to the 2009-2013 tennis "Grand Slam" tournaments, and the World Championships, World Cups, World Leagues and the Olympic Games organised between 2008 and 2014 for volleyball, and virtual bets placed on the outcome of randomly selected samples of 25 matches (using the rand() function in Excel), employing 10 sampling attempts per model (men or women, tennis or volleyball. The bets were either a fixed stake of £2, or a fraction of the current "pot", which was initially set up containing £50. The average returns for each strategy are shown in Table 3 below. In each case, the "better" of our two logistic regression models for that particular sport and gender pair was used.

5 Discussion and Conclusions

It can be observed from Table 1 that, although the success rates of our models vary from tournament to tournament, in most cases both our models perform better than the baseline assumption, namely that the higher ranked player or team will win. In all cases the "better" of our two models outperforms the baseline, but sometimes the model based on official ranking is the better, and other times that using the page ranking is superior. It is not clear that either ranking scheme has an advantage over the other further investigation relating to this would be required.

Strategy:	(ia): back bookie's favorite, scaled odds	(ib): back bookie's favorite, bookie's odds	(ii): back our model's favorite, fixed stake	(iii) fixed bet on more generous odds	(iv) fixed bet on bookie's favorite if odds generous	(v) Kelly bet, variable stake
Men's tennis	+1.56 (5.85)	−0.66 (5.50)	+1.43 (6.13)	−18.53 (30.15)	+8.45 (24.72)	+42.88 (147.98)
Women's tennis	−7.65 (11.05)	−9.45 (10.68)	−7.17 (10.40)	−4.57 (19.12)	+21.34 (23.85)	+13.18 (91.05)
Men's volleyball	−2.67 (6.64)	−5.14 (6.20)	−2.74 (7.56)	−1.62 (17.09)	+17.67 (20.71)	+1.37 (61.66)
Women's volleyball	−0.49 (5.50)	−2.29 (5.35)	+13.38 (16.54)	−27.94 (17.34)	+10.95 (22.77)	+3.67 (55.19)

Table 3: Average profits or losses (with standard deviations in brackets) for the different gambling strategies described above. The tennis results are for random samples of matches in the "Grand Slam" tournaments held between 2009 and 2013, whereas the volleyball results are for random samples of matches from the major international tournaments held between 2008 and 2014.

Regarding the different betting strategies, a simple bet on the bookie's favorite is clearly not a good approach. In most cases, either the Kelly methodology or the "Back the bookie's favorite, but only if the odds are generous" (i.e. predicted as "generous" according to our models) strategy give the best average returns, if with rather high variances (especially for the Kelly method). However, it should be noted that the latter is a conservative gambling scheme – you only bet on the bookie's favorite, and only if our models deem that the odds offered by the bookmakers are generous. It appears that finding the strategy which guarantees a positive return on average is still elusive!

References

[1] T.J. Barnett and S.R. Clarke. Combining player statistics to predict outcomes of tennis matches. *IMA Journal of Management Mathematics*, 16(2):113–120, 2005.

[2] J. Burt. Football Association urged to act swiftly in order to curb corrupt betting patterns on lower league football, 2013. The Daily Telegraph (U.K.), 30th November 2013, `http://www.telegraph.co.uk/sport/football/10486524/Football-Association-urged-to-act-\swiftly-in-order-to-curb-corrupt-betting-patterns-on-lower-league-football.html`.

[3] N.J. Dingle, W.J. Knottenbelt, and D. Spanias. On the (Page) ranking of professional tennis players. In *Proceedings of European Performance Evaluation Workshop/UK Performance Engineering Workshop (EPEW/UKPEW 2012)*, July 2013.

[4] F.C. Duckworth and A.J. Lewis. A fair method of resetting the target in interrupted one-day cricket matches. *Journal of the Operational Research Society*, 49:220–227, 1998.

[5] M. Ferrante and G. Fonseca. Markov chain volleyball. In *Proceedings of 4th International Conference on Mathematics in Sport*, pages 96–107, 2013.

[6] G. Hunter and K. Zienowicz. Can Markov models accurately simulate lawn tennis rallies ? In *Proceedings of 2nd IMA International Mathematics in Sport Conference*, 2009.

[7] J.L. Kelly. A new interpretation of information rate. *Bell System Technical Journal*, 35:917–926, 1956.

[8] V. Mogilenko and G. Hunter. Using probabilistic models to simulate tennis matches, with applications to betting strategies. In *Proceedings of 4th International Conference on Mathematics in Sport*, pages 216–221, 2013.

[9] P. Newton and K. Aslam. Monte Carlo Tennis : A Stochastic Markov Chain Model. *Journal of Quantitative Analysis in Sport*, 5(3), 2009.

[10] J. Polack. A definitive overview of cricket match-fixing, betting and corruption allegations. [online], 2000. `www.espncricinfo.com/ci/content/story/90877.html`.

[11] B. Radford. *Caught Out : Shocking Revelations of Corruption in International Cricket.* John Blake Press, London, 2011.

[12] F. Radicchi. Who is the best player ever ? A complex network analysis of the history of professional tennis. *PLoS ONE*, 6(2):e17249, February 2011.

[13] J.J. Reade and S. Akie. Using forecasting to detect corruption in international football. In *Proceedings of 4th International Conference on Mathematics in Sport*, pages 96–107, 2013.

[14] B. Rumsby. Premier League game 'almost certainly rigged', claims betting expert, 2013. The Daily Telegraph (U.K.), 30th November 2013, `http://www.telegraph.co.uk/sport/football/news/10485329/Premier-League-game-almost-certainly-rigged-claims-betting-expert.html`.

[15] D. Spanias and W. J. Knottenbelt. Predicting the outcomes of tennis matches using a low-level point model. *IMA Journal of Management Mathematics*, 24:311–320, 2013.

[16] R. Sydenham. Prosecution opens with details of illegal betting. [online], 2011. `www.espncricinfo.com/pakistan/content/story/535250.html`.

The World's Biggest Sports Database

Simon Gleave

Infostrada Sports, Netherlands
simon.gleave@infostradasports.com

Abstract

Infostrada Sports has one of the biggest sports databases in the world, covering more than 250 sports. I propose to introduce the content of our database via a presentation outlining the data that we collect, which competitions, the depth of that data, the historic completeness of that data and also the real-time nature of the data collection.

1 Introduction

Infostrada Sports is a Dutch company which was founded in 1995 to collect sports data, both current and historic with the original aim of becoming the world's most comprehensive archive of the history of sports results. The database began with cycling and has since grown into a unique record of sports information which is unmatched in its global and historical coverage. The database now contains more than 250 sports with the earliest result dating back to 1860.

2 Database content

2.1 Results

There are now nearly 24 million sports results from 251 sports currently held in the Infostrada Sports' database with more being added every day as competitions continue to take place all over the world. The majority of the data input is done live, as events happen, in a 24-7 operation coordinated from the head office in the Netherlands and supported by other offices around the world, notably in Harbin, China.

In order to maintain this vast operation, sports have been divided into three types:

Team sports – for example rugby, in which the minimum level of coverage would be the two competing teams and the scoreline. However, the detail can go deeper than this to include the line-ups, scorers, scorer times, disciplinary cards, card times, substitutions and their related times, attendances and even, for some selected leagues, basic performance data.

Head to head sports – for example tennis, in which the minimum level of coverage would be the two participants and the scoreline. However, the detail can go deeper here too in order to include further breakdowns of that performance.

Time Judge sports – for example athletics, in which the minimum level of coverage would be the rank order and performance of each of the participants. Again, the detail can go deeper to include further breakdowns of that performance.

2.2 Actions

Actions are specific things which happen during a team sport, like a goal or a disciplinary card. Those actions will usually have a time associated with them as well as the person or people who were involved. There will often be further supplementary information, such as whether a goal was scored with the left foot, right foot or head for example. There are currently more than 21 million actions present in the database.

2.3 Matches

A match entry in the database contains a date and a scoreline attached to it as well as those participating in it and the competition of which the match was a part. If that competition was split into phases, then that is also recorded. The number of matches in the database is now approaching three million in total.

2.4 People

In 2014, the number of people in the Infostrada Sports database surpassed one million for the first time, and has grown around a further 10% since that moment. People are entered into the database with a date of birth, place of birth, height, weight and various name aliases. In addition, their nationality, second nationality, club and national teams are recorded and linked to date fields in order for changes to these to be accurately recorded. Name changes are also recorded and linked to the dates on which specific names are relevant.

2.5 Competitions

More than 15,000 competitions from 251 sports have been entered into the Infostrada Sports database with many of these containing a complete history of results. For more than a decade now, complete results have been entered for almost all of the competitions so whereas other records of results may only display the medal winners or the finalists, the Infostrada Sports database contains a record for everyone that participated in that event.

		All Sports		
		Team Sports	Head to Head	Time Judge Sports
Sports	251	Football	Tennis	Alpine Skiing
Competitions	15,033	European Championship	Wimbledon	Olympic Games
Editions	115,938	2012	2013	2014
Teams	136,001	Spain	France	Austria
Events	279,100	Men's Final	Women's Singles	Men's Downhill
Persons	1,089,146	Iker Casillas	Marion Bartoli	Matthias Mayer
Matches	2,773,317	Spain - Italy (01 jul 2012) 4-0	Marion Bartoli - Sabine Lisicki 2:0 (6-1, 6-4)	Final
Actions	21,217,755	Goal 1-0: David Silva 14' (header, assist Fábregas)	N.A.	N.A.
Results	23,685,933	1. Spain 2. Italy	1. Bartoli 2. Lisicki	1. Matthias Mayer (2:06.23), 2. Christof Innerhofer (2:06.29), 3. Kjetil Jansrud (2:06.33).

3 Football

Football results are currently collected for all top level domestic leagues in the world that are played in FIFA member countries, which amounts to well in excess of 200 leagues. In addition, goalscorers are entered into the database for many of these competitions. Lower level domestic leagues are also entered into the database at at least result level for a number of selected bigger countries such as England. This part of the data collection can go as low as the fifth level of domestic competition.

Thirty-four of these domestic leagues have full line-ups, scorers, cards, substitutions and attendances as well as the timings of relevant events entered into the database with 11 of these leagues also having goal assists recorded. Eight of those 11 also have basic performance data like shots, saves, corners, fouls, free kicks etc recorded.

Domestic Cups, Super Cups, League Cups and European competition are also entered into the database at varying levels from just the final result in some smaller European nations to basic performance level data for the Champions League.

All national team football is also present in the database, from World Cup and European Championship final tournaments at full performance data level (all ball contacts), to a simple result level for smaller international competitions and some friendlies.

Historic coverage varies but many major competitions are complete at at least the result level. The major international competitions like World Cups, European Championships and Champions League go even deeper than this and are also historically complete at the level of line-ups, cards and goals (with timings). This is also the case with the 23 year history of the English Premier League.

4 All other sports

4.1 Olympic sports

All Olympic sports are administered in the same way as one another with complete rankings being entered for a number of specific groups of competitions. Most of these competitions are complete historically at at least top-3 level. Table 1 shows the current level of data coverage:

Competition	**Level of collection**
Olympic Games	Complete results of all phases and matches including subresults
World Championships	Complete results of all phases and matches including subresults of specific matches
World Cups (or equivalents)	Complete event rankings. Phases, matches and subresults for specific sports
Continental Championships	Complete event rankings. Phases and matches for specific sports
Other relevant competitions	Complete event rankings
Junior/Youth World Championships	Complete event rankings
World Rankings	Monthly update of top-300 (some sports weekly)

Table 1: Data coverage of Olympic Sports at Infostrada Sports

4.2 Paralympic sports

All sports currently included as part of the Paralympic Summer or Winter Games have their complete event rankings entered into the database for both Paralympic Games and World Championships.

4.3 Motor sports

The motor sports in the database are from the sports of auto racing, motorcycle racing and Motocross. Formula 1 and MotoGP include lap times, as well as the complete final results. WRC, DTM, WEC, WTCC, World Championships Superbike/Supercross, and World Championships MXGP/MX2 also have full results recorded.

4.4 Other sports

Two other major sports are collected in reasonable detail. For cricket, scoreboards of top level international Test matches and One Day Internationals are entered in the database. Line-ups, substitutions and key match actions for major rugby union internationals and several domestic leagues in the sport are also recorded. The Rugby World Cup is historically complete at this level of coverage.

Data at World Championship and Continental Championship level for a number of other sports is also entered into the database. These sports are mainly those held at the Asian-, Commonwealth- or Pan American Games.

As well as rugby union, domestic league competitions are also collected for sports like ice hockey, hockey, handball, basketball, volleyball, water polo, bandy and floorball.

4.5 Biographies

In addition to all of the quantitative information, there is also qualitative information collected for athlete biographies in Olympic, Paralympic and non-Olympic sports. There are now in excess of 120,000 of these in the database.

5 Podium, the performance manager

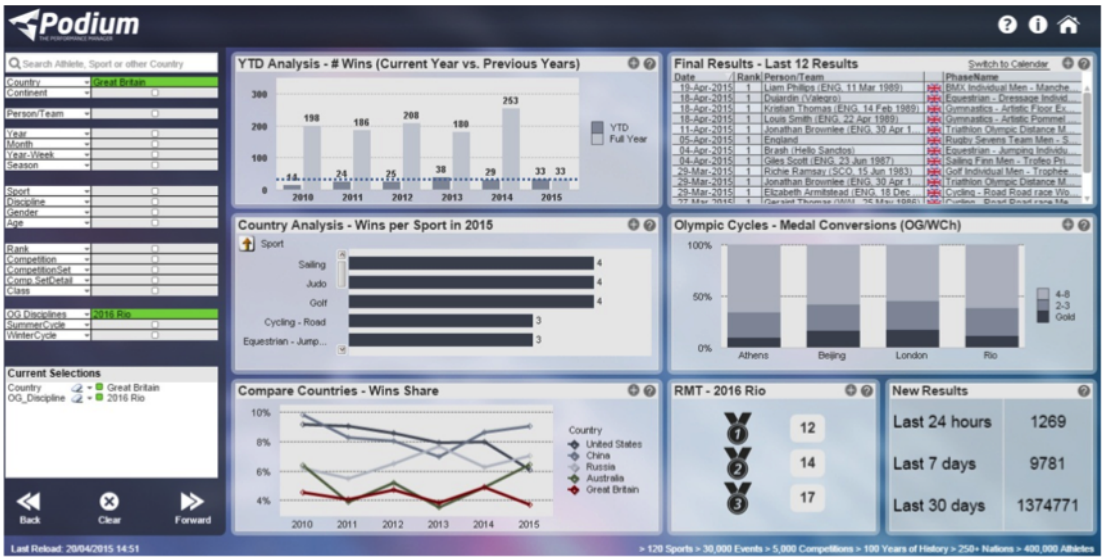
One of the many services provided by Infostrada Sports is an online solution for the presentation and analysis of results and rankings which is made available to National Olympic Committees around the world.

Podium contains a subset of the data held in the Infostrada Sports database and a variety of descriptive analyses which have been built together with various NOCs like UK Sport, NOC*NSF and USOC. The dataset is updated every morning and evening with the new results which have been entered live over the previous 24 hours.

Podium contains all Olympic sports in a dataset described below in Table 2:

Competition	Data depth
Olympics	All placings from 1992. Top-3 for full history
World Championships	All placings from 2004. Top-3 for full history
(World) Cups	All placings from 2008. Top-3 for full history
Continental Championships	All placings from 2012. Top-8 from 2000
World Rankings	Monthly updated top-300 (some sports weekly)
European Games	All placings from 2015
Commonwealth Games	All placings from 2010. Top 8 1994-2006, Top-3 for full history
Pan American Games	All placings from 2015. Top-8 from 2003-2011. Top-3 for full history
Asian Games	All placings from 2006. Top-3 for full history
Youth Olympic Games	All placings for full history
Youth/Junior World Championships	All placings from 2006
Youth/Junior European Championships	All placings from 2012

Table 2: The Podium dataset



Podium will be demonstrated in a separate workshop session in order to provide insight into the data that is contained in the Infostrada Sports database as well as to show some of the descriptive analyses that are made available in this analysis tool.

In order for this demonstration to be as interactive and successful as possible, it is requested that attendees provide some simple questions of the data in advance of the session.

The Ageing Game

Simon Gleave[1] and Mark Taylor[2]

[1] Infostrada Sports, Netherlands
simon.gleave@infostradasports.com
[2] The Power of Goals, UK
taylor.mark1959@yahoo.com

Abstract

Manchester City's ageing squad has been identified in the media as a potential reason for the team to produce a lower level of performance this season but how much truth is there in this assertion and are team age structures changing? The age structure of football teams is an under-utilised area of football analysis and we propose to investigate them thoroughly by means of defining age curves for different positions, describing and analysing the issue of relative age effect and making proposals for the ideal age structure of squads playing in different strata of a top European competition. We also propose to look at whether clubs are moving away from the ideal age structure because of the added pressure from fans and media to deliver immediately on the pitch.

Whilst this analysis concentrates on club football, it is also highly relevant for all team sports as maximising performance and ensuring that younger players gain enough experience are key questions for policy makers at any club.

1 Introduction

As the English top flight has become wealthier, this has had an effect on the age of the players playing, not only in terms of their average age but also the distribution of the ages throughout a starting XI which will be referred to here as the age structure of the team. The average age of English top flight teams gradually increased over the 20 year period prior to the current Premier League being introduced in 1992/1993 but had already reached its approximate current figure of around 27 years old by the time the Premier League started. So, whilst the average age of Premier League teams has only increased slightly since 1992, the age structure within those teams has changed with more of an emphasis on 'peak age' players in particular and less opportunity for players who are still developing towards the age at which they would be expected to produce their peak performance.

The aim of this paper is to investigate how the average age and age structure of English top flight football teams has changed over more than 40 years with a specific focus on those changes within the Premier League era. The age structure of different Premier League clubs will be illustrated by a simple visual illustration. It is also proposed to investigate whether there is a relationship between performance and the age structure of starting line-ups in the Premier League although that is beyond the scope of this first draft.

2 Data

In total there are four datasets being implemented in this work. The first uses 15 complete seasons of Premier League data to define the age curves and includes minutes played as well as players, dates of birth, clubs and seasons. The second and third datasets comprise the 23 seasons of the Premier League from 1992/1993 to 2014/2015 and the fourth uses data from

2001/2002 and 2011/2012 from one of those datasets supplemented by data from 1971/1972, 1981/1982 and 1991/1992 for a greater longitudinal comparison. The first three datasets are from the database maintained by Infostrada Sports and the third is a combination of data from the Infostrada Sports football database and data collected by the authors from the Rothmans Football Yearbooks in the relevant years.

2.1 Dataset 1 – Premier League playing minutes summary

Every player who played in a Premier League match from 1999/2000 to 2013/2014 was collected with his date of birth and minutes played. By using 1 January as a reference date in each season, it was possible to calculate an age for each player for each season.

2.2 Dataset 2 – Premier League starting appearances summary

Every player who played in a Premier League starting line-up from 1992/1993 to 2014/2015 was collected with his date of birth and number of appearances in the starting line-up. By using 1 January as a reference date in each season, it was possible to calculate an age for each player for each season. The data was sourced from the Infostrada Sports football database using the Football Stats Generator, a tool built in Qlikview by the Infostrada Sports Analytics department.

2.3 Dataset 3 – Premier League starting appearances per player per match

In order to provide greater accuracy for the graphics showing team structure, a second dataset covering the same period as the first was generated from the Infostrada Sports football database. This contains a line for each player every time he is named in the starting line-up and provides the opportunity to calculate the player's age on the day of the match.

2.4 Dataset 4 – Top flight starting appearances summary

Three seasons with 10 years between each pair were selected in order to take a longer term longitudinal view at the data to gain insight into how things developed before the Premier League era began. These data were collected by the authors from the relevant Rothmans Football Yearbooks and added to two seasons sampled from dataset 1. The choice of seasons (1971/1972, 1981/1982, 1991/1992, 2001/2002 and 2011/2012) was an arbitrary one driven by the fact that the authors were already in possession of data from the first of those five seasons as that copy of the Rothmans Football Yearbook was digitally available on the internet.

With the exception of dataset 1, the other datasets were all based on players in the starting line-ups only for two reasons. The first is that managers are assumed to select their starting line-ups as being their best chance of winning a match. The second is that those starting players actually play 94% of the minutes played.

3 Defining age curves

In this study, the aim was to produce a simple initial ageing curve in order to have a basic idea of when football players are present in each of three categories – development, peak and decline.

These three categories were then named Talents, Peak Age and Veterans for the purposes of the analyses which followed.

In order to do this, an assumption was made that the managers selecting the players are generally efficient and therefore the number of minutes played by players of different ages would be representative of the age curve of performance. Alternative methods could be to use an on-pitch performance indicator such as shots, a composite performance indicator or ideally, if a source was available, physical data.

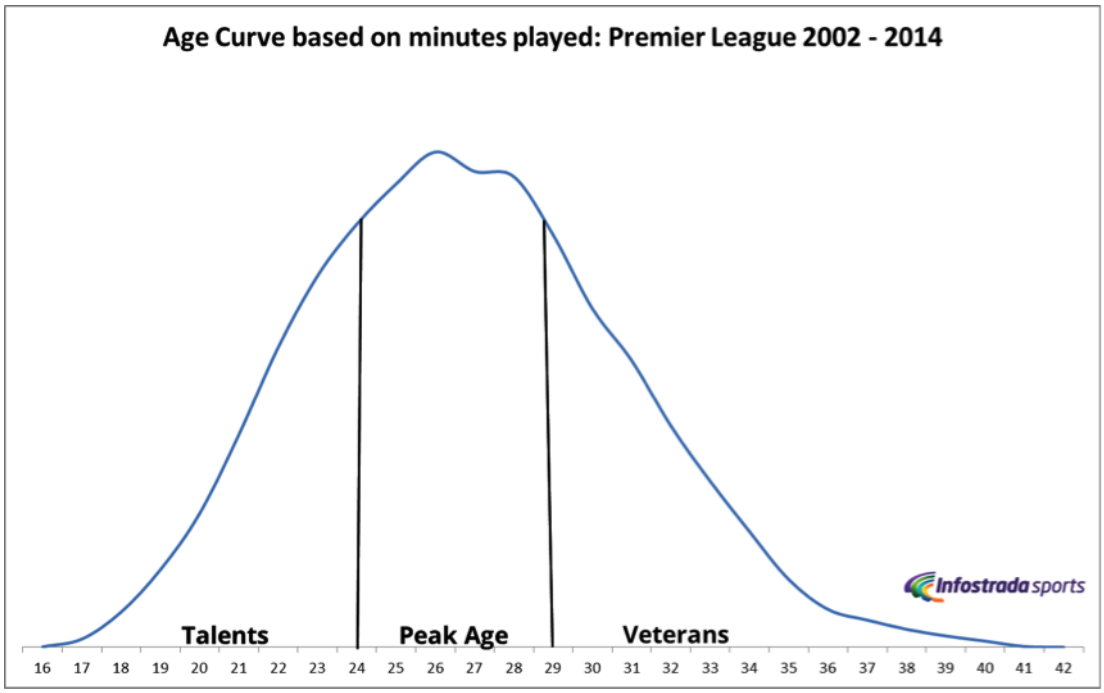


Figure 1: Premier League age curve

Using minutes played and allowing the peak part of the curve (Figure 1) to be a little wider in order to encompass a portion of development and decline allows a six year agespan for the Peak Age group. The vast majority of the other two groups are also contained in age spans of six years (18-23 for Talents and 30-35 for Veterans). Coincidentally, the qualification requirements for the PFA Young Player of the Year also tie in with the Talent definition as players are eligible if they are aged 23 or under at the beginning of the season.

4 Ageing in the English top flight 1971/1972 – 2011/2012

In order to investigate the ageing process of the English top flight over a longer period than just the Premier League era, five data points were taken 10 years apart from each other, beginning with the 1971/1972 season and ending with 2011/2012. As expected, the average age increased in general and a clear trend upwards can be observed in Figure 2.

Two different averages were calculated as a simple average age across all starting players was likely to be biased towards younger players who rarely played. In order to account for

this, a second average age was calculated which was weighted for the number of times each player appeared in the starting line-up. As expected, the weighted average age was always higher than the simple average age but the difference between the weighted average and simple average reduced from about 0.8 at the first three time points to only around 0.3 in each of the two Premier League seasons in the sample.

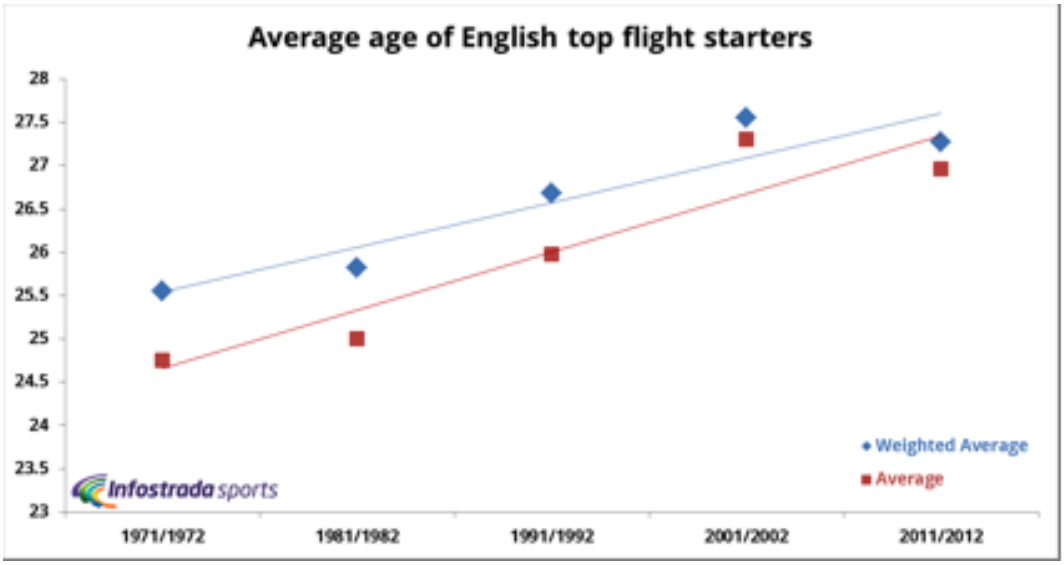


Figure 2: The increase in average age of top flight starters in England 1971 – 2012

Looking at the trend in the number of players aged under 24 starting top flight matches the "Talents" – there has been a declining trend over the same period which is shown in Figure 3. This also helps to illustrate the reason for the narrowing difference between the average age and weighted average age in Figure 2. Quite simply, the "Talents" are receiving less opportunity.

Looking at the Premier League era as a whole, there has been hardly any change in the weighted average age over the 23 year history of the competition with the figures from all but the first season being somewhere between 27 and 27.8. However, the age structure has changed as shown by the decline in the percentage of talents starting matches shown in figure 3.

During the Premier League era alone, this trend means that the average number of players aged under 24 who start for a team has reduced from two to one. This number tended to be between three and four during the 1970s and 1980s allowing a smooth development path for younger players into the professional game at this level. So, although the average age of the players starting matches has remaineed reasonably stable during the Premier League era, the age structure within those teams has not with fewer and fewer young players being selected and more in the peak age groups.

Therefore, using a simple measure of average age is not enough to represent how this is changing. A number of statistics like the gini coefficient or herfindahl-hirschmann index were investigated as possible indicators of team age structure but it was concluded that a visual representation would be better.

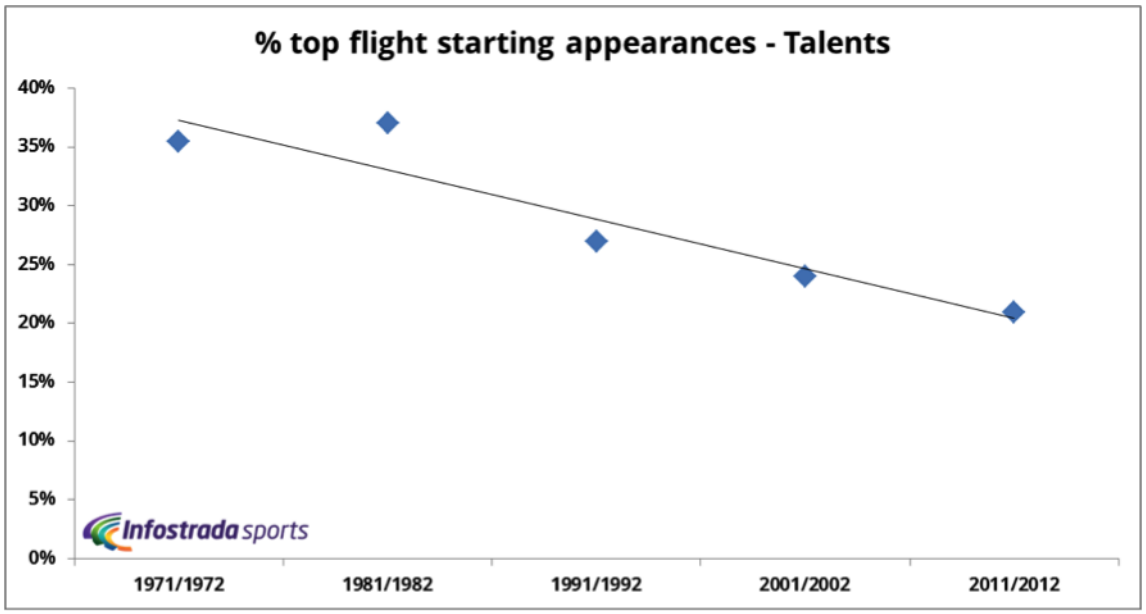


Figure 3: The decline in top flight starting appearances for "Talents"

5 Visualising age structure

An advantage of visualising the age structure of a team and, eventually, showing an optimum age structure is that it is easy to understand for the users of such information. Ages of the starting XIs were tabulated for all matches within a season and then smoothed three times. The equation to do this for each age category i was:

$$y\_\text{smooth}(i) = 0.25 * y(i-1) + 0.5 * y(i) + 0.25 * y(i+1)$$

In Figure 4, these smoothed age distributions are shown for the 2014/15 season in the Premier League overall, and for Manchester City who have attracted a lot of media attention for the age structure of their team, notably the lack of young players and proliferation of players in their late twenties and early thirties.

As Figure 4 shows, Manchester City's starting XIs this season have tended to be older than the Premier League average but also focused heavily on players who are around 28 to 29 years old and therefore coming to the end of peak age in footballing terms. It is arguable that this may already be having an effect on performance but this will certainly become an even bigger issue if not quickly addressed as these players will only age further.

6 Further work

This paper represents work in progress and there is still a lot to achieve, notably to identify a clear link between the age structure of a squad and performance in the league. The theory is that it would be an advantage to try to maximise the number of players of peak age starting matches but in order to avoid periods of rebuilding when those players begin to decline, an

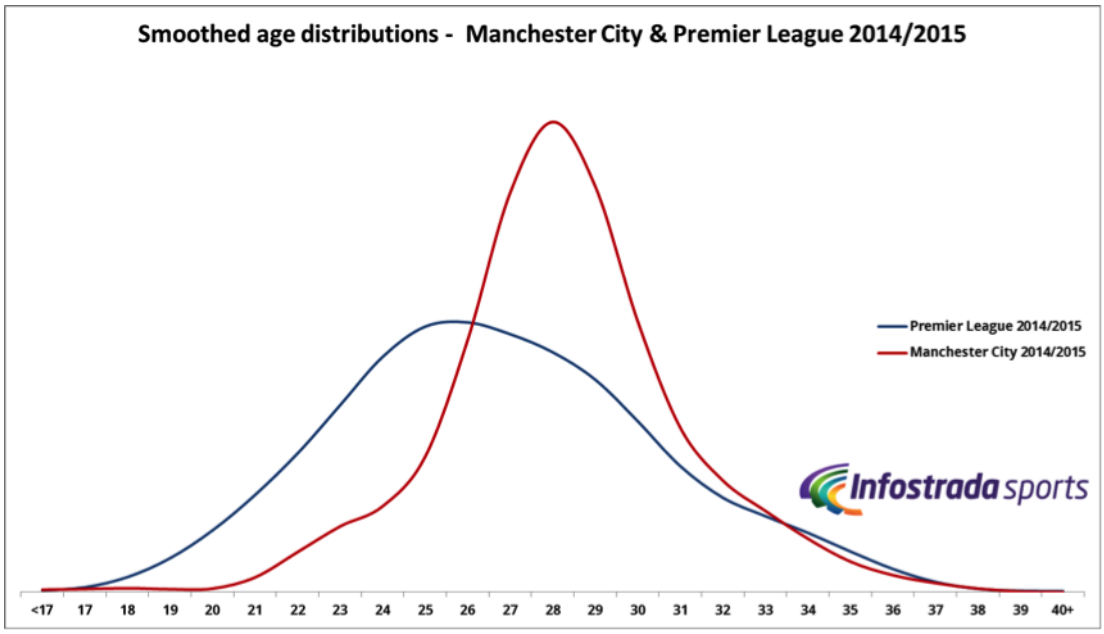


Figure 4: Manchester City's starting XIs in comparison to the Premier League

optimal framework of different ages needs to be found, and, preferably visualised in order to help clubs formulate the players they need to sign to fill in the gaps within their squad.

Mathematical Models That Describe and Predict Performance Change in Mens And Womens Athletic Events: 1960-2014

Ian Heazlewood and Joe Walsh

Charles Darwin University, Darwin, Australia
ian.heazlewood@cdu.edu.au

Abstract

The prediction of athletic performance is a recurring theme as coaches and sports scientists endeavour to understand limits of sports performance. Predictive models based on Olympic data for athletics have derived some accurate predictions of performance in the 2000, 2004, 2008 and 2012 Olympic Games. However, best fitting mathematical functions varied with different events. The aim of this research was to develop descriptive and predictive mathematical models of changing performance over time for track and field events for both men and women to assess if athletes in different events are increasing or decreasing performances. Data was provided by International Association of Athletic Federations (IAAF) and by averaging scores of the top twenty performances for the major track and field events for years 1960 to 2014 for men and women senior outdoor athletes. Mens events were 100m, 200m, 400m, 1500m, 110mH, 400mH, long jump, high jump, shot put, discus, javelin and hammer and womens events were 100m, 200m, 400m, 1500m, 100mH, long jump, high jump, shot put and discus. The mathematical approach utilised regression-curve estimation time series analysis by evaluating data fits to linear, logarithmic, inverse, quadratic, cubic, compound, power, sigmoidal, growth exponential and logistic functions. The calculations were conducted using SPSS Version 22 software. Results indicated cubic functions consistently displayed best fits with the data. Mens track events displayed R-square values of 1500m (.955) to 100m (.791), in field events high jump (.958) to javelin (.514) with increasing performances for 100m, 200m, 110mH and discus. Womens track events displayed R- square values of 400m (.957) to 100mH (.866), in field events high jump (.967) to javelin (.456) and increasing performances for 100m, 100mH and 1500m. Current trends indicated most events performances were decreasing suggesting track and field events have reached limits of performance or training adaptations are declining.

1 Introduction

The prediction of future athletic performance by humans is a recurring theme during Olympic and IAAF World Championship years, as well as forming the basis for some stimulating 'crystal ball gazing' in some of the learned sports science journals and in the mass media and this will continue as sports scientists attempt to predict athletic performances at the 2015 IAAF World Championships in Athletics, Beijing, China and 2016 Rio de Janeiro Summer Olympic Games, Brazil. Mathematics and sport science are predicated on the principles of description and more importantly prediction of relationships. The ability to make substantive and accurate descriptions and predictions of past, current and future trends in elite level sports performance indicates such approaches reflect "good" science. Often these predictions are purely speculative and are not necessarily based upon any substantial evidence or mathematical modelling, rather they are based on an inherent belief that human performances must continue to improve over time and the constant breaking of records in sporting events may suggest this belief. The

accessibility of public domain data in the form of results from Olympic Games, IAAF World Championships, IAAF world records and IAAF world best performances for specific years allow the analysis of performances in all IAAF athletic events. From these analyses, changes in performance over time can be observed, described and predictions of future performance can be made utilising the process of mathematical interpolation and extrapolation.

A number of researchers have attempted to predict future performances by deriving and applying a number of mathematical statistical models based on past performances in athletics. Prendergast [18] applied the average speeds of world record times to determine a mathematical model for world records. The records or data used in the analysis spanned a 10 year period. Following his analysis, Prendergast [18] raised the question of whether any further improvements can be expected or if the limits of human performance have been reached. The sport of athletics [5, 6, 7, 8, 10, 11, 12] has been addressed in this manner. The knowledge of past and future levels of sporting performance has been identified by Banister and Calvert [1] as beneficial in the areas of talent identification, both long and short term goal setting, and training program development. In addition, expected levels of future performance are often used in the selection of national representative teams where performance criteria are explicitly stated in terms of times and distances for example, as required entry standards at 2015 IAAF World Championships [16].

Péronnet and Thibault [2] postulated that some performances, such as the men's 100m sprint is limited to the low 9 seconds, whereas Seiler (referred to by Hopkins [13]) envisages no limits on improvements based on data reflecting progression of records over the last 50 years. According to Seiler improvements per decade have been approximately 1% for sprinting, 1.5% for distance running, 2-3% for jumping, 5% for pole vault, whereas female sprint times may have already peaked. The differences for males and females, where female performances not changing or in some cases regressing is thought to reflect the impact of successful drugs in sport testing on females. For example the 400m womens record in athletics was set in 1985 at 47.6s and the 2013 400m women's time at the 2013 IAAF World Championships was only 49.41s [14].

In athletics for the men's events the mathematical functions based on Olympic performances [6] were 100m inverse, 400m sigmoidal, long jump cubic and the high jump displayed four functions (compound, logistic, exponential and growth). In the womens events the mathematical functions were 100m cubic, 400m sigmoidal, long jump inverse and high jump displayed four functions (compound, logistic, exponential and growth). The functions for the IAAF World Championship data were poor predictors of past, current and future performance in these events with low R^2, non-significant p-values and large residuals or error. The one exception was the womens long jump, which predicted a significant decline in performance. The low predictive ability may be due to a short timeframe of data and insufficient data sets to develop substantive predictive models or excessive variability in athletic cohorts, which mask trends if they occur. The curves that fit the data have also displayed interesting findings as no one curve fits all the data sets. This may indicate that different events are dependent upon different factors that are being trained differently or factors underpinning performance evolving in slightly different ways. This has resulted in different curves or mathematical functions that reflect these improvements in training or phylogenetic changes over time. However, at some point in time how accurately the predictive models reflect reality can be assessed.

The predictions of Heazlewood and Lackey [6] paradoxically predicted the men's 100m to improve to zero by year 5038 and the women's 100m to reach zero by year 2429, which indicates a more rapid improvement over time for women sprinters. In the model [6], the women's times would be faster than men by 2060 where it was predicted the finalist at the Olympic Games would average 9.58s for men and 9.57s for women respectively. A similar crossover effect, where

predicted female performances would exceed male performances, was noted for the 400m and high jump. However, the ability to predict performances at the 2000, 2004, 2008 and 2012 Olympic Games for the mens and womens 100m, 400m, long jump and high jump, based on the 1924 to 1996 data, was very accurate with low percentage error for most of the these events. The aim of this research was to develop descriptive and predictive mathematical models of changing performance using a more substantive data set from IAAF competition years 1960-2014 and analyzing a more substantive data base of events such as for mens of 100m, 200m, 400m, 1500m, 110mH, 400mH, long jump, high jump, shot put, discus, javelin and hammer and womens events of 100m, 200m, 400m, 1500m, 100mH, long jump, high jump, shot put and discus.

2 Methods

2.1 Sample

A substantive data set was provided by International Association of Athletic Federations [15] and by averaging scores of the top twenty performances for the major track and field events for years 1960 to 2014 for men and women senior outdoor athletes. The top 20 athletes to represent the trends in performance were a significant increase over just finalists in major competitions, such as IAAF World Championships and Olympic Games which has been utilised previously in similar research [5, 6, 10, 11, 12]. Men's events were 100m, 200m, 400m, 1500m, 110mH, 400mH, long jump, high jump, shot put, discus, javelin and hammer. The women's events were 100m, 200m, 400m, 1500m, 100mH, long jump, high jump, shot put and discus. The majority of events in this study represented underpinning constructs of speed-power in the 100m, 200m, 400m, jumps and hurdles; strength-power in the shot put, discus, javelin and hammer; and speed-endurance in the 800m and 1500m. Total athletes in the sample were 25,920 where 14,040 were men and 11,880 were women. Times for the track events were measured in seconds and reported to the nearest 0.01s and distances for field events measured in metres to the nearest 0.01m. It is important to note in the mens and womens 100m, 200m, men's 110m hurdles and women's 100m hurdles were competed in with legal wind speeds. These values served as the data set to derive and test the curve estimations in term of model fit. This is a similar method utilised by Heazlewood and Lackey [6] and Heazlewood [12] to curve fit Olympic data for athletic events.

2.2 Mathematical-Statistical Approach

Curve estimation is an exploratory tool in model building and model selection [3, 4] and SPSS Version 22 (2014) where the best mathematical model or function is selected to represent quantitative relationships between the independent or predictor(s) and the dependent or response variable(s). The mathematical solutions and curve estimations were derived using SPSS Version 22 (2014) statistical software. Details of the method for obtaining the regression equation and curve of best fit are given in the accompanying paper [9].

3 Results

Results indicated cubic functions consistently displayed best fits with the data. Table 1 and table 2 indicate information for male and female athletes and the specific track and field events in terms of mathematical function, R^2 value, level of statistical significance, mathematical

equation based on cubic function, parameter estimates for equation constant, coefficients b1, b2, b3 and trends in performance over time for each event.

Mens track events displayed R^2 values of 1500m (.955) to 100m (.791) and in field events high jump (.958) to javelin (.514). Events displaying increasing performances from 1960 to 2014 were 100m, 200m, 110mH and discus. Decreasing performances were observed in the other men's events and trends are highlighted in table 1, specifically, 400m from 2004, 800m from 2010, 1500m from 2009, 400mH from 2001 and high jump from 1999. The throwing events are interesting in performance trends, where shot decreased from 1986 (21.54m) to 2002 (21 20m) and with very marginal increase after 2003 (21.11m) to 2012 (21.28), and discus performance decreasing after 1985 (68.49m) to 2003 (66.83m) and increasing again after 2004 (67.22m) to 2012 (67.74m). The hammer increased up to 2000 (81.17m) and then decreased to current (79.89m). It is interesting to note that current world records in the men's shot put, discus and hammer were set in 1990, 1986 and 1986 respectively.

The javelin displays more interesting trends as in 1986, the men's javelin of 800 grams was redesigned [20], so that the centre of gravity was moved 4 cm forward, the surface areas in front of and behind the centre of gravity were reduced and increased, respectively. The effect was to reduce lift and increase the downward pitching moment and bring the nose down earlier, reducing the flight distance by around 10% but also causing the javelin to stick in the ground more consistently [20]. The outcome was increasing performances in males up to 1986 (91.34m), changes in specification from 1 April 1986 and then increasing performance with new specification 1987 (82.20m) to 1999 (86.91m) and finally slight decreasing performance from 2000 (86.92m) to 2012 (85.32m).

Event	Function	R^2	Sign	Equation Parameter Estimates				Trends
				Constant	b1	b2	b3	
100m	Cubic	.791	P<.001	10.234	-.008	.000	-2.172E-6	Performance increasing
200m	Cubic	.867	P<.001	20.721	-.030	.001	-6.165E-6	Performance increasing
400m	Cubic	.855	P<.001	46.059	-.056	.000	5.237E-6	Performance decreasing after 2004
800m	Cubic	.946	P<.001	107.849	-.190	.003	-1.193E-5	Performance decreasing after 2010.
1500m	Cubic	.955	P<.001	222.348	-.379	.002	2.426E-5	Performance decreasing after 2009.
110mH	Cubic	.891	P<.001	13.809	-.023	.000	-1.381E-6	Performance increasing
400mH	Cubic	.949	P<.001	50.925	-.116	.001	1.328E-6	Performance decreasing after 2005.
Long Jump	Cubic	.933	P<.001	7.816	.021	.000	-2.192E-6	Performance decreasing after 2001.
High Jump	Cubic	.958	P<.001	2.083	.013	.000	-1.509E-8	Performance decreasing after 1999.
Shot Put	Cubic	.908	P<.001	17.816	.308	-.010	9.407E-5	Performance decreasing after 1985. Increasing again after 2003.
Discus	Cubic	.939	P<.001	55.246	1.101	-.033	.000	Performance decreasing after 1986. Increasing again after 2004.
Javelin	Cubic	.514	P<.001	78.387	.950	-.031	.000	Increasing up to 1986. Changes in specification from 1 April 1986. Increasing to 1999. Decreasing 1999 to 2012.
Hammer	Cubic	.928	P<.001	63.228	.811	-.006	-7.130E-5	Increasing to 2000, decreasing 2000 to current.

Table 1. Trends in male track and field events 1960-2014.

Women's track events displayed R^2 values of 400m (.957) to 100mH (.866) and in field events high jump (.967) to javelin (.456). Events displaying increasing performances from 1960 to 2014 were 100m, 100mH and 1500m from 2004. Decreasing performances were observed in the other womens events and specific trends are highlighted in Table 2. In events of the 200m, 400m, 800m, high jump, discus, shot put and long jump performance has been decreasing from 1995, 1997, 1990, 2000 and 1991, 1986 and 1998 respectively. The results were similar in trend to

the men in the shot put where performance is increasing to 1986, decreasing to 2010 and then slight increase to present; and in discus increasing to 1991 and then decreasing 1992 to present. It must be emphasized that current performances in both shot put and discus for women are below the 1986 and 1991 performances. Women's world records for shot put and discus were set in 1987 and 1988 respectively.

The women's javelin displayed more interesting trends as in 1999 the women's javelin of 600 grams was redesigned as in the case of the male javelin [20]. Specifically, performance increasing to 1985, decreasing from 1985 to 1999 with old specification, changed specification 1999, slight decrease from 1999 (64.12m) to 2005 (63.51m) and slight increase 2005 to 2014 (64.97m), however well below 1985 standard (68.18m). Figure 1 indicates a graphical trend for the womens 400m from 1960 to present, as well as extrapolating to 2025.

It is important to emphasize in the majority of men's and women's track and field events few events are trending to increased performance up to 2014 and such as is the situation with men's and women's 100m and 110m hurdles and 100m hurdles. So what factors are causing the majority of events to decline or become essentially static? A number of possible explanations will be provided in the discussion.

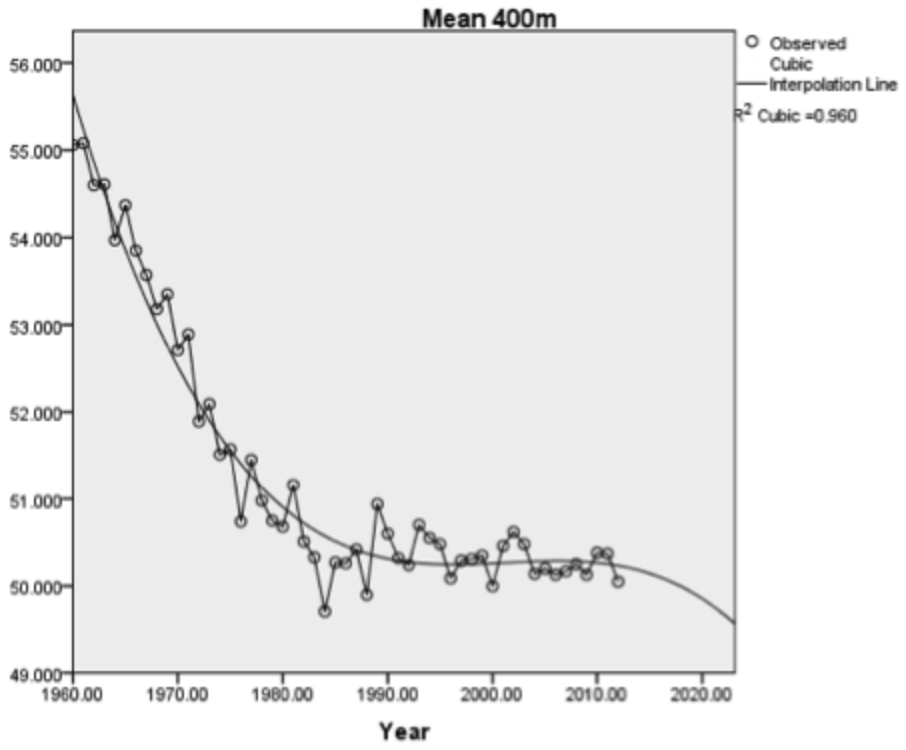


Figure 1: Trends in mean times(s) for women's 400m from 1960 to 2014 and extrapolating to 2025

Event	Function	R^2	Sign	Equation Parameter Estimates				Trends
				Constant	b1	b2	b3	
100m	Cubic	.919	<.001	11.628	-.037	.001	-5.461E-6	Performance increasing
200m	Cubic	.943	<.001	24.149	-.121	.003	-1.808E-5	Performance decreasing after 1995
400m	Cubic	.957	<.001	56.023	-.419	.010	-7.822E-5	Performance decreasing after 1997
800m	Cubic	.933	<.001	129.446	-.914	.023	.000	Performance decreasing after 1990.
1500m	Cubic	.881	<.001	289.659	-4.363	.126	-.001	Performance decreasing after 1987. Increasing again after 2004.
110mH	Cubic	.866	<.001	14.129	-.104	.003	-2.235E-5	Performance increasing
Long Jump	Cubic	.895	<.001	6.111	.045	-.001	8.594E-7	Increasing to 1998, decreasing to present.
High Jump	Cubic	.967	<.001	1.664	.019	.000	1.667E-6	Increasing to 2000, decreasing to present.
Shot Put	Cubic	.890	<.001	14.344	.523	-.015	.000	Increasing to 1986, decreasing to 2010, slight increase to present.
Discus	Cubic	.909	<.001	48.722	1.418	-.035	.000	Increasing to 1991, decreasing 1992 to present.
Javelin	Cubic	.456	<.001	-311.82	.189	.000	.000	Increasing to 1985, deceasing to 2005, slight increase, 2014 well below 1985 standard.

Table 2. Trends in female track and field events 1960-2014.

4 Discussion

Current trends indicated most events' performances were decreasing and suggesting track and field events have reached limits. What factors can be postulated to explain these trends? The review of literature indicated that Prendergast [18] believed the limits of human performance have been reached in athletics and these results support such a proposition, especially where mens world records in throwing events were achieved in the shot put in 1990, the discus in 1986, hammer in 1986 and javelin with new specification in 1996.

The world records in the women's throwing events were achieved for the shot in 1987, discus in 1988 and are consistent with the Prendergast [18] prediction, however javelin in 2008 with the new specification and in the hammer, a new event and first contested in the Olympics in 2000, in 2014 and has scope to improve as women athletes become more technically proficient.

Seiler (referred to by Hopkins [13]) envisaged no limits on athletic improvements based on data reflecting progression of records over the last 50 years although published in 2000. Although Seiler predicted progression in sprints would be minimal for men and static for women, the paradoxical effect is the men's 100m and 200m have improved as has the women's 100m, whereas the men's 400m and women's 200m and 400m have declined. The women's record for the 200m and 400m were set in 1988 and 1985 respectively and at this point in time based on current performances almost unachievable as the women's 400m record of 47.60 is two seconds beyond what best current athletes can sprint.

Three possible theories are presented to explain the decline in performance in the majority of track and field events studied. It is important to note the majority of events studied were speed-power (100m, 200m, 400m, jumps and hurdles) and strength-power (shot put, discus, javelin and hammer) events. Only two events were speed-endurance dependent, the 800m and 1500m.

1. More vigilant drug testing in the sport of athletics, both in and out of competition, as well as public disclosure for first offence for those athletes suspended or banned by the IAAF and therefore dampening or preventing enhancing performance effect of anabolic androgenic hormones. To have a world record ratified by the IAAF the application form has a section indicating the athlete was drug tested at the time of competition and that the athlete has

passed the test, that is no adverse drug findings [15]. No drug test then no world record, no Olympic Gold or IAAF World Champion. The Olympics and IAAF World Championships are where many of the year's top ranked performances occur. The normal suspension/ban for first offence anabolic androgenic hormones is two years. A second offence of this kind now results in a life ban. If found positive all performances, monies paid by IAAF, awards and prizes are forfeit so the punitive outcomes can very significant and are thought to act as a strong deterrent to taking prohibited substances that appear on the World Anti-Doping Agency [19] list. This factor might be one of the most important for the current widespread declines across events.

2. Financial rewards at the top in athletics now permit athletes to be fulltime professional athletes, which means training time is fulltime and training hours can no longer be added to as there are finite hours in a day and days in a week and so on. Fewer athletes are undertaking the throwing events. Human training adaptation is finite and based on the law of diminishing returns and we might be reaching or have reached this biological mathematical limit.

3. The source population providing the sample of potential athletes internationally is getting smaller in terms of motor fitness abilities and thus fewer capable athletes to select from within source population. In the context of Australia as well as in western nations there is an increasing overweight and obesity epidemic/pandemic for males and females. In Australia overweight and obesity for males 18 years and over estimated to be 70% and for females 56% [17]. High performance track and field athletes are generally well over 18 years of age in both sexes. These overweight/obesity rates have increased by five and six percent respectively, when compared to the 1995 results. "People being overweight or obese may have significant health, social and economic impacts, and is closely related to lack of exercise and to diet," [17]. The carry-over effect is reduced motor fitness of which strength, power, speed and endurance are component and the consequence is reduced performances.

5 Conclusion

It is somewhat disconcerting to observe and quantify the declines across many track and field events based on the year by year performances of the top twenty athletes selected to represent ontogenetic and limited phylogenetic trends. Only a very limited number of events displayed improvement, such as 100m and 110mH and 100mH. It is important to highlight these declines might be attributable to three factors, that are, more vigilant drug testing and punitive sanctions for adverse drug tests, the limits of training adaptation in fulltime athletes and finally a decline in the general motor fitness of the global population and the source population for new athletes, especially in the motor fitness factors of strength, power, speed and speed endurance. It will be very interesting to continue to evaluate these trends over the next decade and beyond to asses if we are becoming bigger, stronger faster or just bigger, fatter and slower.

References

[1] E.W. Banister and T.W. Calvert. Planning for future performance: Implications for long term training. *Canadian Journal of Applied Sports Science*, 5:170–176, 1980.

[2] Péronnet. F. and Thibault. G. Mathematical analysis of running performance and world running records. *Journal of Applied Physiology*, 6:453–465, 1989.

[3] D.G. Garson. Curve estimation. statnotes: Topics in multivariate analysis. [online], 2010. `http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm`, Retrieved 1/12/2010.

[4] J.E. Hair, W. Block, B.J. Babin, R.E. Anderson, and R.L. Tatham. *Multivariate Data Analysis (6th Ed.)*. Pearson - Prentice Hall, 2006.

[5] I. Heazlewood. The congruence between mathematical models predicting 2011 athletic world championship performances and actual performance in the 100m, 400m, long jump and high jump, June 2013. *9th International Symposium on Computer Science in Sport (IACSS2013)*. Oral presentation.

[6] I. Heazlewood and G. Lackey. The use of mathematical models to predict elite athletic performance at the olympic games. In deMestre, editor, *Proc. Third Conference on Mathematics and Computers in Sport*, pages 185–206, Amsterdam, 1996. Bond University.

[7] I. Heazlewood and J. Walsh. Mathematical models that predict athletic performance in the men's throwing events at the Olympic Games. In A. Bedford and T. Heazlewood, editors, *Proceedings of the 12th Australasian Conference on Mathematics and Computers in Sport*, pages 48–53, Darwin, Northern Territory, Australia, June 2014. MathSport (ANZIAM).

[8] I. Heazlewood and J. Walsh. Mathematical models that predict athletic performance in the women's throwing events at the Olympic Games. In A. Bedford and T. Heazlewood, editors, *Proceedings of the 12th Australasian Conference on Mathematics and Computers in Sport*, pages 54–60, Darwin, Northern Territory, Australia, June 2014. MathSport (ANZIAM).

[9] I. Heazlewood and J. Walsh. Mathematical models predicting performance in track and field at the 2011, 2013 and 2015 IAAF World Championships. In *Proceedings of the 5th International Conference on Mathematics in Sport*, Loughborough, U.K., 2015.

[10] I.T. Heazlewood and J. Walsh. Mathematical models that predict 2012 London Olympic Games swimming and athletic performances. In *Proc. 3rd International Conference on Mathematics in Sport*, pages 71–78, Salford, UK, 2011.

[11] I.T. Heazlewood and J. Walsh. Mathematical models that predict athletic performance in sprint and jump events for current and future world champions. In *Proc. 3rd International Conference on Mathematics in Sport*, pages 64–70, Salford, UK, 2011.

[12] T. Heazlewood. Prediction versus reality: The use of mathematical models to predict elite performance in swimming and athletics at the Olympic Games. *Journal of Sports Science and Medicine*, 5:541–547, 2006.

[13] W. Hopkins. Limits to performance. *SportScience*, 4, 2000. `http://www.sportsci.org/jour/0002/inbrief.html#limits`, Retrieved 3/4/2006.

[14] International Association of Athletic Federations. 2013 IAAF world championships results. [online], 2013. `http://www.iaaf.org/results/iaaf-world-championships-in-athletics/2013/iaaf-world-athletics-series-4873`.

[15] International Association of Athletic Federations. World record application forms. [online], 2014. `http://www.iaaf.org/about-iaaf/documents/technical#world-record-application-forms`.

[16] International Association of Athletic Federations. Qualification System and Entry Standards. 2015 IAAF World Championships, Beijing, China. [online], 2015. `https://iaafmedia.s3.amazonaws.com/competitioninfo/544c7d0a-0807-4226-a85b-79928c58a097.pdf`.

[17] Australian Bureau of Statistics (ABS). Overweight/ obesity (measured and self-reported body mass index (bmi) 2013 - age standardised 18 years and over. [online], 2014. `http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/4338.0~2011-13~Main%20Features~Overweight%20and%20obesity~10007`.

[18] K. Prendergast. What do world running records tell us? *Modern Athlete and Coach*, 28:33–36, 1989.

[19] World Anti-Doping Agency (WADA). List of prohibited substances and methods. [online], 2014. `http://www.wada-ama.org/en/world-anti-doping-program/sports-and-anti-doping-organizations/international-standards/prohibited-list/`.

[20] Wikipedia. Javelin throw. [online], 2015. `http://en.wikipedia.org/wiki/Javelin_throw`, downloaded 30.4.2015.

Mathematical Models Predicting Performance in Track and Field at the 2011, 2013 and 2015 IAAF World Championships

Ian Heazlewood and Joe Walsh

Charles Darwin University, Darwin, Australia
ian.heazlewood@cdu.edu.au

Abstract

Mathematics and science are based on principles of description and importantly prediction. The ability to make substantive and accurate predictions of future elite sports performances indicates such approaches reflect "good" science. Often these predictions are purely speculative and are not based upon any substantial evidence; however they provide criteria for athletes in terms of performance standards to be achieved to be competitive internationally, resetting international qualifying standards for IAAF sanctioned competitions such as World Championships, to focus on limits of human athletic performance and evaluating long term adaptations of athletes to chronic training. The research aim was to further develop predictive mathematical models using IAAF World Championship data in athletics outdoor championships 1983 to 2009 (previous model) for men's and women's 100m, 400m, long jump and high jump and then evaluate a revised model based on the addition of 2011 and 2013 data to predict the 2015 performances. Data sets consisted of average scores of the top eight performances in the finals of each event. The mathematical approach utilised regression-curve estimation time series analysis by evaluating data fits to linear, logarithmic, inverse, quadratic, cubic, compound, power, sigmoidal, growth exponential and logistic functions. The calculations were conducted using SPSS Version 22 software and goodness of fit by R^2, significance, accuracy of prediction and residuals. The results for mens and womens events indicated cubic functions consistently displayed the best fits with R^2 values for 100m (.723), 400m (.336), long jump (.221) and high jump (.920) for men and 100m (.366), 400m (.526), long jump (.692) and high jump (.496) for women. The inclusion of the 2011 and 2013 data improved the fits very marginally. The outcome was, although some variability existed in R^2 the predictions for events were accurate indicating in the majority of events the cubic functions displayed good model fits.

1 Introduction

The prediction of future athletic performance by athletes is a recurring theme during IAAF World Championships, which are now held biannually in odd calendar years and the next will be Beijing, 2015. Prediction also forms the basis for crystal ball gazing in some of the learned sports and exercise science journals and in mass sports media and this will continue as sports scientists attempt to predict athletic performances at the 2015 IAAF World Championships in Athletics, Beijing, China. The ability to make substantive and accurate descriptions and predictions of past, current and future trends in elite level sports indicates such approaches reflect "good" science, that is if they are successful. Often these predictions are conjecture and "gestimates" and are not necessarily based upon any substantial evidence such as mathematical models accurately mapped with real data. They are based on an inherent belief that human performances must continue to improve over time and the constant breaking of records in sporting events may suggest this belief, even when research may indicate otherwise [6, 7]. The

accessibility of public domain data in the form of results from IAAF World Championships from the Championships' inauguration in 1983, allows the analysis of performances in all IAAF athletic events. From these analyses, changes in performance over time can be observed and described; and predictions of past, recent and future performance can be accomplished utilising the processes of mathematical modelling, interpolation and extrapolation.

A number of researchers have attempted to predict future performances by deriving and applying a number of mathematical statistical models based on past performances in athletics. In the sport of athletics mathematical statistical researchers [4, 5, 6, 7, 8, 9, 10] have applied such a paradigm to successfully predict with varying degrees of accuracy future performances in men's and women's sprint and jump events. The knowledge of past and future levels of sporting performance has been identified by Banister and Calvert [1] as beneficial in the areas of talent identification, both long and short term goal setting, and training program development. It is important to highlight expected levels of future performance are often used in the selection of national representative teams, where performance criteria are explicitly stated in terms of times and distances such as prescribed entry standards for 2015 IAAF World Championships [12].

The predictive models for past IAAF World Championship 1983 to 2009 data were poor predictors of past, current and future performance in these events with low R^2, non-significant p-values and large residuals or error. The one exception was the women's long jump, which predicted a significant decline in performance. The low predictive ability may be due to a short timeframe of data and insufficient data sets to develop substantive predictive models or excessive variability in athletic cohorts, which mask trends if they occur. The curves that fit the data have also displayed interesting findings as no one curve fits all the data sets. This may indicate that different events are dependent upon different factors that are being trained differently or factors underpinning performance evolving in slightly different ways. This has resulted in different curves or mathematical functions that reflect these improvements in training or phylogenetic changes over time. However, at some point in time how accurately the predictive models reflect reality can be assessed.

The research aim was to further develop predictive mathematical models using IAAF World Championship data in athletics outdoor championships 1983 to 2009 (previous model) for men's and women's 100m, 400m, long jump and high jump and then evaluate a revised model based on the addition of 2011 and 2013 data to predict the 2015 performances.

2 Methods

2.1 Sample

The initial model to predict performances in the men's and women's 2015 IAAF World Championships 100m, 400m, long jump and high jump was derived from mean scores for the finalists in the men's and women's 100m, 400m, long jump and high jump from the 1983, 1987, 1991, 1993, 1995, 1997, 1999, 2001, 2003, 2005, 2007 and 2009 IAAF World Outdoor Championships, used as the exemplar for performance in each event for each championship year. Times for the 100m and 400m were in seconds and measured to the nearest 0.01s and distances for the long and high jump were in metres and measured to the nearest 0.01m. It is important to note in the men's and women's 100m all finals were competed with legal wind speeds. These values served as the data set to derive and test the curve estimations in term of model fit. This is a similar method utilised by Heazlewood and Lackey [5] and Heazlewood [10] to curve fit Olympic data for athletics and swimming events.

2.2 Mathematical-Statistical Analysis

Curve estimation is an exploratory tool in model building and model selection [2], as the best mathematical model or function is selected to represent quantitative relationships between the independent or predictor and the dependent or response variable. The mathematical solutions and curve estimations were derived using the SPSS Version 22 statistical software (SPSS Inc., 2015). The revised model to predict the men's and women's 2015 IAAF World Championships 100m, 400m, long jump and high jump was based on the addition of actual performance data from the 2011 and 2013 IAAF World Championships.

The most common curve estimation or model fit approaches are based on the following mathematical functions, which are linear, logarithmic, inverse, quadratic, cubic, power, compound, S-curve, logistic, growth, and exponential models. In terms of statistical approach [2] model fit indices are then applied to test the quality of the model. The general method of determining the appropriate regression models is: 1. start with an initial estimated value for each variable in the equation, 2. generate the curve defined by the initial values, 3. calculate the sum-of-squares where the sum of the squares of the vertical distances of the points from the curve, and 4. then adjust the variables to make the curve come closer to the data points.

There are several algorithms for adjusting the variables. The most commonly used method was derived by Levenberg and Marquardt (often called simply the Marquardt method). Adjust the variables again so that the curve comes even closer to the points. Keep adjusting the variables until the adjustments make virtually no difference in the sum-of-squares. Report the best-fit results and then the precise values you obtain will depend in part on the initial values chosen in step 1 and the stopping criteria of step 5. This means that repeat analyses of the same data will not always give exactly the same results. To investigate the hypotheses of model fit and prediction, the eleven regression models were individually applied to each of the athletic events. The regression equation that produced the best fit for each event, that is, produced the highest coefficient of determination (abbreviated as R^2), was then determined from these eleven equations.

The specific criterion to select the regression equation of best fit was the magnitude of R^2. The significance of the coefficient of determination (R^2) is a measure of accuracy of the model used. A coefficient of determination of 1.00 indicates a perfectly fitting model where the predicted values match the actual values for each independent variable [2, 3, 11]. Where more than one model was able to be selected due to an equal R^2, the simplest model was used under the principle of parsimony, that is, the avoidance of waste and following the simplest explanatory model, as well as the statistical significance of the analysis of variance, the alpha or p-value and size of residuals or error in predictions. Some caution is required to not over interpret a high R^2 as it does not mean that the researcher has chosen the equation that best describes the data. It also does not mean that the fit is unique – other values of the variables may generate a curve that fits just as well.

The experimental hypothesis was that inclusion of the 2011 and 2013 World Championship real data would slightly improve the existing model based on 1983 to 2009 data and slightly improve model fit for 2011 and 2013 performances and improve the prediction for 2015 World Championship performance in the 100m, 400m, long jump and high jump. The quality of the revised model and fit with 2015 World Championship performances can be tested after the championship in August, 2015.

3 Results

Table 1 indicates the actual and predicted track and field performances for men and women at 2011, 2013 and 2015 Athletics Senior World Championships. The actual and predicted scores in the first row of each cell are based on the original model from the 1983 to 2009 data. The value in the second row in each cell in italics is the value based on the revised model using the 2011 and 2013 real data added to derive an updated model. Not unexpectedly the revised model fits slightly better than the initial model. The results for men's and women's events indicated cubic functions consistently displayed the best fits with R^2 values for 100m (.723), 400m (.336), long jump (.221) and high jump (.920) for men and 100m (.366), 400m (.526), long jump (.692) and high jump (.496) for women. The inclusion of the 2011 and 2013 data improved the model fits for each event very marginally. The result was, although some variability existed in R^2 values, the predictions for 100m, 400m, long jump and high jump for both men and women were accurate indicating in the majority of events the cubic functions displayed good model fit, small residual error and accuracy in prediction.

Event Male	2011 Actual/ Predicted	2013 Actual/ Predicted	2015 Predicted Beijing 22-30/8/2015	R^2	Signi-ficance	**Equation Parameter Estimates**			
						Constant	b1	b2	b3
100m	Wind corrected 10.04/9.92 *9.96*	9.98/9.90 *9.95*	9.89 *9.94*	.723	.013	10.48	-.280	.047	-.002
400m	45.09/44.56 *44.68*	44.64/44.53 *44.67*	44.51 *44.66*	.336	.274 ns	45.71	-.52	.07	-.0031
Long jump	8.26/8.27 *8.263*	8.25/8.27 *8.261*	8.27 *8.257*	.221	.122 ns	10.04	-.0009	0	0
High jump	2.32/2.30 *2.316*	2.35/2.298 *2.315*	2.295 *2.313*	.920	.001	3.49	0	0	-1.47E-10
Event Female	2011 Actual/ Predicted	2013 Actual/ Predicted	2015 Predicted	R^2	Signi-ficance	**Equation Parameter Estimates**			
						Constant	b1	b2	b3
100m	Wind corrected 10.98/10.95 *10.961*	10.98/10.94 *10.953*	10.93 *10.945*	.366	.223 ns	11.22	-.01	.01	-.00068
400m	50.69/50.10 *50.24*	50.26/50.10 *50.27*	50.10 *50.29*	.526	.033	44.73	.003	0	0
Long jump	6.64/6.74 *6.747*	6.84/6.72 *6.726*	6.70 *6.706*	.692	.008	28.34	-.011	0	0
High jump	1.97/1.99 *1.982*	1.97/1.993 *1.983*	1.994 *1.984*	.496	.034	1.25	0	0	9.14E-11

Table 1. Actual and predicted track and field performances for men and women at 2011, 2013 and 2015 Athletic Senior World Championships. Times in seconds for 100m and 400m and distances in metres for long jump and high jump. Italic numbers are predicted scores using the revised model for 2011, 2013 and 2015 performances. The mathematical function used in the model is cubic in all cases.

In the 2015 World Championship, Beijing predictions, the first entry per cell is the prediction based on the initial model and the second entry in italics in each cell is the predicted value for each event using the revised model. It is important to note the revised model predicted values marginally different in performance than the initial model values. Specifically, both models indicate improving performance in the mens and womens 100m, improving performance in the

men's 400m, declining performance in women's 400m, declining performance in both the men's and women's long jump and decreasing heights in men's high jump and increasing heights in the women's high jump. It must be emphasised the changes will only be very small in times and distances. Although some of the statistical tests are not significant such as with men's 400m, men's long jump and women's 100m, both models display low prediction error. The 2011 percentage prediction error varied from 0% women 100m to 0.8% men 100m, 1.16% women 400m to 1.19% men 400m, 0.12% men's long jump to 1.50% women's long jump, and 0.86% men's high jump to 1.01% in women's high jump. The 2013 World Championship predictions similarly indicated low percentage error in prediction, specifically 0% women 100m to 0.61% men 100m, 0.22% men 400m to 0.30% women 400m, 0.36% men's long jump to 1.78% women's long jump, and 1.74% men high jump to 1.00% women high jump.

The research is confident that the predicted times and distances for the finalists in the different events will be very close to the actual time, assuming legal wind conditions. If the wind speeds are not legal then correcting for wind speed can be accomplished to remove constant error or bias due to wind speed. This was the case in 2011 for the 100m for both men and women. These findings suggest a high degree of performance consistency in terms of times and distances achieved over the past two championships.

4 Discussion

Five of the model fits were statistically significant, however all models displayed minimal error in prediction. The most consistent model for all events was based on a cubic mathematical function, which displayed the best fit indices. Accuracy of predicting actual event result was very high in 2011 and 2013 due to extremely low performance variability exhibited over time by athletes at the IAAF World Championships. These cubic models can be used with confidence to predict athletic performances in the 100m, 400m, long jump and high jump for both men and women in a near future international event, the 2015 World Championships, in terms of the mean scores for finalists in each event. The mean score represents the overall performance standards for the finalists and indicates that performance standards are marginally improving for men's 100, 400m and declining in the long jump and high jump. In the women's events marginal improvement is expected in the 100m and high jump and a marginal decline in the 400m and high jump. The researcher emphasizes that the changes will be very marginal and be of performance significance to athletes and not statistical significance in terms of changed scores.

References

[1] E.W. Banister and T.W. Calvert. Planning for future performance: Implications for long term training. *Canadian Journal of Applied Sports Science*, 5:170–176, 1980.

[2] D.G. Garson. Curve estimation. statnotes: Topics in multivariate analysis. [online], 2010. `http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm`, Retrieved 1/12/2010.

[3] J.E. Hair, W. Block, B.J. Babin, R.E. Anderson, and R.L. Tatham. *Multivariate Data Analysis (6th Ed.)*. Pearson - Prentice Hall, 2006.

[4] I. Heazlewood. The congruence between mathematical models predicting 2011 athletic world championship performances and actual performance in the 100m, 400m, long jump and high jump, June 2013. *9th International Symposium on Computer Science in Sport (IACSS2013)*. Oral presentation.

[5] I. Heazlewood and G. Lackey. The use of mathematical models to predict elite athletic performance at the olympic games. In deMestre, editor, *Proc. Third Conference on Mathematics and Computers in Sport*, pages 185–206, Amsterdam, 1996. Bond University.

[6] I. Heazlewood and J. Walsh. Mathematical models that predict athletic performance in the mens throwing events at the olympic games. In A. Bedford and T. Heazlewood, editors, *Proceedings of the 12th Australasian Conference on Mathematics and Computers in Sport*, pages 48–53, Darwin, Northern Territory, Australia, June 2014. MathSport (ANZIAM).

[7] I. Heazlewood and J. Walsh. Mathematical models that predict athletic performance in the womens throwing events at the olympic games. In A. Bedford and T. Heazlewood, editors, *Proceedings of the 12th Australasian Conference on Mathematics and Computers in Sport*, pages 54–60, Darwin, Northern Territory, Australia, June 2014. MathSport (ANZIAM).

[8] I.T. Heazlewood and J. Walsh. Mathematical models that predict 2012 london olympic games swimming and athletic performances. In *Proc. 3rd International Conference on Mathematics in Sport*, pages 71–78, Salford, UK, 2011.

[9] I.T. Heazlewood and J. Walsh. Mathematical models that predict athletic performance in sprint and jump events for current and future world champions. In *Proc. 3rd International Conference on Mathematics in Sport*, pages 64–70, Salford, UK, 2011.

[10] T. Heazlewood. Prediction versus reality: The use of mathematical models to predict elite performance in swimming and athletics at the Olympic Games. *Journal of Sports Science and Medicine*, 5:541–547, 2006.

[11] M. Norusis. *SPSS for Windows: Base Systems Users Guide Releases 6.0*. SPSS Inc, 1993.

[12] International Association of Athletic Federations. Qualification system and entry standards. 2015 iaaf world championships, beijing, china. [online], 2015. `https://iaafmedia.s3.amazonaws.com/competitioninfo/544c7d0a-0807-4226-a85b-79928c58a097.pdf`.

Measuring Efficiency of a Set of Players of a Soccer Team and Differentiating Players' Performances by their Reference Frequency

Nobuyoshi Hirotsu[1] and Tohru Ueda[2]

[1] Juntendo University, Inzai, Chiba, Japan
nhirotsu@juntendo.ac.jp
[2] Seikei University, Musashino, Tokyo, Japan
ueda@st.seikei.ac.jp

Abstract

In this paper, we try to evaluate performance of a soccer player from the aspect of importance as a team member. For this purpose, we evaluate a set of 10 field players of the team by introducing a link between teams and players into the model, and rank the individual players by their reference frequency. We illustrate our method using annual data of a J-league team in the 2013 season.

1 Introduction

Data envelopment analysis (DEA) is one of the practical methods for evaluating various sport activities, and has been used mainly for evaluating performances of teams and players separately. In terms of soccer, for the evaluation of teams the management aspect has been analyzed by using such input and output measures as expenditure, revenues, spectator attendance and salaries [1]. The technical aspect of teams has also been analyzed by using such data as the number of goals and passes [4]. For the evaluation of individual players, on the other hand, Hirotsu et al. [2] evaluate the performance of J-league players in order to identify their characteristics by position using data of the 2008 season. They use the CCR model with playing time as input and the numbers of ten basic plays or actions such as goals, passes and fouls as outputs. Santin [5] ranks the Real Madrid players by position using data from the 1946/47 to 2009/10 season. He uses the super-efficiency model with such data as the number of seasons each player played for Real Madrid's first team and the number of official games played. Tiedemann et al.[6] assess the performance of Bundesliga players using data from the 2002/03 to 2008/09 season. They use a metafrontier approach with playing time, the number of goals and assists, tackle ratio (%) and pass completion ratio (%). They show a positive relationship between a team's average player efficiency score and its rank in the league table at the end of the season.

As seen from the above, teams and players have separately been evaluated in most of the literature, and there has been little research from the aspect of importance of a player as a team member, by connecting the performance of the individual players with that of the team, except Tiedemann et al. [6] who analyze the relationship between them. We here try to extend their analysis considerably by introducing a link between teams and players into the DEA model. That is, not just taking an average of the player efficiency scores in a team, we evaluate a set of players of the team by introducing a link into the model, and rank the individual players as a team member by their reference frequency. We illustrate our method using annual player data such as the number of goals and passes of a J-league team in the 2013 season.

2 The Data

For illustrating our method, we use the data of Shimizu S-pulse of J-league. The data are provided by Data Stadium Inc. We picked up 16 field players who played more than 450 minutes in the 2013 season. In general, soccer players are quantitatively evaluated using the frequency of plays or actions in a game. Opta Index Limited [3] gives the ten basic plays or actions: goals, assists, passes, crosses, dribbles, tackles, interceptions, clears, blocks and fouls as measures of evaluation of players, and Hirotsu et al. [2] use them for the evaluation of characteristics of players. In this paper, we use the eight measures except assists and interceptions because their frequency may not be large enough in evaluation of importance of individual players.

Although there is a variety of team concepts for conducting a game, we could say in general that most teams want to increase the number of goals and the relating offensive measures such as crosses, passes and dribbles by keeping a level of defense which is related to the measures such as fouls, tackles, clears and blocks. As playing time will be naturally considered as input, we here define the five input measures (playing time, the numbers of tackles, clears, blocks and fouls), and the four output measures (the numbers of goals, passes, crosses and dribbles). Following this concept, we use output-oriented model for increasing the output values under the condition of keeping the level of the input values.

Table 1 represents the input and output values of the Shimizu S-pulse players. Seeing from the data, we can easily find that, for example, G.OMAE scored the highest number of goals, and his contribution to the team looks significant. T.TAKAGI also made a contribution in crossing. However, by considering their activities from the standpoint of the frequency per 90-minutes, G.OMAE still keeps the highest rate of goals ($0.52 = 7/1207 \times 90$), but the rate of crosses of T.TAKAGI ($1.19 = 22/1670 \times 90$) is less than that of K.MURATA ($1.61 = 9/904 \times 90$). So, there is a variety of aspect to evaluate the individual players. Later in this paper, we try to find other different aspects in evaluation of players using DEA models.

Player	Pos.	Input					Output				CCR	No.	Lambda Value			
		Time	Foul	Tackel	Clear	Block	Goal	Cross	Pass	Dribble	Eff.	50	Model1	Model2	Model3	Model4
T.MURAMATSU	DF	2942	32	77	38	50	4	2	822	2	0.807	1		0.56	1	1
Y.HIRAOKA	DF	2670	24	33	164	77	5	0	1143	1	1	1	1.10	1.94	1	1
K.SUGIYAMA	MF	2602	34	58	75	59	1	0	850	1	0.750	1			1	1
Y.KAWAI	MF	2543	25	37	28	50	1	13	767	13	0.929	1		2.55	1	1
CALVIN J.A.P	DF	2340	48	44	122	37	1	4	892	10	0.981	1		0.91	1	1
Y.YOSHIDA	DF	2155	35	46	51	41	1	20	709	14	0.988	1			1	1
H.ISHIGE	MF	2131	20	37	30	57	2	9	520	20	0.725	1			0.71	
T.TAKAGI	FW	1670	16	14	2	23	6	22	310	38	1			1.20	0.88	1
R.TAKEUCHI	MF	1453	13	19	28	26	0	5	642	7	1		9.67	2.84	1	1
BARE	FW	1436	34	10	16	9	4	9	171	22	1	1				
LEE Kije	DF	1418	19	19	57	33	1	6	415	12	0.748	1			0.27	
T.HONDA	MF	1335	29	44	23	27	0	4	479	3	0.869	1			1	1
S.ITO	FW	1289	15	11	21	22	6	1	236	8	1					
Radoncic	FW	1257	37	7	29	8	6	3	174	13	1					
G.OMAE	FW	1207	18	7	7	25	7	2	285	11	1		2.33		0.14	1
K.MURATA	MF	504	8	9	3	13	1	9	124	22	1		3.53			
Sum. No.		50	21572	300	405	604	440	20	67	6768	98		10	16.63	10	10	10
Ref.Point of Model		1	21572	222	268.0	477.5	440.0	25.3	84.8	8564.7	172.1					
		2	21572	228	312.4	604.0	440.0	22.6	78.4	7642.1	110.7					
		3	21572	276	402.6	568.3	440.0	21.0	75.7	7097.1	103.4					
		4	20917	274	379.0	538.0	415.0	26.0	72.0	6899.0	100.0					

Table 1: Data of Shimizu S-pulse players and the result of calculation

3 Evaluation of individual players

We first begin with the output-oriented CCR model for evaluating an individual player o by comparing n players of his team:

[CCR]

$$max \qquad \phi \tag{1}$$

$s.t.$

$$\sum_{j=1}^{n} x_{ij}\lambda_j \quad \leq \quad x_{io} \quad (i = 1, \cdots, 5) \tag{2}$$

$$\sum_{j=1}^{n} y_{rj}\lambda_j \quad \geq \quad \phi y_{ro} \quad (r = 1, \cdots, 4) \tag{3}$$

where λ_j is a non-negative variable, and the efficiency of player o is given by $1/\phi$. Efficiency scores in the CCR model are shown in Table 1, and the eight players become efficient. They would somehow make a good performance, but this is just relative comparison between the players without any connection with a set of players. Thus, the evaluation of the players by the CCR model may not be reasonable enough for evaluating an importance of a player as a team member. Further, as it will be better to differentiate the efficient players, we here extend the CCR model to evaluate the aggregation of players as a set.

4 Evaluation of a set of players

In order to evaluate a set of players in the team, we connect the performance of individual players and that of the team by summing up each player's input and output measures, as a first step of this research. As the aggregation of players, we here consider a set of 10 field players. Let F be the different set of 10 field players, and the number of element of F is ${}_nC_{10}$. In order to measure efficiency of the set, we first propose the following model:

[Model 1]

$$max \qquad \phi \tag{4}$$

$s.t.$

$$\sum_{j=1}^{n} x_{ij}\lambda_j \quad \leq \quad \sum_{o\in F}^{10} x_{io} \quad (i = 1, \cdots, 5) \tag{5}$$

$$\sum_{j=1}^{n} y_{rj}\lambda_j \quad \geq \quad \phi \sum_{o\in F}^{10} y_{ro} \quad (r = 1, \cdots, 4) \tag{6}$$

In the above expression, the summation of the right hand side is applied to player o (there are 10 players) belonging to F. The efficiency scores as a set of players using the above models are shown in Table 2. The number of different sets of 10 field players chosen from the 16 players is ${}_{16}C_{10}$ =8008. Table 2 represents just a part of the 8008 different sets. Looking at a set of players named "No.50", for example, this set consists of 10 field players who are indicated by "1"s in the row of "No.50". These 10 players are also indicated by "1"s in the column of

No.	Player																	Efficiency			
	T.M	Y.H	K.S	Y.K	CJAP	Y.Y	H.I	T.T	R.T	BARE	LEE	T.H	S.I	Rad.	G.O	K.M	Model1	Model2	Model3	Model4
1	1	1	1	1	1	1	1	1	1	1							0.820	0.935	1	1
2	1	1	1	1	1	1	1	1	1		1						0.825	0.931	1	1
3	1	1	1	1	1	1	1	1	1			1					0.830	0.941	1	1
4	1	1	1	1	1	1	1	1	1				1				0.829	0.931	0.9997	1
5	1	1	1	1	1	1	1	1	1					1			0.829	0.933	0.9986	1
6	1	1	1	1	1	1	1	1	1						1		0.841	0.943	1	1
7	1	1	1	1	1	1	1	1	1							1	0.835	0.934	1	1
8	1	1	1	1	1	1	1	1		1	1						0.803	0.905	0.9976	1
9	1	1	1	1	1	1	1	1		1		1					0.807	0.915	0.9839	1
50	1	1	1	1	1	1	1			1	1	1					0.790	0.886	0.954	0.981
8005						1	1	1		1	1	1	1	1	1	1	0.822	0.824	1	1
8006						1	1		1	1	1	1	1	1	1	1	0.820	0.822	0.9755	1
8007						1		1	1	1	1	1	1	1	1	1	0.877	0.882	1	1
8008							1	1	1	1	1	1	1	1	1	1	0.837	0.841	1	1
															Ratio	of eff.	0	0	0.119	0.726
																Ave.	0.827	0.873	0.969	0.994
																Max.	0.922	0.959	1	1
																Min.	0.737	0.778	0.874	0.915

Table 2: Efficiency scores of a set of players based on Models 1, 2, 3 and 4

"No.50" in Table 1, and the sum of input or output values in terms of No.50 are shown in the row of "Sum. No.50" such that Time and Goal are 21572 and 20, respectively. The efficiency of No.50 is 0.790 under Model 1, as shown in Table 2.

This efficiency is determined by the sum of the product of each player's frequency and lambda value. The lambda values for each player of No.50 are shown in the column of "Model 1" in Table 1. For example, in term of Goal, we can find that a set of virtual players, consisting of Y.HIRAOKA, R.TAKEUCHI, G.OMAE and K.MURATA with their weighted value (lambda), will achieve 25.3 (=5×1.10+0×9.67+7×2.33+1×3.53). As Goal of No.50 is 20, the efficiency is 0.790 (=20/25.3). In other words, No.50 is inferior in the sense that keeping the sum of each input values, it gets less output values than that of the set of virtual players (i.e. "reference set"). In this case, the efficiency of No.50 is also determined by Cross and Pass, such that 67/84.8 and 6768/8564.7, respectively. Summary statistics of the efficiency of the 8008 sets are also shown in Table 2. We can see that in terms of Model 1, the ratio of the efficient sets of 10 players in the 8008 sets is zero. The average of these efficiency scores is 0.827.

Here, as shown in Table 1, the lambda values of each player in the reference set seems to be unreasonable. That is, the sum of the lambda values is 16.63. This is interpreted that "No.50" is compared to the set of players which consists of more than 10 field players on the pitch. Thus, we add the following constraint (i.e.10 field players are on the pitch) to Model 1, and we get

[Model 2] (Model 1 and the following constraint)

$$\sum_{j=1}^{n} \lambda_j = 10. \tag{7}$$

The calculation result under Model 2 is shown in Table 2. In this case, the efficiency of No.50 increased to 0.886. Looking into the lambda values in Table 1, in term of Goal, we can find that the set of virtual players, T.MURAMATSU, Y.HIRAOKA, Y.KAWAI, CALVIN J.A.P., T.TAKAGI and R.TAKEUCHI with their lambda value, will get 22.6 (=4×0.56+5×1.94+1×2.55+1×0.91+6×1.20+0×2.84). As Goal of No.50 is 20, the efficiency is 0.886 (=20/22.6). This efficiency is also determined by Pass and Dribble, such that 6768/7642.1

and 98/110.7, respectively. As shown in Table 2, in terms of Model 2 the average of the efficiency scores increase to 0.873, but there are no efficient sets.

Although the sum of the lambda values equals to 10 in Model 2, there still seems to be a problem with regard to the lambda values such that R.TAKEUCHI's λ is 2.84. This is interpreted that a virtual player who can achieve 2.84 times of R.TAKEUCH's input and output values is on the pitch in the calculation of the efficiency of No.50. As this will be unreasonable, we further add the following constraint to Model 2, and we get

[Model 3] (Model 2 and the following constraint)

$$0 \leq \lambda_j \leq 1 \quad (j = 1, \cdots, n). \tag{8}$$

This is interpreted that no more than each player himself are on the pitch in the calculation. This will be realized by arranging the playing time of each player by substitution in a game. The calculation result under Model 3 has been shown in Table 2. In this case, the efficiency of No.50 increased to 0.954. Looking into the lambda values, in term of Goal, we can find that the set of virtual players, T.MURAMATSU, Y.HIRAOKA, K.SUGIYAMA, Y.KAWAI, CALVIN J.A.P., Y.YOSHIDA, H.ISHIGE, T.TAKAGI, R.TAKEUCHI, LEE Kije, T.HONDA and G.OMAE with their lambda value, will get 21.0(=4×1+5×1+1×1+1×1+1×1+1×1+1×1+2×0.71+6×0.88+0 ×1+1×0.27+0×1+7×0.14), and the efficiency is 0.954 (=20/21.0). This efficiency is also determined by Pass such that 6768/7097.1. In terms of Model 3, there are 952 efficient sets (i.e. 0.119 =952/8008 as shown in Table 2). The average of the efficiency scores increases to 0.969. In Model 3, 12 players are selected as the reference set of "No.50", and the lambda values of H.ISHIGE, T.TAKAGI, LEE Kije and G.OMAE are less than one, which will be realized by arranging their playing time.

In order not to arrange their playing time, we can introduce the following binary constraint instead of (8) to Model 3. That is, each player will be judged to be on the pitch or not throughout the game:

[Model 4] (Model 2 and the following constraint)

$$\lambda_j \in \{0, 1\} \quad (j = 1, \cdots, n). \tag{9}$$

In this case, the efficiency of No.50 increased to 0.981 as shown in Table 2. Looking into the lambda values, in term of Pass, we can find that the set of players, T.MURAMATSU, Y.HIRAOKA, K.SUGIYAMA, Y.KAWAI, CALVIN J.A.P., Y.YOSHIDA, T.TAKAGI, R.TAKEUCHI, T.HONDA and G.OMAE, will get 6899.0 (=822+1143+850+767+892+709+310+642+ 479+285). As Pass of No.50 is 6468, the efficiency is 0.981 (=6468/6899). In this case, we could say that this reference set dominates No.50 with the same implication of the FDH model. The ratio of the efficient sets is 0.726 (=5814/8008) and the average of the efficiency scores is 0.994.

5 Rank players from the aspect of importance to the team

In evaluation of an importance of a player as a team member, the frequency appearing in the reference set will be one of the reasonable measures. Table 3 shows their reference frequency under the CCR model and Models 3 and 4. As shown in Table 3, R.TAKEUCH is referred most frequently, appearing in 7877 reference sets out of the 8008 sets (98%) in Model 3. T.TAKAGI is the second in Model 3 and the top in Model 4. That is, R.TAKEUCH and T.TAKAGI seem to be very important as a team member of Shimizu S-pulse in the 2013 season. In fact R.TAKEUCH and T.TAKEGI are also efficient under the CCR model, but it will be difficult to differentiate the efficient players and rank all players without introducing the aspect of the importance as a team member by means of such models as Models 3 and 4.

Player	Position	CCR	Model3				Model4		
		Eff.	Ref.Freq	Ref.Freq.	Ratio	Rank	Ref.Freq.	Ratio	Rank
R.TAKEUCHI	MF	1	9	7877	0.98	1	6601	0.82	2
T.TAKAGI	FW	1	7	7854	0.98	2	6669	0.83	1
G.OMAE	FW	1	4	7761	0.97	3	6150	0.77	3
Y.YOSHIDA	DF	0.988	0	7666	0.96	4	5562	0.69	5
Y.HIRAOKA	DF	1	4	7564	0.94	5	5363	0.67	6
Y.KAWAI	MF	0.929	0	7514	0.94	6	5832	0.73	4
CALVIN J.A.P.	DF	0.982	0	7116	0.89	7	5099	0.64	8
K.MURATA	MF	1	4	7107	0.89	8	5276	0.66	7
T.MURAMATSU	DF	0.807	0	6435	0.80	9	4802	0.60	9
S.ITO	FW	1	1	5667	0.71	10	4772	0.60	10
T.HONDA	MF	0.869	0	5520	0.69	11	4343	0.54	12
Radoncic	FW	1	3	5217	0.65	12	4526	0.57	11
BARE	FW	1	1	4683	0.58	13	3794	0.47	14
H.ISHIGE	MF	0.725	0	4101	0.51	14	3623	0.45	16
K.SUGIYAMA	MF	0.750	0	3010	0.38	15	3671	0.46	15
LEE Kije	DF	0.748	0	2826	0.35	16	3997	0.50	13

Table 3: Reference frequency

6 Conclusions

In this paper, we attempted to evaluate performance of a soccer player from the aspect of importance as a team member. We evaluated a set of 10 field players of the team by introducing a link between teams and players into the model, and ranked the individual players by their reference frequency. We illustrated our method using annual data of a J-league team in the 2013 season. We found that R.TAKEUCH and T.TAKEGI were very important players as team members and ranked all the 16 players. Although we summed up each player's input and output values in this paper, we could introduce other links into the model.

Acknowledgments

This work was partly supported by Grant-in-aid for Scientific Research (C) of Japan [#21510159] and [#26350434].

References

[1] D.J. Haas. Technical efficiency in the Major League Soccer. *Journal of Sports Economics*, 4:203–215, 2003.

[2] N. Hirotsu, H. Yoshii, Y. Aoba, and M. Yoshimura. An evaluation of characteristics of J-league players using Data Envelopment Analysis. *Football Science*, 9:1–13, 2012.

[3] Opta Index Limited. *The Opta Football Yearbook: 2000-2001*. Carlton Books, 2000.

[4] R. Sala-Garrido, V.L. Carrion, A.M. Esteve, and J.E. Bosca. Analysis and evolution of efficiency in the Spanish soccer league (2000/01 - 2007/08). *Journal of Quantitative Analysis in Sports*, 5(1), 2009.

[5] D. Santin. Measuring the technical efficiency of football legends: Who were Real Madrid's all-time most efficient players? *International Transactions in Operational Research*, 21(3):439–452, 2014.

[6] T. Tiedemann, T. Francksen, and U. Latacz-Lohmann. Assessing the performance of German Bundesliga football players: A non-parametric metafrontier approach. *Central European Journal of Operations Research*, 19(4):571–587, 2011.

Towards in-depth performance analysis for wheelchair rugby with the Australian Steelers

Jason Hunt[1], Nick Sanders[2], Anthony Bedford[1] and Stuart Morgan[3]

[1] RMIT University, Melbourne, Australia
jason.hunt@rmit.edu.au
[2] Australian Paralympic Committee, Melbourne, Australia
[3] Australian Sports Commission, Canberra, Australia

Abstract

This article details the processes which the Australian wheelchair rugby (Steelers) are developing for game performance analysis. In conjunction with RMIT and the Australian Institute of Sport, the Steelers are building the most in depth and detailed game analysis system ever used in the sport. This includes live statistical and video capture as well as post-game sweeps in an effort to create new and insightful metrics for measuring performance in wheelchair rugby. Techniques include qualitative and quantitative statistical comparisons, regression and spatial analysis. Through a combination of technologies (including Sportscode and AIS tracking software), every facet of a game is being scrutinised and analysed in great detail. The processes take inspiration from sports that lead the way in performance analysis such as NBA basketball and football (soccer).

1 Introduction

Wheelchair rugby (formerly Murderball) is a ball sport played on a court with the dimensions of a basketball court (15 metres by 28 metres) where two teams of four players each work together to take a ball from one end of the court to the other. The aim is to carry the ball across the line at the end of the court between the two cones which denote the goal area; the team to achieve the most goals across the match wins. A match is comprised of four quarters of eight minutes each, where the clock only runs whilst the ball is in play.

Wheelchair rugby being is a Paralympic sport and is bound by guidelines which dictate who is allowed to compete in official competition; as such each player who wishes to participate goes through a process which awards them with a classification between 0.5 and 3.5 at intervals of 0.5. The classification procedure is complex, but athletes must have some form of disability which limits the function of both upper and lower limbs. An evaluation of the degree to which an athlete is impaired is then determined by a medical professional and a point value is allocated to that player. The higher-functioning an athlete is, the higher point value he or she will be allocated. The majority of athletes have spinal cord injuries however there are players with neurological disorders and multiple amputations as well.

The total classification points allowed on the court for each team at any time is 8. Since males and females compete in the same arena, for each female on the court a team is allowed an extra 0.5 classification points. A full list of rules and regulations can be found at the International Wheelchair Rugby Federation website[4]. Due in part to its infancy as a sport (wheelchair rugby was first recognised as an official sport in 1993) and in part to its lack of wide exposure, in-depth performance analysis is scarce. There is limited research which explores the physical demands of the sport[5, 8]. The research of Sarro et al. examined one match of wheelchair rugby and used this to quantify the distances which players cover within a match and at what speeds the players typically travel. The information was also used to establish the differences

in demands for higher and lower functioning athletes. There also exists significant research in the area of acceleration/speed monitoring [7, 3, 2]. However research that relates directly to in-game performance analysis does not yet exist for wheelchair rugby which is the motivation for this paper. Analytics specifically incorporating player tracking and in-game statistics are the driving force behind this research, which is taking place in collaboration with the Steelers, Australias national wheelchair rugby team. Although in its early stages, the bulk of set-up work has been done along with some very preliminary results.

2 Methodology

To fully evaluate a game of wheelchair rugby, all aspects of a game should be investigated including team and player-specific metrics. Conventional statistics that are monitored for wheelchair rugby are often limited simply to goal-scorers and penalty recipients (who are then sent to the penalty box, usually until the next goal for the opposition is scored). This not only misses a considerable amount of game information, but it also focusses on the higher functioning players as they typically do most of the scoring and penalty work. For this reason, a more holistic approach had to be taken when creating an infrastructure which encompasses as much of a match as possible.
There are two main streams of technology which create the data required for analysis -

Sportscode Software developed by 'Sportstec' designed specifically for accumulating, databasing and displaying quantitative information in sport.

AIS tracking software Technology that semi-automatically tracks each player's position across a court or field of play.

Two related systems were built within the framework of Sportscode. First, a live coding framework which is possible to use during a game in real time (usually court-side or from a vantage point) which is used to provide coaching staff with live feedback on what is happening within the match. This includes lists of who is scoring goals and being sent to the penalty box for each team, as well as select time-based performance metrics. These vary depending on the coaching focus at the time, but an example of this is 'time to centre' which tracks the time it takes each side to cross the half way line once they receive the ball from an inbound after a goal. Each team must complete this action within 12 seconds, so the motivation for measuring this is to measure the ease with which each team is achieving (or not achieving) this aim. To work in conjunction with the live coding framework, a post game sweep was also built. This system is designed to track as many aspects of the match as possible. With the real-time aspect now removed from the equation, it is possible to keep track of a much larger number of factors within the match by slowing down the playback and replaying aspects of the match as well. This makes it possible to track actions such as individual passes and label metrics in more detail such as describing cause of turnovers.
The player tracking software was developed at the Australian Institute of Sport and it's role is to semi-automatically track the X-Y coordinates of each player at 25 frames per second. Matches were recorded using a network of three or four stationary HD cameras at vantage points which collectively covered the entire 28m by 15m playing area. After recording the match, the video files are synchronised to start at the exact same frame. Following that each camera angle is calibrated so that the vision is projected onto a virtual court. A state-of-the-art player detector [1] generated a set of player detection observations where each observation consisted of an (x, y) ground location, a timestamp t, and a team affiliation estimate. This system has been used

previously to discover adversarial group behaviour in field hockey [6]. The software evaluates every frame from each of the camera angles and subsequently creates detection files which are then collated into one repository. Following that, manual intervention takes place and the automation takes a back seat. Tracklets must be labelled as specific players and joined together as the system often loses track of who is who in congestion which is common in wheelchair rugby. Erroneous detections must be corrected or removed and where the play is simply too congested to effectively detect individual players, manual annotating takes place. Although it is not possible to directly track the ball within this system due to its small size, players are marked to indicate the frames that they possessed the ball. The procedure is then able to produce a text file with X-Y coordinates and frame numbers for every player across an entire match.
Finally, these two sources of data need to be brought into the same context by overlaying the X-Y-t data with the descriptive information provided by Sportscode. This final step provides coaching staff and analysts with the most in-depth and objective evaluation of a match to this point. It will be innovated and developed upon to provide statistical insights into wheelchair rugby that were previously not possible.

3 Results and discussion

The results of this paper are limited mostly to building the infrastructure for the analytical system and examining the results it may be able to obtain in the future.

3.1 Sportscode - Live coding

The live coding window has gone through many iterations and changes dependant on the coaching needs at the time. An example picture of it can be seen in Figure 1.

This code window tracks goals, penalties and turnovers for each player on the court as well as the number of time outs that each team has called. Player time outs are often used in wheelchair rugby to avoid an imminent turnover. An example of this could be that a player in possession is in the back court and about to run out of time to bring the ball into the front court. By calling a time out, that timer is reset and that team will receive another 12 seconds to take the ball into the front court.
This code window also tracks the times between some actions. Most simply, it tracks which team has possession at any given time and for how long overall during a match. Secondly the system tracks the time between a goal being scored and the inbound occurring afterwards (denoted 'time between goals' on the code window). It is hypothesised that this may be a measure of defensive pressure of the team who just scored a goal, since by not allowing the opposition to inbound quickly that team has more time to set up the desired defensive structure. Thirdly the system tracks the time it takes each side to reach the front court after every goal. This can be seen as a measure of defensive pressure for the defending side and a measure of how well the attacking side is flowing on offence.
After using the code window during a match, a time-line storing all of this information is created - this can be seen in Figure 2.

Each instance on the time-line contains codes and labels which pertain to actions that occurred within the time period each instance was open. These labels can then be tallied and summarised to be displayed for coaches, analysts and players both as the match is happening and post-match. This helps to inform their decision making during matches and contributes to player feedback in post-match review sessions.

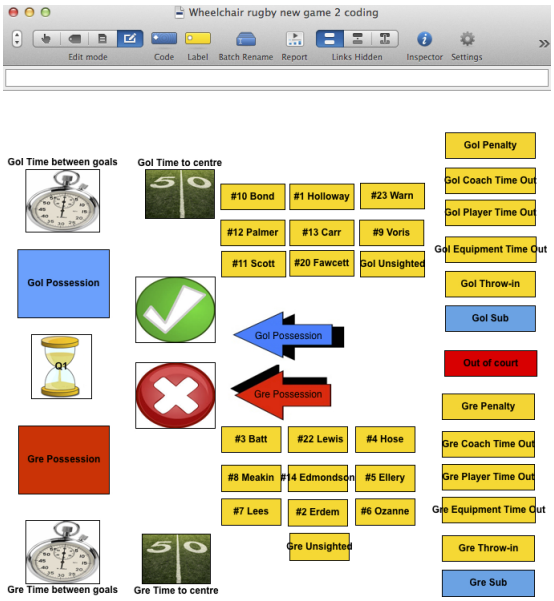


Figure 1: The live code window used for wheelchair rugby.

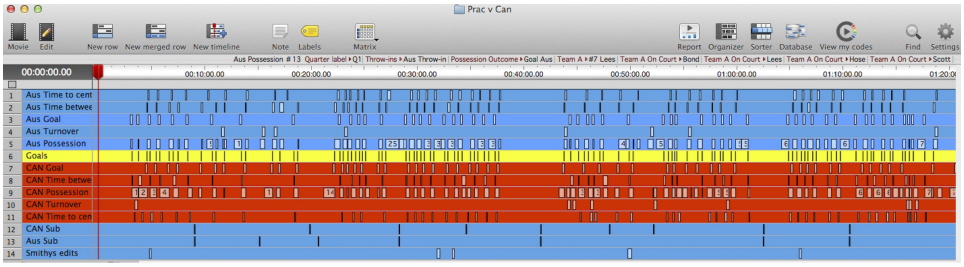


Figure 2: The time-line created by the live coding window.

3.2 Sportscode - Post game

After each match is completed, it is then possible to sweep through the match again, taking the time to log events in greater detail. A portion of the code window can be seen in Figure 3.

This window is used to code details such as when passes occur, the kind of pass that occurred, the pressure being asserted onto offensive player and where on the court the ball was passed to and from. Turnovers are also handled in greater detail as well as how players enter the front court and the nature of penalties that are committed. This creates an exhaustive database of events (which are described in detail by labels within in them) which specifies the vast majority of actions within a match. This information will be used (in conjunction with the positional data) to analyse the entirety of a match and evaluate positive and negative contributions - both at the player and team level. Over time a large database of matches will be accumulated (in 2014, the Australian team only played approximately 15 officially sanctioned matches across two tournaments) and this will allow inferences to be made about effective and ineffective structures that the Australian team plays.

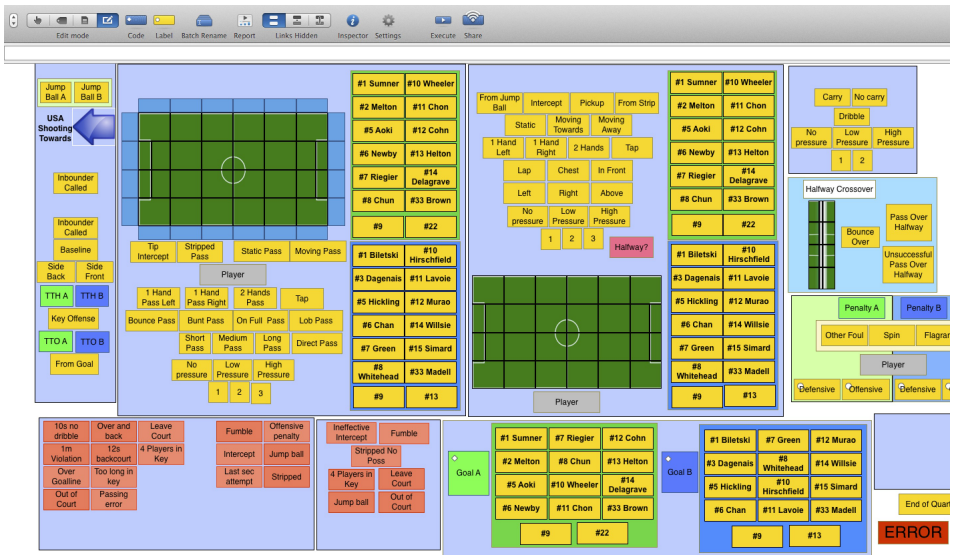


Figure 3: The post-match code window.

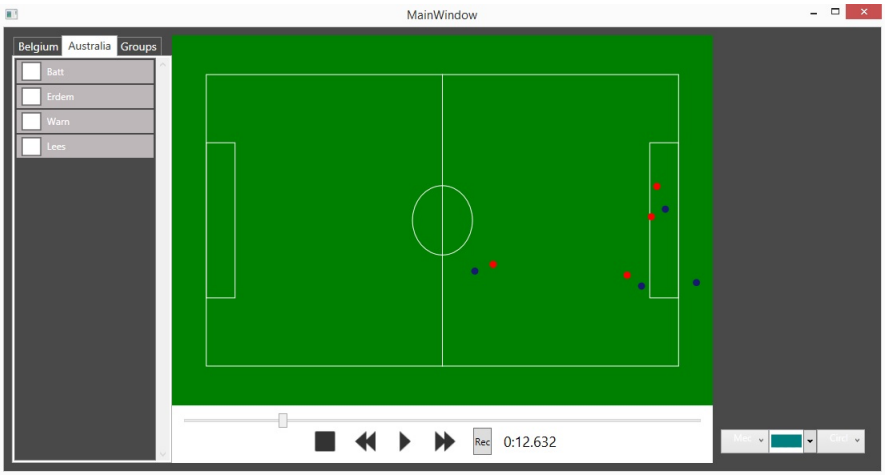


Figure 4: An example of the tracking data visualised.

3.3 Tracking solution

The X-Y-t data obtained from tracking solution can be visualised as shown in Figure 4.

The visualisation tool was coded using the .NET framework and it can be used to examine player movement patterns and evaluate the structural set up of both teams. It is possible to examine a specific player, a particular group of players or all players at once. There are methods built in which allow a coach or analyst to examine the area a group of players are 'covering', the speeds the players are travelling at and the total distance covered by each player at any given time.

3.4 Future work

This body of work is now a foundation from which to begin some analysis on spatial movement in wheelchair rugby and how it relates to in game performance. It also gives the opportunity to analyse the 'time to centre' metric, to see if pressuring teams in their defensive half is an effective tactic that will contribute towards winning more games. I also hope to analyse how good effective pressure on opposition all over the court can be created with effective spatial movement through the court.

4 Conclusion

A system capable of performing in-depth performance analysis for wheelchair rugby has been successfully developed. This will make it possible to evaluate actions and movement patterns with the game and validate new measures of performance. Preliminary testing of the system has been carried out and data collection is under way. This system can be used to assist coaches, analysts and players scrutinize wheelchair rugby both in-game and afterwards. Continued additions and improvement will also take place as coaches change their focus onto other areas of the game.

References

[1] P. Carr, Y. Sheikh, and I. Matthews. Monocular object detection using 3d geometric primitives. In *ECCV*, 2012.

[2] J.J.C. Chua, F.K. Fuss, V.V. Kulish, and A. Subic. Wheelchair rugby: fast activity and performance analysis. *Procedia Engineering*, 2(2):3077–3082, 2010.

[3] F.K. Fuss, A. Subic, and J.J.C. Chua. Analysis of wheelchair rugby accelerations with fractal dimensions. *Procedia Engineering*, 34:439–442, 2014.

[4] IWRF. Wheelchair rugby international rules 2015. `http://www.iwrf.com/resources/iwrf_docs/Wheelchair_Rugby_International_Rules_2015_English.pdf`. Accessed: 2015-02-06.

[5] K.J.Sarro, M.S. Misuta, B. Burkett, L.A. Malone, and R.N.L. Barros. Tracking of wheelchair rugby players in the 2008 demolition derby final. *Journal of Sports Sciences*, 28(2):193–200, 2010.

[6] P. Lucey, A. Bialkowski, P. Carr, S. Morgan, I. Matthews, and Y. Sheikh. Representing and discovering adversarial team behaviors using player roles. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pages 2706–2713, 2013.

[7] B. Mason, J. Lenton, J. Rhodes, R. Cooper, and V. Goosey-Tolfrey. Comparing the activity profiles of wheelchair rugby using a miniaturised data logger and radio-frequency tracking system. *BioMed Research International*, 2014.

[8] M.L. Sporner, G.G. Grindle, A. Kelleher, E.E. Teodorski, and R.A. Cooper R. Cooper. Quantification of activity during wheelchair basketball and rugby at the national veterans wheelchair games: A pilot study. *Prosthetics and Orthotics International*, 33(3):210–217, 2009.

A supervised PCA logistic regression model for predicting fatigue-related injuries using training GPS data

Stylianos Kampakis[1], Dr Ioannis Kosmidis[1], Wayne Diesel[2] and Ed Leng[2]

[1] University College London, London, United Kingdom
stylianos.kampakis@gmail.com,i.kosmidis@ucl.ac.uk
[2] Tottenham Hotspur F.C., London, United Kingdom
waynejdiesel@gmail.com , ed.leng@tottenhamhotspur.com

Abstract

The use of football tracking systems based on GPS technology has become popular over the recent few years in professional football. They allow the tracking of players during training, recording variables such as the number of sprints, the total distance covered or the number of impacts during training, which can subsequently be used as indicators of fatigue and overuse. Successful use of these indicators can help the professional predict fatigue-related injuries before they take place, and hence impact the well-being of the team both in terms of the players health and financially. A challenge for the practitioner is how to use the large numbers of variables produced by modern football tracking systems based on GPS technology in a constructive way.

This research uses a supervised principal component analysis logistic regression model in order to predict fatigue-related injuries. The available data consists of all injuries that took place at the first squad of the Tottenham Hotspur Football Club, and were filtered in order to keep only injuries related to fatigue. The task is to discriminate between weeks were an athlete got injured and weeks where he was not. The performance of the resulting logistic regression model was assessed through 10 rounds of 10-fold cross-validation. A Cohens kappa statistic of 0.3 illustrates that the GPS data can potentially form a useful source of information for the prediction of fatigue-related injuries. Furthermore, the principal components associated with the optimal model are extracted and interpreted in terms of what information each captures. This provides a way to reduce the large number of GPS variables (71 variables in total) to a smaller set that best relates to the occurrence of fatigue-related injury.

1 Introduction

Global positioning systems (GPS) are nowadays commonly used in many sports such as Australian rules football, hockey, rugby and cricket [2, 6]. The benefits of such as a system is the easy collection of variables that are very difficult (or impossible) to acquire otherwise. This includes, for example, the position of a player on the pitch, the acceleration, the deceleration and the highest velocity of a sprint.

From the perspective of injury prediction, a GPS system, used in training, can be particularly useful for getting a more accurate picture of the load of each individual athlete. A great deal of injuries take place as a result of overtraining and fatigue. It is not unreasonable to think that variables collected through a GPS can be used for detecting the onset of overtraining.

Indeed, similar researches have appeared in the last few years. For example, Aughey [1] used GPS data to track the fatigue of players in Australian football. [8] used GPS data to track the physical demands during games in the Australian Football League. [9] used GPS variables to detect muscle damage in Australian football players.

GPS has been used for tracking the physiological demands of rugby players [7]. Similarly, [5] used GPS to track the activities of rugby player during a match, suggesting that the GPS data, used in conjunction with video analysis, can help clarify the mechanism of many injuries.

In soccer, [4] used GPS data as indicators of training load. They discovered that there is a close relationship between the GPS-collected variables and the Rate of Perceived Exertion.

The notable absence of researches regarding GPS and injuries in the premier league is partly due to the banning of the use of GPS units inside the pitch. However, GPS data can be used in training. The performance data gathered from the training might provide good indicators of overuse and fatigue.

The purpose of this experiment was to build a predictive model for overuse and ovetraining injuries based solely on GPS data. It is clear that such a model could benefit the team as a whole and the players individually. Predicting overtraining and increased fatigue at the onset, and adjusting the load accordingly, could prevent a significant proportion of injuries.

The method used to analyze the data was supervised principal component analysis (supervised PCA). Supervised PCA [3] is a technique to produce principal components that are correlated with the response and can improve the performance of a supervised learning task and is particularly effective at dealing with datasets with highly correlated features and a high number of variables.

2 Data and methods

The data consisted of GPS records of 29 professional football players from THFC from the period July – October 2014 and a record of the days during which they were injured. The GPS system was used only during training and not during matches, due to the premier league regulations. The GPS system used was STATSports Viper.

There were in total 71 GPS variables along with the duration of the training session. The large number of variables comes as a result of using zones for measuring a single variable. For example, the following variables are used for describing acceleration: AccelerationsZone1, AccelerationsZone2, AccelerationsZone3, AccelerationsZone4, AccelerationsZone5, AccelerationsZone6.

Each zone describes a speed range relative to the maximum speed for each athlete. There are 6 zones in total. An example is shown at Table 1 below:

Zone	Z1	Z2	Z3	Z4	Z5	Z6
Boundary	<35%	35-45%	45-55%	55-65%	65-75%	>75%

Table 1: Example of different zones defined as regions of the maximum speed.

There were many correlations between the variables. Figure 1 shows a correlogram of the variables, where variables with high correlation have been clustered together. There are many clusters that can be identified throughout the graph. Note that the red lines on the edge of the graph are the names of the variables, which, due to their large number, cannot be displayed in a larger font.

First, all the players that had not been injured were removed. Secondly, out of the players that had been injured, only those that had intrinsic injuries were kept. This left a total of 11 players.

The dataset was then grouped per player and week, so that each variable corresponded to a player i and a week j. A binary 'injury' variable indicated whether the player had been injured

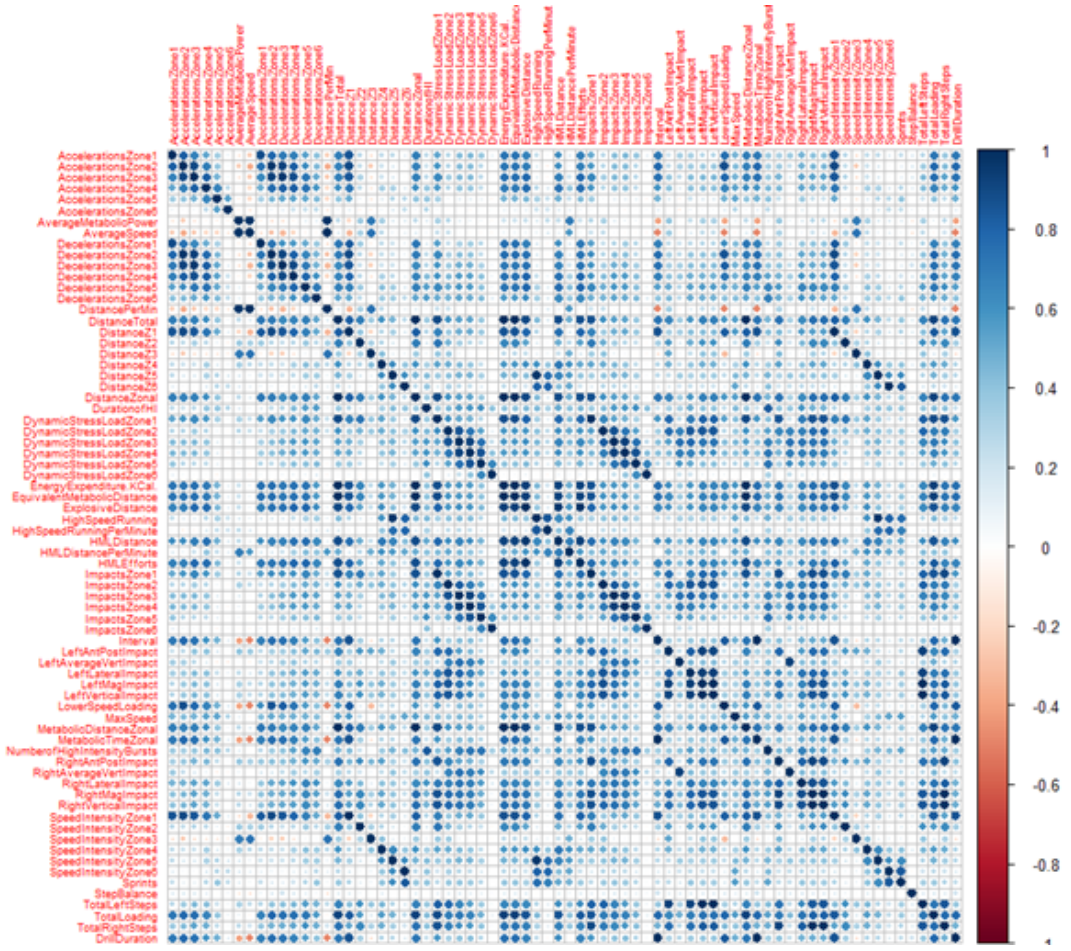


Figure 1: Correlogram for the GPS variables

at a week or not. A week is a pretty common way to break down time in football. Each week usually contains at least one match, during which an injury can take place. Also, the type and intensity of training during each week is affected by the upcoming game.

The mean was used as an aggregation function to aggregate the instances over the week. For each player that had been injured, there were at least 2 training sessions, before the injury took place.

The final dataset consisted of 206 rows and 71 GPS variables, plus the class variable and the duration of the training session. Out of the 206 rows, there were 11 cases of injury (corresponding to the weeks that the player had been injured).

While it can be disputed that the choice to remove all athletes that were not injured can lead to a loss of information, the problem as it stands contains highly unbalanced classes. Adding more athletes that were not injured, would just increase the imbalance, without adding more information for the injured cases.

The aggregation over weeks was chosen for two reasons. First, a week is a natural split

for scheduling in football. Every week is characterized by at least one match, and the training schedule is designed on a weekly basis. Secondly, other natural splits would be daily, or summing up weeks (e.g. the last two weeks). The problem with daily splits (that is, each row in the dataset represents a day of training) are twofold. First, the number of rows would become too large, but the number of injuries would still stay small. Secondly, it is unlikely that the GPS data from a single day are indicative of fatigue. It is more likely that an average over at least a few days is more informative.

In some cases there was more than one game taking place in a week. The participation of each player in games affects the duration of the training. For that purpose, the session duration was used alongside the GPS variables.

Therefore, the problem, as it is posed, is whether it is possible to predict a week during which an overtraining/overuse injury will occur based on GPS data, for professional football players that clearly suffered such injuries.

The data were analyzed using supervised PCA. Supervised PCA is different to regular PCA in that the components derived are correlated to the response variable.

The supervised PCA algorithm works as follows: 1. Fit a univariate regression model for each feature separately. 2. Remove all features with coefficients $|\beta| > \theta$, where θ is some threshold 3. Compute the principal components using the reduced dataset. 4. Keep m components and use them to fit the model.

The performance of the model was assessed through 10 rounds of 10-fold cross-validation.

3 Results

The model achieved an accuracy score of 88.8% (s.d. 1.8%) and a mean kappa statistic of 0.21% (s.d. 0.04%). Figure 2 shows a scree plot for the variance explained. What can clearly be seen is that the first four components explain the bulk of the variance.

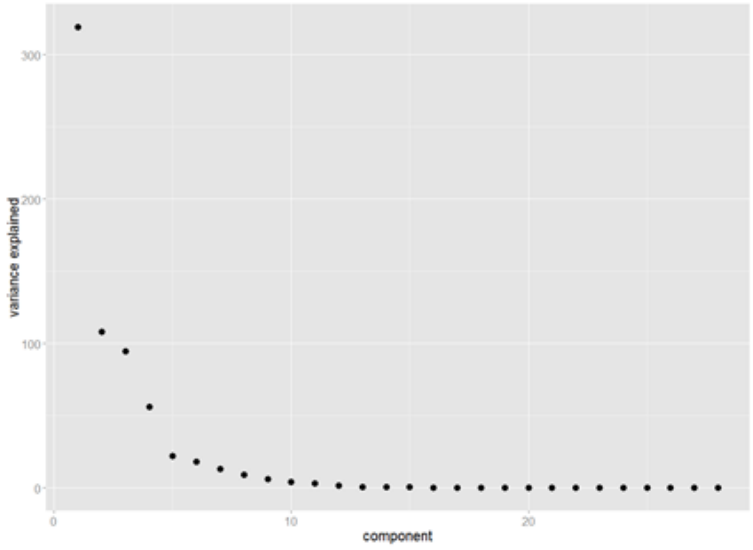


Figure 2: Variance explained for each component

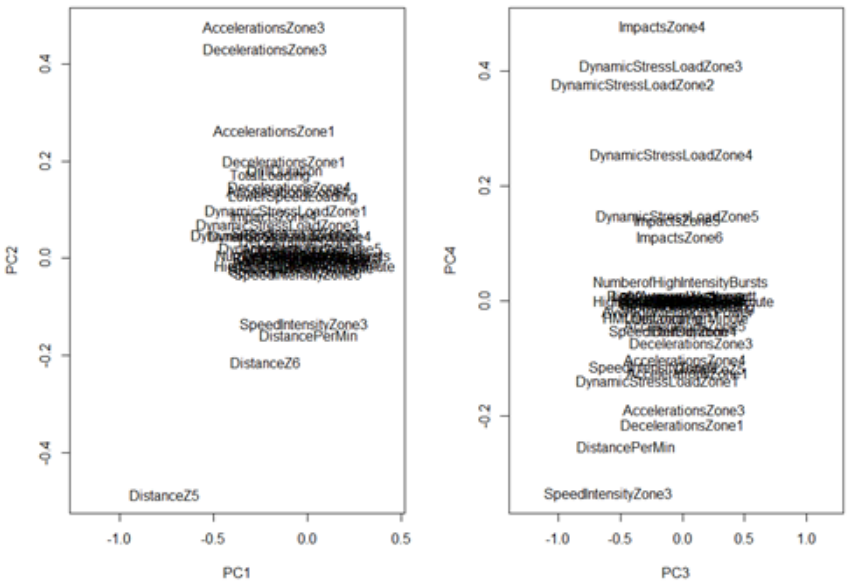


Figure 3: Biplot of the loadings of the first 4 principal components

Figure 3 shows a biplot of the components of the model while Table 2 shows the most important loadings for these components.

	PC1	PC2	PC3	PC4
AccelerationsZone3	-0.22842	0.47640	0.02004	-0.18828
DecelerationsZone3	-0.22305	0.43085	0.07311	-0.07318
DecelerationsZone4	-0.09469	0.14696	-0.00709	-0.00245
DistancePerMin	0.00317	-0.1908	0.43666	-0.19408
DistanceZ5	-0.76398	-0.48592	0.22479	-0.11721
DynamicStressLoadZone2	-0.18703	0.04578	-0.39766	0.37565
DynamicStressLoadZone3	-0.16017	0.06806	-0.17699	0.40768
ImpactsZone4	-0.17954	0.08479	-0.16280	0.47654
SpeedIntensityZone3	-0.02075	-0.13704	-0.59876	-0.33671

Table 2: Plot of the most important loadings for the first four principal components

Based on the table and the graphs the following interpretation can be made. The first component is dominated by a negative component in 'Distance Zone 5'. A larger value in this component indicates that the player covered a large distance running at high speeds. Therefore a good name for this component is 'Distance on High Speeds'.

The second component has a large negative loading on 'distance in zone 5' and a large positive loading for accelerations and decelerations in zone 3. A large value in this component seems to indicate the absence of running in higher speed zones and many sprints in zone 3.

A large negative value indicates the opposite. Therefore, a good name for this component is 'Medium Speed Sprinting'.

The third component has a high loading on 'distance per minute' and a low loading on 'dynamic stress load in zone 3' and 'speed intensity in zone 3'. It also has a moderately large value for 'distance in Z5'. A high value on this component seems to indicate that the player spends a lot of time in higher speed zones, and not in lower zones. Therefore, a good name for this component can be 'High Speed Session'. The fourth component has high loadings on variables that indicate physical stress in the middle speed zones (zones 2-4). Therefore, a good name for this component is 'Medium Intensity Work'.

Figure 4 below shows boxplots of the components having the response variable as the category. It seems that the first two components achieve good separation between the two categories, with the last two components less so.

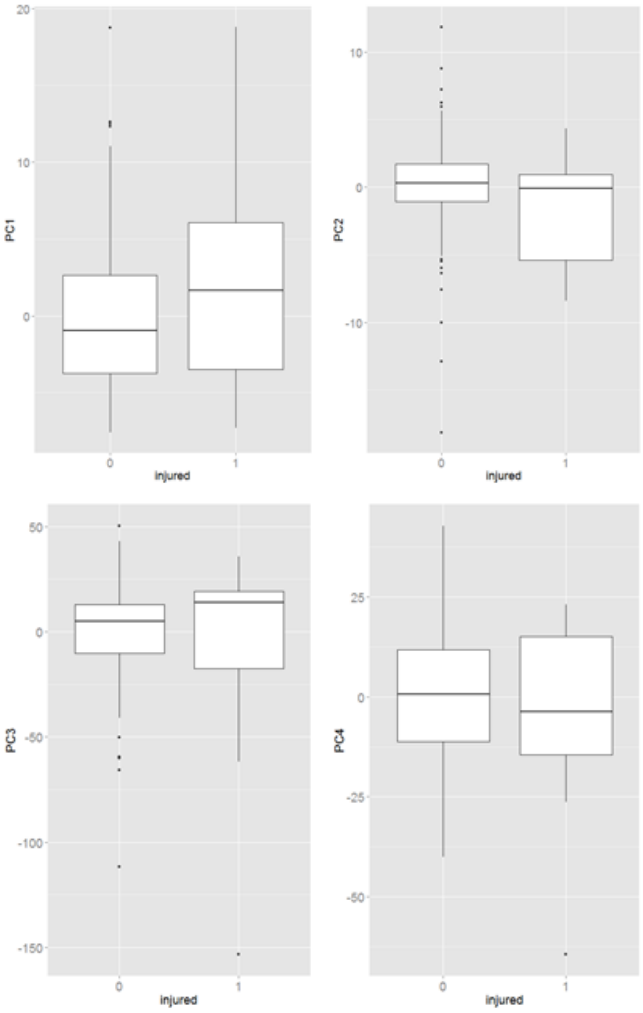


Figure 4: Boxplots for the principal components

4 Discussion and conclusions

This paper studied the problem of predicting fatigue injuries from GPS data gathered from training. The results demonstrated that training GPS data contain information which can be used for predicting injuries. A few points stand out.

First, an issue with this data is that the variables are highly correlated with each other. Any method applied to this problem needs to have a mechanism to deal with this problem. Supervised PCA was by far the best method out of all those that were compared, probably due to the fact that it handles correlated attributes very well.

Secondly, it is possible to reduce the large number of variables to a few meaningful factors that correlate with injury. The factors provide a more parsimonious explanation of the data and can be used more easily by a practitioner.

Obviously, the study suffers from some limitations. First, the total number of fatigue-related injuries in this dataset was small, relative to the number of non-injured instances. Many of the injuries take place during matches, or are not directly related to fatigue. For these cases, GPS data are not very informative. This is the case for traumatic injuries for example, which can come as a result of a single event (e.g. a very fast sprint), and prior sensitivity, which only a physical examination might be able to uncover.

Another issue with this task is the fact that GPS data are not collected from matches. However, matches have a huge impact on the physical condition of an athlete, which is left untracked. This probably poses an upper limit to the performance of the model.

References

[1] R. J. Aughey. Australian football player work rate: evidence of fatigue and pacing? *International Journal of Sports Physiology and Performance*, 5(3):394–405, 2010.

[2] R. J. Aughey. Applications of GPS technologies to field sports. *International Journal of Sports Physiology and Performance*, 6(3):295–310, 2011.

[3] E. Bair, T. Hastie, P. Debashis, and R. Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101:119–137, 2004.

[4] D. Casamichana, J. Castellano, J. Calleja-Gonzalez, J. San Román, and C. Castagna. Relationship between indicators of training load in soccer players. *Journal of Strength and Conditioning Research*, 27(2):369–374, 2013.

[5] G. F. Coughlan, B. S. Green, P. T. Pook, E. Toolan, and S. P. O'Connor. Physical game demands in elite rugby union: A global positioning system analysis and possible implications for rehabilitation. *Journal of Orthopaedic and Sports Physical Therapy*, 41(8):600–605, 2011.

[6] C. Cummins, R. Orr, H. O'Connor, and C. West. Global positioning systems (GPS) and microtechnology sensors in team sports: A systematic review. *Sports Medicine*, 43(10):1025–1042, 2013.

[7] B. Cunniffe, W. Proctor, J. S. Baker, and B. Davies. An evaluation of the physiological demands of elite rugby union using global positioning system tracking software. *Journal of Strength and Conditioning Research*, 23(4):1195–1203, 2009.

[8] B. Wisbey, P. G. Montgomery, D. B. Pyne, and B. Rattray. Quantifying movement demands of AFL football using GPS tracking. *Journal of Science and Medicine in Sport*, 13(5):531–536, 2010.

[9] W. B. Young, J. Hepner, and D. W. Robbins. Movement demands in Australian rules football as indicators of muscle damage. *Journal of Strength and Conditioning Research*, 26(2):492–496, 2012.

Construction methods for sports league schedules

Sigrid Knust

University of Osnabrück, Institute of Computer Science, Osnabrück, Germany
sigrid@informatik.uni-osnabrueck.de

Abstract

We give a survey on different methods to construct schedules for sports leagues. The basic problem is to find a schedule for a single or double round robin tournament in which every team plays against each other team exactly once (or twice), and every team plays one game per round. Additionally, several side constraints have to be respected, e.g. the avoidance of breaks (consecutive home or away games of a team), fairness issues (like opponent strengths, carry-over effects), the consideration of regions or the avoidance of waiting times for a tournament on a single court. After presenting some construction methods based on graph models, general solution approaches are discussed.

1 Introduction

Scheduling problems in sports leagues (cf. the annotated bibliography [10]) may be divided into two main classes: temporally-constrained and temporally-relaxed problems. In the **temporally-constrained** case, the planning horizon consists of the minimum number of periods (so-called rounds) required to schedule all the games and, hence, each team has to play exactly one game in each round. On the other hand, in the **time-relaxed** case the number of periods is generally larger than the minimum number of rounds needed for scheduling all games. In this situation not every team necessarily plays in each round and thus teams may have some periods without a game. While professional leagues often are temporally-constrained, for amateur leagues usually more flexibilities exist, i.e. they are often temporally-relaxed.

The basic temporally-constrained problem for scheduling a sports league may be formulated as follows. The league consists of an even number n of different teams $i = 1, \ldots, n$, where each team has to play against each other team exactly $\ell \geq 1$ times. The number of rounds available to schedule these $\binom{n}{2}\ell = n(n-1)\ell/2$ games is equal to $(n-1)\ell$, where each team has to play exactly one game in each round. Thus, one has to determine which teams $i, j \in \{1, \ldots, n\}$ play against each other in the rounds $t = 1, \ldots, (n-1)\ell$ and, for each of these pairings, whether it is played as a home game for team i or j.

In most cases we have $\ell = 1$ (single round robin tournament, SRRT) or $\ell = 2$ (double round robin tournament, DRRT). For DRRTs the season is often partitioned into two half series, where each pairing has to occur exactly once in each half (with different home rights).

If the league consists of an odd number of teams, in each round one team has a bye, i.e. does not play. This situation may be reduced to the previous case with an even number of teams by adding a dummy team $n+1$. Then, in each round the team playing against $n+1$ has a bye.

Usually, a schedule for a sports league is described by a so-called **opponent schedule** and a so-called **home-away pattern**. An opponent schedule may be represented by an $n \times (n-1)$-matrix where the entry $opp_{it} \in \{1, \ldots, n\} \setminus \{i\}$ specifies the opponent of team i in round t. A home-away pattern (HAP) is defined as $n \times (n-1)$-matrix $H = (h_{it})$, where h_{it} equals "H" (resp. "A") when team i has a home (resp. away) game in round t. If two consecutive entries $h_{i,t-1}$ and h_{it} in a row i are equal for some $t \in \{2, \ldots, n-1\}$, then team i has a so-called **break** in round t (i.e. the alternating sequence of home and away games is broken).

2 Graph-Based Models

Traditional models for constructing sports league schedules are based on graphs. The complete graph K_n on n nodes is used to represent single round robin tournaments. Its nodes $i = 1, \ldots, n$ model the teams, while the edge $i - j$ represents the game between teams $i, j \in \{1, \ldots, n\}$. An edge coloring with $n - 1$ colors, i.e. a partitioning of the edge set into 1-factors $F_1, \ldots, F_{n-1}$ (each consisting of $n/2$ non-adjacent edges), corresponds to the games scheduled in the rounds $t = 1, \ldots, n - 1$. If additionally home and away games have to be distinguished, the edges are directed into arcs (i, j) for a home game of team j or (j, i) for a home game of team i.

De Werra [9] considered the problem of determining the minimum number of breaks for a single round robin tournament with n teams and constructed corresponding break-minimal schedules. Since only two different patterns without break exist (HAHA ... H and AHAH ... A) and all teams must have different patterns (since two teams with the same pattern cannot play against each other), at most two teams do not have any break. Thus, in each feasible schedule at least $n-2$ breaks occur. It has been shown that this lower bound for the number of breaks can be achieved by constructing corresponding schedules. One class of such schedules is based on the so-called **canonical 1-factorization** $(F_1, \ldots, F_{n-1})$ of the graph K_n. In this factorization the factors F_i are defined by the edge sets

$$F_i = \{\{n, i\}\} \cup \{\{i + k, i - k\} | k = 1, \ldots, n/2 - 1\},$$

where the numbers $i + k$ and $i - k$ are taken modulo $n - 1$ as one of the numbers $1, \ldots, n - 1$. A corresponding oriented 1-factorization $(\vec{F}_1, \ldots, \vec{F}_{n-1})$ may be constructed by orienting the edges in $\vec{F}_i$ into

- (i, n) if i is odd, and (n, i) if i is even, and
- $(i + k, i - k)$ if k is odd, and $(i - k, i + k)$ if k is even.

An example with $n = 6$ teams is illustrated in Figure 1 where the corresponding schedule and home-away pattern are shown. In the schedule, the entry $+j$ in row i and column t means that in round t team i plays at home against j, the entry $-j$ means that the game takes place at j. The oriented 1-factors are given by $\vec{F}_1 = \{(1,6),(2,5),(4,3)\}$, $\vec{F}_2 = \{(6,2),(3,1),(5,4)\}$, $\vec{F}_3 = \{(3,6),(4,2),(1,5)\}$, $\vec{F}_4 = \{(6,4),(5,3),(2,1)\}$, $\vec{F}_5 = \{(5,6),(1,4),(3,2)\}$. In this schedule $n - 2 = 4$ breaks occur, they are underlined.

	round				
team	1	2	3	4	5
1	-6	$+3$	-5	$+2$	-4
2	-5	$+6$	$\underline{+4}$	-1	$+3$
3	$+4$	-1	$\underline{-6}$	$+5$	-2
4	-3	$+5$	-2	$+6$	$\underline{+1}$
5	$+2$	-4	$+1$	-3	$\underline{-6}$
6	$+1$	-2	$+3$	-4	$+5$

$$H = \begin{pmatrix} A & H & A & H & A \\ A & H & \underline{H} & A & H \\ H & A & \underline{A} & H & A \\ A & H & A & H & \underline{H} \\ H & A & H & A & \underline{A} \\ H & A & H & A & H \end{pmatrix}$$

Figure 1: Canonical 1-factorization for $n = 6$

The canonical 1-factorization has some additional structural properties. For example, for $i = 1, \ldots, n - 2$ the set $F_i \cup F_{i+1}$ forms a hamiltonian cycle and the $n - 2$ breaks occur in pairs on the odd-numbered days $b_i := 2i + 1$ $(i = 1, \ldots, n/2 - 1)$. For the reversed schedule

$(\overrightarrow{F}_{n-1}, \overrightarrow{F}_{2n-2}, \ldots, \overrightarrow{F}_1)$ breaks occur in pairs on the even-numbered days $2i$. Furthermore, it can be shown that for any odd number of teams the canonical 1-factorization leads to a schedule without any break.

If a league contains teams which share the same stadium, two such teams cannot play at home simultaneously. Thus, it is required that these teams have complementary home-away patterns (i.e., when one team plays at home, the other plays away). De Werra [7] has shown that in each oriented 1-factorization of K_n with $n-2$ breaks the teams can be partitioned into $n/2$ pairs having complementary patterns. For example, in the canonical 1-factorization teams 1 and n have complementary patterns without any break, teams $2i$ and $2i+1$ ($i = 1, \ldots, n/2-1$) have complementary patterns with exactly one break in round $b_i = 2i+1$.

In [9] additionally so-called **binary factorizations** were introduced, which have applications in leagues divided into subdivisions. In such a situation it is required that in certain rounds only teams belonging to the same division play against each other and that in the other rounds each team plays against a team belonging to the other division. In [6] it has been shown that for such a situation schedules with $n-2$ breaks exist, where the "internal" and "external" days alternate regularly. Later in [8] this result was generalized to leagues with more than two subdivisions.

For **double round robin tournaments** with two half series at least $2n-4$ breaks occur (cf. [7]). Unfortunately, schedules with $2n-4$ breaks have the often undesired property that the first round of the second half series contains the same pairings as the last round of the first half series, e.g. in a "mirrored" schedule $(\overrightarrow{F}_1, \ldots, \overrightarrow{F}_{n-1}, \overleftarrow{F}_{n-1}, \ldots, \overleftarrow{F}_1)$ (where $\overleftarrow{F}_i$ denotes the 1-factor which is obtained from $\overrightarrow{F}_i$ by orienting all arcs into the opposite direction). Often, schedules of the form $(\overrightarrow{F}_1, \ldots, \overrightarrow{F}_{n-1}, \overleftarrow{F}_1, \ldots, \overleftarrow{F}_{n-1})$, are preferred, where the pairings in rounds t and $n-1+t$ are the same. De Werra [7] proved that for such schedules at least $3n-6$ breaks occur and constructed corresponding schedules with $3n-6$ breaks by changing the orientation of some edges $n-i$ in the canonical 1-factorization. For $n > 4$ in the first half series all breaks occur in rounds $3, \ldots, n-2$, i.e. in the schedule no team has two consecutive breaks (which would correspond to three consecutive home or away games).

Fairness issues with respect to different opponent strenghts may be taken into account by partitioning the teams into $g \geq 2$ different strength groups. Then, the objective is to avoid that a team plays against extremely weak or extremely strong teams in consecutive rounds. Two concepts have been introduced in [3] assuring a certain degree of fairness with respect to such groups. A schedule for a SRRT is called **group-changing** if no team plays against teams of the same group in two consecutive games. It is called **group-balanced** if no team plays more than once against teams of the same group within g consecutive games. The second property implies the first one, since each group-balanced schedule (where each team plays against each group exactly once within $g \geq 2$ consecutive games) is also group-changing. In [3] and [4] for several special cases group-balanced or group-changing schedules were constructed based on graph models.

Another fairness-related concept deals with so-called **carry-over effects** (cf. [18]). For a given single round robin schedule with n teams, it is said that team i gives a carry-over effect to team j if some other team plays consecutively against teams i and j, i.e. some team plays against i in round t and against j in round $t+1$ for some $t \in \{1, \ldots, n-1\}$, where the rounds are considered cyclically (i.e. round n corresponds to round 1). If team i is a very strong or very weak team and several teams play consecutively against teams i and j, team j may be handicapped or favoured compared with other teams. In an "ideal" schedule with respect to carry-over effects, no teams i, j, u, v should exist such that teams u and v both play against

team j immediately after playing against team i. In [18] the carry-over effects $n \times n$-matrix $C = (c_{ij})$ has been introduced in order to measure the balance with respect to this criterion. Each entry c_{ij} of this matrix counts the number of carry-over effects given by team i to team j. The quality of a schedule with respect to carry-over effects is measured by the carry-over effects value $coe = \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}^2$. An ideal or balanced schedule is one in which the lower bound value $n(n-1)$ is achieved, i.e. all non-diagonal elements of the corresponding matrix C are equal to 1. In [18] it was shown that balanced schedules exist for all $n = 2^m$ by giving a construction scheme based on the Galois fields $GF(2^m)$. It was conjectured that balanced schedules exist if and only if n is a power of 2. Later on in [1] this conjecture was shown to be false by providing balanced schedules for $n = 20$ and $n = 22$.

Special requirements regarding the venues (courts, fields, stadiums, locations) have to be taken into account for some sports scheduling problems. In the so-called **balanced tournament problem** (cf. [14]), n teams (or players) have to play a single round robin tournament in $n-1$ rounds using $n/2$ different courts in each round. In order to balance the effects of the different courts, it is required that no team plays more than twice on any court (i.e. each team plays twice on $n/2-1$ courts and once on the remaining court).

Another sports scheduling problem dealing with special court restrictions is considered in [11]. The goal is to construct a schedule for an odd number n of teams and $k = \frac{n-1}{2}$ rounds where each team plays exactly two games in each round. Furthermore, all n games of a round have to be played consecutively on a single court. The objective is to minimize waiting times for the teams. This problem can again be modeled by the complete graph K_n where the nodes $1, \ldots, n$ model the teams and the edge $i - j$ corresponds to the game between teams i and j. Since in every round each team has to play twice, each round corresponds to a 2-factor in K_n (i.e. a collection of node-disjoint cycles). Moreover, since the games in the rounds have to be different, all 2-factors have to be disjoint, i.e. they constitute a 2-factorization $F = (F_1, \ldots, F_k)$. Furthermore, the assignment of the games to the n slots may be modeled by assigning the slot numbers $1, \ldots, n$ to the n edges for each 2-factor. Based on special 2-factorizations (hamiltonian cycles, solutions of certain Oberwolfach problems), for two variants (zero waiting times allowed or not) schedules for each odd number of teams are constructed minimizing the number of long waiting times and the total waiting time simultaneously.

Several applets visualizing the construction methods mentioned above can be found at the website `http://www2.inf.uos.de/knust/sportssched/webapp/`.

3 General Solution Approaches

While the graph-based construction methods introduced in the previous section mainly focus on one special side constraint (e.g., minimizing the number or breaks or the consideration of opponent strengths or carry-over effects), for sports scheduling problems in practice often much more side constraints have to be respected and the problems become more complex. For this reason, such problems are often decomposed into subproblems which are solved sequentially by exact or heuristic algorithms. Mainly, three approaches can be distinguished:

1. "**First-schedule, then-break**": At first only the pairings are determined for each round (i.e. which teams play against each other in this round). Afterwards for this opponent schedule a corresponding home-away pattern is calculated.
2. "**First-break, then-schedule**": At first a feasible home-away pattern is determined, afterwards the pairings for the corresponding pattern are fixed. In this case often at first keys are used to determine the pairings and specific teams are assigned to the keys afterwards.
3. "**First-assign-modes, then-schedule**": At first for each game the home team is determined, afterwards a corresponding schedule is constructed.

All three approaches have been studied in the literature for different problem settings. The first two are often used in combination with integer linear programming or constraint programming formulations (cf. [19]).

In a "first-schedule, then-break" approach the subproblem of the second stage is to find a home-away pattern corresponding to the pairings determined in the first stage. Although such an approach seems to be very natural, it has some disadvantages if breaks are important. In this situation it is difficult to ensure that the final schedule has a small number of breaks since opponent schedules exist which cannot be scheduled with less than $\frac{1}{4}n(n-2)$ breaks (see [17], [5]). For example, for $n = 16$ teams opponent schedules exist where the minimum number of breaks is 56 (in contrast to the minimum number 14). The complexity status of the break minimization problem is still unknown, but it is conjectured that it is NP-complete. On the other hand, it can be decided in polynomial time whether the given pairings can be scheduled with at most n breaks or not ([16]).

In a "first-break, then-schedule" approach, the subproblem of the second stage consists in finding a feasible opponent schedule for a given home-away pattern. A HAP is called feasible if a corresponding opponent schedule exists which is compatible with this HAP. The pattern set feasibility problem is to determine whether a given HAP is feasible or not. The complexity status of this problem is also still not known. In [15] necessary conditions for feasibility based on checking all possible subsets of teams were provided. For pattern sets with $n-2$ breaks these conditions can be checked in polynomial time. By computational experiments it was shown that in this case the conditions are also sufficient for problems with up to 26 teams. In [2] stronger conditions for feasibility based on a linear program were provided.

The "first-assign-modes, then-schedule" approach has been suggested in [13]. A balanced home-away assignment is a collection of modes determining the home and away teams for each game such that the difference between the number of home games and the number of away games is at most one for each team. By graph theoretical considerations it was shown how balanced home-away assignments can be constructed, how they can be changed according to a connected neighborhood structure, and how mode preassignments can be handled with network flow techniques. Later on, in [12] such an approach was successfully applied to construct schedules for non-professional table-tennis leagues.

References

[1] I. Anderson. Balancing carry-over effects in tournaments. In F.C. Holroyd, K.A.S. Quinn, C. Rowley, and B.S. Webb, editors, *Combinatorial Designs and their Applications*, pages 1–16. Chapman & Hall/CRC Research Notes in Mathematics, 1999.

[2] D. Briskorn. Feasibility of home-away-pattern sets for round robin tournaments. *Operations Research Letters*, 36:283–284, 2008.

[3] D. Briskorn. Combinatorial properties of strength groups in round robin tournaments. *European Journal of Operational Research*, 192:744–754, 2009.

[4] D. Briskorn and S. Knust. Constructing fair sports leagues schedules with regard to strength groups. *Discrete Applied Mathematics*, 158:123–135, 2010.

[5] A.E. Brouwer, G. Post, and G.J. Woeginger. Tight bounds for break minimization. *Journal of Combinatorial Theory A*, 115:1065–1068, 2008.

[6] D. de Werra. Geography, games, and graphs. *Discrete Applied Mathematics*, 2:327–337, 1980.

[7] D. de Werra. Scheduling in sports. In P. Hansen, editor, *Studies on Graphs and Discrete Programming*, pages 381–395. North Holland, Amsterdam, 1981.

[8] D. de Werra. On the multiplication of divisions: The use of graphs for sports scheduling. *Networks*, 15:125–136, 1985.

[9] D. de Werra. Some models of graphs for scheduling sports competitions. *Discrete Applied Mathematics*, 21:47–65, 1988.

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

[11] S. Knust. Scheduling sports tournaments on a single court minimizing waiting times. *Operations Research Letters*, 36:471–476, 2008.

[12] S. Knust. Scheduling non-professional table-tennis leagues. *European Journal of Operational Research*, 200:358–367, 2010.

[13] S. Knust and M. von Thaden. Balanced home-away assignments. *Discrete Optimization*, 3:354–365, 2006.

[14] E.R. Lamken. Balanced tournament designs. In C. J. Colbourn and J. H. Dinitz, editors, *Handbook of Combinatorial Designs*, pages 333–336. CRC Press, 2006.

[15] R. Miyashiro, H. Iwasaki, and T. Matsui. Characterizing feasible pattern sets with a minimum number of breaks. In E. Burke and P. de Causmaecker, editors, *The 4th International Conference on the Practice and Theory of Automated Timetabling*, volume 2740 of *Lecture Notes in Computer Science*, pages 78–99. Springer, 2003.

[16] R. Miyashiro and T. Matsui. A polynomial-time algorithm to find an equitable home-away assignment. *Operations Research Letters*, 33:235–241, 2005.

[17] G. Post and G.J. Woeginger. Sports tournaments, home-away assignments, and the break minimization problem. *Discrete Optimization*, 3:165–173, 2006.

[18] K.G. Russell. Balancing carry-over effects in round robin tournaments. *Biometrika*, 67:127–131, 1980.

[19] M. A. Trick. A schedule-then-break approach to sports timetabling. In E. K. Burke and E. Erben, editors, *The 3rd International Conference on the Practice and Theory of Automated Timetabling*, volume 2079 of *Lecture Notes in Computer Science*, pages 242–252. Springer, 2001.

Comparative performance of models to forecast match wins in professional tennis: Is there a GOAT for tennis prediction?

Stephanie A. Kovalchik

RAND Corporation, Santa Monica, CA
skovalch@rand.org

Abstract

Sports forecasting models—beyond their obvious significance to bettors—are important resources for sports analysts and coaches. Like the best athletes, the best forecasting models should be rigorously tested and judged by how well their performance holds up against top competitors. Although a number of models have been proposed for predicting match outcomes in professional tennis, their comparative performance is largely unknown. In this paper, I compare the accuracy and discriminatory power of ten published forecasting models for predicting the outcomes of 2,443 singles matches during the 2014 season of the Association of Tennis Professionals Tour. It is shown that regression models generally outperformed the predictions of point-based models, which use an algebraic expression for match win probability under the assumption that points are independent and identically distributed. Only one model had excellent accuracy (overall predicted/actual wins = 1.01) and discrimination (AUC = 0.90): a probit regression model whose predictors included prize money, stage of tournament, player age, ranking points, head-to-head results, career match wins for matches played on the same surface, and tournament tier. This study reveals significant variation in forecasting performance among published models and suggests that factors beyond rank and serve/return ability can improve pre-match predictions of match outcomes.

1 Introduction

Multiple statistical models have been proposed for predicting tennis wins, yet few have been rigorously tested or compared against alternative approaches. Just as in sport itself, a high standard of performance should be the ultimate goal of a prediction model. However, there is currently little known about the validity and comparative utility of existing tennis forecasting models.

The purpose of the present paper was to study the performance of published models that predict singles match outcomes in professional tennis. The accuracy and discriminatory power of ten different forecasting approaches were evaluated in a large dataset of singles match outcomes for the 2014 season of the Association of Tennis Professionals (ATP) Tour, and differences in performance by type of surface and tournament level were investigated. By demonstrating the comparative validity and utility of existing prediction models, the results of this study aim to provide insights into the major determinants of win ability in professional tennis and ways to improve the performance of existing models.

2 Methods

2.1 Models

A literature review was conducted to identify published models for forecasting wins in tennis. Seventeen articles underwent a detailed reading to determine their eligibility. Articles were excluded that did not present enough information to predict match wins (2 articles), did not present a forecasting model (2 articles), presented a previously published model (2 articles), or required within-match updating (1 article). Ten articles remained after applying these exclusion criteria.

Shorthand	Model Description (Source)
Point-Based	
Basic iid	Match win probability based on iid points [9]
Opponent-adjusted iid	iid model with opponent-adjusted serve and return probabilities [1]
Low-level iid	iid model with low-level service win probability [10]
Common opponent iid	iid model with serve and return conditional on common opponents [6]
Regression-Based	
Logistic	Logistic model with player ranks as predictors [5]
Basic probit	Probit model with difference in seedings as predictor [2]
Probit plus	Probit model with difference in ranks and player demographics [3]
Prize probit	Probit model with prize earnings and head-to-head outcomes [4]
Other	
Bradley-Terry	Bradley-Terry model with likelihood-based estimates of player abilities [8]
BCM	Bookmaker consensus model (BCM) of winning odds [7]

iid = independent identically distributed

Table 1: Summary of Published Models for Forecasting Outcomes in Tennis

The models included in the validation study are described in Table 1. The majority of the approaches were either based on a regression model or a point-based model, where match outcomes are predicted from algebraic formula under an assumption that points are independent and identically distributed (iid). Most of the regression models were in the probit family and included player rankings as a predictor. Differences in the predictors among these models are shown in Table 2.1. The point-based models extend the basic approach introduced by Newton and Keller [9] by refining the approach to the estimation of the probabilities of winning a point on serve and return. Two models—the Bradley-Terry model of McHale and Morton [8] and the bookmaker consensus model of Leitner et al. [7]—do not fit into either of these categories.

Variable	Logistic	Basic probit	Probit plus	Prize Probit
Difference in seeds	+			
Difference in ranks		+	+	
Previous tournament result			+	
Former top 10 player			+	
Difference in age			+	+
Difference in height			+	
Handedness			+	
Potential prize earnings				+
Head-to-head wins				+
Head-to-head losses				+
Difference in rank points				+
Difference in career wins				+
Rounds remaining			+	+
Grand Slam indicator			+	+
Masters 1000 indicator			+	

Table 2: Predictors for Regression Models of Tennis Match Wins

Technical details about each can be found from their source papers.

2.2 Parameter estimation

Parameter estimates of the prediction models were limited to a 52 week period, with the exception of model predictors that specifically called for a longer look-back period (e.g. career wins). The specific calendar period of inclusion was determined with respect to the date of the match whose outcome was being forecasted to ensure that the most recent year of data was considered. All Grand Slam matches and ATP Tour matches above the Challenger level were included.

Data on model predictors and match outomces were gathered from publicly available websites using the author's R package `deuce` (`https://github.com/skoval/deuce`).

2.3 Validation data

The performance of the models was tested against outcomes for 2,443 ATP singles matches played during the 2014 season. Among the matches in the independent validation dataset, 20% occurred at a Grand Slam, the majority were on hard court surfaces, and slightly more than half included a top 30 player. Higher ranked players won 68% of matches.

2.4 Performance

Two properties of model predictive performance were evaluated: calibration and discrimination. Calibration measures the accuracy of model predictions as is measured by the ratio of predicted match wins for the higher ranked player to their actual match wins. Models with a calibration ratio greater than 1 overestimate match wins for higher ranked players, models with a ratio less than 1 underestimate. Discrimination refers to a model's ability to assign a higher win probability to the player who wins the match than the player who loses. Discrimination was

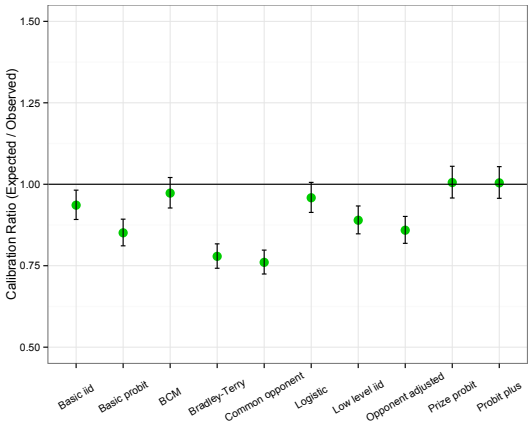


Figure 1: Model Overall Calibration, 2014 ATP Validation Data

measured in two ways: the area under the receiver operating characteristic curve (AUC) and the mean probability difference (MPD) of winners and losers.

3 Results

Several models had excellent overall calibration (Figure 1). Three were the match-based probabilistic models—Probit plus, Prize probit, and Logistic—and the BCM. Point-based models tended to underestimate the actual match wins of higher ranked players. The two models that condition their parameters on head-to-head performance (Bradley-Terry) or performance against common opponents (Common opponent iid) underestimated match wins the most.

The calibration of the four models with the best overall accuracy had good calibration across surfaces; and the Probit plus and Prize probit models had consistently excellent accuracy regardless of player rank or tournament tier (data not shown).

There was a wide range in discriminatory performance across models. Two models had good discriminatory power with AUCs above 0.7: Probit plus (AUC = 0.77) and BCM (AUC = 0.78). Two models had excellent discriminatory power with AUCs at or approaching 0.9: Prize probit (AUC = 0.90) and the Bradley-Terry model (AUC = 0.89). All other models had AUCs of approximately 0.6 or less, indicating poor discriminatory ability.

The discrimination of the Probit plus and Prize probit models as measured by the MPD were clearly superior than the alternative models (Figure 2). These models were the only that had an average difference in the predicted win probabilities of match winners and losers greater than 10% (MPD = 25% for the Probit plus model; MPD = 33% for the Prize probit model).

4 Discussion

This is the first study to compare the performance of proposed models for forecasting tennis match wins. Using a large, independent set of recent ATP singles matches, it was found that regression models generally outperformed point-based models in terms of predictive accuracy and discriminatory power. A probit model that included predictors for player rankings, de-

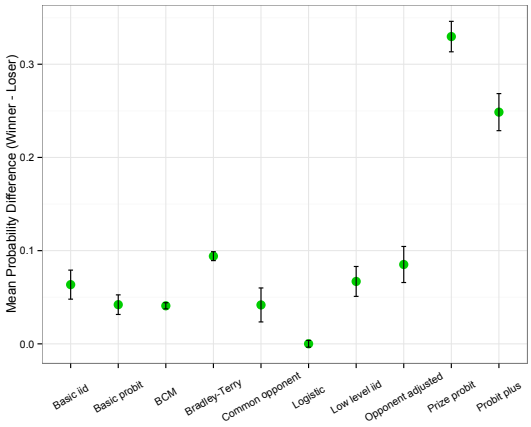


Figure 2: Model Mean Probability Difference, 2014 ATP Validation Data

mographics, and tournament prize money yielded highly accurate predictions and had greater discriminatory power than any alternative model.

An unexpected finding of the present paper was that the accuracy of predictions of regression models were generally better than those of point-based models. The primary difference between the two classes of approaches is that regression models directly propose a statistical relationship between predictors and match win probability whereas point-based models derive an expression for the chance of winning a match as a by-product of models for winning a point on serve and return. Moreover, proposed models for points won on serve or return have been fairly naive with respect to the predictors considered. No point model employed regression techniques that could draw on systematic relationships between player characteristics and serve and return performance. Instead, simple summaries of a player's historical performance have been used as inputs for point-based models with possible stratification by common opponents and/or surface. Point-based models were found to consistenly underestimate actual wins of higher-ranked players, which raises doubt about the ignorability of dependence between points and the adequacy of prior models of serve and return ability.

Another key finding of this paper was that the two models that did the most to account for head-to-head results or difficulty of opponents faced—the Bradley-Terry and Common-opponent models—underestimated the match wins of higher-ranked players more than any competing model. A consequence of focusing on either common opponents or head-to-head results is that the sample size for the serve and return stats decreases, which could result in unstable estimates. Increasing the time period of included matches can help to address this but runs counter to a typical desire to emphasize more recent match results. Another strategy would be a shrinkage estimator that could draw on information about the performance of similar competitors to get more stable estimates of serve and return ability for individual players.

The superior performance of the Gilsdorf and Sukhatme [4] probit model in both accuracy and discrimination has several implications. First, it suggests that factors beyond player rank provide added value for forecasting. These additional factors include monetary incentives, player age, stage and tier of tournament, and career head-to-head, and it would be interesting for further research to clarify the relative importance among these factors. Second, it suggests that a regression model can predict pre-play match outcomes with high accuracy and discrimination

without information about surface or expert opinion (e.g. bookmakers), though it remains unclear whether such information could improve on the performance of the Prize probit model.

Strengths of the present study include the use of an independent validation dataset with many matches on all surfaces and stages of the ATP tour above the Challenger level. However, by focusing on only one year of data it is unclear whether these findings can be generalized to past or future generations of players. Another limitation in the generalizability of the findings is that the paper did not consider performance for the Women's Tennis Association. Further, the present paper focused on pre-match predictive performance and did not investigate the advantages of within-match updating, which is a distinctive potential strength of point-based models.

Although it would be premature to claim any model as the 'greatest of all time', the present comparative study of models to forecast match outcomes of men's professional tennis did reveal some clear winners. The range in performance observed also highlights the importance of comparative validation studies, and supports the broader conclusion that publication doesn't guarantee that a model is useful. Further research is needed to understand the causes of the variation in performance of prediction models for winning in tennis that was observed in this study and ultimately improve the performance of current models.

References

[1] T. Barnett and S. R. Clarke. Combining player statistics to predict outcomes of tennis matches. *IMA Journal of Management Mathematics*, 16(2):113–120, 2005.

[2] B. L. Boulier and H. O. Stekler. Are sports seedings good predictors?: an evaluation. *International Journal of Forecasting*, 15(1):83–91, 1999.

[3] J. Del Corral and J. Prieto-Rodriguez. Are differences in ranks good predictors for grand slam tennis matches? *International Journal of Forecasting*, 26(3):551–563, 2010.

[4] K. F. Gilsdorf and V. A. Sukhatme. Testing Rosen's sequential elimination tournament model incentives and player performance in professional tennis. *Journal of Sports Economics*, 9(3):287–303, 2008.

[5] F. J. Klaassen and J. R. Magnus. Forecasting the winner of a tennis match. *European Journal of Operational Research*, 148(2):257–267, 2003.

[6] W. J. Knottenbelt, D. Spanias, and A. M. Madurska. A common-opponent stochastic model for predicting the outcome of professional tennis matches. *Computers & Mathematics with Applications*, 64(12):3820–3827, 2012.

[7] C. Leitner, A. Zeileis, and K. Hornik. Is Federer stronger in a tournament without Nadal? An evaluation of odds and seedings for Wimbledon 2009. *Research Report Series / Department of Statistics and Mathematics*, 94, 2009.

[8] I. McHale and A. Morton. A Bradley-Terry type model for forecasting tennis match results. *International Journal of Forecasting*, 27(2):619–630, 2011.

[9] P. K. Newton and J. B. Keller. Probability of winning at tennis i. theory and data. *Studies in Applied Mathematics*, 114(3):241–269, 2005.

[10] D. Spanias and W. J. Knottenbelt. Predicting the outcomes of tennis matches using a low-level point model. *IMA Journal of Management Mathematics*, page dps010, 2012.

Raw Rating Systems and Strategy Approaches to Sports Betting

George Kyriakides[1], Kyriacos Talattinis[2] and George Stephanides[3]

[1] University Of Macedonia
it1137@uom.edu.gr
[2] University Of Macedonia
ktalattinis@uom.edu.gr
[3] University Of Macedonia
steph@uom.gr

Abstract

The process of rating and ranking an object yields its position relative to a set of items with similar attributes. This allows the observer to determine which item between two or more, is better, simply by checking their positions. In turn, this can allow quite accurate predictions in case two items of the given set are put against one another. Items can be for example web pages (which one is more relevant), marathon runners (who will win the race) or soccer teams (which one will win the championship). Because many systems have been developed to rate and rank objects, using the correct system for a given situation is of great importance. Also, a careful interpretation of the system's result is of paramount importance For example PageRank is quite efficient in ranking webpages, so is the elo system to rank chess players. The other way around is not guaranteed (and is also not expected) to work. In this paper, we are examining the efficiency of such systems on the context of sports, with very promising results. The possibilities seem quite wide, both in academic and industry applications, making the topic more attractive than ever to researchers around the globe

1 Introduction

Since the invention of money, people try to win as much as possible, with the least possible effort and time invested. The notion that time equals money and vice versa, completely justifies this. Gambling gives the false hope that money can actually be won with very little effort. Although false, people continue to gamble, often getting addicted and developing neurotic behaviors [1]. The size and growth of the industry clearly indicates the fact that people are buying hope, rather than winning money. In Europe (EU28) the gambling industry recorded a Gross Gaming Revenue of 82.36bn euros[3]. Still, as with financial markets, there are certain individuals who try to rationally and scientifically study, analyze and predict the sports and markets. A relatively simple way of doing so, is by rating each participant, determining the strongest and betting on him/her. In this paper we aim to present some linear algebra rating systems, used to rank teams from the English Premier League. We study their overall accuracy and profitability in betting and finally we propose some simple strategies that improve raw mathematical models, by taking into account the nature of the markets and bookmakers.

2 Overview and Soccer's Nature

Although rating systems can be a valuable tool in any sports bettor's arsenal, they can never fully model the complexity of other factors involved in betting. Human psychology, injuries,

fatigue, weather and shady factors play role in the outcome of a match. This calls for more complex, multi-stage systems in order to accurately and effectively place sports bets. In this paper, we briefly present some linear algebra rating systems and compare them in terms of accuracy and profitability. Accurate models are not always profitable [7]. We then try to improve them with simple strategies, aiming to show that strategies can greatly improve the performance of such systems.

Soccer is quite different from other sports, as a tie (draw) is a realistic and possible outcome, whereas in other sports, for example basketball or American football, special rules apply in order to avoid ties. In the English Premier League for example, during the seasons 2008-2014, more than a quarter of the matches (26%) concluded in a draw, as shown in Figure 1.This clearly indicates that soccer has a ternary rather than binary result. An uninformed guess would have a 33.3% chance of guessing the correct result, rather than 50%. Thus, forecasting is a lot harder.

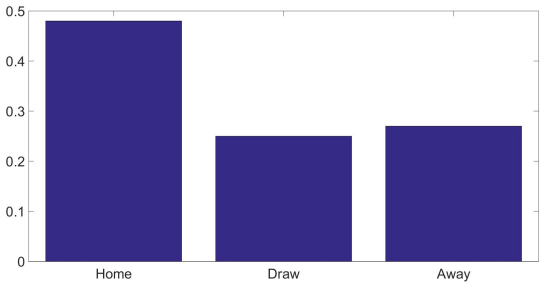


Figure 1: Outcome Percentages. English Premier League 2008-2014

3 Methods

In this paper, we study four rating systems, namely Massey's[8], Colley's[2], Elo's[4] and mHITS[6], as well as an aggregation, which consists of the average of their normalized means. We would like to thank `http://www.football-data.co.uk/` [5] for their freely available historical data. All the data used originate from there. All systems were implemented in MATLAB by The MathWorks, Inc.[9]

3.1 Massey's Method

Originally developed by Kenneth Massey during his final year project in 1997, this method uses a least squares approach to solve the system of linear equations $Mr = p$, yielding a rating vector as a result. Where M is the Massey matrix, with each element M_{ij} representing the number of games played between teams i and j. The diagonal (elements M_{ii}) represent the total number of games played by team i. Vector p contains the net difference in score for each team. . Finally, by solving for r, we get the vector containing the teams' ratings.

3.2 Colley's Method

Proposed by astrophysicist Dr. Wesley Colley in 2001, this method is a variation of the popular method

$$r_i = \frac{w_i}{t_i} \tag{1}$$

Where r_i is the rating of team i, w_i is the total wins of team i and t_i is the total games of team i. The actual method consists of creating a square matrix, C and a vector b and solving the system $Cr = b$ for r.

The matrix C_{ij} contains the total number of games played between the teams i and j, multiplied by -1. The diagonal contains the total number of games played by each team, plus 2. The vector b is calculated as follows:

$$b = 1 + \frac{w_i - l_i}{2} \tag{2}$$

3.3 mHITS

Proposed by Anjela Govan, as a part of her PhD thesis, this system is inspired by the popular Hyper-Induced Topic Search (HITS) algorithm. The main concept of the algorithm is the use of offense and defense rating for each team, and the final rating is calculated as the ratio between these two ratings. Assuming o_i is the offensive and d_i is the defensive rating for each team, high o values signify a strong offense and low d values signify strong defense. Using the matrix A, where A_{ij} is the score generated from team j against team i, and by initializing d as a vector of ones and o as

$$o = A^T \frac{1}{d} \tag{3}$$

the algorithm continuously refines o,d as follows (on k'th iteration)

$$d^{(k)} = A\frac{1}{o^{(k-1)}}, o^{(k)} = A^T \frac{1}{d^{(k)}} \tag{4}$$

3.4 Elo System

Originally developed by Arpad Elo in order to rank chess players, this system has been adopted by quite a lot of sports and organizations. For each participant i, their rating after a match against j is calculated as follows:

$$r_i(new) = r_i(old) + K(E(i,j) - T(i,j)) \tag{5}$$

Where K is a constant, determined by the sport and organization, T is a ternary variable, taking values 0, 0.5, 1, if team i lost, tied, or won, respectively. Finally, the expected outcome:

$$E(i,j) = \frac{1}{10^{-(r_i - r_j)/400} + 1} \tag{6}$$

The actual system for soccer was taken from `http://www.eloratings.net/system.html`

3.5 Aggregation

An aggregation of the methods was also used, by normalizing each of the model results (mean=0, standard deviation = 1) and taking their average. This is done in an effort to reduce variance in the results. Although Arrow [1] argued that it is not possible to correctly aggregate a number of different opinions (in our case, ranks), we do not aim to make better predictions, rather than show that improvements can be made even if a system is not efficient on its own to begin with. Thus,showing that an inefficient system can be improved using strategies in order to filter its preferences.

4 Application

4.1 Accuracy Comparison

We now proceed to use the presented systems in order to predict matches in the English Premier League. The first two weeks of each season were used as a training phase, so no predictions take place during this time. After that, ratings are computed after each team plays one match (there are twenty teams, so after ten matches). Seasons from 2008 through 2014 were used, a total of six. Contestants with a higher rating are expected to be winners, so each model bets on the highest rated contestant. The raw models, although for the most part achieved a better accuracy than an uninformed guess, did not achieve a better performance than playing only the home team as shown on chapter 2 (47% accuracy).

Method	Colley	mHITS	Massey	ELO	Aggregate
Accuracy	0.386	0.491	0.485	0.364	0.386
Games	2160	2160	2160	2160	2160

Table 1: Raw Model Accuracies

Given that most models are not capable of predicting draws on their own, we filtered out the matches concluding in a draw. Although there is a great improvement, Table 2 shows that a lot of games are filtered out. Elo, as well as the aggregation of the results are still below 47% accuracy. Also, this method is not really suitable for gambling, as we are in no position to determine if a match will end up in a draw or not beforehand.

Method	Colley	mHITS	Massey	ELO	Aggregate
Accuracy	0.519	0.662	0.653	0.456	0.445
Games	1604	1604	1604	1604	1604

Table 2: No-Draw Model Accuracies

4.2 Profitability

The next step is to try our raw models performance in gambling. Using the average odds from the biggest online bookmakers, we simulate the course of our models. Each model is given 1,000 money units as a starting capital. Each model bets 50 units on each match, and receives the profits dictated by the average odds.As we can see, none of the models achieve any profit.

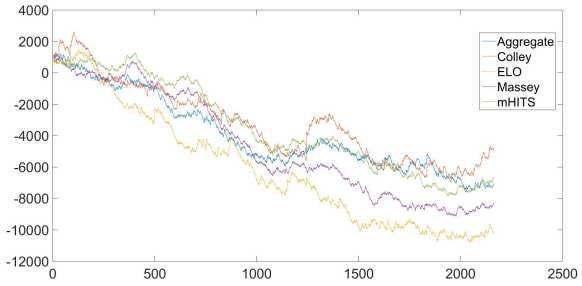


Figure 2: Raw Model Profits

4.3 Strategies

In order to improve our models' performance, we try to pick the safest matches, and bet only on them, in an effort to cut our losses. By looking at the bookmakers' odds, we can determine whether the selected outcome can be considered safe or not. Outcomes with high odds are considered unsafe. This is because bookmakers increase the odds of an outcome unlikely to happen, in order to lure in more bets and balance their portfolio and risk. By applying this strategy, a lot of matches are filtered out, as shown on Table 3, but the ones remaining are a lot safer.

Method	Colley	mHITS	Massey	ELO	Aggregate
Accuracy	0.822	0.792	0.788	0.768	0.802
Games	180	327	322	164	243

Table 3: Low Odds Strategy

This strategy almost doubles the accuracy in some models. It is important to note that we include all matches in the selection process, not only the ones resulting in home/away wins. By applying the strategy to our betting simulation we notice that Colley's method achieves a profit, and none of the models end up with a negative profit. It is a really great improvement over the raw models as shown in Figure 3.

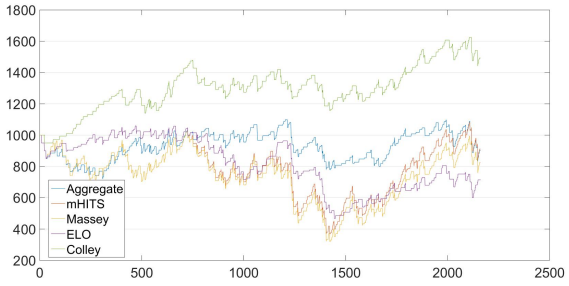


Figure 3: Low Odds Strategy Profits

Still, there is room for improvement. By reducing our exposure to risk, we can further improve the model. Instead of staking 50 units per match, we now stake the required amount

in order to win 50 units. In other words $S_i = 50/O_i$. S_i are the stakes for game i and O_i are the odds of the predicted outcome. Using Colley's model as an example, Figure 4 shows the increase in profits, throughout the simulation.

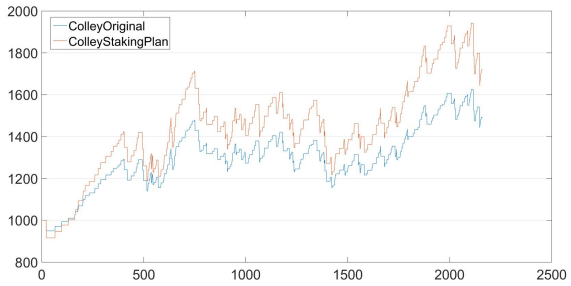


Figure 4: Staking Plan Colley Profits

5 Conclusions

Rating and ranking a team in a league does not guarantee an effective or profitable prediction of future matches. Being a market, sports betting needs more than just fundamental analysis of the product being traded (match outcomes). Correct model and match selection, as well as a sound staking plan is needed in order to be profitable in these markets. In this paper we tried to prove these statements by comparing raw models against the same models with match filtering strategies applied. Given the many factors that determine the outcome of a match, it is better to pick matches with less randomness in these factors, than trying to model them and then pick the right outcome.

References

[1] Kenneth J Arrow. A difficulty in the concept of social welfare. *The Journal of Political Economy*, pages 328–346, 1950.

[2] Wesley N Colley. Colley's bias free college football ranking method: The Colley matrix explained. *Princeton University*, 2002.

[3] Egba.eu. EGBA, 2014.

[4] Arpad E Elo. *The rating of chessplayers, past and present*, volume 3. Batsford London, 1978.

[5] Football-data.co.uk. Football betting and soccer results, 2015.

[6] Anjela Y Govan, Amy N Langville, and Carl D Meyer. Offense-defense approach to ranking team sports. *Journal of Quantitative Analysis in Sports*, 5(1), 2009.

[7] George Kyriakides, Kyriacos Talattinis, and Stephanides George. Rating systems vs machine learning on the context of sports. In *Proceedings of the 18th Panhellenic Conference on Informatics*, pages 1–6. ACM, 2014.

[8] Kenneth Massey. Statistical models applied to the rating of sports teams. *Bluefield College*, 1997.

[9] Mathworks.com. MathWorks - MATLAB and Simulink for technical computing, 2015. `http://www.mathworks.com/`.

The Age Advantage in Association Football

Steve Lawrence

The Football Analytics Lab, Milk Studios, 121 College Road, London, NW10 5EY
steve.lawrence.ata@gmail.com

Abstract

I propose the concept of a measurable Age Advantage in football. A concept of Home Advantage is well established and this paper aims to explore an analogous Age Advantage. Using data from 24 top-level competitions this paper explores the relationship between average age and results. The simple analytical tool of a mean age calculation shows that for younger cohorts, fielding a team with an older average age is advantageous and conversely, for older cohorts, fielding a team with a younger average age is advantageous. Home Advantage statistics provide a comparator against which to test the extent of the Age Advantage.

1 Introduction

To test the age advantage a simple measure of the age of any competing team was established, it's referred to as the average team age and abbreviated as ATA. It is calculated as the mean of the chronological ages of all the players who make up the team at the time of the competition. For the sake of uniformity and simplicity the study is restricted to the ATA of only the starting line-up in each case on a given match day, no allowance is made for substitutions on or off. (Across the top 6 European Leagues the average playing time in a match for a player present in the starting line-up is 84.55 minutes (94% of match playing time), it can be said therefore that the players in the starting line up will tend to be making the most significant contribution to results.)

To give extreme examples in order to illustrate the point, a team with an ATA of 15 will tend to win against a team with an ATA of 10, whereas, a team with an ATA of 45 will tend to lose to a team with an ATA of 30.

Three interesting questions arise out of consideration of the hypothesis:

1. Does a relationship between the ATA of competing cohorts and performance outcomes exist at the highest competitive levels of the sport where gambling on outcomes may take place?

2. Does such a relationship between the ATA of competing cohorts and performance outcomes exist at a micro level i.e. within age banding where scouting and elite selection is taking place?

3. Is there an identifiable ATA tipping point, across populations as cohort age increases, where the advantage passes from older to younger teams?

To explore this hypothesis a dataset was analysed consisting of 6,389 matches played during 2013/14 at the highest competitive levels organised by national leagues in Europe, USA and South America and by UEFA and FIFA and involving 12,778 teams exhibiting average ages (ATAs) from under 17 to over 28.

2 Home Team Advantage

The baseline comparator of *'Home Advantage'* is variously held to be due to the familiarity of the home ground, home bias of referees, the larger number of home supporters and disadvantage of long travel times for the away team. League competition is broadly constructed on a duplicate home and away match basis to attenuate *'Home Advantage'*.

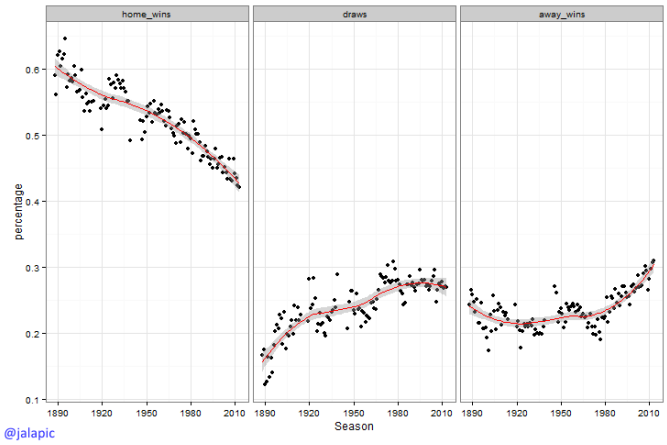


Figure 1: The development of the % of home wins, draws & away wins 1890 to 2010. ©James Curley.

This analysis of 120 years of English First Division/Premier League football shows the relationship between the chance of a home win, a draw or an away win. Percentages of home wins, draws and away wins are fairly consistent across contemporary football competitions with 46% home wins, 24% draws and 30% away wins. It is possible to translate these percentages into points accrued for the home team and the away team based on 3 points for a home win, 1 point for a draw and 0 points for a loss. For example the ratio of 46% - 24% - 30% translates into **1.62** points per game for the home team and **1.14** points per game for the away team. This creates a convenient ratio for analysis and this technique of translating home/draw/away win percentages into points per game (PPG) is used for analysis of *'Age Advantage'*.

3 Methodology

The accumulated data was compiled into sets for each competition, the percentages of wins, draws and losses accruing to each team according to various parameters were calculated and then converted into PPG for comparison. The parameters against which PPG scores were accumulated for comparison were:

1. Home Team v Away Team (baseline).
2. Higher ATA v Lower ATA (older v younger).
3. Higher RAEi v Lower RAEi (more biased v less biased).

(The RAE parameter is devised as a simple measure to express the ratio of early-born to late-born players in any given team. It is termed the 'relative age index' or 'RAEi' and it is calculated as the number of early-born players divided by the total number of players and is expressed as a decimal proportion between 0 and 1.)

4. Proximity to a notional 'Optimum ATA' (close or remote).

   (The 'Optimum ATA' is calculated as that ATA, proximity to which, gives the highest possible PPG score.)

4 Comparing Home Team Advantage with Age Advantage and Relative Age Advantage

Figure 2, below, shows how data from 20 of the 24 competitions in our sample indicates fairly consistent Home Team/Away Team PPG characteristics when charted against cohort age (single location competition data is excluded because no venues constituted a home ground).

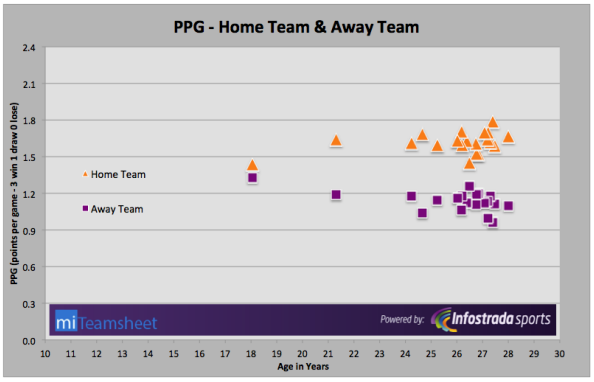


Figure 2: Graph of PPG accrued by the **Home** Team (orange) or the **Away** Team (purple).

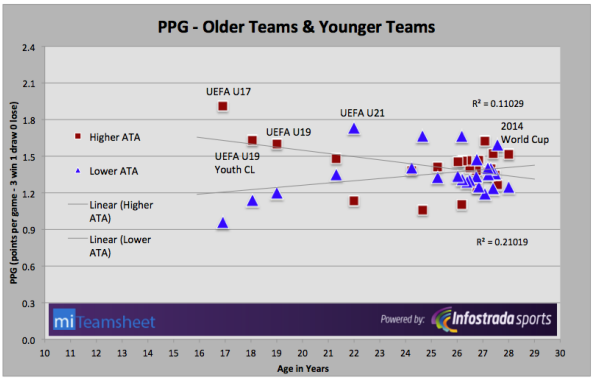


Figure 3: Graph of PPG accrued by the **Older** Team (red) or the **Younger** Team (blue).

In Figure 3 however we see a pattern of diminishing advantage for older teams and increasing advantage for younger teams.

At younger ages the *'Age Advantage'* accruing to an older team is evident as a PPG score of **1.9** at U17 and **1.65** at U19 and these are high values exceeding *'Home Advantage'* scores. In the Netherlands U23 competition for 2012/13 the average PPG scores are reduced almost to parity and within the Netherlands Eredivisie where the league average age is under 25 and the Belgium Jupiler Pro League where the mean ATA is 26.18 it seems that younger teams have gained an advantage. The PPG scores for most of the other top leagues are close to parity on an ATA basis with Italy, the MLS and the UEFA Champions League showing a continuing advantage to older teams. The World Cup stands out because the age advantage has passed quite significantly from the older teams to the younger teams.

In examining carefully the profiles of ATA distributions for the various competitions we observe restricted standard deviations lower down the age ranges. In the UEFA U17 and UEFA U19 competitions in particular we observe a very restricted standard deviation and if we look at the relationship between the Inter Quartile Range and the mean age of the cohorts as shown in Figure 4 we see a pattern of fielding teams within narrower age ranges at younger ages. Within the youngest of our samples for the UEFA U17's the mean age is very close to the extreme upper limit of the age eligibility range and the IQR is 0.125. In the context of an eligibility range which is 2 years the IQR for the U17 cohort is inordinately small.

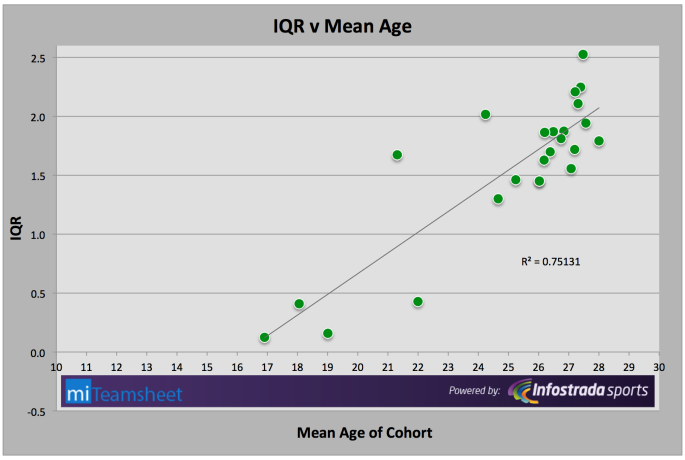


Figure 4: A graph of IQR of the ATA distribution v Cohort Age.

In order to achieve such a high mean age, close to the age eligibility cut-off date, it is logically necessary to recruit, into the roster, a high number of players born as early as possible within the competition year giving rise to a 'relative age effect'.

The motivation for achieving high PPG scores is obviously very strong and if an *'Age Advantage'* does exist then it is reasonable to expect a relationship between achieving high PPG scores and fielding older teams within the age bands of youth competition. If this is the case we should see a correlation between teams exhibiting high relative age bias and performance outcomes and Figure 5 shows that this is the case.

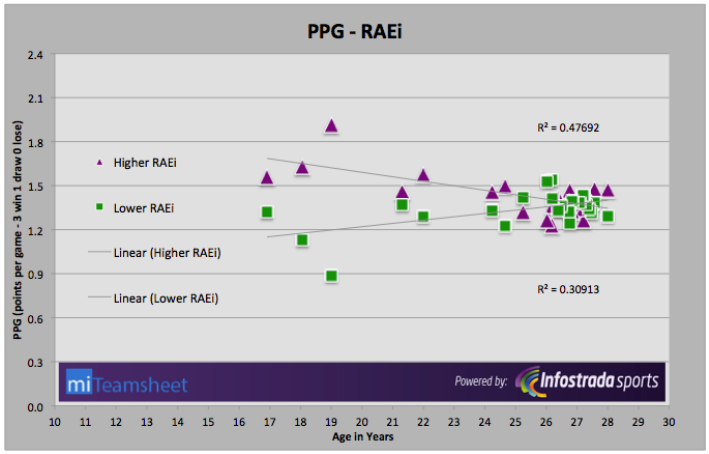


Figure 5: Graph of PPG accrued by the **Higher RAEi** Team (purple) or the **Lower RAEi** Team (green).

5 Optimum ATA

In Figure 4 it was apparent from the IQR values that at younger ages the ATAs of competing teams were very close to the upper limit of age eligibility ranges and that the IQR values increased with increasing cohort age. If at younger ages, older teams tend to prevail and at older ages, younger teams tend to prevail then it is natural that an ATA evident at younger ages will be close to the eligibility limit, in effect the optimum ATA is the eligibility limit. We might also expect that at older ages, where no age eligibility limit exists, the ATA will also tend towards an optimum and for much older cohorts this will be lower than the average age of the cohort. Its more difficult to analyze this because the attrition of older players sustains ATA's at optimum levels. When the performance levels of older players diminish, younger players replace them.

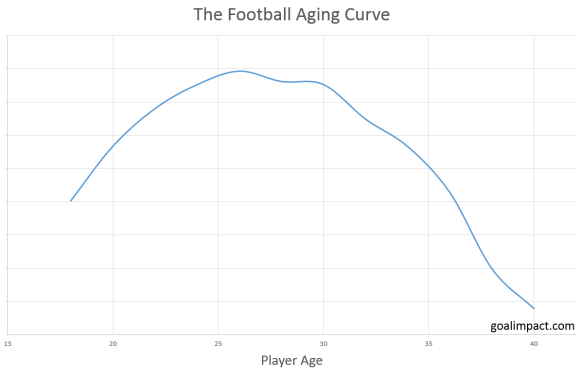


Figure 6: The Goal Impact rating of professional players v. age. ©Jörg Seidel @goalimpact

Figure 6 shows the goal impact rating (an index of player effectiveness) for professional players against age and illustrates this process very well.

The ruthless efficiency of the system, a kind of natural selection, sustains the optimum but this, of course, implies that the optimum will always tend towards the cohort average. At younger ages the optimum is the age eligibility limit and at older ages it tends towards the mean of the ATAs.

It is possible to calculate an optimum ATA, i.e. an ATA which produces the best outcome, by calculating back from the competition outcomes. An iterative goal seeking process can determine the optimum ATA where the team with an ATA closest to the optimum achieves the highest PPG. The derived optimum ATA chart is shown in Figure 7. The data points seem to converge on age 27.

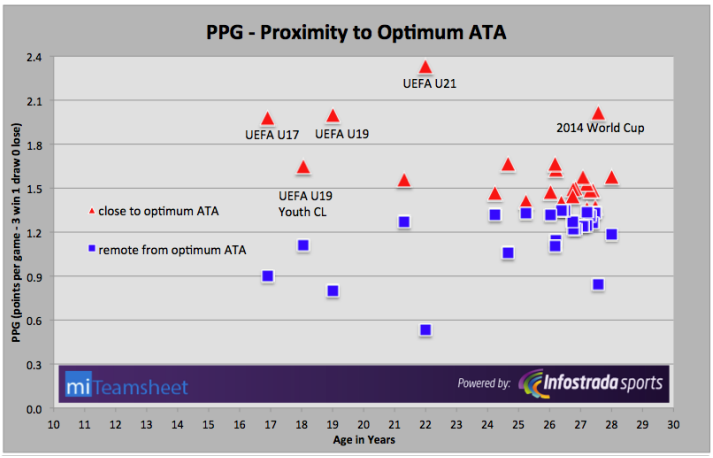


Figure 7: A chart of PPG against proximity to an Optimum ATA.

The 2014 World Cup, as the pinnacle of global soccer competition, provides very interesting data. The high mean ATA of 27.57 is coupled with a tight range of just 5.71. The younger team won 32 times, there were 9 draws and the older team won 23 times.

During the final knock-out phase when it might be reasonably expected that performance levels had increased, the mean ATA remained at 27.55 but the range was further reduced to 3.63 implying a tendency towards an optimum. During the knock-out phase 13 out of the 16 matches were won by the team with an ATA closest to the optimum.

The optimum can only be calculated after the event based on actual outcomes however the argument that the optimum tends towards the mean due to the attrition of older players seems to hold because charting the derived optimum ATA against cohort average age produces an R^2 value of 0.84.

6 Conclusions and Further Work

The initial proposition that at younger ages, older teams have an advantage and that at older ages, younger teams have an advantage is both logical and un-contentious. This study shows that the phenomenon is quite easily analysed and when examined more closely reveals consistent predictive relationships. The study has introduced the concepts of average team age

(ATA), relative age index (RAEi), optimum ATA and cohort age tipping point. Furthermore it has shown that the technique of translating match outcomes across competitions into a single comparator of points per game (PPG) and then comparing PPG scores according to the different parameters such as ATA, RAEi and optimum ATA provides insight into factors driving player recruitment and relative age effects.

The study confirms that a relationship exists between ATA's and match results at the highest competitive levels which may have implications for gambling, also that a relationship exists within year groups with implication for scouting, youth development, recruitment and retention and that there is an identifiable ATA tipping point, across populations as cohort age increases, where the advantage passes from older to younger teams and this would seem to have a relationship with the cohort mean and a hypothetical optimum.

There would therefore seem to be some merit in developing the tools described in this study and pursuing further research into age advantages in football using them.

An Integer Programming Model to provide Optimal Line-up for Master Swimming Relay

Simona Mancini

Politecnico di Torino
simona.mancini@polito.it

Abstract

Masters swimming is a special class of competitive swimming for swimmers aged over 20 years who compete within age groups of five years. In addition to individual competition, Master meetings programs host relay events. Providing an optimal relays line-up is not a trivial issue and cannot be easily managed by hand, even for problems involving a small number of athletes. Master meetings differs in many aspects from professional ones, because given the ludic and recreational nature of the master competitions, it is also important to meet athletes wishes and to allow more swimmers to take part of the competition if this does not imply a strong decrease of the team line-up competitiveness. Another crucial issue in master meetings is that, while there is a very high competition in the lower age categories, and an excellent chronometric performance is necessary to win a medal, in the older age categories competitiveness is not so high and it is possible to win a competition even with a medium level squad. In this work an Integer Programming formulation, taking into account all these aspects, is presented and computational tests on real data, showing the effectiveness of the proposed approach, are reported.

1 Introduction

Operations research has been extensively used in sports, both to determine winning tactics and to schedule sport leagues. Many papers that address tactical and strategic questions for both team sports like baseball [3], cricket [8, 2] and ice hockey [16] and for single player sports like tennis [7], long jump [13] and pole vaulting [4], can be found in literature. The literature on sports league scheduling is somewhat large and covers all team sports; a complete survey can be found in [5]. Both integer programming (for instance, [15] for basketball, [11, 14] for American football) and heuristics ([10] for baseball, [1] for cricket and [12], who address sport league scheduling for different sports) have been proposed. A complete overview of the application of OR methods in sport can be found in [17].

Masters swimming is a special class of competitive swimming for swimmers aged over 20 years who compete within age groups of five years. In addition to individual competition, Master meetings programs host relay events. In official championships, both 4×50m medley and freestyle relays are scheduled. Medley relay consists of four different swimmers, each swimming 50 meters in one of the four strokes, the order of strokes being fixed. In a freestyle relay all the four participants swim in the freestyle stroke. In addition to men's and women's relay events also mixed relay events (2 men/2 women) are planned. Age groups for the relay events are determined by the combined age of the team participants, allowing for swimmers of very different ages to compete together. Each team may line-up one relay for each age category. Each athlete may compete in in up to 4 different relays, (medley-freestyle, mixed medley and mixed freestyle), but in only one age category for each type of relay. Furthermore, while at a professional level the only goal of the coach is to reach the best results possible, at the Masters level, also swimmers' exigencies and satisfaction should be taken into account, and coaches

aim to obtain competitive line-ups while trying to involve in the relay events as many athletes as possible. Moreover, while in the lower aged categories competition is very strong and it is possible to obtain a good placement only with a top level line-up, in older categories it is easier to gain a medal even with a poor chronometric performance. The averaged line-up quality in the lower age categories is very high, and the number of teams with a top-level foursome is large while in older age categories the number of top level line-ups is very small. A coach must take into account all these aspects while deciding the team line-ups. In this case, the goal is not only to provide the fastest line-ups possible but to maximize the team global competitiveness while trying to involve the highest number of athletes.

A similar problem has been addressed in [9], where dual meet events were considered and a model capable of maximizing team's score, supposing to know opponents' team scores and line up, was provided. In [6] this problem has been generalized to multi-team meetings, but the supposition that opponents' line-ups are known in advance still holds.

The novelty of the approach proposed is that no information on the opponents is required, because the competitiveness of a line-up is computed taking into account the personal bests of the involved athletes and a reference time (which could be the national record or the world record) for the category in which the foursome compete. For each category a multiplier parameter is also considered to take into account the competitiveness level in the category. In fact, in lower age category, a performance which is 20% slower than the reference time, probably will relegate the team to the middle ranking positions, and certainly will not allow the team to win a medal while in an older age category it gives great chance of victory. Furthermore the objective function to be maximized is defined as a linear combination of competitiveness of the team and satisfaction of the athletes (expressed in terms of number of athletes of the team involved in the relays line-ups).

The paper is organized as follows. A detailed description of the problem is given in Section 2, while the integer programming model is briefly summarized in Section 3. The computational results obtained testing the model on different real cases are reported and a detailed analysis of these results is provided in Section 4. Finally, Section 5 is devoted to the conclusions and possible future developments.

2 Problem Description

The problem treated in this paper addresses the relay line-ups optimization for a Masters swimming meeting. There are six kind of relay: freestyle men, freestyle women, medley men, medley women, mixed freestyle and mixed medley. Each relay line-up is composed by 4 athletes; for the "men" and "women" events, the quartet is composed of only men and of only women, respectively. In the "mixed" events, the quartet must be composed of 2 men and 2 women. Five different age categories are defined:

- Category A: sum of the ages of the participants in the relay must be between 80 and 120
- Category B: sum of the ages of the participants in the relay must be between 120 and 160
- Category C: sum of the ages of the participants in the relay must be between 160 and 200
- Category D: sum of the ages of the participants in the relay must be between 200 and 240
- Category E: sum of the ages of the participants in the relay must be between 240 and 280

Where applicable further categories may be added operating in the same way (all the categories have an age range of 40 years).

Each team can line-up at most one relay for each category for all the 6 kinds of relays. Each athlete may compete in up to 4 relays but in only one relay for each kind. Since the age limit is on the sum of the participants' ages and not on the single athlete ones, a quartet may be composed of athletes with heterogeneous ages. For instance a line-up for a Category B relay may involve one athlete aged 20 years, one 30 years, one 40 years and one 50 years. The same athlete may compete in different categories in different kinds of relay. We defined an athlete "involved" in the competition only if he takes part in at least one relay.

The competitiveness C of a line-up in a given kind of relay is computed as follows:

$$C = 1 - \gamma \frac{T - \tau}{\tau}$$

where T is the sum of personal bests of the athletes involved in the relay, τ is the reference time for the corresponding kind of relay and category and γ is a parameter, such that $0 \leq \gamma \leq 1$, which takes into account the level of averaged quality and competitiveness for the correspondent category. The category A holds the highest value of γ, $\gamma = 1$, while this value is non linearly decreasing with the increment of age categories. Basing on experimental data, the following values of γ have been defined:

$$\gamma_A = 1; \gamma_B = 0.9; \gamma_C = 0.7; \gamma_D = 0.4; \gamma_E = 0.2.$$

We assume, basing on real data analysis, that $T < (1 + \frac{1}{\gamma})\tau$, for each possible line-up. In this way, the competiveness of a relay line-up C is always $0 < C \leq 1$. This gives an actual description of the reality because, even inscribing a very slow quartet, the competitiveness in the event is always greater than zero. If the team does not inscribe any quartet for a given event, its competitiveness in that event is assumed to be equal to zero.

The goal is to maximize a function which is a linear combination of the total competitiveness of the team and the number of involved athletes.

3 Model formulation

An Integer Programming Model to define the optimal relay line-ups is proposed. For brevity, the model is not explicitly reported here, but it is described in detail at the conference.

The objective function is defined as the weighted sum of two terms: the total competitiveness of the team, which is computed as the sum of the competitiveness of the team in each event, and the number of involved athletes. The multiplying parameter for the second term (the number of involved athletes) in the objective function is denoted β.

4 Computational Results

4.1 Instances Description

Three different real instances, related to an Italian Masters team, Torino Nuoto, have been addressed. Each one of them show different characteristics. The first one deals with the 2014 Piedmont Regional Championships (Reg2014), in which Torino Nuoto lined-up 14 male and 10 female athletes. The greatest part of the athletes are in the range between 30 and 40 years for men and between 20 and 30 for women (with a 59 years outsider). Based on the ages of athletes

it would be possible to line-up relay only for categories A,B and C. Personal bests are quite homogeneous which make more complex the line-up decision problem. In a regional championship only four kinds of relays are scheduled: men freestyle, men medley, women freestyle and women medley. The second instance is related to the 2014 Italian Championship (Ita2014). In this meeting also mixed relays are planned. For Torino Nuoto, 15 athletes have been involved (9 men and 6 women). In this case the team may compete in categories up to C for men and mixed relay and up to B for the women ones. The last instance is the larger one and it is related to the 2011 Piedmont Regional Championships (Reg2011). The Torino Nuoto roster was composed of 26 men and 11 women. The men's roster is very heterogeneous, varying from 21 to 58 years in age. Athletes' personal bests are also very different, varying from 30.84 seconds to 47.93 seconds for the backstroke, from 31.52 to 48.61 for the breaststroke, from 28.52 to 44.08 for the butterfly, and from 24.82 to 34.97 for the freestyle. The women's roster is more homogenous both for age (from 20 to 30) and for performances with an older outsider of 46 years with slightly worse personal bests. Even in this meeting, as in the Reg2011, no mixed relays are planned. In this case the team may compete in categories up to D for men relay and up to B for the women ones.

4.2 Results analysis

Each instance has been solved by the model firstly considering the parameter $\beta = 0$. This means that we are only trying to maximize competitiveness. Secondly the model is run with $\beta = 0.1$. In this case, all the athletes have been involved in the line-up. This implies that this objective has been completely fulfilled and so, even increasing the value of β, the optimal solution would not change. The model is able to find the optimal solution in 0.1 seconds for all the tested instances.

In Reg2014, the model with $\beta = 0$ obtains a sensibly better solution with respect to the real line-up, but the number of involved athletes, (20 out of 24) is lower with respect to the actually involved athletes with the real line-ups (23 out of 24). Running the model with $\beta = 0.1$ it is possible to obtain a solution which is slightly worse than the optimal, but still better than the real one, and in which all the 24 athletes are involved. This result shows that, acting with a proper tool, it is possible to provide very good line-up even allowing all the athletes to participate creating a good balance between team performances and athletes' satisfaction which is a crucial issue in an amateur context such as the master one. In Ita2014, the optimal solutions with $\beta = 0$ (involving 14 out of 15 athletes) and $\beta = 0.1$ (involving all the 15 athletes) have similar objective functions values and they are both strongly better that the actually provided line-up. Finally, in Reg2011, the optimal solution obtained with $\beta = 0$ is better and involves more athletes than the real one (28 vs 26 out of 37), but this solution is not so fair because 9 athletes would not take part in the event. Fixing $\beta = 0.1$ it is possible to obtain a solution which is very near to the optimal one (and better than the real one) and in which all the 37 athletes are involved.

5 Conclusion and Future Developments

This paper addresses the relay line-up optimization for Master Swimming. Master competitions are amateur events, therefore the goal of a coach is not only to maximize team performance and competitiveness but also to involve in the event as many athletes as possible. An Integer Programming Model is proposed in which the objective function to be maximized is given by a linear combination of team competitiveness and athletes' satisfaction (An athlete is satisfied

if he/she is involved in at least one relay). The model has been tested on real instances and is able to solve them to the optimality in 0.1 second. Results obtained show that, with a smart line-up, it is possible to involve a larger number of athletes (in the tested cases, all the athletes) without a significant loss of competitiveness. Further developments in this field could address an extensive computational testing on larger instances. Moreover, a similar approach could be adopted to address other sports in which team line-up must be scheduled, among which track and field events, artistic and rhythmic gymnastics, ski, cross country ski and snowboard.

References

[1] J. Armstrong and J. Willis. Scheduling the cricket world cup: a case study. *Journal of the Operational Research Society*, 44(11):1067–1072, 1993.

[2] S. Clarke. Dynamic programming in one-day cricket optimal scoring rates. *Journal of the Operational Research Society*, 49(4):331–337, 1998.

[3] R. Freeze. An analysis of baseball batting order by monte carlo simulation. *Operations Research*, 22(4):728–735, 1974.

[4] M. Hersh and S. Ladany. Optimal pole-vaulting strategy. *Operations Research*, 37(1):172–175, 1989.

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

[6] S. Mancini. Assignment of swimmers to events in a multi-team meeting for team global performance optimization. In *Proceedings of Mathsport2013, Leuven, Belgium*, June 2013.

[7] J. Norman. Dynamic programming in tennis: When to use a fast serve. *Journal of the Operational Research Society*, 36(1):75–77, 1985.

[8] J. Norman and S. Clarke. Dynamic programming in cricket: Optimizing batting order for a sticky wicket. *Journal of the Operational Research Society*, 58(12):1678–1682, 2007.

[9] M. Nowak, M. Epelman, and S. Pollock. Assignment of swimmers to dual meet events. *Computers & Operations Research*, 33:1951–1962, 2006.

[10] R. Russell and J. Leung. Devising a cost effective schedule for a baseball league. *Operations Research*, 42(4):614–625, 1994.

[11] R. Saltzman and R. Bradford. Optimal realignments of the teams in the national football league. *European Journal of Operational Research*, 93(3):469–475, 1996.

[12] J. Schonberger, D. Mattfeld, and H. Kopfer. Memetic algorithm timetabling for non-commercial sport leagues. *European Journal of Operational Research*, 153(1):102–116, 2004.

[13] G. Sphicas and S. Ladany. Dynamic policies in the long jump. In *Optimal Strategies in Sport*, pages 101–112. North Holland, 1977.

[14] T. Urban and R. Russell. Scheduling sports competitions on multiple venues. *European Journal of Operational Research*, 148(2):302–311, 2003.

[15] T. Van Voorhis. Highly constrained college basketball scheduling. *Journal of the Operational Research Society*, 53(6):603–609, 2002.

[16] A. Washburn. Still more on pulling the goalie. *Interfaces*, 21:59–64, 1991.

[17] M. Wright. 50 years of OR in sport. *Journal of the Operational Research Society*, 60:161–168, 2009.

Effects of Rule Changes in NHL

Patrice Marek

University of West Bohemia, Plzeň, Czech Republic
patrke@kma.zcu.cz

Abstract

There are two main reasons for changing rules in ice hockey. The first reason is the safety of players and spectators and the second reason is the attractiveness of matches. We study effects of rule changes that were made because of the second reason, e.g. allowing the two-line pass, narrowing the neutral zone, overtime in the case of a tie game. We analyse these rule changes from two perspectives. The first one is how many goals are scored in a match and the second one is how many games are tied after the 60-minute regulation time. We use the data from the 1979–1980 season to the 2014–2015 season, i.e. data after the last big expansion of NHL in 1979.

1 Introduction

Ice hockey is a dynamic sport that evolves over time. One reason is the advance in technology, e.g. composite sticks, goaltender masks (introduced in the National Hockey League (NHL) by Jacques Plante during the 1959–60 season), or the availability of information about every single player. Usually a change caused by an advance in technology is slow and with no clear starting point (in 1970s, there still were some goaltenders without a mask).

On the contrary, rule changes take effect rapidly and it is easy to identify their starting point as the rules are changed only between seasons. There are two main reasons to change the rules.

- The first reason is the safety of players and spectators.
- The second reason is the attractiveness of matches, i.e. how many goals are scored.

We study effects of rule changes that were made because of the second reason, e.g. allowing the two-line pass, narrowing the neutral zone, overtime in the case of a tie game. We analyse these rule changes from two perspectives. The first one is how many goals are scored in a match and the second one is how many games are tied after the 60-minute regulation time. Although, the data from the first NHL season (1917–1918) can be obtained we use only the data from the 1979–1980 season to the 2014–2015 season. The last big expansion of the NHL was made between the 1978–79 season and the 1979–1980 season when four teams joined the league and set the number of teams to 21. Each further expansion included only one or two teams and since the 2000–2001 season the NHL consists of 30 teams.

2 Data Processing

Match results can be obtained from numerous online sources. In this work, we used [1] which offers all results since the 1942–1943 season. As mentioned before the data from the 1979–1980 season to the 2014–2015 season are used. Due to the cancellation of the 2004–2005 season we have 35 seasons for the analysis where only regular season matches are used (play-off matches are played with different rules for the overtime).

The list of rule changes in the NHL since the 1979–1980 season follows. Mentioned seasons are those when a new rule took effect.

(A) 1983–1984 season: Five-minute overtime period was introduced if the score was tied after the end of regulation time. If either team scored during the overtime period, the game ended and the scoring team received two points and the losing team received no point. If no goal was scored, each team received one point [2, pp. 178].

(B) 1999–2000 season: Number of skaters in the overtime was reduced from five to four, plus a goalkeeper and both teams automatically received one point for games that went into the overtime, i.e. if either team scored during the overtime period, the game ended and the scoring team received two points and the losing team received one point. If no goal was scored, each team received one point [2, pp. 178-9].

(C) 2003–2004 season: Maximum length of goaltenders' pads was set at 38 inches [7, pp. 11].

(D) 2005–2006 season: If no goal was scored during the overtime period, a shootout was used to decide the game's winner, i.e. the team that receive the extra point [2, pp. 179-180]. The NHL salary cap was introduced [2, pp. 124]. The neutral zone was reduced from 54 feet to 50 feet, center red line was eliminated for two-line passes, size of goaltenders' equipment was reduced by approximately 11 percent and some more minor changes to rules were made [7, pp. 11].

(E) 2010–2011 season: Goaltenders' pads shall not exceed eleven inches. Size of pads is set to be anatomically proportional [6, pp. 18].

(F) 2013–2014 season: More reduction to the size of pads due to a change in the calculation of anatomically proportional size [8, pp. 19].

All analysed seasons are depicted in Figure 1 and Figure 2; vertical lines in these figures show rule changes mentioned in the previous list.

3 Hypotheses Formulation

In this section, hypotheses that will be tested in the next section are formulated. The first group of hypotheses concern average number of goals in the regulation time and the second group concern relative number of ties in the regulation time.

3.1 Average Number of Goals

The hypothesis for each pair of consecutive seasons is

- H_0 : Average number of goals scored in the regulation time in a season X+1/X+2 is the same as in a season X/X+1.

We expect that the rule changes (A) and (B) from the list in Section 2 had no effect on the average number of goals in a season and that the rule changes (C)–(F) resulted in an increase of the average number of goals in a season. Therefore, we use one-tailed alternative hypothesis for (C)–(F) and two-tailed hypothesis for (A) and (B) defined as

- H_1 : Average number of goals scored in the regulation time in a season X+1/X+2 is higher (different) than in a season X/X+1.

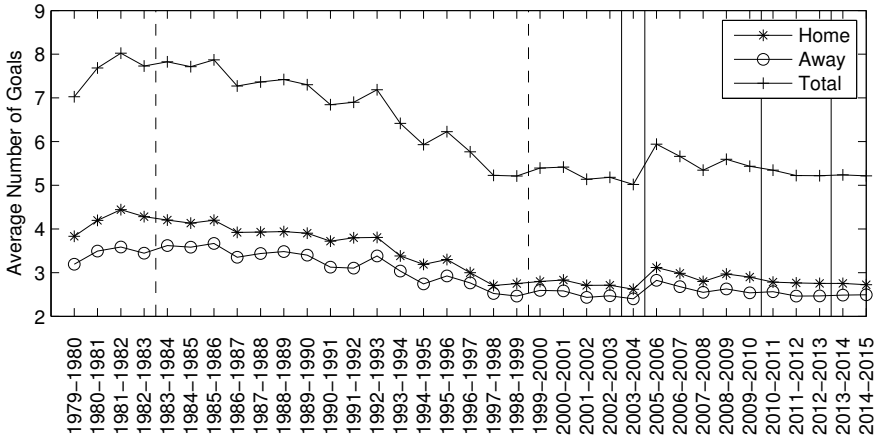


Figure 1: Average number of goals in NHL seasons (the dashed vertical line is used for rule changes where no effect is expected and the full vertical line is used for rule changes where increase is expected)

For consecutive seasons where no major rule change was made a test with two-tailed alternative hypothesis is performed.

3.2 Relative Number of Ties

The hypothesis for each pair of consecutive seasons is

- H_0 : Relative number of ties in the regulation time in a season X+1/X+2 is the same as in a season X/X+1.

We expect that the rule change (A) from the list in Section 2 had no effect on the relative number of ties, the rule change (B) resulted in an increase of the relative number of ties and rule changes (C)–(F) resulted in a decrease of the relative number of ties. Therefore, we use appropriate one- or two-tailed alternative hypothesis defined as

- H_1 : Relative number of ties in the regulation time in a season X+1/X+2 is lower (higher for the rule change (B), different for the rule change (A)) than in a season X/X+1.

For consecutive seasons where no major rule change was made a test with two-tailed alternative hypothesis is performed.

4 Hypotheses Testing

First, a test whether the number of goals in a game is Poisson distributed is performed. Since [3], Poisson distribution is widely used for the modelling number of scored goals and usually, papers use this as a fact. To test whether the number of goals is Poisson distributed one would need i.i.d. repetitions of a match between two teams; clearly, this is impossible to achieve. In reality, each match could have different parameter of Poisson distribution. It means that there are some differences between parameters in each match but usually not large in one season; we recall that rules are changed between seasons. Therefore, in one season, a mixture of Poisson

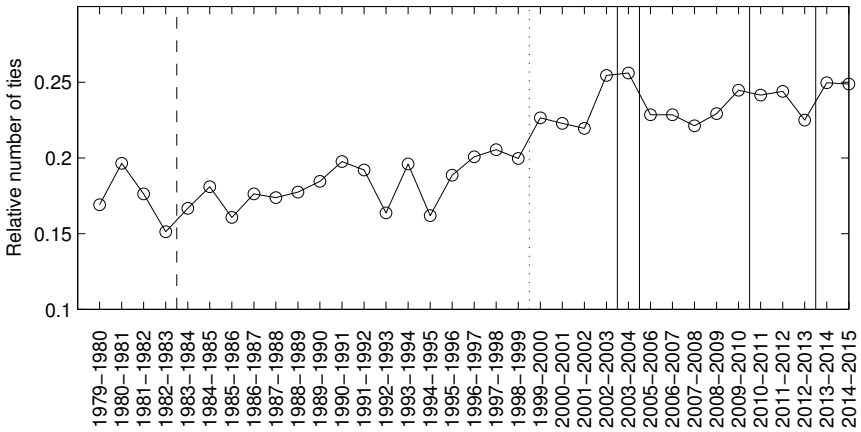


Figure 2: Relative number of ties in NHL seasons (the dashed vertical line is used for rule change where no effect is expected, the dotted vertical line is used for rule change where increase is expected and the full line is used for rule changes where decrease is expected)

distributions with similar parameters is obtained. This is the reason why lower 0.01 significance level is used in the test of the following hypothesis.

- H_0 : Number of goals scored in a game during a season is Poisson distributed.

Alternative hypothesis is defined as

- H_1 : Number of goals scored in a game during a season is not Poisson distributed.

Classical Pearson's chi-squared test is used for each season. This means that family of 35 tests (one for each season) is obtained and therefore, a correction for the significance level has to be made. Bonferroni correction [5, pp. 103-107] is chosen to maintain the familywise error less than 0.01. After Bonferroni correction, the significance level for each hypothesis is $\alpha^* = 0.01/35 \approx 2.8751 \cdot 10^{-4}$. The null hypothesis cannot be rejected at the 0.01 level of significance as the p-value of the whole test is 0.017.

4.1 Average Number of Goals

To test hypotheses defined in Section 3.1 the Conditional test (C-test) is used (see [4]). C-test uses simple fact that conditional distribution of X_1 (number of goal scored in the first season) conditionally given $X_1 + X_2 = k$ (sum of number of goals scored in both seasons) is binomial with the number of trials k and success probability

$$p = \frac{\frac{n_1}{n_2}}{1 + \frac{n_1}{n_2}}, \tag{1}$$

where n_1 is number of matches in the first season and n_2 is number of matches in the second season.

Table 1 confirms our expectation that for the first two rule changes (A) and (B) we cannot reject the null hypothesis that average number of scored goals in consecutive seasons is the same. In the case of rule changes (C)–(F), null hypothesis was rejected only for the rule change (D),

Season	p-value	Season	p-value	Season	p-value
2014–2015	0.791	2001–2002*	0.003	1989–1990	0.373
2013–2014 (–,x)	0.427	2000–2001	0.834	1988–1989	0.686
2012–2013	0.962	1999–2000	0.064	1987–1988	0.487
2011–2012	0.200	1998–1999	0.910	1986–1987*	<0.001
2010–2011 (–,x)	0.839	1997–1998*	<0.001	1985–1986	0.263
2009–2010	0.092	1996–1997*	<0.001	1984–1985	0.431
2008–2009*	0.008	1995–1996*	0.018	1983–1984	0.489
2007–2008*	0.001	1994–1995*	<0.001	1982–1983*	0.032
2006–2007*	0.004	1993–1994*	<0.001	1981–1982*	0.014
2005–2006* (–,x)	<0.001	1992–1993	0.018	1980–1981*	<0.001
2003–2004 (–,x)	0.964	1991–1992	0.685	1979–1980	—
2002–2003	0.657	1990–1991*	<0.001		

Table 1: P-values for the test of the hypothesis *Average number of goals scored in the regulation time in a season X+1/X+2 is the same as in a season X/X+1*; first p-value 0.791 is for the test between seasons 2014–2015 and 2013–2014, etc.; (–,x) indicates one-sided alternative hypothesis, i.e. average is higher; * indicates rejection of null hypothesis

the largest rule change that was introduced for the 2005–2006 season. There are several seasons where the test rejected null hypothesis and no major rule change was made that season. This suggests that rule changes have no effect on the average number of scored goals and some essential rule change has to be made.

4.2 Relative Number of Ties

To compare the relative number of ties we use the test for the difference in two population proportions as defined in [9, pp. 486-491]. This test uses the approximation of the binomial distribution by the normal distribution and it allows to test hypotheses defined in Section 3.2.

Table 2 contains only p-values lower than 0.2. It confirms our expectation that for the first rule change (A) we cannot reject the null hypothesis that the relative number of ties in consecutive seasons is the same. It was expected that the rule change (B) should lead to the increase of the relative number of ties but this hypothesis can be confirmed only at the level of significance 0.1. The rule changes (C)–(F) should lead to the decrease of the relative number of ties but this hypothesis can be confirmed only for the largest rule change that was introduced in the 2005–2006 season when the level of significance 0.1 is used. As can be seen, the only season where the null hypothesis can be rejected at the level of significance 0.05 is 2002–2003 when no major rule change was done.

5 Conclusion

In this paper, there were presented major rule changes in the NHL since 1979. First, the test of hypothesis that the number of goals scored in a game during one season is Poisson distributed was performed; this hypothesis cannot be rejected on the level of significance 0.01.

The effect of rule changes was tested from two perspectives - their effect on the average number of goals and relative number of ties. As expected, the rule change (A), five-minute overtime for a tie game, had no effect from both perspectives. Rule change (B), one point

Season	p-value	Season	p-value	Season	p-value
2005–2006* (x,–)	0.055	1995–1996	0.167	1992–1993	0.107
2002–2003**	0.041	1994–1995*	0.079	1982–1983	0.166
1999–2000* (–,x)	0.060	1993–1994*	0.055	1980–1981	0.146

Table 2: Table contains the lowest obtained p-values for the test of the hypothesis *Relative number of ties in the regulation time in a season X+1/X+2 is the same as in a season X/X+1*; second p-value 0.041 is for the test between seasons 2002–2003 and 2001–2002, etc.; (–,x) (or (x,–)) indicates one-sided alternative hypothesis, i.e. relative number of ties is higher (or lower); * indicates rejection of null hypothesis with the level of significance 0.1 (** for the level of significance 0.05)

guaranteed in the overtime, had no effect when the 0.05 level of significance is used but for the 0.1 level of significance null hypotheses are rejected and their alternatives that the average number of goals and relative number of ties are higher are confirmed.

For rule changes (C)–(F), mostly involving equipment reduction, an increase in the average number of goals and decrease in the relative number of ties was expected. The null hypothesis was rejected only for the biggest rule change that included two-line pass legalization and reduction of neutral zone, i.e. (D). In this case, using significance level 0.05, average number of goals is higher after the change and, using significance level 0.1, the relative number of ties decreased.

Using the significance level 0.05, the null hypothesis for the average number of goals was rejected in numerous seasons where no major rule change was made. The same was recorded only for the 2002–2003 season in the case of the relative number of ties.

Learning from these results we can investigate effect of three-point system in different leagues. For the moment, it can be seen that to test it from the perspective of the relative number of ties could offer good results. On the contrary, testing average number of goals does not offer good results.

References

[1] Hockey History. Hockey History - Birthplace, Ancestors, NHL, 2015. `http://hockey-history.com/nhl/historical-results/`.

[2] L. H. Kahane and S. Shmanske. *The Oxford Handbook of Sports Economics, Volume 1: The Economics of Sports*. Oxford University Press, New York, 2012.

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

[4] J. Przyborowski and Wilenski H. Homogeneity of Results in Testing Samples from Poisson Series. *Biometrika*, 31(3/4):313–323, 1940.

[5] N. J. Salkind. *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: SAGE Publications, Inc, 2007.

[6] The National Hockey League. *Official Rules 2010–2011*. Dan Diamond and Associates, Inc., Canada, 2010.

[7] The National Hockey League. *Official Guide & Record Book 2013*. Dan Diamond and Associates, Inc., Canada, 2012.

[8] The National Hockey League. *Official Rules 2013–2014*. Dan Diamond and Associates, Inc., Canada, 2013.

[9] J. M. Utts and Heckard R. F. *Mind on Statistics*. Cengage Learning, Stamford, 2015.

ShotLink and Consistency in Golf

Roland Minton

Roanoke College,
Salem, Virginia, USA
minton@roanoke.edu

Abstract

An exploration of the role of consistency in golf utilizes ShotLink data. The metric used for the quality of golf shots is Strokes Gained, which compares the golfer's performance on each shot to an average performance. ShotLink data from the PGA Tour provides several characteristics of every shot taken on the PGA Tour, with distances measured to the inch. These data are used to compute Strokes Gained for twenty different types of shots (putts, bunker shots, shots from the fairway between 50 and 100 yards, and so on). With this metric, consistency can be correlated with quality. The strongest effects are shown for tee shots. For approach shots, distance to the hole after the shot can be decomposed into distance off-line and distance long or short. While conventional wisdom is that distance control is more important for golfers, the data indicate that the two errors have equal importance.

1 Introduction

A repeatable swing is essential in golf. While there are almost as many swing types as professional golfers, to reach the professional level a golfer must achieve an impressive level of consistency to reliably produce effective shots. Given this high level of performance, there remains a question of whether (relative) consistency is an advantage in professional golf. Are the top golfers characterized by large swings of performance to reach high peaks, or are they characterized by such consistency that their worst performances still leave them in contention?

For professional golfers, the issue may be more about strategy than swing mechanics. That is, is it necessary to play a high-risk, high-reward game to succeed? Or is golfing excellence more a product of ruthless efficiency? In the past, a "yes" answer to either question could be backed up with examples of Arnold Palmer or Jack Nicklaus, Greg Norman or Nick Faldo, or Phil Mickelson or Tiger Woods. In this paper, the question is explored using data from the PGA Tour.

A theoretical investigation of consistency (or riskiness in strategies) has been based on a model proposed by, of all people, the great mathematician G.H. Hardy [4]. Work on this model has been continued in recent years [2, 6, 9]. The consistent golfer has been shown to have a lower mean score but, under many circumstances, the less consistent golfers are more likely to win the tournament [7].

2 Strokes Gained

To address the role of consistency in a single element of the game such as putting, we need a statistic that reliably measures putting effectiveness. The classic statistic used to evaluate putting has been the number of putts per round. The flaws in this statistic are obvious: a golfer whose longest putt is from 4 feet will likely have fewer putts than one whose shortest first putt is from 30 feet, regardless of how well either one putts. A new statistic that has quickly become widely accepted is Strokes Gained Putting.

Strokes Gained is a concept proposed by a number of people. It has been popularized by Mark Broadie [1], who was instrumental in getting the PGA Tour to officially adopt the statistic. The name comes from Fearing et al. [3], whose work was featured in a *Wall Street Journal* article that helped introduce the idea to the public [8]. The version of Strokes Gained used in this paper extends on the ideas presented by Minton [7].

Conceptually, Strokes Gained is simple. Suppose that, for every location on the golf course, the average score for professional golfers is known. Then, for each shot, Strokes Gained equals the average score after the shot minus the average score before the shot. For example, suppose that the average score on a par 4 is 4.2. A golfer booms a long drive down the middle of the fairway. From that position, the average score is 3.8. Then Strokes Gained for the drive is $3.8 - 4.2 = -0.4$; the excellent drive saved the golfer 0.4 strokes. If the golfer does this on every drive, then we can say that the golfer's driving ability is 0.4 strokes better than average. This will be denoted -0.4, following the golfing tradition of negatives meaning under par. If the golfer's second shot is pulled left into a sand trap from which the average score is 4.6, then Strokes Gained for the approach shot is 4.6-3.8=0.8 strokes (over par, or worse than average). A sand shot to a foot from the hole, with an average score of 4.0, saves 4.0-4.6=0.6 strokes, and the tap-in is worth 4.0-4.0=0.0 strokes. Note that the golfer had a one-putt green, but gets no credit for good putting because, in fact, no credit is deserved.

The practical issue is how to compute the average scores. ShotLink is a system administered by the PGA Tour, which the Tour graciously allows researchers to access for academic purposes. Every shot on the PGA Tour (this does not include any of the four major tournaments) is recorded with 38 characteristics, including distances measured to the inch. A variety of lies (tee, fairway, green, greenside bunker, intermediate rough, primary rough, water, and so on) are recorded. Line-of-sight blockages due to trees and mounds are not recorded, but we can, for example, compile all of the shots taken from the fairway 101-102 yards out and find their average scores. Here, shots are categorized based on fairway, intermediate rough, primary rough, or miscellaneous lies and distances in chunks of 50 yards. Tee shots are considered separately for par 3s, par 4s, and par 5s. Putts and shots from the fringe are separated. It should be noted that the ShotLink data does not record the club used, or the type of shot attempted.

Another issue to consider is course effects. Clearly, putting can be easier or harder depending on the course and its setup. It is at least plausible that all other types of shots could be affected by the course setup and other factors such as the weather. An easy fix is to compute the Strokes Gained for a round relative to the average of the field in that round. This creates a logical problem, however. A course may look easy because the best players are playing in that tournament, and then the course correction would make those players look worse. In fact, they should get credit for playing in the toughest tournament. A technique of Larkey [5] is adapted to untangle this mess. Course averages are computed, then used to adjust the player ratings, which then feed back to recompute course averages, and so on. Player and course ratings converged, in most cases, by the fifth iteration.

Some examples may provide useful context, as well as being interesting to golf fans. In 2014, the leader in putting saved 1.03 strokes per round compared to the average PGA Tour putter. The leader in drives on par fours rated -0.72 strokes per round (under par, or better than average). Other leading values for Strokes Gained per round are for tee shots on par threes (-0.40), fairway shots from 0-50 yards (-0.36), fairway shots from 150-200 yards (-0.35), and tee shots on par fives (-0.32). The number one golfer for 2014, in terms of total Strokes Gained, is Rory McIlroy at -1.95 strokes per round better than average on the PGA Tour (majors not included). Other year-end number ones are 2013: Justin Rose (-1.84), 2012: Rory McIlroy (-2.92), 2011: Luke Donald (-2.44), and 2010: Steve Stricker (-2.04). Lest you think

that McIlroy's 2012 was historically dominant, here are the values for Tiger Woods, who is number one for each of the previous five years: Tiger 2009: -3.06 strokes, Tiger 2008: -4.24 strokes, Tiger 2007: -2.96 strokes, Tiger 2006: -3.37 strokes, Tiger 2005: -2.68 strokes.

3 The Importance of Putting

When asked to identify the most important skill in golf, many people will automatically identify putting. Strokes Gained gives us a way to evaluate such statements. Previous research shows that between 2004 and 2012 the top 40 golfers in total Strokes Gained on the PGA Tour averaged 1.13 strokes per round better than average [1]. They were a mere 0.15 strokes per round better than average putting. The top 10 included 5 golfers (McIlroy, Singh, Els, Garcia, and Scott) who were actually *worse* than average at putting. On the other hand, winners of tournaments averaged 3.7 strokes per round better than average total, and 1.3 strokes per round better than average putting [1]. By percentage, putting seems to be more important to success in a tournament than it does to long-term career success.

As part of the current study, Strokes Gained for each of twenty skills was computed for every round of every golfer in the 2011 ShotLink data set. The correlation of each skill with round score was then computed. The largest correlation, by far, was for putting (0.567). Combining all fairway shots between 50 and 200 yards gave the second highest correlation at 0.383. Then came penalty strokes (0.317), par 3 tee shots (0.309) and par 4 tee shots (0.291).

A different way to slice the data is to look at individual players over the full year. The 50 players with the most rounds in the 2011 data set were analyzed in the same way. For 49 of the 50 players, the skill with the highest correlation to score is putting. The remaining player (Troy Matteson) had a correlation of 0.51 for shots from the fairway (150 to 200 yards), with putting second at 0.46.

Various explanations may be advanced for the apparent differences in the importance of putting. One has to do with consistency. Suppose that all aspects of a player's game were completely consistent except putting. Research has shown that the bumpiness of greens introduces a fair amount of luck into putting results [7, 10]. If only putting varies significantly from round to round, then putting controls the change in score from round to round, whether or not that putting is especially good. The hypothesis, then, is that if a skill is unusually inconsistent then it could be unusually influential to a golfer's score.

4 Consistency

For each of the ten years from 2005 to 2014, the 230 golfers with the most rounds in the ShotLink data set are evaluated on each of 20 skills: tee shots on par 3s, 4s, and 5s; bunker shots; fairway shots from 0-50 yards, 50-100 yards, ..., 200-250 yards; shots from other lies (intermediate rough, primary rough, miscellaneous) from 0-50 yards and 50-200 yards; shots from the fringe; putts; recovery shots; layup shots; and penalties. The overall rating is simply the sum of the individual ratings; one of the major advantages of having skill ratings in units of strokes per round is the ability to meaningfully sum the ratings.

The mean of each skill over the 2300 golfer-years is essentially zero. This is not surprising since the units are strokes above or below average. The standard deviations give some indication of the typical spread of the ratings. While "consistency" is not the best description of what this measures, it does give a sense of which skills tend to separate golfers the most in the overall ratings. Putting has the largest standard deviation at 0.364. Par 4 tee shots is second largest

at 0.208, then par tee shots at 0.169. The largest standard deviation for fairway approach shots is 0.148 for the range 150-200 yards. Thus, putting ratings have the largest spread of values, and contribute the most to separating the player ratings.

Out of the 2300 player-years in the original study, in 514 cases the same player is represented in consecutive years. Correlations of consecutive-year ratings gives a direct measure of consistency. The two largest autocorrelations were for tee shot ratings on par 5s ($\rho = 0.80$) and par 4s ($\rho = 0.77$). For the players in the study, then, driving is the single skill that carries over the most from year to year. This makes some sense, as tee shots are not subject to bad lies and tight pin placements. The autocorrelation for overall ratings was $\rho = 0.63$, so players' overall ratings in one year explain about 40% of the variation in overall ratings the next year. The autocorrelation for bunker play was $\rho = 0.49$, while the autocorrelation for putting was $\rho = 0.44$. Other autocorrelations of note are $\rho = 0.41$ for par 3 tee shots and $\rho = 0.40$ for fairway approach shots from 100-150 yards.

In sports analytics circles, the autocorrelation is used to measure the persistence of a statistic. Large values give evidence that the statistic is measuring a repeatable skill, while small values can cast doubt on the significance of the statistic. The statistic may be more a measure of luck than skill if the autocorrelation is too small. For golf, the transience of skills is experienced by everyone. However, if it is equally true that there are consistently good/bad putters and good/bad drivers, then the lower autocorrelation for putting may be due to a significant luck factor.

There are 57 players who appear in the top 230 in 8 of the 10 years from 2005 to 2014. For each of these 57 players, means and standard deviations were computed for each of the 20 skills. Interestingly, the mean over the entire set of player-years for each skill is essentially zero. However, the mean of the overall rating over all player-years is 0.41 strokes better than average. There is no paradox here, just that the sum of enough 0.02 and 0.03 values can equal a large number. For most of the skills, the standard deviation is approximately 0.10. The three main exceptions are the overall rating (mean -0.41 and standard deviation 0.72), par 4 tee shots (mean -0.02 and standard deviation 0.20) and putting (mean -0.06 and standard deviation 0.34). This is evidence that Strokes Gained putting is more variable than Strokes Gained for other skills. The cause could be some combination of the luck involved in putting, the skill of good putting being harder to maintain, and the potential being greater for large values of Strokes Gained putting compared to other skills.

The main purpose of organizing these data is to examine whether consistent players are better players at any or all of the skills. For most skills, there is little evidence for this hypothesis. The correlation between player means and player standard deviations for the 57 players is between 0.1 and 0.2 for many skills. This is not very high, but the positive sign indicates that more consistent players (smaller σ) are better players (smaller means, especially more negative means, are better). The putting correlation is 0.19. By far, the largest correlation is for par 4 tee shots, with $\rho = 0.64$. Consistent drivers tend to be better drivers. The next highest correlation is for penalty strokes with $\rho = 0.47$ (this is not exactly a skill, however) and then par 3 tee shots, with $\rho = 0.34$. Interestingly, the correlation for overall rating is only $\rho = 0.15$.

5 Distance versus Accuracy

The title of this section could apply to any type of golf shot, but the focus here is on approach shots. The ShotLink data lists, for each shot, the distances to the hole in inches before and after the shot, and the length of the shot. These three values can be triangulated to give the number of inches long or short, and the number of inches off-line. However, it cannot be determined

whether the ball is off-line to the left or right.

For this study, all shots from the fairway at a distance of 100-200 yards in 2014 were compiled and grouped into 10-yard groupings. Strokes Gained for each shot provides a measure of the quality of the shot, and six other statistics were computed: whether the shot finished on the green, whether the shot finished on the green more than 30 feet away, whether the shot finished off the green less than 30 feet away, the number of inches off-line, the number of inches long (positive) or short (negative), and the absolute value of the distance long or short. The idea of the second and third statistics is that a conservative strategy of aiming for the middle of the green could produce numerous shots on the green but not close to the hole. A more aggressive strategy of aiming at the pin could produce shots close to the hole but not on the green.

Given that the distances were only divided into ten-yard intervals, there could be large correlations in performances. For example, a golfer who is especially good from 120-130 yards could reasonably be expected to be good from 130-140 yards. This is not supported by the data. Correlations between Strokes Gained in the shorter half (e.g., between 120-130 yards and 130-140 yards) hover around 0.1. In the longer half (e.g., between 160-170 yards and 180-190 yards), the correlations are approximately 0.2, larger but not impressively so. The only correlations above 0.2 are between larger distances, such as the correlation of 0.23 between Strokes Gained from 170-180 yards and Strokes Gained from 190-200 yards. This, in fact, is the largest of the correlations. Thus, performance in the different distance intervals is not very consistent.

The same lack of crossover occurs when correlating a given skill to Strokes Gained. For example, the correlation between Greens Hit from 100-110 yards and Strokes Gained from 100-110 yards is -0.71, while the correlation between Greens Hit from 100-110 yards and Strokes Gained from 110-120 yards is -0.02. Other correlations are similar, with slightly larger values for the longer distances. The correlations show that golfers who hit more shots on greens score better, but do not reveal anything surprising.

For the conservative statistic, the correlations at the same distances are in the 0.3 range up to 150 yards, and essentially 0 beyond 150 yards. For the aggressive statistic, the correlations at the same distance are approximately 0.4 for distances up to 120 yards, and in the 0.15 range beyond 120 yards. In both cases, correlations for different distances are essentially zero. The fact that the correlations are positive indicates that a likely conclusion is simply that it is better to be on the green *and* close to the hole.

The amount off-line correlates to Strokes Gained in the 0.5-0.6 range for the same distances. The correlations for different distances are similar to the correlations between Strokes Gained at different distances. For the signed distance control statistic, the correlations tend to be small and negative at the same distances, and zero at different distances. However, for the absolute difference in distances, the correlations are in the 0.4-0.5 range for the same distances. The correlations for distance control and angle control are very similar. There is no evidence here that distance control is more important than angle control.

Correlating the Strokes Gained for a specific distance range with overall score could shed light on which distance is the most important. The highest correlation is 0.62 for the distance range 170-180 yards. Other correlations are in the 0.4-0.5 range. A multiple regression of distance range Strokes Gained values to overall score showed similar results. All distances are significant at the 0.05 level, but the largest coefficient is for the 170-180 yard range. At fixed distances, regressions of Strokes Gained as a function of the other six statistics give r-squared values ranging from 0.69 for 100-110 yards to 0.83 for 170-180 yards. In each case, the signed distance control was not a significant contributor to the model.

6 Conclusions

Golf is a game requiring several discrete skills, including driving, hitting irons off the ground, and putting. Strokes Gained gives a method of isolating performance for each shot in a round, leading to a measure of proficiency at individual skills. However, the skills do not show large values of autocorrelation, or persistence. This may be due to transient levels of proficiency or large noise levels. Driving shows the greatest persistence among golf skills for PGA Tour professionals. Putting is by almost any measure the most important skill (although its importance has been overstated in the past), but it shows a low level of persistence. This inconsistency could, in fact, help contribute to its role in determining results. Approach shots show lower levels of importance and persistence than driving and putting. Performance from different distances is surprisingly inconsistent. Evidence points to the 170-180 yard range as being the most influential distance range for PGA Tour professionals.

In the last five years, golf has seen several different players win major tournaments and attain the number one world ranking. No one person has dominated golf as Tiger Woods did in the early and mid 2000s. The low level of persistence of golf skills, whether due to skill fluctuations or random elements, helps explain why it is difficult to stay at the top.

References

[1] M. Broadie. *Every Shot Counts.* Gotham Books, New York, 2014.

[2] G.L. Cohen. On a theorem of G.H. Hardy concerning golf. *The Mathematical Gazette*, 86:120–124, 2002.

[3] D. Fearing, J. Acimovic, and S.C. Graves. How to catch a tiger: Understanding putting performance on the PGA tour. *Journal of Quantitative Analysis in Sports*, 7(1), 2011.

[4] G.H. Hardy. A mathematical theorem about golf. *The Mathematical Gazette*, 29:226–227, 1945.

[5] P.D. Larkey. Comparing players in professional golf. In *Science and Golf II: Proceedings of the World Scientific Congress of Golf*, pages 193–198. E&FN Spon, 1994.

[6] R.B. Minton. G.H. Hardy's golfing adventure. *Math Horizons*, pages 5–9, April 2010.

[7] R.B. Minton. *Golf By the Numbers.* The Johns Hopkins University Press, Baltimore, 2012.

[8] J.P. Newport. A stat is born: Golf's new way to measure putting, 2010. *The Wall Street Journal*, `http://www.wsj.com/articles/SB10001424052748703791704575114071142473884` Accessed on April 22, 2015.

[9] A.H.G.S. Van Der Ven. The Hardy distribution for golf hole scores. *The Mathematical Gazette*, 96:428–438, 2012.

[10] F.D. Werner and R.C. Greig. *How Golf Clubs Really Work and How to Optimize Their Designs.* Origin, Inc., Jackson, Wyoming, 2000.

Some Interesting Sports League Models

Kimmo Nurmi and Jari Kyngäs

Satakunta University of Applied Sciences, Pori, Finland
cimmo.nurmi@samk.fi, jari.kyngas@samk.fi

Abstract

Professional sports leagues are big businesses. It is vital to generate league schedules that secure revenues for the teams and at the same time increase transparency to secure fairness for the teams. This paper presents some interesting sports league formats, extra game models, special constraints and other features that make it a very challenging task to generate an acceptable schedule. Examples of such are incomplete round robins, mini-tournaments, local rival games, back-to-back games, leveling games, traveling issues and venue restrictions. A set of simplified instances based on the presented features is made available for the sports scheduling community as benchmarks.

1 Introduction

Professional sports leagues have become big businesses. At the same time, the quality of the schedules has become increasingly important; the schedule has a direct impact on revenue for all involved parties. For instance, the number of spectators in the stadiums and the traveling costs for the teams are influenced by the schedule, and TV networks that pay for broadcasting rights want the most attractive games to be scheduled at commercially interesting times. Furthermore, a good schedule makes a season more interesting for the media and the fans, and fairer for the teams, so that all teams play under the same conditions. A good overview of sports scheduling can be found in [2] and an extensive bibliography in [3].

In a sports tournament, n teams play against each other over a period of time according to a given timetable. The teams belong to a *league*, which organizes *games* between the teams. Each game consists of an ordered pair of teams, denoted (i, j) or $i - j$, where team i plays at *home* – that is, uses its own *venue* (stadium) for a game - and team j plays *away*. Games are scheduled in *rounds*, which are played on given *days*. A *schedule* consists of games assigned to rounds. A schedule is *compact* if it uses the minimum number of rounds required to schedule all the games; otherwise it is *relaxed*. If almost all teams play on a certain round, the team that does not play on that round is said to be on *bye*.

If a team plays two home or two away games in two consecutive rounds, it is said to have a *break*. In general, for reasons of fairness, breaks are to be avoided. However, a team can prefer to have two or more consecutive away games if its stadium is located far from the opponents venues, and the venues of these opponents are close to each other. A series of consecutive away games is called an *away tour*.

In a *round robin tournament* each team plays against each other team a fixed number of times. Most sports leagues play a double round robin tournament *(2RR)*, where the teams meet twice (once at home, once away), but quadruple round robin tournaments *(4RR)* are also quite common. A *mirrored* double round robin tournament *(M2RR)* is a tournament where every team plays against every other team once in the first $n - 1$ rounds, followed by the same games with reversed venues in the last $n - 1$ rounds. In an incomplete round robin some teams do not play against each other.

Constructing a single, double or quadruple round robin tournament is quite an easy task nowadays, but when we introduce restrictions, requirements and requests, the problem becomes

intractable. Furthermore, being able to produce an acceptable schedule is not only about first defining the restrictions, requirements and requests and then developing a suitable solution method. An essential part of the problem is the process of consulting with the various league parties. The scheduling process can easily take several months. First, the possible improvements to the format must be brainstormed. Then, the format has to be accepted by the team CEOs and the broadcasting companies. Next, all the restrictions, requirements and requests by the various parties are gathered. After the importance of the constraints has been decided, some test schedules have to be generated to find out whether it is possible to cope with the given framework. Finally, the schedule can be optimized.

Sports scheduling involves three main problems. First, the problem of finding a schedule with the *minimum number of breaks* is the easiest one. De Werra [1] has presented an efficient algorithm to compute a minimum break schedule for a *1RR*. If n is even, it is always possible to construct a schedule with $n-2$ breaks. For a *mirrored 2RR*, it is possible to construct a schedule with $3n-6$ breaks [1]. The computational results by Rasmussen and Trick [10] indicate that for a *non-mirrored 2RR*, it is possible to construct a schedule with $n-2$ breaks.

Second, the problem of finding a schedule that *minimizes the travel distances* is called the *Traveling Tournament Problem (TTP)* as defined by Easton, Nemhauser and Trick [2]. Its a 2RR tournament where the teams do not return home after each away game but instead travel from one away game to the next. No team plays more than three games in a row at home, or three games in a row away. In addition, no team plays the same opponent in two consecutive rounds. Thielen and Westphal [11] have shown the *TTP* to be strongly NP-complete.

Third, most professional sports leagues introduce many additional requirements in addition to minimizing breaks and travel distances. We call the problem of finding a schedule that *satisfies given constraints* [7] the Constrained Sports Scheduling Problem (CSSP). The goal is to find a feasible solution that is the most acceptable for the sports league owner – that is, secures the interest of media, TV network and fans, increases revenues for the owners/shareholders and increases transparency to secure fairness for the teams.

Table 1 shows a schedule for a compact double round robin tournament with six teams. The given schedule has many drawbacks if we consider it as a CSSP. For example, it has no breaks for team 5, but four-in-a-row home games for teams 2 and 4 and a four-game away tour for team 2. Note also, that the teams play against each other on consecutive rounds 8 and 9. These match-up games are called *back-to-back games*.

R1	**R2**	**R3**	**R4**	**R5**	**R6**	**R7**	**R8**	**R9**	**R10**
1 – 2	3 – 2	4 – 3	1 – 5	1 – 4	2 – 1	2 – 4	1 – 3	3 – 1	1 – 6
4 – 6	4 – 1	5 – 2	3 – 6	2 – 3	3 – 5	5 – 1	2 – 6	5 – 4	2 – 5
5 – 3	6 – 5	6 – 1	4 – 2	5 – 6	6 – 4	6 – 3	4 – 5	6 – 2	3 – 4

Table 1. A schedule for a compact double round robin tournament with six teams.

2 Interesting Sports League Models

This section presents some sports league models that differ from the standard round robin tournaments. The presented additional features make it an even more challenging task to generate an acceptable schedule. The features are classified to incomplete round robins, media and fan games, traveling issues and venue restrictions. The set of simplified test instances introduced in Section 3 is based on this classification.

Most of the features presented seek to increase the revenues for the owners. Therefore, it is very important to consider the fairness issues at the same time. At least the following should

be acknowledged:

- For each team and at any point in the tournament, the difference between home and away games played should be minimized.
- The difference in the number of games played between different teams should be minimized at any point in the tournament.
- The teams' number of games played should be as close to equal as possible at any point in the tournament.
- The preferred minimum for the number of rounds between two games featuring the same teams should be big enough.

2.1 Incomplete round robins

In an incomplete round robin some teams do not play against each other. The teams can still play a complete single round robin (1RR) or a complete double round robin (2RR). Only the last round robin is incomplete.

In the Australian Football League (AFL) eighteen teams play against each other once - i.e., a single round robin. In addition, each team plays five additional games. This adds up to 22 games, 11 home and 11 away games, for each team. The schedule consists of 23 rounds, and each team has one bye during rounds 11-13.

Building the schedule is a process comprising three phases. First, the home teams of the single round robin tournament and the five additional games are decided. The second phase includes building the actual schedule. In the last phase the exact weekdays and venues of the games are decided. These decisions are mostly constrained by various revenue and broadcasting issues and as well as venue requirements. Since each team meets five teams twice and twelve teams only once, this causes fairness, revenue and media issues mainly considering the decision of the home teams of the games. The schedule should secure fairness for the teams. It should also increase revenues for owners/shareholders by increasing the number of spectators and decreasing traveling costs. Finally, it should secure the interest of media, TV network and fans and optimize matches considering their needs. The problem is to find such home teams that all teams play under the same conditions. The league authorities define some rules for the selection of the five additional games, for example all the local teams and "big" clubs should meet each other twice. Kyngäs et al. [4] give a detailed discussion of scheduling the Australian Football League.

A minitournament is a series of games involving a subset of the teams in the league which are all played consecutively at the home venue of one organizing team. In games that do not feature an organizing team, both opponents play an away game. The main issue with minitournaments is to find a schedule such that at the end of the season, the number of home games is balanced over all teams. The main advantage of minitournaments is that they bring several matches to one venue and e.g. one weekend, which makes it easier for fans and players to enjoy the sport. They can see and play many different teams without having to travel several times. This is not only relevant in large geographical regions, but also for youth tournaments, amateur competitions, or wheelchair sports, where traveling may not be obvious, even for shorter distances.

In the Finnish National Youth Ice Hockey League fifteen teams play against each other twice - i.e., a double round robin. In addition, a third round robin is played introducing five minitournaments each with six teams. The minitournaments are played on weekends so that

each team plays on Friday, Saturday and Sunday. Therefore, the host team has three home games in a row; the other teams effectively have three away games in a row. The teams play 45 games in minitournaments, so 60 games still remain to be considered in the third round robin. The construction of the 3RR includes four main phases. First, we should find a series of minitournaments that balances the number of teams in common. Second, we should select the teams that will play against each other in the minitournaments, considering that two teams should not meet more than once in the minitournaments. In the third phase, we will settle the home advantage of each match. We should decide, for each pair of teams, which team will have a home game more than the other over their three confrontations. Finally, we should solve the generated 3RR problem. Nurmi et al. [8] give a detailed discussion of scheduling the Finnish National Youth Ice Hockey League.

2.2 Media and Fan Games

In Section 2.1. we described that the Australian Football League selects additional games to the schedule based on the interest of media and fans. Also the Finnish Major Ice Hockey League uses various game setups to secure the interest of media and fans. The fourteen teams of the league play a quadruple round robin tournament. In addition, each team plays eight additional games totaling 60 games per team.

The teams are divided into two groups of seven teams which play a single round robin tournament. The so-called "January leveling" adds two extra games for each team. In January, in the middle of the season, the last team on the current standings selects an opponent against which it plays once at home and once away on two consecutive days on Friday and on Saturday. The opponent selects the day for its home game. Then, the second last team (or the third last if the second last was selected by the last team) selects its opponent from the rest of teams and so on. The teams can choose to select their opponents either by maximizing the winning possibilities or by maximizing the ticket sales.

The schedule should maximize the number of Saturday games between local rivals. The schedule should also include a weekend when seven pairs of local rivals play against each other on consecutive rounds on Friday and on Saturday, i.e. the home teams of Friday become the away teams on Saturday. These match-up games are called "back-to-back games". Finally, each team must play at least one Friday or Saturday home game against the two "big clubs". Nurmi et.al. [9] give a detailed discussion of scheduling the Finnish Major Ice Hockey League.

The league accepted a new team for the 2015-2016 season. The league continues to play a quadruple round robin now with the 15 teams. The 4RR adds up to 56 games per team. The teams are divided into five groups of three teams. Each group plays a 2RR thus ending up with the same 60 games per team as in the previous seasons. The 2RRs will be played as back-to-back games within eight days: Friday, Saturday, Tuesday, Wednesday, Friday and Saturday.

2.3 Traveling issues

Many leagues have to consider traveling issues due to the distances of the teams. For example, in Australian Football League the teams from Western Australia have to travel about 70,000 kilometers per season. Unfortunately, only to a certain extent can be done to reduce the total traveling distance since no away tours (two or more consecutive away games) can be arranged due to a regular game being scheduled in each city per week.

In the Finnish Major Ice Hockey League some teams prefer to play two away games in a row. This has a serious drawback. We cannot schedule an away tour to the rounds in the

consecutive days since that would imply two home games in a row for another team, at least not, if we want to maximize the number of compact rounds. Therefore, the more away tours we have the more rounds we have. It is obvious that the more we introduce away tours the more we will have breaks in the final schedule. Actually, traveling issues very often compete with the smooth flow of the total schedule. This is the key point in the Traveling Tournament Problem presented in Section 1.

2.4 Venue Restrictions

The break minimization is distracted not only by away tours but also by venue restrictions. Venue restrictions have a major impact to the difficulty of generating the schedule. For example, in the Finnish Major Ice Hockey League the usual number of venue restrictions is close to one hundred.

Mainly four issues can prevent the team playing a home game. First, the team can share a home venue with another team. Second, two teams cannot play at home on the same day because they share the same (businessmen) spectators. Third, the team might not want to play at home if another team in the same region plays at home in another league. Finally, the team cannot play at home if the venue is reserved for some other event, such as rock concerts. A venue can also be under renovation which can delay the start of the teams home games.

3 A Set of Simplified Test Instances

We introduce a set of simplified test instances based on the different league features and issues discussed in Section 2. We invite the sports scheduling community to try their methods on these instances. The basic problem is to minimize the number of breaks for a compact 2RR with 14 teams. As noted in Section 1, it is possible to construct this schedule with 12 breaks.

The basic problem is complicated by introducing the following constraints:

1. There must be at least six rounds between two games with the same opponents.
2. Team #14 cannot play at home in rounds divisible by three: 3, 6, 9, 12, , 30.
3. The schedule should include six away tours for team #1. The team should meet teams #2 and #3, #4 and #5, #6 and #7, #8 and #9, #10 and #11, and #12 and #13 in consecutive rounds away.
4. Teams #4 and #6 and teams #9 and #11 cannot play at home at the same time.
5. Seven extra games must be played: 1-2, 3-4, 5-6, 7-8, 9-10, 11-12 and 13-14.

The problem is to find the minimum number of breaks satisfying any combination of these constraints. This gives us 23 different test instances. The best solutions found are available online [6]. We solved the instances using the PEAST algorithm [5] which is used to generate the schedule for the Finnish Major Ice Hockey League.

References

[1] D. De Werra. Scheduling in sports. In Amsterdam and Hansen, editors, *Studies on graphs and discrete programming*, pages 381–395, 1981.

[2] K. Easton, G. Nemhauser, and M. Trick. The traveling tournament problem: description and benchmarks. In *Proc of the 7th. International Conference on Principles and Practice of Constraint Programming*, pages 580–584, 2001.

[3] S. Knust. Sports scheduling bibliography. [online]. `http://www.inf.uos.de/knust/sportssched/sportlit_class/`, (Last update 02.04.2015).

[4] J. Kyngäs, K. Nurmi, N. Kyngäs, G. Lilley, and T. Salter. Scheduling the Australian football league. In *Proc. of the 10th Conference on the Practice and Theory of Automated Timetabling (PATAT)*, pages 309–317, 2014.

[5] N. Kyngäs, K. Nurmi, and J. Kyngäs. Crucial components of the PEAST algorithm in solving real-world scheduling problems. In *Proc of the 2nd International Conference on Software and Computer Applications*, 2013.

[6] K. Nurmi. Sports scheduling test instances. [online], 2015. `http://web.samk.fi/public/tkiy/SS/`, (Last access April 2015).

[7] K. Nurmi, D. Goossens, T. Bartsch, F. Bonomo, D. Briskorn, J. Duran, G. Kyngäs, J. Marenco, C.C. Ribeiro, F.C.R. Spieksma, S. Urrutia, and R. Wolf-Yadlin. A framework for scheduling professional sports leagues. In Sio-Iong Ao, editor, *IAENG Transactions on Engineering Technologies*, volume 5. Springer, USA, 2010.

[8] K. Nurmi, D. Goossens, and J. Kyngäs. Scheduling a triple round robin tournament with mini-tournaments for the Finnish national youth ice hockey league. *Journal of the Operational Research Society*, 65(11):1770–1779, 2014.

[9] K. Nurmi, J. Kyngäs, D. Goossens, and N. Kyngäs. Scheduling a professional sports league using the PEAST algorithm. In *Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists*, pages 1176–1182, 2014.

[10] R.V. Rasmussen and M. A. Trick. A Benders approach for the constrained minimum break problem. *European Journal of Operational Research*, 177(1):198–213, 2007.

[11] C. Thielen and S. Westphal. Complexity of the traveling tournament problem. *Theoretical Computer Science*, 412(4–5):345–351, 2011.

Assessing the effect of player ability on scoring performance for holes of different par score in PGA golf

Bradley O'Bree, James Baglin, and Anthony Bedford

RMIT University, Melbourne, Australia
bradley.obree@rmit.edu.au, james.baglin@rmit.edu.au, anthony.bedford@rmit.edu.au

Abstract

This work looks to serve as foundation research into the effect scoring performance on holes of different pars has on corresponding final round scores in PGA golf. When conducting predictive modelling to estimate round scores it can be argued that each of the different hole structures, par threes, fours and fives, should be player dependent because players can have different strengths and weaknesses. The purpose of this study was to investigate the degree to which performance is player dependent by analysing the variation of par totals (the total score for all hole scores for each par type) using multilevel regression modelling. An intercept only multilevel regression model allows the decomposition of the variance of scores into two parts, a par total based and a player based. The separate components are used to calculate the Intraclass Correlation Coefficient (ICC), which describes the proportion of variation in par totals that can be attributed to the player. Smaller ICC values indicate smaller player components and imply performance does not vary across players. The complete 2014 US PGA tour was analysed, with results indicating that when all players were considered together that each ICC was significant ($p<.00$), however each relatively small at .012, .046 and .055 for par threes, fours and fives respectively. Controlling for player strength showed clear differences in ICC estimates and levels of significance, with a consistent trend of decreasing player strength associated with increasing ICC values and improved statistical significance. The latter result indicates that the proportion of variation in scores explained using the player component increases as player skill decreases, suggesting it is more important to consider player ability at the bottom of the leaderboard than at the top. These results suggest that when modelling par totals, and by extension round scores, some level of player-specific modelling should be incorporated.

1 Introduction

The initial motivation to analyse player-based components of golf score prediction was the result of previously completed research into simulation of tournament outcomes. Appropriate and accurate simulation inputs require player based score simulation, and as a result we need know to what degree a model input should be influenced by player-based characteristics. Different players will have different strengths. Players who can hit further would be more likely to perform better on longer holes, such as par fives. So it is logical to account for different strengths when simulating hole scores, or any scores for that matter. One way to assess if player based measures should be included is to determine whether some characteristic of scores are dependent on player based measures.

The aim of this research was to explore this concept by analysing the variance of score measures and attempting to identify where player-based influence is non-zero, hence justifying including player-based characteristics. We take a simple approach so as to not overcomplicate data calculation and focus on identifying trends. We don't want to standardise scores, make

adjustments, or introduce anything of a complex measure of player ability. The player-based score metric in question is par total (PT), the total score for all hole scores for each par type. For each standard tournament round we observe three PTs, one each for par threes, fours and fives. Analysis was restricted to tournaments from the 2014 PGA Tour season.

2 Methods

2.1 Multilevel Linear Regression

Multilevel analysis is applied to data structures that have a clear hierarchy, meaning outcomes and their predictors can be separated into distinctly different levels. In this case, we can separate the entire database of scores into a par total score level within a player level. For each player we can conduct a regression analysis to determine predicted PT values. The benefit in applying multilevel analysis is that these regressions can be analysed concurrently across players within the player-level to describe general results. We begin with a basic intercept-only regression model to predict PT for individual players.

$$Y_{ij} = \beta_{0j} + e_{ij} \tag{1}$$

Where Y_{ij} is the PT for tournament round i for player j, β_{0j} is the player-level intercept, and e_{ij} is the PT-level error term.

Equation 1 shows how we structure the intercept-only model, which determines a unique constant for the player of focus. Multilevel analysis allows us to concurrently assess this basic model across all players.

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij} \tag{2}$$

Where γ_{00} is the intercept estimate across all players, and u_{0j} is the player-level error term.

Essentially we graduate from player-specific models to a general model that now determines player-level errors from a global intercept estimate, as opposed to unique intercept estimates for each player. The intercept-only model does not explain any variance in Y[2], in this case, par total. It only decomposes the variance of the errors into two independent parts, player-level variance $\sigma^2_{u_0}$ and PT variance σ^2_e. The sum of the two components gives the total variance in PTs, and as a result can be used to calculate the Intraclass Correlation Coefficient ρ (ICC), which describes the proportion of variation in PT that can be attributed to the player (Equation 3).

$$\rho = \frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_e} \tag{3}$$

Smaller ICC values indicate smaller player components and imply performance does not vary across players. Reporting the ICC is the simplest way to show the player-based contribution to score variance.

	Variance Estimate		
Hole Par	Intercept	Residual	ICC
Three	2.11	0.03	0.01
Four	19.25	0.94	0.05
Five	18.31	1.06	0.06
All Holes	10.14	1.70	0.14

Table 1: Estimates for intercept and residual variance, with corresponding ICC values

2.2 Grouping Players

To measure how the ability of the player can influence the player-based effect on variance we needed to devise a way of grouping players. We must be careful to not enforce a metric that is too complex, because it is not necessary to observe trends and itself would require statistical justification outside of the scope and purpose of this paper. We used a very simple metric of player ability, as determined by overall performance during the 2014 PGA Tour season, where players were ranked using a function of their tournament round scores throughout the year. For each round, we took the ratio of a player's round score to the average round score across the field. We did this to attempt to account for variations in course difficulty, playing field skill and individual day effects such as poor weather conditions, while keeping the metric itself in a somewhat standardised form. Of course this does not completely capture the external effects, but the simplicity outweighs the crudeness of the metric for its use in this research. Taking the full list of competing players for the season, we ranked players by average score ratio across the entire season. We must note that this method can not account for fluctuations in player form, it only ranks players based on a broad measure of performance across the entire season. Players were ranked and split into 10 groups of equal sample size, where the number of groups was chosen arbitrarily.

3 Results

The results comprise of two parts, the first focussed on analysing PT variance independently of player ability, followed by an analysis dependent on the player ability metric mentioned in Section 2.2.

3.1 Neglecting Player Ability

Table 1 summarises the regression results for PT for par threes, fours and fives, as well for total round scores, where all players were considered together. Note, in each case the ICC was significant ($p<.00$). There were three separate findings from the initial analysis that provide valuable insight.

1. The size of the variance components for total round scores was clearly less than the sum of the components for separate par totals.
   This shows there is no easy transition between PT and total round variance, where one could assume the latter would comprise the sum of the former.
2. The ICC values increased in size as the hole par for each PT increased.
   This result indicates player-specific characteristics of performance become more prominent as the hole length increases. One could assume this stems from the fact that the longer

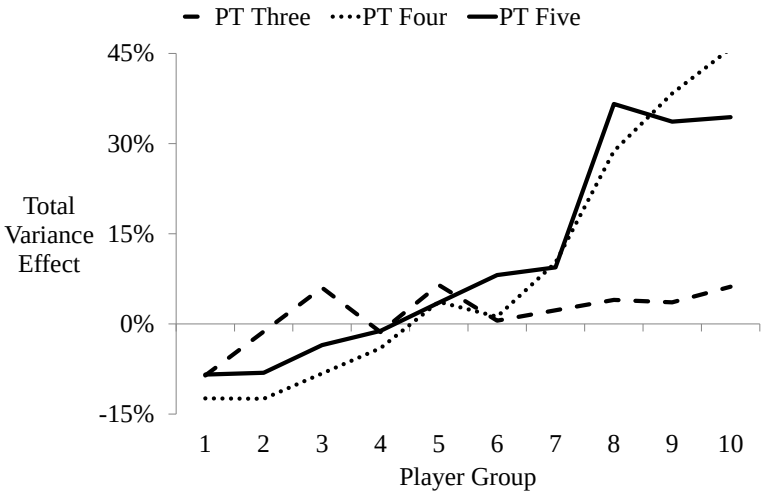


Figure 1: The effect of player ability on total PT variance

holes by default require more shots to be played making it more likely differences in ability would become clearer when comparing scores.

3. The ICC value for total round score is clearly larger than the ICC for any PT.
   If we assume that total round score is in some way always a function of corresponding PT components, this result would suggest it is more appropriate to account for player-specific characteristics at the round score level. The first result however provides evidence against such an assumption.

3.2 Integrating Player Ability

Before we can measure the effect of player ability on variance components and ICC values, we must first identify how the variance of par totals changes as player ability varies. Identifying if a trend exists is important because it can provide a context to the findings. Figure 1 shows the difference between observed PT variance by player group and the average variance across all groups. It is clear that the total variance increases as the player group increases, which follows naturally from the fact that worser players are more likely to have greater variation in scores.

Table 2 displays the ICC values for each player group. We note in some cases that the final Hessian matrix was not positive definite, despite convergence criteria being met. This typically indicates a linear dependency between predictors, or a large difference in the scale of predictors, which are remedied through predictor removal and variable standardisation respectively. In this situation however we are dealing with simple intercept-only models, so such a result suggests the variance of the random effect (player-level) is zero.

Controlling for player strength showed clear differences in ICC estimates and levels of significance, with a consistent trend of decreasing player strength associated with increasing ICC values and improved statistical significance. This improvement in ICC strength comes despite

	Group Number									
Hole Par	1	2	3	4	5	6	7	8	9	10
Three	0.01^{A}	$0.00^{\#}$	0.00	0.00	0.00	0.00	$0.00^{\#}$	0.05^{B}	0.04	0.12^{A}
Four	0.00	0.00	$0.00^{\#}$	0.01	0.02^{A}	0.04^{A}	0.07^{A}	0.20^{B}	0.32^{B}	0.38^{B}
Five	0.00	0.01	0.01	0.02^{B}	0.08^{B}	0.11^{B}	0.11^{B}	0.31^{B}	0.43^{B}	0.34^{B}

A - Significant at the 0.10 level. B - Significant at the 0.05 level.
Final Hessian matrix was not positive definite, though convergence criteria were satisfied.

Table 2: ICC values by player group

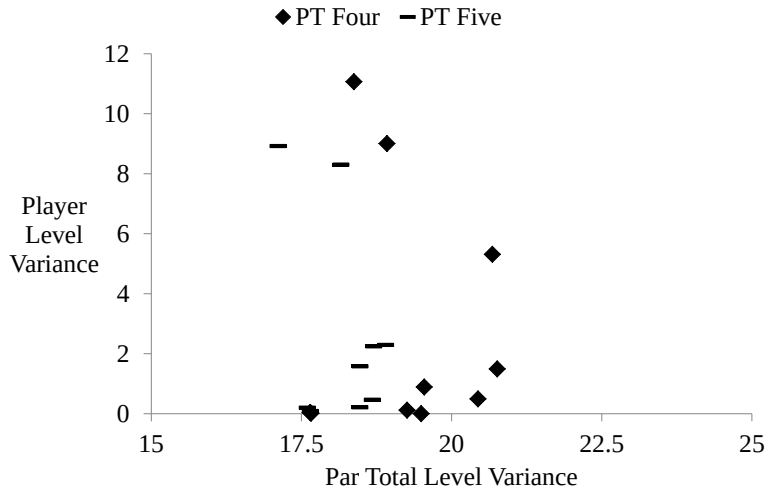


Figure 2: Scatter plot of player-level variance by PT-level variance

the size of total PT variance increasing with decreasing player strength. The latter result indicates that the proportion of variation in scores explained using the player component increases as player skill decreases, suggesting it is more important to consider player ability at the bottom of the leaderboard than at the top.

As an aside to the ICC tables, we've included a simple scatter plot (Figure 2) to attempt to show a relationship between ICC values and the actual size of the variance components. Were ICC values independent of the size of the variance components the plot would form a diagonal. The fact it doesn't suggests the trends observed in Table 2 can be attributed to player strength rather than larger variance components, despite there being a clear association between decreases in player strength and relative increases in total PT variance.

4 Discussion and Future Work

The objective of this study was to determine whether the variation in golf scores can be attributed to individual player characteristics. The results indicated that while ICC values were low, they were significant and certainly non-zero. While the values for individual par totals were clearly less than the value for round scores, we can still justify looking to predict par totals when simulating tournament outcomes because there is value in modelling player ability individually. These results don't give an indication into how well a score model that includes player-based predictors would perform, they simply demonstrate that score variance, and therefore some explanatory power, can be found by partitioning the variance in observed player scores.

When a simple metric for grouping players based on apparent ability was included in the analysis it demonstrated clearly that scoring performance varies dramatically between players. The identified trends indicated the partitioning of variance is somewhat dependent on player ability. The exploratory nature of this work meant we introduced only a crude, simple metric, but we are confident when more methodical, thorough metrics are included the effects will become more pronounced.

For this research to persevere in the future we would need to incorporate standardised measures of player scores. Findings from this research provide evidence that score variation is dependent in part on player-specific characteristics. They should however should be considered preliminary given golf courses have different course pars, meaning scores should be standardised, and we use a crude method for assigning players a metric for player strength. Nevertheless the analyses presented in this work provide a foundation from which to develop future research into player effects on score prediction.

5 Conclusion

These results suggest that when modelling par totals, and by extension round scores, some level of player-specific modelling should be incorporated.

References

[1] Ronald H. Heck, Scott L. Thomas, and Lynn N. Tabata. *Multilevel Modelling of Categorical Outcomes Using IBM SPSS*. Routledge, New York, NY, 2012.

[2] Joop J. Hox. *Multilevel Analysis: Techniques and Applications*. Routledge, New York, NY, 2010.

[3] J. W. R. Twisk. *Applied Multilevel Analysis: A practical guide*. Cambridge University Press, Cambridge, UK, 2010.

Can betting in PGA golf be profitable? An analysis of common wagering techniques applied to outright winner market prices

Bradley O'Bree and Anthony Bedford

RMIT University, Melbourne, Australia
bradley.obree@rmit.edu.au, anthony.bedford@rmit.edu.au

Abstract

Golf is a difficult sport to master in the physical domain, and it is possibly even harder to accurately model tournament outcomes. Betting markets for golf contain the same variety of exotic outcome wagering seen in most sports, but one shouldn't focus on these sources of complexity when even the simplest outcome, who will win the tournament, is so tough to predict. The objective of this work was to demonstrate that making a profit is difficult and explain why this is the case. Using publicly available outright winner market prices from Bet365.com for seasons 2012, 2013 and 2014 of the US PGA tour, the analysis was separated into two parts. Initially the focus was exploratory, looking at the general aspects of the data to gain insight into how prices are set. Following this, two common wagering techniques, Kelly and fractional Kelly criterion systems, were examined to determine whether they can be used to make a profit. Further, regardless of the level of success of their application, determine why it was the case for each system. We report the metrics typically used to describe each system's effectiveness, such as Return on Investment and prediction strike rates, where relevant. Results indicated that the systems considered were not profitable for a variety of reasons. Rather consistently, the major problem existed in the poor win conversion for players within striking distance of the current leader.

1 Introduction

Golf is generally considered amongst sports fans a difficult sport to become expertly skilled in. Data analysts too can speak about the issues involved in modelling outcomes and attempting to predict the winner. Bookmakers can take advantage of the volatility of player performance also by setting markets with, in a sporting context, unusually large overround figures. In this work we look to conduct a small exploratory analysis into the complexities that need to be considered when betting in golf. We use the common knowledge Kelly system to establish a basic betting system, report performance metrics and demonstrate where difficulties can arise when striving for betting success.

2 Methods

2.1 Data

We used two forms of data in this research. The first was tournament outcome data, where for each competing player we knew their rank and shots behind the leader following each round, and the corresponding final rank for the tournament. This player data spanned all regular match play tournaments from the 2007 through 2014 PGA Tour seasons. The second was outright win market prices obtained from Bet365.com for seasons 2012 through 2014 prior to

	Samples (Round)				Overround (Round)			
Season	1	2	3	4	1	2	3	4
2012	9	6	6	8	139%	139%	129%	124%
2013	28	24	29	26	145%	137%	129%	122%
2014	41	40	39	38	141%	136%	127%	120%
All Seasons	78	70	74	72	142%	137%	128%	121%

Table 1: Sample size and average overround by round and season

the commencement of each round of a tournament. The market price data does not span all tournaments for each season. An important consideration when looking at market prices is the overround across a set of round prices. Overround is calculated as the sum of the reciprocal of all prices in a set, and represents the edge the bookmaker has, where the expected profit to be made is equal to the $(Overround - 100)\%$ of the total market volume (in the long term). The overround for these sets of round prices is rather large, and summarised by season in Table 1.

2.2 Calculating Winning Probabilities

The approach we took in this research was to place bets on each player who had a current rank of no higher than 5, or was no more than 4 shots behind the current leader. This was to ensure we only placed bets on players who realistically had a chance of winning. Further, market prices typically are not provided without direct request for players outside of these ranges. The amount wagered was a function of the winning probability assigned to the player through empirical outcome distributions and the corresponding market price for that player.

2.2.1 Empirical Probability Distributions

We use two separate methods of determining the probability a player will win the tournament. The first uses the current player ranks, and the second uses the total shots behind the current leader, as displayed in Figure 1 and Figure 2 respectively. We use both the rank and absolute shot difference methods because we are ignoring the specific abilities of the competing players for simplicity. The empirical probability distributions are generated using data from seasons 2007 through 2011.

2.3 Wagering Techniques

We use two methods for determining wager size, the Kelly Criterion and the Fractional Kelly Criterion. The Kelly Criterion is a strategy used to determine the size of bets based on believed win probability to maximise the growth rate of a pot over the long-term. The equation output, described in Equation 1, represents the fraction of the total pot to bet on any given wager. Fractional Kelly is a variation where only a fixed fraction of the Kelly Criterion output is wagered. The advantage in doing so is that it further protects the better from going bust due to streaks of lost bets and in general reduces volatility in pot fluctuations.

$$F = p - \frac{q}{c} \tag{1}$$

Where F is the pot fraction to bet, c is the market price to win, p is the probability of winning, and $q = 1 - p$.

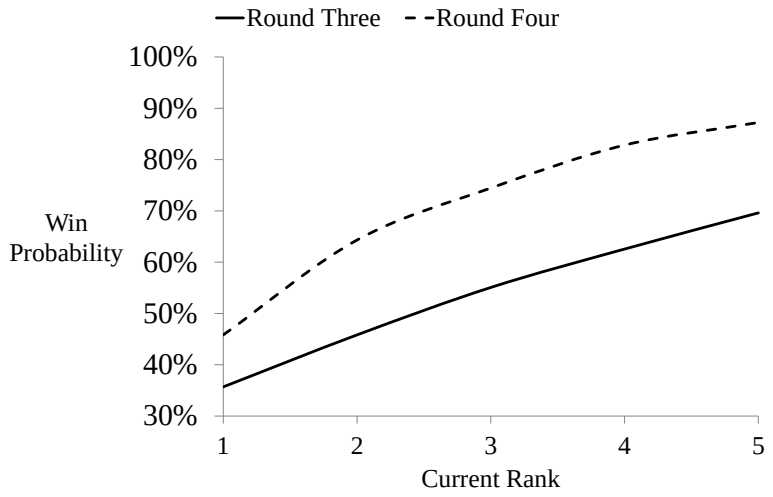


Figure 1: Tournament win probability by current rank pre rounds three and four

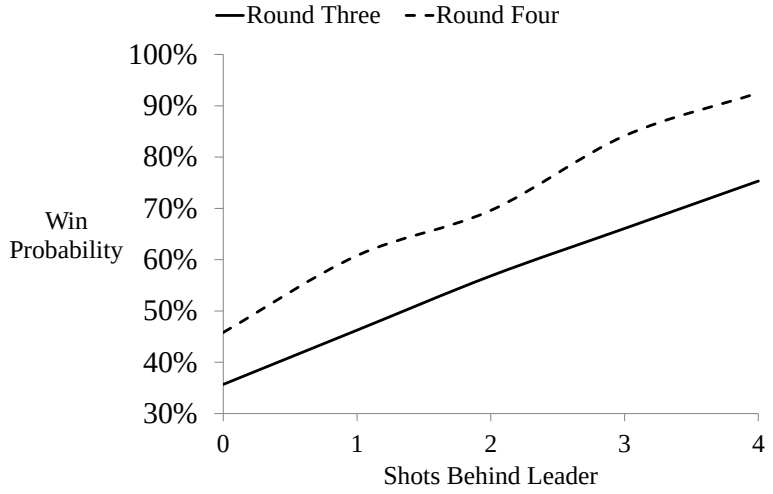


Figure 2: Tournament win probability by current shots behind leader pre rounds three and four

	Rank					Shots Behind				
Metric	1	2	3	4	5	0	1	2	3	4
Samples	91	129	157	175	184	244	411	562	866	0
ROI	16%	-25%	-18%	-15%	-12%	-22%	-42%	-52%	-68%	-
Strike Rate	34%	7%	6%	4%	3%	35%	7%	4%	3%	-

Table 2: Betting performance metrics ROI and strike rate, with corresponding number of bets placed

3 Results

The results are divided into two separate sections. The first analysis looks at the performance of the standard Kelly Criterion system, while the second focusses on the influence the Fractional system had on the performance metrics.

3.1 Standard Kelly

The performance of the system was seen to be poor for virtually all possible betting scenarios that have been included in the analysis. ROI statistics were negative for each scenario, with the exception of betting on the current (or joint) leader. Predictive strike rates were very poor for all scenarios, again with the exception of betting on the (current or joint) leader. There is little doubt the poor predictive strike rates provide reasoning as to why ROI metrics were negative. Noticing the number of bets placed by scenario, we gain an understanding as to why system performance was poor. It is likely this is a result of too many players being within striking distance of the leader who never actually win, attracting bets because good value market prices are provided. The summary of performance metrics is provided in Table 2.

3.2 Fractional Kelly

The ability to control the size of bets has the benefit of potentially limiting the variability of return results. This has been the case in this situation, where ROI statistics are consistently negative where the imposed fraction of the Kelly equation output is varied between 0.01 and 1. We have arbitrarily chosen four of the betting scenarios to display in Figure 3.

4 Discussion and Future Work

The objective of this work was to show the difficulty involved in betting successfully in professional golf using a simple Kelly betting system. The system was rather unsuccessful, with results attributing much of the lack of success to the large number of players within striking distance of the leader, which makes selecting the winner difficult.

There is however a need to consider the efficiency of the betting markets in conjunction with these results. Typically overround margins are less than 7% across an entire market in the sports realm, but in the case of the Bet365 golf market prices these margins were substantially greater. Given the margin represents the edge of the bookmaker we are essentially fighting a losing battle. There is so much volatility within performance and outcomes in a golf tournament that it is simply too difficult to predict with certainty, or to bet with success.

The progression of this research would benefit from including market prices from other bookmakers, given their overround margins may differ. It would also benefit from exploring

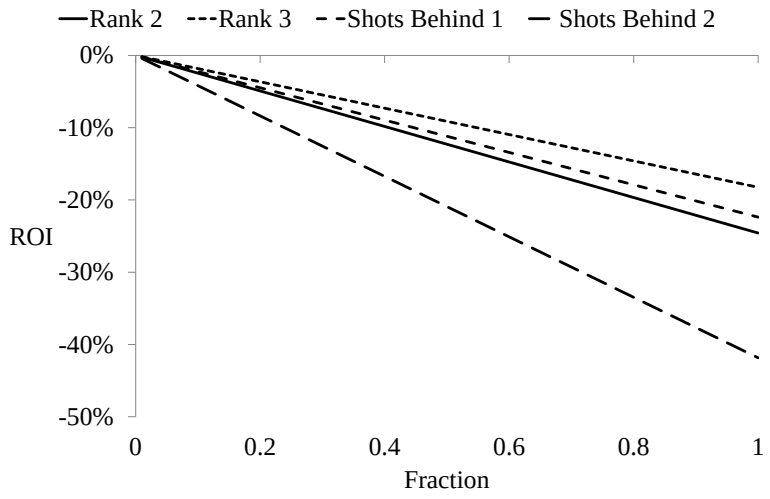


Figure 3: The impact of the Fractional Kelly system on ROI

live betting. This work only considers betting between rounds, in particular prior to rounds three and four. Perhaps the value in betting in golf lies within live betting, where physical factors can be assessed and included in the system, rather than simple scores at round by round intervals.

5 Conclusion

Findings from this work indicate that a basic application of the Kelly Criterion to wagering in PGA Tour golf is unsuccessful using win probability distributions generated using empirical outcome data.

References

[1] J.L. Kelly. A new interpretation of information rule. *Bell System Technicial Journal*, 35, 1956.

[2] Edward. O. Thorpe. The Kelly criterion in blackjack, sports betting, and the stock market. In *The 10th International Conference on Gambling and Risk Taking*, Montreal, June 1997.

Assistive Sports Video Annotation: Modelling and Detecting Complex Events in Sports Video

Aled Owen[1], David Marshall[1], Kirill Sidorov[1], Yulia Hicks[1], and Rhodri Brown[2]

[1] Cardiff University, Cardiff, UK
[2] Welsh Rugby Union

Abstract

Video analysis in professional sports is a relatively new assistive tool for coaching. Currently, manual annotation and analysis of video footage is the modus operandi. This is a laborious and time consuming process, which does not afford a cost effective or scalable solution as the demand and uses of video analysis grows. This paper describes a method for automatic annotation and segmentation of video footage of rugby games (one of the sports that pioneered the use of computer vision techniques for game analysis and coaching) into specific events (*e.g.* a scrum), with the aim to reduce time and cost associated with manual annotation of multiple videos. This is achieved in a data-driven fashion, whereby the models that are used for automatic annotation are trained from video footage. Training data consists of annotated events in a game and corresponding video. We propose a supervised machine learning solution. We use human annotations from a large corpus of international matches to extract video of such events. Dense SIFT (Scale Invariant Feature Transform) features are then extracted for each frame from which a bag-of-words vocabulary is determined. A classifier is then built from labelled data and the features extracted for each corresponding video frame. We present promising results on broadcast video for a international rugby matches annotated by expert video analysts.

1 Introduction

Video analysis in sport is a growing and important field within the professional sporting environment, providing varying levels of assistance and insight depending on the sport or even the team. However, sports video analysis is currently a laborious manual process. This drastically limits the volume and quality of annotation, especially when considering the number of games that are played at a professional level. Other means of instrumentation are available at a professional level, such as relative GPS tracking. However, these are invasive (affixing physical devices to the players). These other sources often provide higher fidelity data but their invasiveness limits data capture in team based sports to only a home teams' players, thus limiting the information that you can learn from the game. This is especially true in sports such as rugby where accurate measurement of relative positions of players and constellations of players is crucial to the game's analysis.

Computer Vision in sports analysis, compared to many other fields, is in its relative infancy. Computer vision techniques have, however, been incorporated in an assistive role into broadcast packages, such as the BBC's usage of the pitch tracking to overlay pitch information [7].

Other efforts in sports analysis have focused on the tracking and identification of individual players to varying levels of success and largely fall into two categories according to the model of the camera motion. The Scale Invariant Feature Transform (SIFT) [4] has shown promise in distinguishing players form both fixed cameras [3] and from a pan-tilt-zoom camera [5]. However, in both cases these methods are applied to sports that lack the complex occlusions and structures which are present in a rugby game.

Figure 1: Images before (left) and after (right) background subtraction.

The challenge addressed in this paper is the automatic classification of video events from rugby footage. We have annotated data from past rugby matches where key events such as *scrum*, *lineout*, *ruck*, *maul*, *etc.* have been labelled. The main contribution of this paper is method to take previously unseen footage and automatically annotate each frame as one of these classes.

2 Method

In order to achieve the above classification of rugby events we have devised a pipeline that comprises the following stages:

Background Removal: The input video is broadcast footage which often contains imagery of the crowd, advertising hoardings and other periphery. Removal of such imagery is required as features we extract and use in the later classification process are confounded by their presence. The pitch is relatively easy to robustly detect as it has a constant colour and texture. We, therefore, detect the pitch area and remove all objects not bounded by it. Following this, the pitch itself is removed leaving only the players.

In order to detect and remove the pitch we employ a technique similar to chroma-keying to remove the green component from the image. Our method here closely follows the approach of [7] for pitch tracking over multiple frames for broadcast purposes. Within each image we histogram the hue values of pixels and remove the pixels that belong to a small neighbourhood around the largest peak of the histogram. This technique is sufficient for our purposes even though it fails to remove other artefacts, such as sponsor logos on the pitch, we are not interested in and have no effect on subsequent classification. Examples of the background removal are shown in Figure 1. An exemplar histogram of the hue values in a typical image is shown in Figure 2, with a distinct peak in the green section of the colour space.

As the peak is so distinct within the colour space no advanced technique is employed to extract the relevant area. Instead a simple walk from the maximum point is used, terminating when a point of inflection is detected. Once the area determined to be the pitch is decided, pixels outside of the convex hull of the pitch are also rejected. This ensures that the crowd is removed that would not be detected by the chroma-keying.

PHOW Image Description: The next stage is to compute a suitable description of the shapes of the scrums *etc.* that has sufficient discriminative power for classification. We have explored a variety of image description techniques, initially starting with the popular SIFT feature descriptor and, in particular, the pyramid histogram of words (PHOW) [1]. It essentially

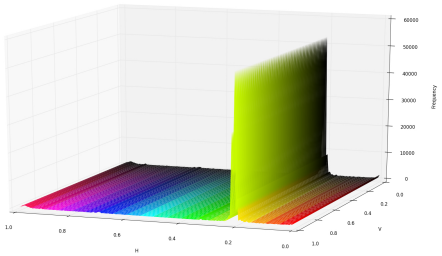


Figure 2: Histogram of hue values within the HSV colour space. A fixed saturation value of 1 is chosen for visualisation.

uses dense SIFT descriptors over a variety of scales to achieve higher scale independence. Each video frame is described in terms of these PHOW features. We then cluster these features into a fixed number of "visual words", as in [1, 6]. This quantisation yields a compact summary which is also more invariant to noise. In our experiments, we use 300 visual words to describe the scene in order to ensure we are not too coarsely quantising the large number of SIFT features that we are computing, this is a similar number to those used classifying on the Caltech 101 dataset [2].

Once a suitable vocabulary is generated, the visual words are then spatially binned to produce a spatial histogram of the given image. Using a coarse 2×2 grid within the image, we aim to describe the key regions of the scene, while minimising the number of spatial bins used to increase the pipeline's ability to deal with minor changes in scale that cover the majority of eventualities occurring in the data. A diagram summarising the key stages of the PHOW pipeline is shown in Figure 3.

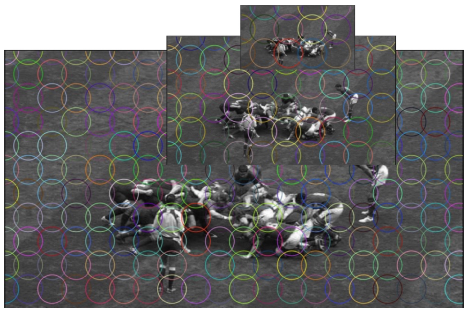
Frame-by-Frame Event Classification: Having produced a feature description of each frame we finally proceed to classification. We use a Support Vector Machine (SVM) in order to find the separating plane(s) between the classes. The SVM has become a popular workhorse supervised classifier in computer vision. We have observed a relatively low performance of a classic linear SVM which suggests that our feature vectors are not linearly separable. We therefore make use of an explicit feature map using the χ^2 kernel as described in [8]. This combined approach of using explicit feature maps and linear SVMs is more efficient than using a non-linear SVM for the large corpora of available data.

3 Results

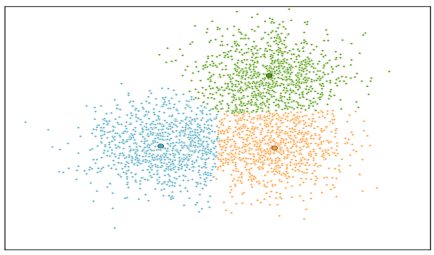
We have a performed an experimental validation of our approach using broadcast footage as described below.

Video Footage Data To train and test the classifier we took a collection of annotated games from the 2012 Six Nations rugby tournament. This data was annotated with high level labels showing the timestamps of scrums and lineouts by expert rugby video analysts.

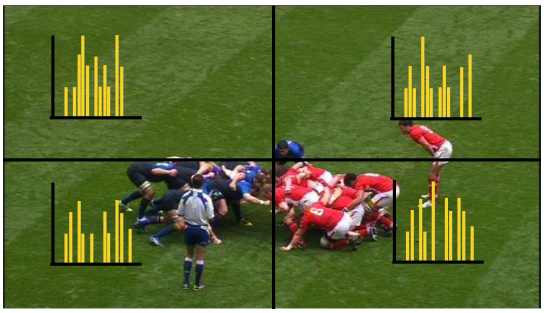
(1) Original Image



(2) Scale Space Pyramid and Computation of Features



(3) Clustering into Visual Words



(4) Computation of Spatial Histogram

Figure 3: Example progress through the PHOW pipeline highlighting key stages.

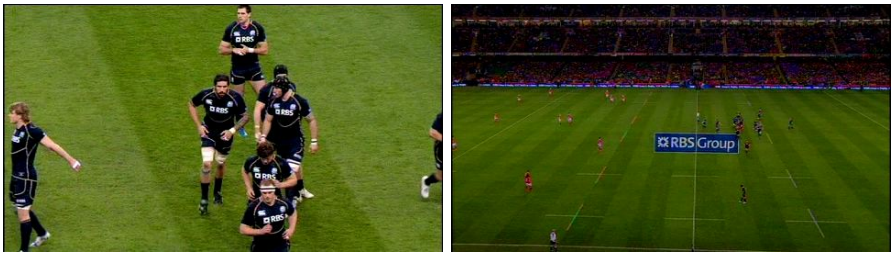


Figure 4: Example frames from a typical footage, close zoom (left) and far zoom (right).

In our preliminary experiments (the focus of this work) we focused on two games, Wales V Scotland for training and Vales V France for testing. Due to the broadcast nature of the footage, the game is captured from several cameras. However, we use only footage from a single camera that covers the primary action rather than focusing on specific players or wide-angle views (see Figure 4).

This camera zooms tightly to the specific events on the pitch at a raised angle. Hence, it is likely that to a certain extent our classification is influenced by the actions of the cameraman and how they choose to frame the shot, although from our experience cameramen are fairly consistent in framing the action across games.

Experiment We have a carried a basic experiment to validate our approach. From the video data we extracted all examples of scrums and lineouts. The training data was exclusively extracted from the Scotland game. The test data was obtained from the France game. Features were extracted as described above and a classifier built to discriminate between the two classes: scrums and lineouts. There were 750 frames per class (1500 in total) in the training set and 200 frames for each class in the test set, randomly chosen from the entire corpus. The results in Table 1 demonstrate the performance characteristics of the trained classifier.

	True	False
Positive	115 (29%)	0 (0%)
Negative	200 (50%)	45 (11%)

Table 1: Classification performance.

These are encouraging, showing an 88.75% classification accuracy overall. Close examination of failure cases revealed that some misclassifications are a result of coarse annotations where breaks or resets of actions (e.g. when a scrum is reset) are not clearly delineated in normal class boundaries. They actually look more like other plays (e.g. broken play).

4 Conclusion and Future Work

We have developed a practical solution to a novel problem of rugby event classification that has great potential to automate the current manual analysis process. Our current method does not take advantage of any temporal information. We intend to incorporate this in the future. Some initial investigation in using temporal models such Hidden Markov Models has shown promise. More refined domain specific features are also under investigation.

References

[1] Anna Bosch, Andrew Zisserman, Xavier Mu, and Xavier Munoz. Image classification using random forests and ferns. In *IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[2] Li Fei-Fei. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

[3] Haopeng Li and Markus Flierl. SIFT-based multi-view cooperative tracking for soccer video. *Speech and Signal Processing*, pages 1001–1004, 2012.

[4] David G Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[5] Wei-lwun Lu, Jo-anne Ting, James J Little, and Kevin P Murphy. Learning to track and identify players from broadcast sports videos shot segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1704–1716, 2013.

[6] Joseph Sivic and Andrew Zisserman. Video Google: a text retrieval approach to object matching in videos. In *IEEE International Conference on Computer Vision*, pages 1470–1477, 2003.

[7] Graham Thomas. Real-time camera tracking using sports pitch markings. *Journal of Real-Time Image Processing*, 2(2-3):117–132, 2007.

[8] Andrea Vedaldi and Andrew Zisserman. Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(3):480–492, 2012.

A new algorithmic methodology for assessing, and potentially predicting, racing cyclists' fitness from large training data sets

Andreas Petrides, Anthony J. Purnell and Glenn Vinnicombe

Department of Engineering, University of Cambridge, United Kingdom
ap650@cam.ac.uk, ajp203@cam.ac.ukgv@eng.cam.ac.uk

Abstract

This paper introduces a new algorithmic methodology for assessing, and potentially predicting, the fitness of an athlete in sports where large training data sets containing exogenous power and endogenous heart rate are available. The proposed algorithm seeks, in the first instance, to provide an additional computational tool in facilitating optimised individualistic fitness training for racing cyclists. In addition to introducing the algorithm and its underlying discrete-event non-linear model, we propose a new way of treating power output in similar sports; power output is abstracted as a requirement to which the body must respond accordingly, rather than as the output of a physiological process. Consequently, the output of the overall system is the body's response, as illustrated by the dynamic response of the heart rate and more specifically by time constants of the underlying dynamic model. The effectiveness of those parameters as fitness indicators is examined in comparison with established predictive techniques, indicating that future development of the method may overcome current limitations.

1 Introduction

Fitness is regarded as the ability to perform at the maximum of an athlete's potential at a given race event, given the conditions of the event, their preparation and of course, talent. Training is the procedure of unlocking one athlete's maximum potential, aiming a peak performance in a race. Training is a complex art where firstly the overall body's ability to perform and secondly the athlete's freshness at the day of the event are both carefully balanced; thus this procedure can be referred to as fitness optimisation.

The difficulties of quantitative fitness optimisation result from the plurality and variety of parameters that affect the cardiovascular and skeletomuscular control mechanisms and hence the efficiency of supplying the muscles with the required oxygen, both for aerobic respiration as well as for the breakdown of anaerobic wastes. Current techniques revolve about either reaching body capacity or avoiding fatigue on the day of the event. Reaching maximum body capacity is obtained by training in such a way as to best tune the respective variable parameters that can affect his performance to his benefit. The usual approach therefore is tuning these parameters using different types of training, so as to best adapt the athlete's body to the demanding needs of the race (increased plasma volume, increased lactate threshold, increased muscle capillarization, increased VO2 max, increased muscle high energy phosphate stores, increased stroke volume and several others [4]). The accepted view, fully compatible with the underlying physiology [11], is that oxygen supply is the input and power is the output. Actually this is exactly the relationship investigated by most formal contemporary measurement tests in this industry [3].

But assessing the current fitness of the athlete, even though it can certainly be related, it is not the same as predicting his fitness on the day of the event. Therefore any proposed technique

must incorporate a model which can take into account both the capacity as well as the fatigue of the athlete on that day. The most important model that is currently widely used in the cycling industry and is found built-in in cycling power software as well, such as TrainingPeaks and CyclingPeaks, where post-ride analysis is easily done, is the *Performance Manager* of Hunter Allen and Andrew Coggan[4]. Its concept lies in the basic principle of training: *breakdown* of the organism due to exercise, is followed by rest (*recovery*), which results in an improvement in performance (adaptation). An imbalance in the training load and recovery time can result in symptoms of fatigue. If the imbalance between training and rest persists, the athlete may develop serious symptoms of fatigue that will affect his ability to sustain a high training volume and will have a negative effect on his performance [9]. Therefore the athlete can be perceived as *fresh* for the event when the *recovery* at least *matches* the *breakdown*. The main advantage of the Performance Manager is that it captures the empirical knowledge of coaches regarding training, overtraining and fatigue.

The current purely physiological or semi-quantitative empirical techniques, taken so far, despite the respective advantages they offer, require a tool to bridge the gap between the two worlds: the endogenous physiological one, with the exogenous systemic, more empirical, one. This is exactly what we are trying to offer: an algorithmic systems approach whose parameters do not exclusively depend upon either ability or freshness, taking into account both endogenous as well as exogenous information.

2 The power-heart rate dynamic relationship and its significance

Power has so far been thought to be the output of the system under investigation. We, however, are going to regard it as the input, the requirement applied on the combined cardiovascular and skeletomuscular system, whereas heart rate can now be thought of as the response (the output) of this system to this new input (Figure 1). Heart rate is associated with the efficiency of oxygen delivery and waste removal to and from the muscles respectively, thus this approach is also compatible with the underlying physiology, just seen from a different viewpoint.

For training under carefully controlled conditions, the relationship between steady-state power and heart rate is known to be a good indicator of fitness. However, the dynamics and time constants of this relationship can also give useful information. The current work is predicated on the idea (which needs be validated) that these time constants are less dependent on the type of training being undertaken (intervals, sustained efforts, long rides etc) than the absolute value of heart rate required for a given power. To this end the tools we develop concentrate solely on these dynamics.

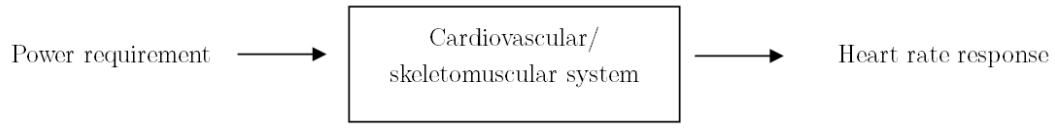


Figure 1: New approach proposed with power as the input requirement of the system rather than as the output. Heart rate response is the output of the system instead.

This approach seems to be a good technique in relating endogenous and exogenous information relating to both the athlete's ability and freshness. This dynamic response tries to capture the

current state of the body by looking closely on how the latter responds to the varying stress the athlete exerts on it. As illustrated by Figure 2, there is a lag introduced between the power step input and the heart rate response. This lag applies both to step increases and step decreases in power, yet probably it has a different size. As it can be seen from Figure 2, the time constant related to the increase of heart rate, referred later as tc_1, is shorter than the corresponding constant for heart rate decrease, tc_2. This illustrates the asymmetric property of the heart rate.

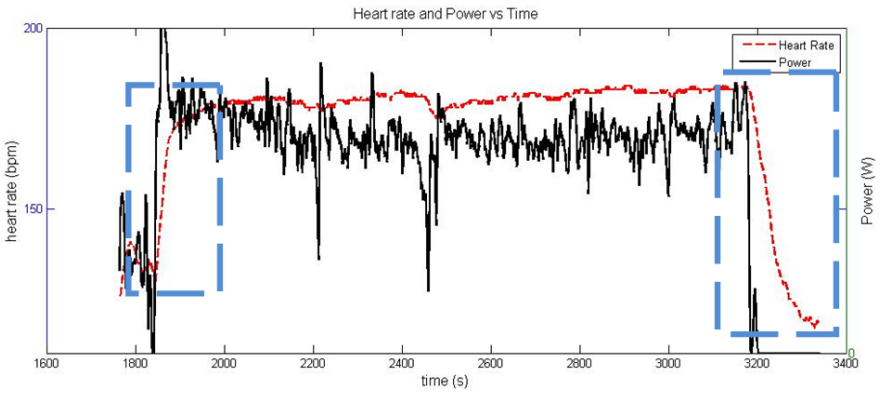


Figure 2: Power and heart rate dynamic relationship. The asymmetric property of the heart rate (the fact that heart rate increase is faster than heart rate decrease) is also observed.

3 Capturing the dynamics from large sets of training data

3.1 The underlying model: a discrete event system specification approach for continuous state systems using a simple dynamical model

The main challenge on developing an underlying model for the algorithm is that the heart rate response is non-linear due to heart rate assymetry. After serious consideration and testing of various models [10], predominantly using several non-linear combinations Infinitite Impulse Response filters as also suggested previously [7], it was found that no combination of linear filters, even if the combination were non-linear, would provide the necessary requirements regarding symmetry and/or asymmetry of the heart rate. Therefore, instead of approaching the problem top-down, trying to find the right combination of linear IIR filters to produce an adequate response, a bottom-up approach was chosen.

The system could be thought of in the same way as a second-order mechanical system with non-linearities responding to external forces. This can be obtained by setting different conditions for selecting the value of the time constant T such that $T = tc_1$, when power (pow) $>$ predicted heart rate (phr), and $T = tc_2$ when $pow < phr$. In other words, there is an addition of discontinuity in the time constant component. This is a simple example of simulation for continuous state systems using discrete event system specification [6]. Equation (1) illustrates the final obtained equation for this model.

$$T^2(\ddot{phr}) + 2T(\dot{phr}) + (phr) = pow, \begin{cases} T = tc_1, pow > phr \\ T = tc_2, pow < phr \end{cases} \quad (1)$$

The new non-linear discrete-event model equation was subsequently discretised and was turned in a form which could be used to obtain the predicted heart rate at each instant (equation (2)).

$$phr(a) = \frac{pow(a) + 2Tphr(a-1) + T^2(2phr(a-1) - phr(a-2))}{(1+T)^2} \quad (2)$$

This model was put under extensive filter response testing by plotting the response of the model to a particular pulse of power (10-30 seconds), with different values of tc_1 and tc_2, as illustrated by Figure 3. As the figure demonstrates, the filtered power, which can be thought of as the precursor of predicted heart rate, exhibits the expected responses according to physiology. If the two constants are equal, then there is symmetry in the response, yet when they are not equal, the two *tails* of the response exhibit lags analogous to their inequality. The response is sensitive to even small changes in tc_1 or tc_2, as required.

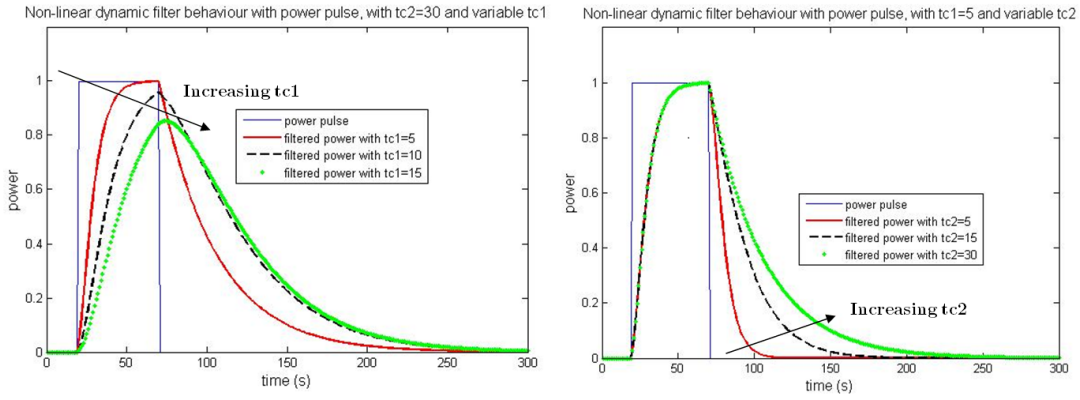


Figure 3: tc_1 and tc_2 effect on non-linear dynamic discrete-event filter. The output's behaviour is sensitive to tc_1 and tc_2, as required. Response agrees relatively well with heart-rate physiology.

3.2 The algorithm: core and extended

The core algorithm refers to the algorithm which, given the entire set of training data, can output the two parameters, tc_1 and tc_2. The methodology used involves an initial filtering of power (function $g(pow, tc_1, tc_2)$) using the model described in section 3.1, followed by least squares analytical fitting between filtered power, time and heart rate. The model parameters, found using optimisation methods, were those which together yield the smallest root mean squared error (RMSE) between predicted (from the model) and true heart rate. This extends a similar algorithm of previous work [7]. A very important aspect of the algorithm, which is illustrated in the flow diagram of Figure 4, is related to the use of the least squares fitting; unknown constants A, B, C, D and E allow for variation of absolute heart rate with power and time without affecting the estimation of the underlying time constants, but are not used in this analysis, as discussed in Section 2. As it was found by running on real athletes' data (amateur and semi-professional), it was clearly shown that both tc_1 and tc_2 were very reasonable

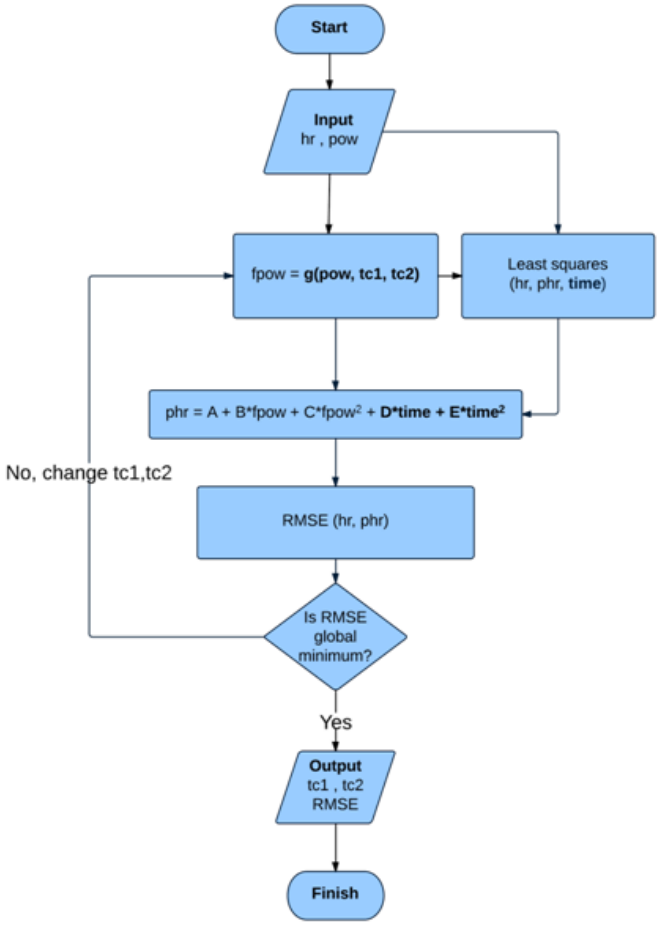


Figure 4: Core algorithm flowchart.

and agreed with physiology. Nevertheless, for best results, the confidence interval for each model parameter (i.e. the time constant tc_1 or tc_2), should be minimised. Assuming the error is Gaussian distributed, a 2D likelihood grid [8] was constructed, from where marginal distributions were calculated [10]. The confidence intervals using the entire data sets were found to be relatively large, therefore a way to narrow them down had to be found. Training sessions which have clearly defined pulses (i.e. are more structured) like the one shown in Figure 2, were found to have narrower confidence intervals. Indeed, as Figure 5 demonstrates, the 2-D probability surface of such training session is delta-like, which also corresponds to shorter confidence interval widths in the marginal distributions and better prediction for the data. It was therefore decided that an extented algorithm which could select those regions of

higher structure and thus provide more accurate results should be developed. This was done in practice by rewarding large heart rate variations on the sides of the region, but penalising both frequent as well as intra-step variations. Due to the possibility of having several local optima, the optimimisation algorithm is initialised at several points in the data set, thus trying to find a global optimum for values tc_1 and tc_2 with respect to RMSE. It should be stressed that the optimum objective function of this second optimisation is still under investigation, yet its aim is clear.

The full algorithmic tool (the core together with the extended algorithm) was firstly developed and tested in MATLAB and then converted into C++, being also compatible with the open-source Golden Cheetah framework [2].

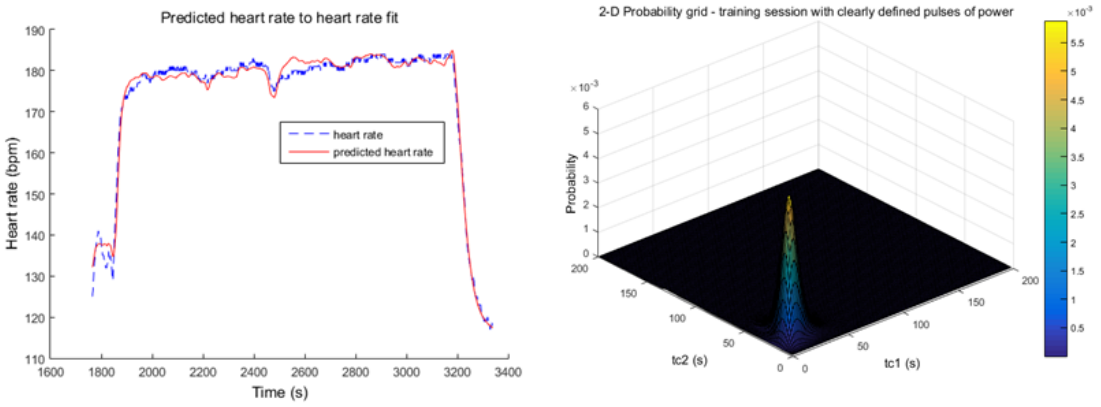


Figure 5: The data fit and and the 2-D Probability grid surface plot of training session with a clearly defined pulse of power, showing a more delta-like peak of probability at estimated values of tc_1 and tc_2 (corresponding to maximum likelihood). The grid surface is normalized to 1.

4 Comparison with established predictive techniques

Despite the fact that the aim of this paper is to illustrate the development of a new algorithmic tool and not to present a complete analysis of the inferences that could be made from the algorithm's output, it was deemed important to illustrate an example of how the model parameters which are provided as output relate to the dynamic macroscopic state of the athlete's body. Therefore, the effectiveness of model parameters as fitness indicators was examined mainly for two athletes (for whom large enough data sets were available) in comparison with the Performance Manager concept of Hunter Allen and Andrew Coggan [4], which is a simplified version of the Banisters Impulse Response Model [5] and is based on the widely known Training Stress Score (TSS). TSS can be thought of as a measure of the effort the cyclist put during a training session. TSS takes into account both the intensity (by using the Intensity Factor) and the duration of each training session. It can be viewed as a predictor of the amount of glycogen utilised in each workout. TSS is then used to calculate the Chronic Training Load (CTL) and Acute Training Load (ATL), which are usually the 42 and 7 day exponentially weighted moving averages of daily TSS. CTL is thought to provide a measure of how much an athlete has been training historically, whereas ATL is a measure of how much an athlete has been training recently, or acutely. The difference between these two values is call the Training Stress Balance

and provides a measure of "freshness" [4].

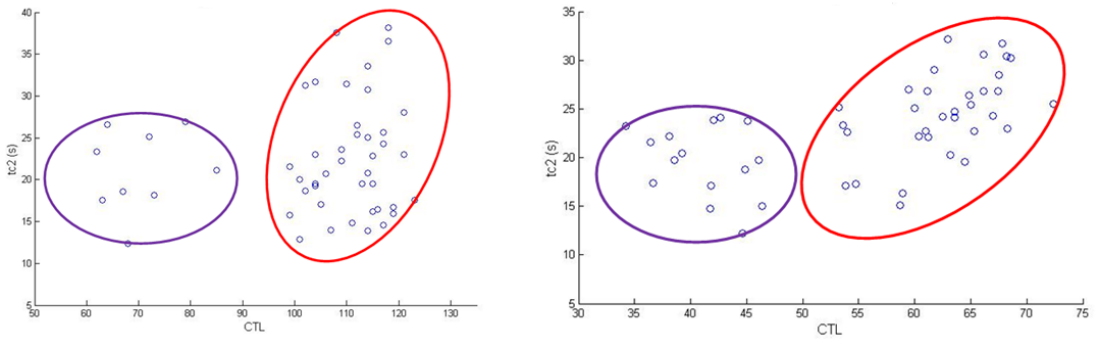


Figure 6: Model parameter tc_2 plotted against Chronic Training Load (CTL) for Athletes 1 and 2. tc_2 illustrates a possible linear relationship with CTL after a particular threshold of CTL. The threshold is higher for the younger athlete (20 years old) when compared with the older athlete (50 years old), as expected.

Using this information in conjuction with the tc_2 output of the algorithm for those two athletes, it was found that tc_2 might be used as an indicator of overtraining. CTL is thought to provide a measure of how much an athlete has been training historically. This has traditionally been used as a measure of training adaptation of the athlete. A high value of CTL is considered as having a positive impact to the athlete, whereas ATL has a negative impact. This, however, assumes that fatigue is solely captured in Acute Training Load (i.e. due recent acute training), neglecting any *chronic* fatigue due to long-term accumulation of fatigue. Figure 6 illustrates a particularly interesting potential finding, using data from the two athletes. tc_2, up to a particular threshold in CTL is not dependent of CTL. However, after this threshold, tc_2, a possible measure of fatigue, illustrates a possible *linear* relationship with CTL.

This finding, if true in general (more testing is required with data from professional athletes), could have a serious impact in the future design of training schedules by coaches. This is because this observation might mean that there is an upper finite limit of long-term training for each individual due to accumulated chronic fatigue, which cannot be counterbalanced by one or two week rest. Very few contemporary models assume that fatigue can be caused by chronic training, despite the fact that the difficulty of winning Tour de France and Giro d'Italia in the same year is common knowledge [1].

5 Conclusion

This paper lays the foundations of the creation of an open-source individualistic fitness predictor tool using a simple, algorithmic methodology, compatible with both physiological as well as empirical models. There are indications that long-term effects of training related to fatigue can be captured nicely in just one a set of model parameters and potentially this could also be extended to capture effects related to the performance of the athlete that are independent of training schedule, but have an impact on the athlete. Of course, further statistical analysis with professional athletes with larger data sets is required before any final conclusions can be drawn.

References

[1] Hall of Fame Results, Cycling Hall of Fame. http://www.cyclinghalloffame.com. Accessed: 2015-04-10.

[2] Golden cheetah. http://www.goldencheetah.org. Accessed: 2015-04-10.

[3] Hunter Allen and Stephen Cheung. *Cutting-edge Cycling*. Human Kinetics, 2012.

[4] Hunter Allen and Andrew Coggan. *Training and Racing with a Power Meter*. VeloPress, Boulder, CO, 2010.

[5] Thomas W Calvert, Eric W Banister, Margaret V Savage, and Tim Bach. A systems model of the effects of training on physical performance. *Systems, Man and Cybernetics, IEEE Transactions on*, 6(2):94–102, 1976.

[6] François E Cellier and Ernesto Kofman. *Continuous system simulation*. Springer Science & Business Media, 2006.

[7] William Collins. Modelling of the cardiovascular/respiratory system from data. Master's thesis, Department of Engineering, University of Cambridge, United Kingdom, 2011.

[8] David M Glover, William J Jenkins, and Scott C Doney. *Modeling methods for marine science*. Cambridge University Press, 2011.

[9] Michael Lambert and Jill Borresen. A theoretical basis of monitoring fatigue: a practical approach for coaches. *International Journal of Sports Science and Coaching*, 1(4):371–388, 2006.

[10] Andreas Petrides. A fitness predictor for racing cyclists. Master's thesis, Department of Engineering, University of Cambridge, United Kingdom, 2014.

[11] Jack H. Wilmore and David L. Costill. *Physiology of sport and exercise*. Human Kinetics, 1994.

Adapting the Elo rating system for multi-competitor sports

Pim Ritmeijer

Infostrada Sports, Netherlands
pim.ritmeijer@infostradasports.com

Abstract

Arpad Elo introduced the Elo rating system as a method for calculating the relative skill levels of players in competitor-versus-competitor games. This rating system originated in chess and is still used by the FIDE today to rank chess players. A number of variations on this method exist such as the Glicko rating system proposed by Mark Glickman which also includes a variance parameter. Microsoft's TrueSkill uses a Bayesian approach to rank multi-competitor sports.

Elo systems are already in common use in competitor-versus-competitor games or team sports. I propose to use a simple variant of the Elo rating system to rate multi-competitor sports where each competitor competes against one 'average' competitor. This gives a more transparent approach to rating changes and is computationally much easier to work with than to treat multi-competitor sports as a large group of head-to-head pairings. Such a simple system also makes it possible to put different aspects of a sport into separate rating parts (e.g. classic style or freestyle ratings in cross country skiing). Combining these different parts into one rating makes it much easier to make sport specific ratings.

1 Introduction

There are many rating systems for head-to-head sports, where the Elo rating system is perhaps the most well-known. This model will be discussed briefly in section two. Rating systems for multi-competitor sports are less common. In section three we will introduce an adaption of the Elo rating system for multi-competitor sports, as well as some extensions that can improve the forecasting performance of this model. In section four we will apply this rating system to athletics, where the forecasting performance for the Olympic Games and World Championships will be tested. This paper concludes with a discussion of future work in section 5.

2 Elo rating system

The Elo rating system to calculate skill level for competitor-vs-competitor games was introduced in the 1960s by Arpad Elo. Nowadays this rating system is still used by the World Chess Federation (FIDE) today and has many applications in online video game rankings.

The main principle behind these ratings is that the player rating is adjusted after each match or competition. If a player wins against a stronger opponent, then his rating before the match was probably too low and will be adjusted upwards. If a player loses against a weaker opponent, then his rating before the match was probably too high and will be adjusted downwards. Eventually all players should converge to a rating indicating their 'true' skill.

2.1 Elo variants

Many Elo rating systems variants have since been created, most notably the Glicko and Glicko-2 ratings [1] and TrueSkill [2]. The Glicko rating introduces a 'ratings deviation', which can be seen as a value for the reliability of the ratings. The longer a player is inactive, the higher his rating deviation will be. It turns out that the Elo rating system is a special case of the Glicko rating system where the rating deviation is a constant. TrueSkill is an extension of the Glicko rating system to multiplayer games.

3 Multi-competitor rating system

An Elo rating system for multi-competitor events is less straightforward. You could see an event ranking with N participants as a combination of $N(N-1)/2$ different head-to-head 'matches', but translated to the regular Elo rating system this means that the winner of the event automatically wins $N-1$ matches at once, and similarly that the last ranked participant loses $N-1$ matches at once. If a tennis player wins a lot of matches in a row it has to mean that he has performed well over a longer period of time. In contrast, for many multi-competitor sports the difference between winning and losing can be really small. Think of gymnastics for example, where one mistake could potentially be the difference between first and last place.

Another problem with the head-to-head approach is that there exists a trade-off between strength of competition and the number of participants. Suppose a player has a rating of 2000, and plays N opponents with a lower rating. The points gain for defeating 100 players with a rating of 1200 is roughly the same as defeating 34 players with a rating of 1400, or defeating 12 players with a rating of 1600, as can be seen in Figure 1.

This is why I propose to see a multi-competitor event as an event where each participant plays one 'match' against the 'average' opponent. With this approach the main influence on the rating changes will be the performance and strength of competition, instead of the number of participants.

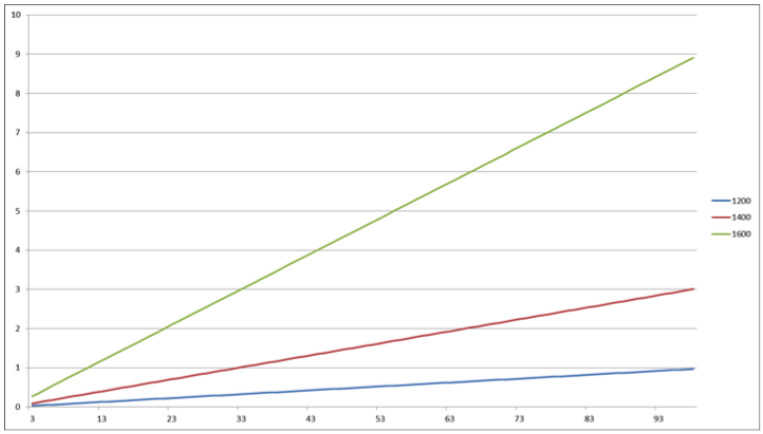


Figure 1: The amount of rating points won with when you regard an event ranking as a combination of head-to-head pairings (with a K-factor of one) is shown on the y-axis. The x-axis shows the number of opponents.

3.1 Methodology

Per event ranking an average rating is calculated. This is the average rating of all participants with a rank in that event ranking:

$$R_{\text{avg}} = \frac{1}{n}\sum_{i=1}^{n} R_i$$

Just like in the regular Elo model we calculate an expected score for all participants, where the average rating for that event ranking is taken as the opponent rating:

$$E_i = \frac{1}{1 + 10^{-(R_i - R_{\text{avg}})/400}}$$

The actual score can no longer be a simple win, draw or loss. Instead, we take the percentile in which the participant finished. This means that the winner of the event still gets an actual score of one, and the participant in last place still gets an actual score of zero, and everyone in between will get a score of between zero and one. The actual score can be calculated in the following way:

$$S_i = 1 - \frac{\text{rank}_i - 1}{N - 1}$$

The new rating for the participants is then calculated by applying the Elo updating method.

$$R_i^{new} = R_i^{old} + K(S_i - E_i)$$

3.2 Extensions

Speciality ratings

In some sports there might be a specific speciality involved which holds extra information for a result. Think of the influence of the surface in tennis, or the difference between classic- and freestyle cross country skiing for example. In these cases we can split a players rating into two parts: an overall rating and two or more speciality ratings. By taking the average of the players overall rating and the relevant speciality rating and using this average to forecast the results we get a more sport specific forecast.

Combined discipline ratings

In other sports there is some overlap between different disciplines. Think of the 100m and 200m in athletics. Both of these disciplines can be seen as sprint disciplines. If an athlete is world class at the 100m that means he is probably also good at the 200m. In this case we take the average of the discipline specific rating and the combined discipline rating to forecast results.

Discounting the ratings

In the Elo model there is no variance parameter for player ratings. But when a player is new to the system he gets the start rating, which basically means there is no information for that player yet. So we might consider 'having a rating' as information and 'having no rating' as a lack of information. When a player is inactive for a longer period of time we can see this as losing information. To account for this in the model we can define two cases: one where a player has a rating above the start rating, and one where a player has a rating below the start rating:

$$\begin{cases} R^* = d(t) * \text{Start rating} + (1 - d(t)) * R & \text{when } R > \text{Start rating} \\ R^* = R & \text{when } R > \text{Start rating} \end{cases}$$

Where $d(t)$ is a function of time passed since the last result that lies between zero and one. We only discount the ratings of players above the start rating, because it doesn't make sense to improve some players rating because of a lack of results.

Disappointment factor

Due to the fact some multi-competitor sports are very prone to random variation it might not be sensible to punish a disappointing result too severely as there are many ways in which this disappointing result could have resulted, some of which might have been entirely out of the control of the player himself. We can counter this by introducing a 'disappointment factor' which lowers the impact of a disappointing result:

$$\begin{cases} R_i^{new} = R_i^{old} + DF * K(S_i - E_i) & \text{when } S_i < E_i \\ R_i^{new} = R_i^{old} + K(S_i - E_i) & \text{when } S_i \geq E_i \end{cases}$$

4 Multi-competitor Elo model applied to athletics

In this section we will look at how well this method works for athletics. Using data from 2004 until April 2015, consisting of all Olympic Games, World Championships and Continental Championships as well as some of the bigger competitions from outside those (like the Diamond League) we can forecast all results. As all competitors start at the same rating we use the period 2004-2008 to let the ratings converge to sensible values and then use the data from 2009-2015 to calculate several Root Mean Squared Errors (RMSE), where the errors are defined as the actual score minus the expected score. We will also use the ratings to see how well this model does in forecasting results at the Olympic Games and World Championships.

First we need to find the right value for the K-factor. We do this by minimising several RMSE. The first RMSE consists of all the data in the model. The second looks at only data from Olympic Games and World Championships. And the next ones filter this even further, by looking only at the data of the actual gold medal winners, all medal winners, and all top8 results. The results for the multi-competitor Elo model without extensions are shown in Table 1. We

Measure	**K = 250**	**K = 350**	**K = 500**	**K = 700**
RMSE	0.26028	0.26096	0.26539	0.27601
RMSE WCh/OG	0.22608	0.22988	0.23811	0.25590
RMSE WCh/OG gold	0.21909	0.20561	0.19092	0.20045
RMSE WCh/OG medal	0.24160	0.22610	0.20421	0.21486
RMSE WCh/OG top8	0.26683	0.25908	0.25342	0.26182

Table 1: RMSE values for different values for the K-factor.

can see from Table 1 that the lowest RMSE occurs with a K-factor of 250. But when we only look at athletes who finished in the top8 at either the Olympic Games or World Championships the K-factor should be twice as high to get the optimal RMSE. This is also what we find when we do include the extensions in the model, as can be seen in Table 2. To see how this model does at forecasting results at the Olympic Games and World Championships we have defined nine measures. These measures look at the actual results and compare these with the results forecast. The results for the model without extensions are shown in Table 3. For seven out of nine measures we obtain the best forecast results for the model without extensions with a K-factor of 500. This value for the K-factor minimised the RMSE when only looking at top8 results at the Olympic Games and World Championships and also gives the best forecasts for

Measure	**K = 250**	**K = 350**	**K = 500**	**K = 700**
RMSE	0.25794	0.25574	0.25624	0.26104
RMSE WCh/OG	0.22182	0.22194	0.22703	0.23811
RMSE WCh/OG gold	0.24000	0.22111	0.20635	0.20004
RMSE WCh/OG medal	0.25519	0.23488	0.21680	0.20581
RMSE WCh/OG top8	0.27047	0.25966	0.25173	0.24940

Table 2: RMSE values for different values for the K-factor when extensions are included in the model.

Measure	**K = 250**	**K = 350**	**K = 500**	**K = 700**
gold = gold	37.87%	39.15%	39.57%	39.15%
gold = medal	69.36%	71.49%	71.91%	70.64%
gold = top8	84.26%	83.83%	84.68%	83.40%
gold = top16	94.89%	95.32%	95.32%	95.32%
medal = medal	52.47%	52.61%	52.89%	52.33%
medal = top8	79.13%	79.83%	79.41%	78.56%
medal = top16	92.67%	93.09%	93.51%	93.23%
top8 = top8	59.69%	60.01%	60.75%	60.33%
top8 = top16	83.01%	83.64%	83.48%	82.85%

Table 3: Percentage correctly forecast results for the model without extensions.

the measures we have chosen. Almost 40% of the actual gold medal winners were also forecast to win gold, and 95% of them were forecast to finish in the top16. When we look at the medallists we can see that more than half of them were also forecast in the medals and almost 80% of them were forecast within the top8.

In Table 4 the forecast results for the model with extensions are shown. For most measures the model with extensions performs a bit better at forecasting the results than the model without extensions. This difference is the biggest when we look at the actual gold medal winner forecast in the top8.

Measure	**K = 250**	**K = 350**	**K = 500**	**K = 700**
gold = gold	40.43%	41.70%	39.57%	37.87%
gold = medal	68.51%	68.09%	67.23%	68.51%
gold = top8	85.96%	87.23%	88.09%	88.94%
gold = top16	96.60%	97.02%	95.74%	95.32%
medal = medal	52.26%	52.59%	52.11%	52.11%
medal = top8	80.79%	82.05%	82.25%	82.72%
medal = top16	93.50%	94.39%	93.94%	94.10%
top8 = top8	60.95%	61.61%	61.76%	62.67%
top8 = top16	84.75%	84.99%	85.02%	85.37%

Table 4: Percentage correctly forecast results for the model with extensions.

5 Discussion and future work

In this paper, we developed a method to adapt the Elo rating system to multi-competitor sports. As many multi-competitor sports are a lot more complex than their head-to-head counterparts some extensions of the model could provide useful to make the forecast more sport specific. The optimal parameters in this article are not yet optimised per discipline. For future work it would make sense to have a different K-factor for the 100m than for the Marathon.

We have shown that the use of this model depends on your specific goal. If you want to make an 'honest' Elo model without any extensions a lower K-factor should be used than when you want to use the model to forecast possible future medallists.

One of the uses of this model that has not yet been discussed is simulation. Simulation is less straightforward than in the head-to-head case. Depending on the spread of ratings and number of participants robust simulations can be done that respect the underlying win probabilities in the head-to-head case.

References

[1] M. Glickman. Example of the Glicko-2 system. [online], 2013. `http://www.glicko.net/glicko/glicko2.pdf`.

[2] R. Herbrich, T. Minka, and T. Graepel. TrueSkill^{TM}: A Bayesian skill rating system. In *Advances in Neural Information Processing Systems*, volume 20, pages 569–576. MIT Press, 2007.

Batting at Three in Five-Day Cricket: a Descriptive Analysis

Jonathan Sargent

Stats On, Melbourne, Australia
jon@statson.com.au

Abstract

The number three batsman in test match cricket is commonly the most accomplished in the side, required to provide stability should the number one or number two batsmen lose their wicket early in the innings, or to capitalise on a solid opening partnership. This paper descriptively analyses individual test match batting performance to identify important characteristics associated with the number three batsman, referencing a recent deficiency within the Australian test cricket team at this position. Exponential distributions are fitted to the innings-specific run distributions of a random sample of recognised number three batsmen from test cricket history (the "Bat Pack"), yielding pre-match and in-play statistical benchmarks (ability, adaptability and continuity) to help cross-examine number three batsmen. The statistical outcomes from this descriptive stage of the research will become inputs for the next, predictive stage where a classification algorithm will scientifically identify misplacement within a team's batting order. National and domestic team selections can benefit from such an approach while probabilities of particularly ordered batsmen passing particular milestones will aid wagering on individual pre-match and in-play score markets.

1 Introduction

The "one-on-one" nature of cricket – that is, a specialist bowler competing directly with a specialist batsman – offers countless directions for the analysis of the teams' and players' performances and the pragmatic application of statistical models. Lewis [6] went so far as to describe the game of cricket as a "sporting statistician's dream", sweetened by the inclusion of limited-overs cricket because observations and predictions can occur within a discrete match frame. While modern formats shorten with the general audience attention span, test cricket, played over a maximum of five days, remains the purest form of the sport, demanding severer levels of endurance and strategy than its 50 and 20 over-per-side counterparts. As in baseball, logical strategy places the best batsmen towards the top of the order so a team is allowed the best possible start, but the optimal ordering of cricket batsmen remains complex, dependent on a team-specific blend of quantitative and qualitative measures [7]. The opening batsmen ideally lay a scoring platform upon which the remainder of the team can build over the ensuing days; a successful opening stand reduces the probability of loss [8]. The opening batsmen are each required to be among the most capable in the side, however, the number three batsman is often quoted as being the best and most important in the team. Australia's longest ever serving number three batsman, Ricky Ponting, stated that a team's most accomplished batsman should always bat at number three, not only to accept the greatest batting challenge ahead of less talented team-mates but also to present the most positive front to the opposition [3]. Since Ponting's retirement, Australian selectors have been unable or unwilling to settle on a "specialist" number three batsman. In the 51 test match innings since Ponting's retirement, Australia has trialed six different number three batsmen for three or more innings, averaging just 30.2

runs (total runs / total innings), whereas South Africa, another dominant cricket nation, after 30 innings in the post-Ponting period, has averaged 45.7 runs at number three, mostly through the efforts of specialist batsman, Hashim Amla. It is worth noting that South Africa has had a superior winning record to Australia, winning 12 of 18 tests in the period while Australia has won 12 of 26 tests; potential evidence of the importance of an established number three batsman.

While literature exists on the value of opening partnerships in cricket [8, 9], the significance of the number three position seems yet to have been adequately documented. The research outlined in this paper was motivated by the capability to identify an effective number three batsman with respect to three quantifiable categories: proven run scorer (ability), ability to build his innings (continuity) and reaction to match state (adaptability). Ability and continuity were modelled with exponential distributions where the probability of the batsman achieving a certain score was estimated prior to an innings and after passing 0, 10, 30 and 50 runs. Conditional run estimates reveal that accomplished number three batsmen possess the necessary qualities to score runs for extended periods. The metrics generated from these categories were compared with the opening and number four and five batsmen's to provide a statistical context for the importance of the number three. The match state under which any number three batsman commenced his innings – good, reasonable or poor start – was established after investigating the distribution of opening batsmen's combined runs (partnership) prior to the first wicket falling in the sampled matches. It was hypothesised that adaptable number threes should be able to rescue their team from a poor start (first dismissal for fewer than ten runs) and capitalise on a strong opening partnership (first dismissal after 50 runs scored). The statistical benchmarks from the descriptive and predictive stages of the research will ultimately reveal the occurrence of historical batting order misplacement, for example, that the controversy surrounding current Australian number three, Shane Watson can be statistically justified.

2 Innings and Batting Orders

Test match cricket allows each team two innings to score their runs over five days, divided into three sessions of approximately 30 overs[1] ($t = 1, \ldots, T$ where T is the completion of the last session on day five). Deliveries follow a continuous distribution: prior to the match, it is unknown how many overs will be bowled in one of the possible four innings; the twenty wickets required by a team to win the match are the only discrete resource. The batting team (p) in the first innings ($i = 1$) must score as many runs (R) as possible before one of two terminal points: the bowling team (q) dismisses 10 first innings batsmen ($w_{q1} = 10$) from team p or the captain of team p "declares" their first innings over ($w_{q1} < 10$) and sends team q into bat, believing they have compiled sufficient runs to take 10 team q wickets while ensuring $R_{p1} > R_{q1}$. Teams p and q bat in consecutive innings in the majority of matches; the "follow-on" scenario is the exception but is outside the focus of this research. The following match results are possible for team p:

$$result_p = \begin{cases} R_{p1} + R_{p2} > R_{q1} + R_{q2}, & w_p = 20 & win \\ R_{p1} + R_{p2} < R_{q1} + R_{q2}, & w_q = 20 & loss \\ R_{p1} + R_{p2} = R_{q1} + R_{q2}, & t = T & tie \\ t = T, & w_p < 20 & draw \end{cases} \quad (1)$$

For some wins, team p is not required to bat a second time as $R_{p1} > R_{q1} + R_{q2}$ while some losses only require one innings from team q as $R_{p1} + R_{p2} < R_{q1}$. Each case assumes $w = 20$.

[1] 180 deliveries + re-bowling of illegitimate deliveries

Each innings $(p1, \ldots, q2)$ varies in nature from match to match but as a rule, $p1$ is for building, $p2$ (if required[2]) is for consolidation, $q1$ for building toward R_{p1} and $q2$ (if required) is a "chase to win" or "survive to draw" innings. For these reasons, the parent sample in this research consisted of match data only from $p1$ and $q1$ where the innings strategy is consistent and the researcher was guaranteed their existence from each completed match. Figure 1 reveals interesting characteristics of team runs (R) through the four innings where $(1 = p1, 2 = q1, 3 = p2$ and $4 = q2)$: R_{p1} and R_{q1} distributions have an expectedly higher mean for wins and draws than losses, but also a wider dispersion, suggesting losses are more predictable in the first two innings of the match because of the smaller range. The higher mean of R_{q2} losses is logical; the target for victory is too high so team q decides to bat for the draw. Conversely, the dense left tail and right skew of R_{q2} wins reflects matches where the team q must only achieve a relatively small amount of runs to win the match, sometimes achieved without the loss of any wickets, for example, the West Indies (q) only requiring one run for victory at the commencement of their second innings against India (p) in a test match dated April, 1983: $R_{p1} = 209, R_{p2} = 277, R_{q1} = 486, R_{q2} = 1$. The R_{q2} win distribution also shows a 400-run target has been successfully chased for victory on very few occasions.

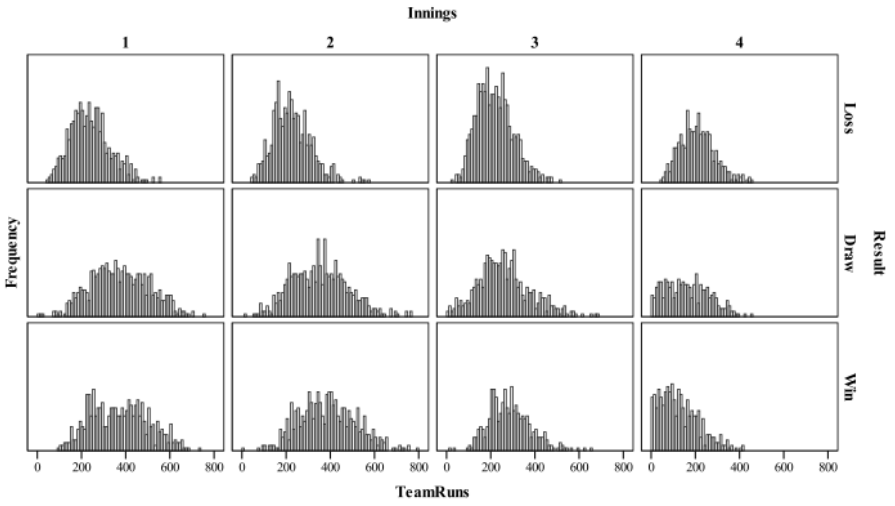


Figure 1: Team run distributions through innings and match result

Regardless of test match or limited overs cricket, allocating the right batsmen in the right orders is crucial for a team's success [7]. This is a complex task and is usually the result of "shuffling" players around the order until a productive solution is found. Former England captain, Nasser Hussain believes the top three batsmen have to provide the right momentum for the rest of the team, stressing that Ian Bell, a senior player with 97 tests to his name, should have batted at number three in the 2013 Ashes campaign, using experience to control the tempo of the innings in the case of opening partnership failures [4].

Selectors must decide the right mix of specialist batsmen, specialist bowlers and all-rounders. Figure 2 reveals a difference in the average runs of the top five batsmen, O_x and the remaining orders in completed test matches, most noticeably for p_1 and q_1. Although O_4 is historically

[2]Where the "follow-on" is enforced, a team may have sufficient runs from their first batting innings to win the match.

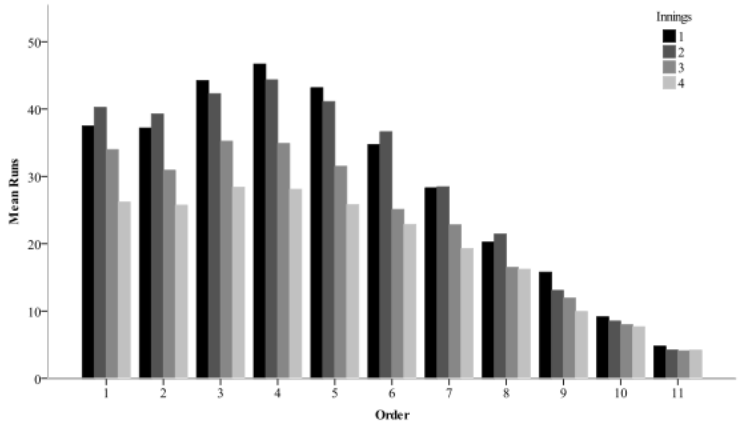


Figure 2: Average runs through batting order and innings.

a higher scorer than O_3 (see Figure 2), this is not necessarily a reflection of superior talent; averages alone do not tell the full story [2, 6]. Typically, at the start of the first innings when the cricket ball and pitch are untouched, the ball behaves more erratically meaning wickets occur more regularly. Assuming a reasonable opening partnership, by the time the number four batsman comes in to bat, the condition of the ball and pitch should have deteriorated sufficiently to minimise unpredictable deliveries and accommodate longer periods of batting. The opening batsmen do not have this luxury, hence lower averages for O_1 and O_2. The first order batsman faces the first delivery so is more susceptible to dismissal than any other because he possesses the least knowledge of match factors, for example, pitch condition and the immediate form of the opening bowler; this accounts for the higher frequency of single-figure scores for O_1 than other batsmen. Nerves are another explanation. Established O_3 batsmen are renowned for their defensive approach to an innings, especially when the opening partnership is not fruitful which is reflected in the smallest frequency of 0-20 runs in comparison to other orders. This O_3 quality will be explored further in the following sections.

3 Methods and Results

This research argues an established number three batsman is required to be a proven run scorer ("ability"), react to the value of the opening partnership ("adaptability") and bat productively for as long as possible ("continuity"). In measuring ability, a batsman's career average and strike rate were calculated by:

$$\overline{r} = \frac{1}{n}\sum_{j=1}^{n} r \qquad SR = \frac{1}{n}\sum_{j=1}^{n} r/d \tag{2}$$

where r is a batsman's runs and d are his deliveries faced in innings j. The universal calculation subtracts "not out" innings, or cases where the batsman has not been dismissed at the completion of innings i, from the denominator. Recognizing not-outs was viewed as unfairly sweetening lower-order averages because those players were more likely to be batting at an innings-end. Table 1 reveals the top ten members of the Bat Pack, or all number three batsmen

that have achieved 30 or more innings ($p1, q1$), retiring after 1980. The table has been ranked in descending order of average (Equation (2)i). The three most sampled players at number three, Dravid ($n = 130$), Sangakkara ($n = 118$), Ponting ($n = 112$) are fittingly consecutively ranked (3, 4 and 5), and, along with Amla (9), are regarded as some of the best batsmen of the modern era (Sangakkara and Amla are still playing), all averaging above 50 runs per innings. This suggests that teams settle on successful number threes.

Batter	Team	n	Ave	SR	Batter	Team	n	Ave	SR
BC Lara	W. Indies	37	70.70	0.6753	Younis Khan	Pakistan	46	55.82	0.5234
IVA Richards	W. Indies	41	66.51	0.5851	SP Fleming	N. Zealand	39	54.72	0.5085
KC Sangakkara	Sri Lanka	118	62.25	0.5594	DI Gower	England	32	53.50	0.5632
RT Ponting	Australia	112	60.96	0.6236	HM Amla	S. Africa	67	50.94	0.5260
R Dravid	India	130	55.95	0.4394	RB Richardson	W. Indies	67	49.24	0.4979

Table 1: Top ten number three batsmen in either of first innings ($n \geq 30$) since 1980.

Numerous efforts have been made over the last century to retrospectively fit statistical models to batsmen's scores. Early work by Elderton [5] proposed test match cricket batting scores followed a geometric distribution, but subsequent research revealed this distribution provided an inadequate fit for zero scores [10], as detailed in the previous section. For this research, a single parameter (λ) exponential distribution with mean $1/\lambda$ and variance $1/\lambda^2$ was an adequate fit for O_3 batsmens scores but still failed to solve the zero-run anomaly. This is apparent in the over-fit far left interval (0-9 runs) in Figure 3i and the difference in the observed ($\sigma = 55.02$) and expected standard deviation ($\sigma = 50.42$).

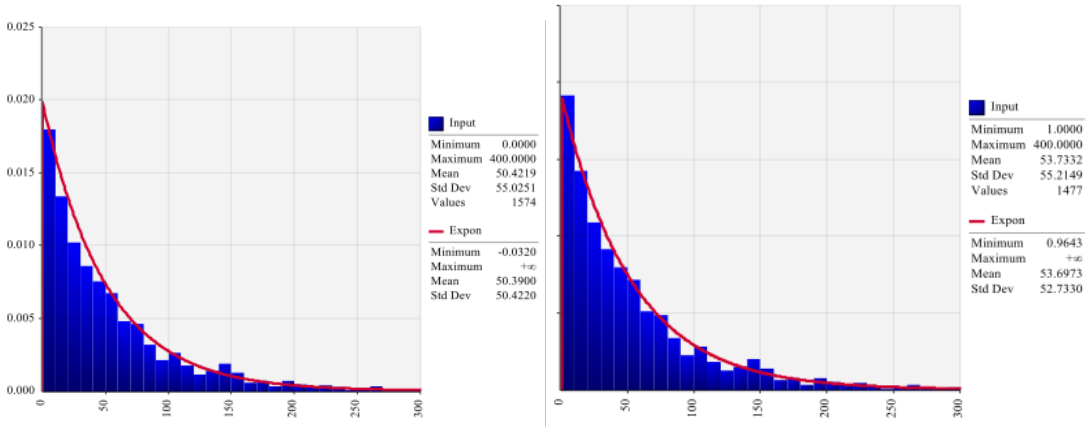


Figure 3: Histograms of O_3 runs and exponential distributions; zero scores included (i) and excluded (ii).

The fitted distribution also seems to be an inadequate fit for major scoring milestones; unlike the model, O_3 batsmen do not seem to suffer from the "nervous nineties" exhibiting a noticeably low frequency of scores between 90 and 99 than neighboring intervals. This is an insight into the patience of successful number three batsmen. There are clear spikes in the interval prior to the batsman reaching 150 and 200 runs – probably mental fatigue – which

must be flagged if making inferences from the exponential distribution. Removing zero-run innings from the sample considerably improves the fit for single-figure scores and the standard deviation discrepancy (see Figure 3ii), however, the run milestone problem remains.

Previous sections have mentioned a number three batsman's responsibility with respect to the productivity of the opening partnership ($r_1 + r_2$), that is, to stabilize the innings after an early dismissal or capitalize on a good start. An O_3 batsman's "adaptability" was determined through his contribution to the innings given a good ($r_1 + r_2 > 50$), reasonable ($10 \leq r_1 + r_2 \leq 50$) or poor ($r_1 + r_2 < 10$) opening stand. Contribution, C is a batsman's standardized performance, expressing his runs in innings j as a percentage of total team runs in innings i, simply denoted as $C_j = r_j/R_i$. A score of 20 by a top batsman may be viewed as a low score on its own, but if the total team score was also low, say 120, due to difficult batting conditions (for example, a cracked pitch or humidity), the batsman's score may be deemed as satisfactory. Conversely, standardizing a large value of r with respect to a large R, perhaps because of a weak bowling attack, provides r with further context. Several other factors impact a player's contribution, but this approach was considered suitable for this stage of the research. Tabulating the opening stand categories against the completed match result provided further evidence of the importance of adaptable O_3 batsmen. From Table 2, the most significant O_3 team contributions, with respect to the first five batsmen, are wins after poor starts ($C = 0.1540$) and draws and wins after reasonable starts ($C = 0.1512$ and $C = 0.1500$). Contributions through the top five from losses after poor starts are negatively related to wins after good starts: C gradually increases through to O_5 with losses from poor starts with the top three only contributing 30.2% of the runs, compared to 49.9% of the runs when a team wins from a good start. The remaining orders have been included in Table 2 for referencing purposes.

		Opening Partnership								
		Poor			OK			Good		
		Loss	Draw	Win	Loss	Draw	Win	Loss	Draw	Win
Order	1	0.0929	0.1335	0.1435	0.1081	0.1472	0.1412	0.1288	0.1606	0.1854
	2	0.0978	0.1283	0.1386	0.1015	0.1291	0.1392	0.1183	0.1521	0.1722
	3	0.1115	0.1438	0.1540	0.1124	0.1512	0.1500	0.1149	0.1347	0.1414
	4	0.1205	0.1385	0.1415	0.1183	0.1366	0.1331	0.1151	0.1542	0.1344
	5	0.1277	0.1319	0.1207	0.1184	0.1220	0.1335	0.1126	0.1151	0.1088
	6	0.1104	0.1082	0.1104	0.1030	0.1066	0.1086	0.0981	0.0956	0.0978
	7	0.0923	0.0871	0.0901	0.0907	0.0877	0.0876	0.0837	0.0779	0.0777
	8	0.0720	0.0669	0.0651	0.0716	0.0687	0.0669	0.0651	0.0597	0.0572
	9	0.0513	0.0491	0.0503	0.0504	0.0495	0.0466	0.0453	0.0425	0.0383
	10	0.0370	0.0294	0.0291	0.0344	0.0281	0.0286	0.0357	0.0279	0.0238
	11	0.0166	0.0158	0.0171	0.0192	0.0138	0.0170	0.0163	0.0130	0.0135

Table 2: Adaptability to match state through contribution to team runs.

The term, "continuity" has been adopted to describe the ability of a batsman to continue scoring runs, having accumulated particular run totals during his innings. For example, a top order batsman might aim to achieve 50 runs having reached 20 runs, then 100 runs having reached 50. From the exponential distribution (see Figure 3), the probability of passing a certain score, $\hat{r}$ given accumulated runs, X was calculated with the cumulative probability function:

$$F(\hat{r}; \lambda | r > X) = e^{-\lambda \hat{r}}\,. \tag{3}$$

Table 3 lists the probabilities of passing 50 and 100 runs when $X = (1, 10, 30, 50)$ for the top five batsmen. Commencing at $X = 1$ removes the zero-score anomaly. O_3 batsmen are more likely to achieve the two milestones than any other order through all X. There is a distinct difference in the probability of O_3 reaching 100 runs once achieving 50, again highlighting the

patience of the number three. One could argue that O_3 batsmen have more time at the crease than lower orders to score their runs, remembering the openers are the most susceptible to dismissal in the top five, however, this would be difficult to deduce from our data.

	O_1		O_2		O_3		O_4		O_5	
X	$r = 50$	$r = 100$	$r = 50$	$r = 100$	$r = 50$	$r = 100$	$r = 50$	$r = 100$	$r = 50$	$r = 100$
1	34.9%	11.9%	35.8%	12.6%	39.5%	15.3%	38.8%	14.8%	37.4%	13.8%
10	43.2%	15.1%	43.6%	15.5%	47.7%	18.9%	47.2%	18.4%	45.5%	17.0%
30	66.9%	24.4%	65.9%	23.3%	69.6%	28.0%	68.7%	26.4%	67.2%	24.9%
50	-	36.1%	-	36.4%	-	41.0%	-	37.5%	-	35.6%

Table 3: Probabilities of reaching 50 and 100 given accumulated runs (X) through orders 1-5.

The next stages of this research will involve the classification of batting orders with a discriminant function (Mahalanobis distance measure) using the metrics outlined in this paper as inputs. The anticipated level of complexity in the predictive stage commands a separate research paper, offering scientific conclusions about misplacement within batting orders, specifically, the worth of Shane Watson batting at number three for Australia and potential replacements. Analysis of international experience prior to settling at number three might be valuable, as well as whether domestic batting metrics correlate with international.

4 Conclusion

Although not always statistically significant, this paper has provided sufficient evidence that number three batsmen exhibit characteristics that distinguish them from opening and number four and five positions. Such characteristics should be distinguished further with the addition of key factors, omitted from this stage of the research, for example, home ground advantage, chemistry between certain batsmen in a partnership and the strength of the opposition bowling attack. Fitting an exponential distribution to batsmen's pre-match scores did not account for the high frequency of zeros, however, fitting after a batsman scores a run is a distinct improvement and would be beneficial for in-play wagering on batsmen head-to-head run markets, as covered by Bailey and Clarke [1]. A non-parametric approach may another option, that is, modelling the form of the batsman by smoothing his last x innings, rather than his observed scores, thereby dampening zero and large score effects.

References

[1] M.J. Bailey and S.R. Clarke. Market inefficiencies in player head-to-head betting on the 2003 cricket world cup. In P.Pardalos and S. Butenko, editors, *Economics, Management and Optimization in Sports*, pages 185–201, New York, 2004. Springer.

[2] S.R. Clarke and P. Allsopp. Fair measures of performance: the World Cup of cricket. *Journal of the Operational Research Society*, 52:471–479, 2001.

[3] Cricinfo, October 2013. `http://www.espncricinfo.com/australia/content/story/682351`.

[4] Dailymail, December 2013. `http://www.dailymail.co.uk/sport/cricket/article-2529383/NASSER-HUSSAIN-Ian-Bell-No-3-Johnny-Bairstow-proper-chance-now.html#ixzz3Ykf4yeOC`.

[5] W.P. Elderton. Cricket scores and some skew correlation distribution. *Journal of the Royal Statistical Society, Series A*, 108:1–11, 1945.

[6] A.J. Lewis. Towards fairer measures of player performance in one-day cricket. *Journal of the Operational Research Society*, 56:804–815, 2005.

[7] T.B. Swartz, P.G. Gill, D. Beaudoin, and B.M. deSilva. Optimal batting orders in one-day cricket. *Computers & Operations Research*, 33:1939–1950, 2006.

[8] A. Tan and D. Zhang. Distribution of batting scores and opening wicket partnerships in cricket, 2001. `http://ms.appliedprobability.org/data/files/Abstracts%2034/34-1-5.pdf`.

[9] J. Valero and T.B Swartz. An investigation of synergy between batsmen in opening partnerships. *Sri Lankan Journal of Applied Statistics*, 13:87–98, 2012.

[10] G.H. Wood. Cricket scores and geometrical progression. *Journal of the Royal Statistical Society, Series A*, 108:12–22, 1945.

Scheduling of Sport League Systems with Inter-League Constraints

Jörn Schönberger

Technische Universität Dresden, Dresden, Germany
joern.schoenberger@tu-dresden.de

Abstract

We investigate the simultaneous schedule determination of several leagues in a championship. Beside limited availability of venues, the substitution of players among teams of a club requires the consideration of mutual match slot exclusions associated with teams playing in different leagues during schedule construction. We propose a mathematical optimization model for this complicated decision task and report results from initial computational experiments.

1 Introduction

A set of sport leagues is called a championship. While in professional sports, the schedules of the leagues in a championship are determined consecutively starting with the most important league (e.g. the Premier League in soccer is scheduled first, followed by the Second League and so on) non-commercial (leisure) sport typically cannot realize a sequential schedule determination due to additional constraints establishing interdependency between the schedules of different leagues. This paper addresses the simultaneous determination of schedules from several leagues called the *championship scheduling problem*. We have reported elsewhere on the championship scheduling problem of a regional table tennis association in Germany that addresses the simultaneous determination of schedules for several leagues. Two types of inter-league constraints are to be considered: limited venue capacity as well as player substitution opportunities between several teams of a club. The first mentioned issue requires the coordination of available match slots usage among all teams of a club even if these teams are members of different leagues. The second mentioned issue requires that two teams of a club assigned to different leagues (the receiving as well as the providing team of a substitute player) do not have a match in the same slot.

A survey of the state-of-the-art in sport league scheduling is given by [4]. [1] proposes and compares several mathematical models for the sport league scheduling problem. Applications of sport league scheduling approaches in the context of table tennis are reported in [4, 6]. [2] as well as [5] consider a variation of the sport league scheduling setup in which one venue from a set of available venues has to be selected for each match from the competition program. To the best of the author's knowledge these are the only attempts to incorporate scarce venue issues into sport league schedule determination and there are no contributions that report research on the simultaneous scheduling of several leagues.

An informal problem description is provided in Section 2. Afterwards, Section 3 reports the development of a mathematical optimization model. The proposed model is verified in initial computational experiments whose results are given in Section 4.

2 The Championship Scheduling Problem

We consider a set CP of leagues formed by teams from the set $\mathcal{T}$ which are delegated by clubs collected in the set $\mathcal{C}$. The binary parameter $TC(m;c)$ is equal to 1 if and only if team $m \in \mathcal{T}$ is associated with club $c \in \mathcal{C}$. A *match* in the league $L \in CP$ is the ordered pair $(i,j) \in L \times L \setminus \{(i,i) \mid i \in L\}$. We call team i the home-team and team j the away team. It is assumed that the venue of the home team hosts the match $(i;j)$. If $(i;j)$ is a match then we call the set $\{i;j\}$ a *meeting* with involvement of i and j.

The competition program of league L contains all matches specified by competition rules. Without loss of generality, we assume that a match $(i;j)$ among two teams of league L is required exactly once. In the championship scheduling problem it is necessary to schedule all matches in all leagues of the championship. For each of these matches a time slot $s_{ij} \in S$ has to be selected as date of match $(i;j)$ from the set of available time slots S. The set S is partitioned into the set S^F of slots of the *fall round* and the set S^A of slots of the *autumn round*. We assume that all slots from S^F precede the slots of S^A. A feasible schedule of an individual league in the considered championship requires the fulfillment of the following 5 conditions.

- Exactly one slot $s_{ij} \in S$ must be selected for each match in the competition program (R_1).
- At most one match per team can be assigned to a slot from S (R_2).
- If match $(i;j)$ is scheduled in S^F then and only then match $(j;i)$ is scheduled in S^A (R_3).
- For each team, home with away matches are alternately scheduled over the season (R_4).
- In order to achieve a regular distribution of all matches of a team over the season it is necessary to ensure that at least N^{DIST} slots can be found between two consecutive meetings with participation of this team (R_5).

To achieve feasibility of all generated schedules in the championship it is necessary to respect the following inter-league conditions. Two teams for which a substitution opportunity is specified must not have a meeting (with any other team) in the same slot (R_6). The capacity of club c's venue $VC(c)$ must not be exceeded by the home matches of all teams associated with club c in any slot (R_7).

The conditions R_1 to R_3 are *obligatory* since the table tennis competition rules require the fulfillment of these conditions. Also the fulfillment of R_6 as well as R_7 is obligatory in order to get a feasible schedule. The conditions R_4 as well as R_5 are imposed in order to meet fairness aspects. They are not induced by any rule set. They are called *soft* conditions. We are going to minimize the number of violations of the soft conditions while all obligatory conditions are fulfilled without any exception.

3 A Mathematical Optimization Model

We first derive mathematical constraints whose consideration ensures the feasibility of an individual league schedule for a single league with respect to R_1 – R_5. Afterwards, we extend this model in order to meet the conditions R_6 as well as R_7.

Let L be a league. Team k meets each other team i twice. Let $\bar{L}$ be the set that contains two copies of each team $k \in L$, say k^+ and k^-. Team k meets each other team from $\bar{L}(k) := \{i \in \bar{L} \mid i \neq k^+ \wedge i \neq k^-\}$ exactly once. All teams from $\bar{L}(k)$ labeled by $^+$ are collected in the set $\bar{L}^+(k)$ and a match against team $j \in \bar{L}^+(k)$ takes place in the venue of k (home-match of team k). All $^-$-labeled teams are put into the set $\bar{L}^-(k)$ accordingly. A match between k and and $i \in \bar{L}^-(k)$ is conducted at the venue of team i (away-match of team k). As an example,

we consider the league $L = \{1, 2, 3, 4\}$ and get $\bar{L} := \{1^+, 1^-, 2^+, 2^-, 3^+, 3^-, 4^+, 4^-\}$. For team $k = 2$ we have the sets $\bar{L}(2) := \{1^+, 1^-, 3^+, 3^-, 4^+, 4^-\}$ as well as $\bar{L}^+(2) := \{1^+, 3^+, 4^+\}$ and $\bar{L}^+(2) := \{1^-, 3^-, 4^-\}$. The meetings to be scheduled with participation of team 2 are $\{\{1^+; 2\}, \{3^+; 2\}, \{4^+; 2\}, \{1^-; 2\}, \{3^-; 2\}, \{4^-; 2\}\}$.

For team $k \in L$, a Hamiltonian path goes through all teams from $\bar{L}$. This path originates from k^+, terminates in k^- and visits each other member of $\bar{L}$ exactly once. It determines the sequence in which the meetings with participation of k are carried out. For the coding of such a sequence, we declare the binary decision variable x_{ijk} $(i, j \in \bar{L}, k \in L)$. If $x_{ijk} = 1$ then the meeting of k with i precedes the meeting of k with j and k has no other meeting between these two meetings. For example, the Hamiltonian path $(3^-, 1^-, 2^+, 1^+, 3^+, 2^-)$ defines the match pattern for team 2: first, 2 has an away match (AM) against 3, followed by an AM against 1 and a home match (HM) against 2. The next matches are against 1 (HM), 3 (HM) and 2 (AM).

Beside the determination of the match pattern for each team $k \in L$ it is necessary to select the slot into which a match falls. We declare the integer-valued decision variable z_{ik} to carry the selected slot of the meeting with participation of k and i.

$$\sum_{j \in \bar{L}(k)} x_{k^+jk} = 1 \; \forall k \in L \wedge \sum_{j \in \bar{L}(k)} x_{jk^+k} = 0 \qquad \forall k \in L \tag{1}$$

$$\sum_{j \in \bar{L}(k)} x_{jk^-k} = 1 \; \forall k \in L \wedge \sum_{j \in \bar{L}(k)} x_{k^-jk} = 0 \qquad \forall k \in L \tag{2}$$

$$\sum_{j \in \bar{L}(k)} x_{ijk} = 1 \qquad \forall k \in L, i \in \bar{L}(k) \tag{3}$$

$$\sum_{j \in \bar{L}(k)} x_{ijk} = \sum_{j \in \bar{L}(k)} x_{jik} \qquad \forall k \in L, i \in \bar{L}(k) \tag{4}$$

As proposed for the well-known capacitated vehicle routing problem [3] (1)–(4) determines a set of Hamiltonian paths through the sets $\bar{L}(k)$ fulfilling condition R_2. (1) ensures that a first opponent is defined but prevents that the initial dummy opponent k^+ has a predecessor. Similarly, (2) ensures that team k's Hamilton path through $\bar{L}(k)$ terminates in k^-. Constraint (3) takes care that each meeting of team k has a predecessor as well as a successor and that a continuous path is achieved (4).

$$z_{k^+k} = 0 \; \forall k \in L \wedge z_{k^-k} = |S^F| + |S^A| + 1 \; \forall k \in L \tag{5}$$

$$z_{ik} + 1 \leq z_{jk} + (1 - x_{ijk}) * \mathcal{M} \; \forall k \in L, i, j \in \bar{L}(k), i \neq j \tag{6}$$

$$z_{l^+k} = z_{k^-l} \; \forall k, l \in L, k \neq l \tag{7}$$

$$z_{ik} \neq z_{jk} \; \forall k \in L \; \forall i, j \in \bar{L}(k), i \neq j \tag{8}$$

$$z_{ik} + 1 + N^{DIST} \leq z_{jk} + (1 - x_{ijk}) * \mathcal{M} + e^{DIST}_{ijk} \; \forall k \in L, i, j \in \bar{L} \tag{9}$$

$$u^F_{ik} \cdot \mathcal{M} \geq |S^F| + 1 - z_{ik} \forall k \in L \; \forall i \in \bar{L}(k) \; \wedge \; u^A_{ik} \dot{\mathcal{M}} \geq z_{ik} - S^F \; \forall k \in L \; \forall i \in \bar{L}(k) \tag{10}$$

$$u^F_{i^+k} + u^F_{i^-k} \leq 1 \; \forall k, i \in L, k \neq i \; \wedge \; u^A_{i^+k} + u^A_{i^-k} \leq 1 \; \forall k, i \in L, k \neq i \tag{11}$$

$$x_{opk} + x_{pqk} \leq 1 + e^H_{opqk} \cdot \mathcal{M} \; \wedge \; x_{opk} + x_{pqk} \leq 1 + e^A_{opqk} \cdot \mathcal{M} \; \forall k \in L, o, p, q \in \bar{L}(k)^- \tag{12}$$

The constraint families (5)–(9) ensure the selection a feasible slot for each match. (5) determines the slot for the dummy meeting that initiates the Hamiltonian path and the slot for the dummy meeting that terminates the Hamiltonian path associated with team k. Let

$\mathcal{M}$ be a sufficiently large number. In case that team k meets team j immediately after team i then the meeting of j with k is scheduled at least 1 slot later than the meeting of i with k (6). Constraint family (7) synchronizes the slots for the meetings scheduled for each pair of two different teams k and l. At most one meeting per slot is scheduled for team k (8). The binary indicator decision variable e_{ijk}^{DIST} is 1 if the required number N^{DIST} of free slots between the two consecutive meetings of k remains unsatisfied (9).

In order to meet constraint R_3 it is necessary to decide for each meeting $\{i;k\}$ if it is scheduled in the fall season S^F or in the autumn season S^A. We introduce the binary decision variable families u_{ik}^F as well as u_{ik}^A as indicators for the selected half season of a meeting between two teams. (10) enforces u_{ik}^F to be 1 if i and k meet in the fall season and enforces u_{ik}^A to be 1 if i and k meet in the spring season. Either the meeting of both teams at the venue of k or the meeting of both teams in the venue of i can be assigned into the same half season (11).

In order to fulfill R_4 we state the constraint (12). It ensures that at most two consecutive meetings with the involvement of team k take place in the venue of k (in the venue of k's opponent). The binary indicator decision variables e_{opqk}^H (e_{opqk}^{AA}) is enforced into the value 1 if the three consecutive meetings $\{o;k\}$, $\{p;k\}$ and $\{q;k\}$ are all home (all away) matches for team k.

$$Z := \sum_{k\in L} \sum_{i,j\in \bar{L}(k)} e_{ijk}^{DIST} + \sum_{k\in L} \sum_{o,p,q\in \bar{L}(k)} e_{opqk}^H + e_{opqk}^A \tag{13}$$

$$\bar{Z} := \sum_{m\in CP} \sum_{k\in L_m} \sum_{i,j\in \bar{L}_m(k)} e_{ijk}^{DIST} + \sum_{m\in CP} \sum_{k\in L_m} \sum_{o,p,q\in \bar{L}_m(k)} e_{opqk}^H + e_{opqk}^A \tag{14}$$

The sport league schedule determination problem for a double round robin competition program is now represented by the mixed-integer-linear program (MILP) consisting of the constraints (1)-(12) and the objective function (13) to be minimized. This model follows a significantly different representation logic compared to the modeling approaches proposed in [1].

Assume now, that the competition comprises several leagues collected in the set CP. Let $m \in CP$ and L_m denote the set of the teams forming league m. We can replace L by L_m in the previously presented constraints (1)–(12) and activate these constraints for all leagues $m \in CP$. Solving the resulting MILP consisting of the constraints (1)–(12) as well as the objective function (14) determines feasible schedules for each individual league in the championship but the inter-league constraints remain unconsidered. To incorporate the inter-league conditions we extend the previously described MILP.

$$z_{ji} <> z_{lk} \qquad \forall i,j,k,l \in \bigcup_{m\in CP} L_m \tag{15}$$

$$z_{ji} = \sum_{s\in S} y_{ijs}^{HOME} \cdot t \wedge \sum_{s\in S} y_{ijs}^{HOME} = 1 \qquad \forall m \in CP, i \in \bar{L}_m^+, j \in \bar{L}_m^+ \tag{16}$$

$$\sum_{t\in\mathcal{T}} TC(t,c) \cdot \sum_{i\in L_m} \sum_{j\in L_m} y_{ijt}^{HOME} \leq VC(c) \qquad \forall c \in \mathcal{C}, s \in S \tag{17}$$

Let $s(i,j,k,l)$ be a binary parameter that is fixed to 1 if and only if the two matches (i,j) and (k,l) have to be scheduled in different slots in order to allow player substitution by a player from team k to complete team i. Constraint (15) ensures that matches with the participation of two teams with a substitution relation are not assigned into the same time slot. This fulfills condition R_6.

schedules proposed without consideration of inter-league constraints

team	slots																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1(1)	—	-5	—	—	—	—	—	-6	—	3	2	-4	—	—	—	4	—	—	-2	—	6	-3	5	—
2(2)	—	—	—	—	—	—	—	-5	6	4	-1	3	—	—	—	—	—	-4	1	-3	—	-6	—	5
3(3)	6	—	—	—	—	—	—	—	5	-1	4	-2	—	—	—	—	—	—	—	2	-5	1	-6	-4
4(4)	—	6	—	—	—	—	5	—	—	-2	-3	1	—	—	—	-1	-5	2	-6	—	—	—	—	3
5(1)	—	1	—	—	—	—	-4	2	-3	6	—	—	—	—	-6	—	4	—	—	—	3	—	-1	-2
6(2)	-3	-4	—	—	—	—	—	1	-2	-5	—	—	—	—	5	—	—	—	4	—	-1	2	3	—
7(1)	—	—	—	-11	—	—	9	—	—	-8	12	10	-12	11	—	—	—	-9	—	—	—	8	-10	—
8(2)	-12	—	—	—	11	—	—	—	-10	7	9	—	-11	12	—	—	—	—	—	—	10	-7	-9	—
9(3)	—	—	—	—	—	10	-7	—	—	12	-8	-11	—	—	—	11	—	7	-10	—	—	—	8	-12
10(4)	—	—	11	—	—	-9	—	12	8	—	—	-7	—	—	—	—	—	—	9	—	-8	-12	7	-11
11(1)	—	—	-10	7	-8	—	—	—	-12	—	—	9	8	-7	12	-9	—	—	—	—	—	—	—	10
12(2)	8	—	—	—	—	—	—	-10	11	-9	-7	—	7	-8	-11	—	—	—	—	—	—	10	—	9

schedules proposed after activation of inter-league constraints

team	slots																							
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1(1)	—	—	-6	—	3	2	-4	—	—	—	—	-5	—	6	—	—	-3	—	5	-2	4	—	—	—
2(2)	—	6	—	—	—	-1	—	-3	—	5	4	—	-4	—	—	—	—	—	—	1	—	-6	-5	3
3(3)	—	—	4	5	-1	—	—	2	—	—	—	6	-5	—	-6	—	1	—	-4	—	—	—	—	-2
4(4)	6	5	-3	—	—	—	1	—	—	—	-2	—	2	—	-5	—	-6	—	3	—	-1	—	—	—
5(1)	—	-4	—	-3	—	—	—	—	6	-2	—	1	3	—	4	-6	—	—	-1	—	—	—	2	—
6(2)	-4	-2	1	—	—	—	—	—	-5	—	—	-3	—	-1	3	5	4	—	—	—	—	2	—	—
7(1)	-9	—	12	—	—	—	-11	-10	—	—	8	—	—	-12	—	—	—	10	—	—	-8	9	—	11
8(2)	—	—	—	-11	10	—	9	-12	—	—	-7	—	12	—	—	—	—	—	11	-9	7	—	—	-10
9(3)	7	12	—	—	—	—	-8	—	11	-10	—	—	—	10	—	-11	—	-12	—	8	—	-7	—	—
10(4)	—	—	—	—	-8	11	—	7	—	9	—	-12	—	-9	—	—	—	-7	—	-11	—	12	—	8
11(1)	—	—	—	8	—	-10	7	—	-9	12	—	—	—	—	—	9	—	—	-8	10	—	—	-12	-7
12(2)	—	-9	-7	—	—	—	—	8	—	-11	—	10	-8	7	—	—	—	9	—	—	—	-10	11	—

Figure 1: Generated schedules: The table entry gives the opponent. If the opponent number is ≤ 0 then the meeting will take place at the opponent's venue. The number in brackets right of the team number indicates the team's club.

To determine the number of home matches of the teams of club c in a slot t we use the binary indicator decision variables y_{ijs}^{HOME} which is 1 if and only if match $(i; j)$ is scheduled in slot s (16). Constraint (17) then limits the number of matches for each slot to the venue capacity $VC(c)$ fulfilling R_7.

4 First Computational Experiments

In order to verify the proposed model we have setup an artificial championship with two leagues ($CP = \{1; 2\}$). League 1 is composed by six teams $\{1; 2; 3; 4; 5; 6\}$ and league 2 is composed by the other six teams $\{7; 8; 9; 10; 11; 12\}$. The teams are delegated by four clubs. Teams 1, 5, 7 and 11 are associated with club 1 while teams 2, 6, 8 and 12 are formed by players from club 2. Team 3 and 9 are associated with club 3 and team 4 and 10 belong to club 4. We have the venue capacity vector $VC = (1; 1; 1; 1)$, i.e. each club allows one home match in a slot. Substitution opportunities must be considered between the following pairs of teams: (1;5), (5;7), (7;11), (2;6), (6;8), (8;12), (3;9) as well as (4;10). The complete season comprises 24 slots and the fall round is formed by the first twelve slots. The minimal number of slots between

two consecutively scheduled matches of a team is $N^{DIST} = 0$, so that two adjacent slots can be used for matches in which the same team is involved.

We use CPLEX to solve the resulting model instance. In a first experiment, we do not consider any inter-league constraints. The solver is granted 15 minutes processing time. After approx. 3 minutes an optimal solution is found. In a second experiment we activate the two inter-league constraints. The solver returns an solution after 15 minutes. This solution was not proven to be optimal. Fig. 1 shows the generated schedules for the championship. In the upper table, the gray shaded entries contribute to conflicts due to venue capacity excess or disregarded substitution opportunities. There are neither home-away alternation errors nor regularity errors. The lower table contains the schedules with activated cross-league constraints. Nearly all matches are re-scheduled and shifted into another slot in order to meet the requirements of the inter-league constraints. However, seven home-away-alternation errors cannot be eliminated within 15 minutes computational time (the gray shaded entries contribute to these failures).

5 Conclusions

We have proposed and validated a mathematical optimization model of a complex decision problem from non-commercial sports management. For the first time, the simultaneous determination of schedules for several round-robin competitions sharing scarce venues is addressed. Furthermore, it is the first time that substitution options are considered in the schedule determination. The next research steps comprise the conduction of further computational experiments even with larger championships in order to identify the most critical factors for the simultaneous schedule determination. Due to the quite high complexity of the proposed model we are going to develop a heuristic approach that is able to handle championships of realistic size. In this context, we can use comprehensive historized real planning data. Furthermore, additional regularity measures will be evaluated.

References

[1] Dirk Briskorn. Alternative ip models for sport leagues scheduling. In Jörg Kalcsics and Stefan Nickel, editors, *Operations Research Proceedings 2007*, volume 2007 of *Operations Research Proceedings*, pages 403–408. Springer Berlin Heidelberg, 2008.

[2] E.K. Burke, D. De Werra, J.D.L. Silva, and C. Raess. Applying heuristic methods to schedule sports competitions on multiple venues. In *Proceedings of the Fifth International Conference on Practice and Theory of Automated Timetabling*, pages 441–444, 2004.

[3] B.L. Golden, S. Raghavan, and E.A. Wasil, editors. *The Vehicle Routing Problem: Latest Advances and New Challenges*. Springer, New York, 2008.

[4] S. Knust. Scheduling non-professional table-tennis leagues. *European Journal of Operational Research*, 200:358–367, 2010.

[5] R. Ramli, C.-J. Soong, and H. Ibrahim. A sports tournament scheduling problem: Exploiting constraint-based algorithm and neighborhood search for the solution. In *Proceedings of the World Congress on Engineering 2012, Vol. 1*, 2012.

[6] J. Schönberger, D.C. Mattfeld, and H. Kopfer. Memetic algorithm timetabling for non-commercial sport leagues. *European Journal of Operational Research*, 153:102–116, 2004.

Dynamic Forecasting Model Based on Ranking System for Hockey Match

Blanka Šedivá

University of West Bohemia, Plzeň, Czech Republic
sediva@kma.zcu.cz

Abstract

The ranking of sports teams is of significant importance to those who are involved with or interested in the various professional and amateur leagues that exist around the world. We present a modified ranking algorithm based on the Perron-Frobenius theorem and the application of this algorithm in the context of hockey match prediction. To examine the behavior of our approach we implemented the method for prediction of results of matches for the highest-level ice hockey league in the Czech Republic from the 2008 through 2015.

1 Introduction

Ranking and selection are related statistical problems with numerous applications and a good algorithm for ranking has many potential uses. Ranking methods can provide key information not only for sports fans and betting companies and their customers but also in other different areas. Although most publications are focused on applications of ranking models in team sports and lottery markets, e.g. [4], [2], [11], examples from medical research, school effectiveness [6] or a website's ranking [7] can be found.

In this paper four different ranking schemes and methods of estimation of probabilities of winning for home or away team are described. In section 2.1 is formulated the well known Keener's ranking model as a linear eigenvalue problem, the first method used for an estimation of probabilities the last ranking score. The second method is based on a mean of ranking. In section 2.2 are described two advanced methods based on Keener's model including time delays.

In section 4 we will test and compare predictive abilities of our models on the set of data that contains results of regular seasons matches from the 2008/2009 season to 2014/2015 season of the highest ice hockey league in the Czech Republic.

2 Ranking Models

A number of articles exist in the statistics literature dealing with the ranking of sport teams. Two common terms associated with sports models are ranking and rating. A ranking of N teams places them in order of relative importance, with the best team receiving rank one. A rating of the same teams describes the degree of relative importance of each team. The ratings carry more information than the ranks, they provide us with the degree of relative importance and can be used for estimation of expected probabilities of home and away teams winning and drawing.

Analysts have proposed several different approaches for ranking sports teams, a survey of basic and advanced ranking models is described in detail for example in [3].

2.1 Keener's Ranking Methods

James P. Keener [5] formulated eigenvalue problem and provided ranking existence and uniqueness conditions that are compatible with the Perron-Frobenius theorem; a nonlinear version of the model is also described. He proposed his ranking model based on dominant eigenvalue of nonnegative matrix in 1993 [5]. The Keener ranking model makes two fundamental assumptions. His first assumption is that the strength of teams expressed in scores depends on both the outcome of the interaction and the strength of its opponents. He defines the strength of teams i to be

$$s_i = \frac{1}{n_i} \sum_{j=1}^{N} a_{ij} r_j, \tag{1}$$

where a_{ij} is non-negative value that depends upon the outcome of the match between i and j, r_j is rating of team j, n_i is number of matches played by team i and N is total number of teams. The matrix A with entries a_{ij} is often called a preference matrix. Keener assigns the value of a_{ij} as follows

$$a_{ij} = h\left(\frac{S_{ij}+1}{S_{ij}+S_{ji}+2}\right), \tag{2}$$

where S_{ij} is the number of points i scored against j and h is the strictly increasing nonlinear function $h(x) = 1/2 + 1/2\,\mathrm{sgn}\,(x-1/2)\sqrt{|2x-1|}$ that minimizes the incentive of the winning team to "run up" the score.

Keener's second assumption is that the strength of a team should be proportional to a team's rating. In equation form

$$\mathbf{s} = A\mathbf{r} = \lambda \mathbf{r}, \tag{3}$$

where A is an $N \times N$ matrix with a_{ij} as components, $\mathbf{r}$ is a column vector of N ratings, and $\mathbf{s}$ is a column vector of N strengths. By the Perron–Frobenius Theorem, this eigenvalue-eigenvector equation will have a unique, up to a scalar multiple, and positive solution, provided that the matrix A is non-negative and irreducible. More information about linear and also extended nonlinear Perron–Frobenius theory can be found in e.g.[8].

2.2 Dynamic Ranking Methods

Team strength changes over time, so the Keener model can be used and ranking vector $\mathbf{r}$ can be recalculated several times during the season. The outcome of each match modifies matrix A and implies changes in the ranking vector. We use a similar approach as [10] and estimated ranking vector for each time t when some match is played. Let us denote the time t_1 of the first date, t_2 of the second date etc. ($t_1 < t_2 < \dots$). The ranking vector depends on stability of matrix A and it is more variable at the beginning of the season when entries a_{ij} that are computed as a sum of scores are changed more rapidly. For reduction of the unpleasant properties we have used for estimation of probabilities of home team win, draw and away team win two different approaches: (a) to calculate expected outcomes of matches at time t_k based on ranking at previous time t_{k-1} and (b) to calculate expected outcomes of matches at time t_k based on the simple mean ranking from 0 to t_{k-1}.

The articles [1] and [10] was our inspiration for adding model parameters reflecting time variability between played matches. One possibility is a direct method that assumed a current ranking of team as a weighted combination of previous rankings. The weighting function in time of estimation is $w_\tau = \exp(-\alpha\tau)$, where α is a positive parameter and τ is time delay.

Estimations of expected outcomes of matches at the time t_k is determined on basis of weighted ranking

$$r(t_k) = r(t_1)\ \exp\left(-\alpha(t_k - t_1)\right) + r(t_2)\ \exp\left(-\alpha(t_k - t_2)\right) + \cdots + r(t_{k-1})\ \exp\left(-\alpha(t_k - t_{k-1})\right), \tag{4}$$

where $r(t_1), r(t_2), \ldots, r(t_k)$ are values of team rankings at times $t_1, t_2, \ldots, t_k$ and $t_k - t_1, t_k - t_2, \ldots, t_k - t_{k-1}$ are time delays.

The second method that takes into consideration time difference between matches focuses on weighting of matrix A. The innovation is that the process of calculation of scores S_{ij} used in equations (2) is complemented by adaptive coefficient λ. We suppose that each team played with each other several times in one season; we denote $S_{ij}(t)$ as the cumulative number of points that team i scored against opponent team j in matches played at times $\leq t$. The adaptive cumulative score is calculated using relationship

$$S_{ij}(t_k) = S_{ij}(t_{k-1}) \cdot \exp\left(-\lambda(t_k - t_{k-1})\right) + s_{ij}, \tag{5}$$

where $S_{ij}(t_{k-1})$ is cumulative score at time t_{k-1} and s_{ij} is the count of scored points from time t_{k-1} to time t_k. The adaptive approach reflects the time factor and the points scored in recent time are more influential to the strength of the team.

3 Assessing the Probability of Winning

One of the most common applications of ranking models is to predict the probability that team i will beat team j and to predict the outcome of matches between rivals. Suppose the ranking vector $\mathbf{r}$ is defined so that probability π_{ij} that team i beats team j is $\pi_{ij} = \frac{r_i}{r_i + r_j}$. This approach allows us to estimate the probability of win of home teams and win of away teams. To estimate the probability of a draw modified rules can be used,

$$\begin{aligned} p^H &= \frac{r_i}{r_i + r_j} - \vartheta_1, \\ p^A &= \frac{r_i}{r_i + r_j} - \vartheta_2, \\ p^D &= \vartheta_1 + \vartheta_2, \end{aligned} \tag{6}$$

where r_i, r_j are rankings of home and away teams and $\vartheta_1, \vartheta_2 \in [0, 1/2)$ are parameters of model relevant in situation when strength of teams are similar ($r_i \approx r_j$) and draw result can be expected.

The relationship between probabilities and ranking vector can be also derived from the idea that a team has an actual performance that is a random variable with mean r_i resp. r_j and variance σ_i^2 resp. σ_j^2. Then the probability of a home team win can be estimated as

$$\begin{aligned} p^H &= 1 - F\left(r_i - r_j + \vartheta_1; \boldsymbol{\Theta}\right), \\ p^A &= F\left(r_i - r_j - \vartheta_2; \boldsymbol{\Theta}\right), \\ p^D &= 1 - \left(p^H + p^A\right), \end{aligned} \tag{7}$$

where r_i, r_j are rankings of home and away teams and $F(\cdot)$ is appropriately chosen cumulative distribution function (e.g. normal cumulative distribution function) with parameters $\boldsymbol{\Theta}$ and parameters ϑ_1, ϑ_2 supplement the probability rules so that draw is also included.

4 Empirical Application

4.1 Data Description

We analyze a time series of 7 years of ice hockey match results from the highest level ice hockey league in the Czech Republic. In this paper we have used data from websites LiveSport s.r.o. [9] where data including regular season matches, play off matches, prerelegation matches and relegation/promotion matches are available. Our models have been calibrated and tested only for results after 60 min, e.g. a game can end with a tie. Each season 14 teams played Extraliga and each pair of teams played with each other exactly four times, two games as a home team and two games as a visiting team. In total, each team played 52 matches in a season. The ten best teams of the regular season advance to the play off and the remaining four teams play the pre-relegation where all points from the regular season are counted. Later, the worst team from the prerelegation competes in the relegation/promotion with the winner of the second-level ice hockey league in the Czech Republic for the place in the Extraliga. The Extraliga was played by 14 teams in each season but thanks to the possibility of relegation and promotion, the Extraliga and relegation/promotion were played by 18 different teams in total in mentioned seasons.

For testing of quality of predictive models and ability of betting strategy the data are divided into two sets. The first set contains regular season matches from the 2008/2009 season to the 2013/2014 season and the second set contains regular season matches from the last season, the 2014/2015 season.

4.2 Specification of Testing Models

The following algorithms were used for estimation of ranking vectors and probabilities of winning home team, draw and winning away team:

KM-LR the ranking vectors are calculated by Keener's approach, described in section 2.1, for each day when at least one match was played, probabilities of winning home team, away team and draw are estimated using equations (6) (version (a)) or equations (7) (version (b)), the last available ranking score was used;

KM-MR the algorithm for ranking computation is the same as in the model KM-LR but mean ranking score is used;

KM-WR the algorithm for ranking computation is the same as in the model KM-MR but weighted ranking score defined in equations (4) is used;

AKM-LR the algorithm for ranking is based on adaptive computing of matrix A with time delay factor defined in (5), for each day when at least one match was played, probabilities of winning home team, away team and draw are estimated using equations (6) (version (a)) or equations (7) (version (b)), the last available ranking score was used.

The predictive performance of each model was compared to other models, as well as to the bookmakers by function

$$C = \sum_{m=1}^{M} \left(\delta_m^H \ln p_m^H + \delta_m^D \ln p_m^D + \delta_m^A \ln p_m^A \right), \tag{8}$$

where δ_m is an indicator function, e.g. $\delta_m^H = 1$ if match m ends with home win, and p_m^H, p_m^D, p_m^A are probabilities estimated by each model of home team win, draw and away team win [10]. Function C can be considered as a measure of prediction error of outcomes because for outcomes with low predicted probability it gives a large negative number whereas for outcomes with high predicted probability it gives a negative number close to zero.

4.3 Obtained Results

The optimal parameters of models were obtained by maximization of the function C and the best model according to the criterion C was selected for each season. The values of function C for the bookmakers and for the selected model for each season are shown in Table 1. The percentages of correctly estimated matches, e.g. matches for which the real result of the match corresponded with the highest estimated probability, are shown here too.

Season	Bookmakers		Model		Fitting parameters
	Value of C		Value of C	Corr. est.	
2008/2009	-377.8	KM-WR (b)	-374.7	51.4%	$\Theta_1 = 0.253, \Theta_2 = 0.234$
2009/2010	-380.8	KM-WR (a)	-385.1	46.2%	$\vartheta_1 = 0.038, \vartheta_2 = 0.198$
					$\alpha = 4.010$
2010/2011	-364.0	KM-LR (a)	-367.2	51.4%	$\vartheta_1 = 0.000, \vartheta_2 = 0.180$
		KM-WR (a)			$\vartheta_1 = 0.000, \vartheta_2 = 0.179$
					$\alpha = 2.922$
2011/2012	-380.6	AKM (a)	-378.8	47.5%	$\vartheta_1 = 0.002, \vartheta_2 = 0.224$
					$\lambda = 0.237$
2012/2013	-377.5	KM-WR (a)	-378.2	49.2%	$\vartheta_1 = 0.009, \vartheta_2 = 0.212$
					$\alpha = 4.276$
2013/2014	-359.2	KM-LR (a)	-367.7	51.7%	$\vartheta_1 = 0.014, \vartheta_2 = 0.193$
		KM-WR (a)			$\vartheta_1 = 0.015, \vartheta_2 = 0.183$
					$\alpha = 4.501$

Table 1: Comparison of bookmakers and the selected best model for each season.

The results included in the table show that the most successful algorithm is KM-WR model that used the calculation of ranking vector by Keener's methods, computing probabilities as a ratio of home team ranking and sum of home and away team ranking with modification involving probability of draw, see equations (6). The estimation of probabilities is based on weighted historical value of rankings with respect to time delay.

4.4 Model Evaluations: in Sample and out of Sample

The last part of the article deals with an application of testing models for predictions of outcomes of matches in season 2014/2015. The model KM-WR was selected pursuant the results of previous experiments. The parameters of the optimal model were fitted by using data from previous seasons, resp. from seasons 2008/2009-2013/2014. The optimization process of model KM-WR(a) based on data from 2008 to 2014 that included results of 2 512 matches has given us parameters $\vartheta_1 = 0.011, \vartheta_2 = 0.209$ and $\alpha = 3.455$.

The value of function C for the selected model is -364.33 and the percentage of correctly estimated matches, i.e. when the real result of the match corresponds with the highest estimated probability was 52.20%. These results are comparable with probabilities derived on bookmakers bets, in this case the value of function C is -355.19. Average margin of bookmakers for our data is 7.00% and comparison of probabilities for the season 2014/2015 are shown in Table 2.

	Home win	Draw	Away win
Real results	52.47%	16.21%	31.32%
Bookmakers	46.94%	22.01%	31.05%
KM-WR (a) model	49.06%	22.00%	28.94%

Table 2: Comparison of bookmakers and the KM-WR(a) model from 2008 to 2014.

5 Conclusion

The two basic and two advanced models based on ranking methods were analyzed and were applied to real data sets. Our results show that extension of model specification with time delay factors can improve the prediction qualities of models. The use of time delay weighted ranking score comes to light as an auspicious experiment. For anyone serious about the business of game prediction more variables must be taken into account, but the estimation of probabilities of winning home and away teams built on ranking models with respect to time delay between matches can be used for developing betting strategy.

References

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

[2] Anjela Y Govan, Amy N Langville, and Carl D Meyer. Offense-defense approach to ranking team sports. *Journal of Quantitative Analysis in Sports*, 5(1), 2009.

[3] Anjela Yuryevna Govan. *Ranking Theory with Application to Popular Sports* -. ProQuest, Ann Arbor, 2008.

[4] Donald B. Hausch and W.T. Ziemba. *Handbook of Sports and Lottery Markets* -. Elsevier, Amsterdam, 1. aufl. edition, 2011.

[5] James P Keener. The Perron-Frobenius theorem and the ranking of football teams. *SIAM review*, 35(1):80–93, 1993.

[6] Nan M Laird and Thomas A Louis. Empirical bayes ranking methods. *Journal of Educational and Behavioral Statistics*, 14(1):29–46, 1989.

[7] Amy N. Langville and Carl D. Meyer. *Google's PageRank and Beyond - The Science of Search Engine Rankings*. Princeton University Press, Kassel, 2011.

[8] Bas Lemmens and Roger Nussbaum. *Nonlinear Perron-Frobenius Theory*, volume 189. Cambridge University Press, 2012.

[9] LiveSport s.r.o. 2015. Odds portal: Odds comparison, sports betting odds. `http://http://http://www.livesport.cz/`, last viewed April 2015.

[10] Marek Patrice, Šedivá Blanka, and Ťoupal Tomáš. Modeling and prediction of ice hockey match results. *Journal of Quantitative Analysis in Sports*, 10(3), 2014.

[11] Robert P Schumaker, Osama K Solieman, and Hsinchun Chen. *Predictive modeling for sports and gaming*. Springer, 2010.

Statistical Modelling and Forecasting Association Football Results

José Sellitti Rangel Junior

University of Salford, Salford, U.K.
j.sellittirangeljunior@edu.salford.ac.uk
PhD Supervisor: Prof. Philip Scarf

Abstract

This paper provides a brief summary of the literature review of statistical modelling in forecasting association football results and a critique of the methodology used up to date. When analysing time series data to study the dynamics of team strengths, it is important to test whether the process is nonstationary and if so, how it violates stationarity. Additionally, omitting variables in the time series model leads to biased estimators that will affect predictions for team strengths and consequently, match results. These two issues have not been investigated in the literature: unit-root testing of the ARIMA process for the dynamics of team strengths and testing for the existence of trends and structural breaks in the time series. A model is presented which addresses these issues and its predictive performance is compared with a benchmark model which does not.

1 Introduction

If a statistical model for predicting football results was able to predict better than the bookmakers' lines, it should generate positive returns. Therefore there has been an interest from statisticians to constantly improve predictive models in the literature. Additionally, football is the most popular sport on the globe and there is an added incentive in investigating a topic about which some authors are passionate. However, when it comes to generating positive returns against the bookmakers, other issues come into play, such as betting strategies and money management techniques such as the Kelly criterion [8]. Simply put, the statistical model is only one of the ingredients involved.

Several authors have proposed modelling team strength dynamics in different ways. However, the dynamic models proposed in the literature have always been simple; typically $AR(1)$. Thus, investigating the existence of trends and the possibility of higher order autoregressive processes is important to make sure that there is no omitted variable bias in the coefficients. Additionally, no author has decided to test for unit-root tests of the time series of team strengths. This is imperative to determine firstly whether the process is stationary and if not, how it violates stationarity so we can model the time series for forecasting purposes.

This paper addresses two challenges to the literature: i) how to model the dynamics of the strength parameters and ii) testing for the existence of trends and structural breaks in the series. Firstly the team strength parameters in the sample are estimated using a rolling window. Then, an adequate unit root test which accounts for structural breaks is carried out to investigate whether parameter estimates in the ARIMA modelling change after the transfer window. This is motivated through the findings in [11].

Firstly, a brief literature review describes how other authors have modelled team strength dynamics, and then describes developments in unit root tests which account for structural breaks. In Section 3, I present the proposed methodology. Since this is a work in progress, results will only be presented at the MathSport International Conference 2015.

2 Literature Review

2.1 Model Estimation

The first paper that really discussed modelling and forecasting football results was [15], which showed that football results follows a negative binomial distribution. The double-Poisson model presented by [13] demonstrated that a double-Poisson model provided a better fit. Maximum likelihood estimation (MLE) of the log-likelihood function was used to estimate attack and defence parameters for the teams in the sample, as well as a home advantage parameter. [4] estimate a pseudolikelihood by enhancing the double-Poisson estimated in [13] by estimating a dependence parameter. Additionally, they added more weight to recent results in the likelihood function. [7] also consider MLE by using a Newton-Raphson approach and inflating the probability of observing draws. [14] add flexibility to modelling by using certain copula functions which allow negative dependence to be captured, applicable to international football.

2.2 Dynamic Models and Unit Root Testing

[21] not only model the strength parameters of each team, but use Brownian motion to allow the strength to vary between two particular points in time. [10] also allows team strengths to evolve in a random walk process - a nonstationary model. [2] develop an $AR(1)$ in a state space model to model strength dynamics. [17] utilises Bayesian dynamic models and allows the attack and defence parameters to vary stochastically over time, attributing a normal distribution to them. He estimates what he calls the "evolution" variance σ^2 by maximising the one-step ahead predictive probability of the model within a football season in the Scottish Premier League. This is the volatility parameter in the strength dynamics. A higher evolution variace indicates a higher probability of observing a deviation from last period's strength value. [11] on the other hand model a stationary $AR(1)$ process.

Our paper provides a methodology which allows flexibility in the autoregressive process used to model the dynamics of strength parameters. Firstly it allows for an $AR(p)$ process, where p is determined after analysing the sample autocorrelation function (SACF) and partial autocorrelation functions (PACF) of the autoregressive process. If the dynamics follow an $AR(p)$ where $p \neq 1$, omitting a variable would lead to a biased estimator of the $AR(1)$ coefficient and in turn, a biased prediction of the one-step ahead forecast for the parameter.

Additionally, our paper investigates the existence of trends in the autoregressive process. By observing plots of team strengths over time in several papers (for example:[4], [10], [11]), we can observe that the data generating process (DGP) does not exhibit the properties attributed to a stationary process. Nevertheless, the approach to which nonstationary processes are modelled depends on whether a unit root is present in the autoregressive parameter. Essentially, it becomes a question of whether the time series process is trend stationary (TS) or difference stationary (DS); a topic very much discussed in macroeconomic and financial time series econometrics. However, this is the first time it is being implemented in a football forecasting context.

As a result, the literature on unit root testing is quite extensive, and there are several issues which must be considered before determining what test should be carried out. One of the issues is structural breaks, where there is a breakdown in the statistical relationship between variables in a model through a change in the parameter coefficient at some breakpoint t.

Traditional unit root tests such as the Augmented Dicky-Fuller (ADF) test [22], and the PP test [20] have been criticised for having low power [3]. This means they tend to accept the null hypothesis of unit root more frequently than they should, leading to an increase in type

II error (accepting the null hypothesis when it is false). [23] presents Monte Carlo evidence of size distortions, since by adding more parameters, the critical value on the 5% level increases and it becomes harder to reject the null hypothesis of a unit root. For more details about these problems and for test alternatives, refer to [12].

2.3 Unit Root Testing and Structural Breaks

[11] implement an innovative approach to address changes to the strength parameters through augmenting the variance of the random shocks within the $AR(1)$ process at specific time periods - transfer windows. They argue that since the autoregressive coefficient is quite close to 1, the shocks will be sufficiently persistent to model a change in strength of the team. This finding is one of the main motivations behind this paper. Rather than capturing dynamics through an enhanced volatility in the shock during the summer transfer window, structural breaks in the parameter coefficients of an autoregressive equation is considered. This has an impact in unit root testing. [18] says that if the time series has a structural break, traditional unit root tests cannot reject the unit root null hypothesis. This is backed by other papers such as [9]. This could become a problem if the time series has multiple breaks, which arguably the case in this paper (there should be a break at the end of every season). A good survey of the literature is made by [5], where several tests are carried out on the [16] macroeconomic dataset. However, they provide no conclusion as to which is the most adequate unit root test when accounting for structural breaks. [5] mention there are criticisms of [18] in that the break in trend is known *a priori*.

[6] solve two important problems: how to estimate the break point when a break actually occurs and how the break point estimator behaves when no break occurs. Even though the test proposed only allows for a single break in trend, [1] allows for two or more. Any more than two becomes quite problematic. Therefore, given the desirable finite sample properties of [6], this paper intends to follow the methodology proposed by them to make sure that, if a break point does exist, it can be estimated correctly. This test can then be used across the sample for a single team several times.

3 Proposed Methodology

3.1 The Model

Match results are available from the `www.football-data.co.uk` website for the past 22 seasons of the top 4 divisions in UK football, although for the purpose of this paper, only some seasons from the Premiership will be used. This is directly available in the website for download as a csv file.

This paper uses very similar notation found throughout the literature. The final score of a football match between the home team i and away team j at time t is defined to be (X_{it}, Y_{jt}), for $i, j = 1, \ldots, N$ teams and $t = 1, \ldots, T$. Thus

$$P(X_{it}, Y_{jt}) = \tau_{(\lambda,\mu)}(x, y)\frac{\lambda^x \exp(-\lambda)}{x!}\frac{\mu^y \exp(-\mu)}{y!} \tag{1}$$

$$\lambda = \alpha_i \beta_j \gamma \tag{2}$$

$$\mu = \alpha_j \beta_i \tag{3}$$

The expected means of λ and μ at time t depend on the attack parameters α_{it} and α_{jt}, the defence parameters β_{it} and β_{jt} of the home and away teams respectively as well as the home advantage γ. The τ function, like in [4]

$$[\tau_{(\lambda,\mu)}(x,y) = \begin{cases} 1-\lambda\mu\rho & \text{if } x=y=0 \\ 1+\lambda\rho & \text{if } x=0,\ y=1 \\ 1+\mu\rho & \text{if } x=1,\ y=0] \\ 1-\rho & \text{if } x=y=1 \\ 1 & \text{otherwise} \end{cases} \tag{4}$$

Since (X_{it}, Y_{jt}) follow a bivariate Poisson, the expected means must be non-negative, and this applies for the covariance term as well. This is the case in league football, as [13], [4], and [7] all find a positive correlation between goals.

In order to estimate these parameters, the constraint on the attack parameters is set

$$n^{-1}\sum_{i=1}^{n}\alpha_i = 1 \tag{5}$$

The likelihood is downweighted in the same way as [4] in order to give more weight to recent match results

$$L_t = \prod_{i,t=1}^{J}[\tau_{(\lambda,\mu)}(x,y)\frac{\lambda^x \exp(-\lambda)}{x!}\frac{\mu^y \exp(-\mu)}{y!}]^{\psi(t-t_k)} \tag{6}$$

where t_k represents the match played at time $t_k < t$, number of matches in the sample is $J = \frac{NT}{2}$, and the same downweighted exponential function

$$\psi(t) = \exp(-\xi t) \tag{7}$$

This would generally require estimation of the ξ parameter. In order to speed estimation it makes sense to fix the parameter $\xi = 0.0065$, which is the estimate in [4], so a similar result is expected.

By MLE, we can obtain estimates for α_i and β_i for all teams at time t. In vector form, this can be denoted as $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N)'$.The same applies for $\boldsymbol{\beta}$.

In order to construct a time series to study the dynamics of the strength parameters, a rolling window is required. Then the different vectors can be stacked together in a matrix to generate a panel dataset for α_{it} and β_{it}

$$A_{i,t} = \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} & \cdots & \alpha_{1,T} \\ \alpha_{2,1} & \alpha_{2,2} & \cdots & \alpha_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{N,1} & \alpha_{N,2} & \cdots & \alpha_{N,T} \end{pmatrix}, B_{i,t} = \begin{pmatrix} \beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,T} \\ \beta_{2,1} & \beta_{2,2} & \cdots & \beta_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{N,1} & \beta_{N,2} & \cdots & \beta_{N,T} \end{pmatrix}$$

The next step requires building an autoregressive time series model for α_{it} and β_{it} in order to carry out the appropriate unit root test.

A time series process for α_t would be generated by

$$\alpha_{it} = cons + \delta t + \theta DT_t(\tau_0) + u_t, t = 1, \ldots, T, \tag{8}$$

$$u_t = \phi_T u_{t-1} + \epsilon_t, t = 2, \ldots, T \tag{9}$$

where $DT_t(\tau_0) = (t > [\tau_0 T])(t - [\tau_0 T])$ with $[\tau_0 T]$ being the potential trend break point with associated break fraction τ_0.

The details on how the test works is given in Sections 2, 3 and 4 of [6]. Suffice to say that it is an extension of the QD de-trended ADF-type unit root test in [19].

3.2 Forecasting

After building the model and obtaining estimates, the next step involves analysing its forecasting power. Both in-sample and out-of-sample forecasting will be carried out and the percentage of successful predictions for the outcomes (win-draw-loss) reported. Given the parameter estimates, we can easily obtain $\hat{\alpha}_{it+1}$ and $\hat{\beta}_{it+1}$ to predict the outcome of match $J + 1$. Another measure of forecasting would be to see if the model can produce positive returns by implementing a betting strategy, but that is beyond the scope of this paper.

4 Closing Remarks

This is a novel application of unit root testing outside of the field of macro and financial econometrics. Motivated by findings in [11], unit root testing with a possible break in trend is carried out. If a structural break exists, it will open up the research into why they exist, whether they are deterministic in order to attempt to predict growth patterns in team strength dynamics.

References

[1] Josep Lluís Carrion-i Silvestre, Dukpa Kim, and Pierre Perron. Gls-based unit root test with multiple structural breaks under both the null and the alternative hypotheses. *Econometric Theory*, 25(6):1754–1792, 2009.

[2] Martin Crowder, Mark Dixon, Anthony Ledford, and Mike Robinson. Dynamic modelling and prediction of english football league matches for betting. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 51(2):157–168, 2002.

[3] David N. DeJong, John C. Nankervis, N. E. Savin, and H. Whiteman, Charles. The power problems of unit root test in time series with autoregressive errors. *Journal of Econometrics*, 53(1–3):323–343, 1992.

[4] Mark J. Dixon and Stuart G. Coles. Modelling association football scores and inefficiencies in the football betting market. *Applied Statistics*, 46(2):265–280, 1997.

[5] John Glynn, Nelson Perera, and Reetu Verma. Unit root tests and structural breaks: a survey with applications. *Revista de Métodos Cuatitativos para la Economía y la Empresa = Journal of Quantitative Methods for Economics and Business Administration*, 3(1):63–79, 2007.

[6] David Harris, David I. Harvey, Steven J. Leybourne, and A. M. Robert Taylor. Testing for a unit root in the presence of a possible break in trend. *Econometric Theory*, 25(6):1545–1588, 2009.

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

[8] J. L. Kelly Jr. A new interpretation of information rate. *The Bell System Technical Journal*, 35(4):917–926, 1956.

[9] In-Moo Kim. *Structural Change and Unit Roots*. PhD thesis, University of Florida, 1990.

[10] Leonhard Knorr-Held. Dynamic rating of sports teams. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(2):261–276, 2000.

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

[12] G. S. Maddala. *Introduction to Econometrics*. Wiley, California, 3 edition, 2001.

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

[14] Ian McHale and Philip Scarf. Modelling soccer matches using bivariate discrete distributions with general dependence structure. *Statistica Neerlandica*, 61(4):432–445, 2007.

[15] Michael Joseph Moroney. *Facts from figures*. Penguin, London, 3 edition, 1956.

[16] Charles Nelson and Charles Plosser. Trends and random walks in macroeconomics time series: some evidence and implications. *Journal of Monetary Economics*, 10(2):139–162, 1982.

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

[18] Pierre Perron. The great crash, the oil price shock, and the unit root hypothesis. *Econometrica*, 57(6):1361–1401, 1989.

[19] Pierre Perron and Gabriel Rodriguez. Gls detrending, efficient unit root tests and structural change. *Journal of Econometrics*, 115(1):1–27, 2003.

[20] Peter C. B. Phillips and Pierre Perron. Testing for a unit root in time series regression. *Biometrika*, 75(2):335–346, 1988.

[21] Havard Rue and Oyvind Salvesen. Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(3):399–418, 2000.

[22] Said E. Said and David A. Dickey. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3):599–607, 1984.

[23] G. William Schwert. Tests for unit roots: A monte carlo investigation. *Journal of Business & Economic Statistics, American Statistical Association*, 7(2):147–159, 1989.

Analysis of the 2014 Formula 1 Constructors Championship using a modified Condorcet Method

João Carlos Correia Baptista Soares de Mello, Catarina Loureiro de Oliveira Mourão, Lidia Angulo Meza and Silvio Figueiredo Gomes Junior

Federal Fluminense University, Niteroi, Rio de Janeiro, Brazil
jcsmello@producao.uff.br, Catarina.mourao@hotmail.com, lidia_a_meza@pq.cnpq.br

Abstract

The ordinal methods are widely used to establish rankings in sports competitions because sporting results are on an ordinal scale. However, these methods assume that the decision maker is highly rational and provide a full ranking. This paper analyses the case in which the decision maker is weakly rational. In this case, the decision maker respects the transitivity of preferences but not indifference. It also shows how to adapt the Condorcet method to deal with this situation, using the results of the 2014 Formula 1 Constructors World Championship.

1 Introduction

A tournament is composed of a set of various games, or rounds, the results of which are aggregated to establish the final result of the competition, according to the description given by Haigh [2]. In some cases there is a complete aggregation, in others, each result indicates the next games to be held. In either of the two cases, if each game or round is interpreted as a criterion, or a decision maker, the final result of the championship is a multi-criteria problem, normally ordinal, and the decision-maker is considered to be strongly rational.

The aim of this research is to study a case in which we cannot assume the decision maker to be strongly rational. We shall remember that the difference between a strongly rational decision maker and a weakly rational one is that while both respect transitivity for the preference relation, the weakly rational decision maker does not have the obligation of respecting the property of transitivity in relation to indifference [3].

This article presents a variation of the Condorcet method with weakly rational decision makers. One of the situations in which this occurs is when each alternative is evaluated not by itself, but by some components called sub-alternatives. In order to illustrate this situation, a comparison will be made of the teams which participated in the 2014 Formula 1 Constructors World Championship.

2 The Condorcet Method

The Condorcet method requires each decision maker to rank all of the alternatives according to their preferences. Then it must be checked which alternative out of each pair of alternatives is preferred by the majority of the decision makers. In this case, we can say that this alternative is preferable in relation to the other. Graphs can be drawn representing these preference relations, in which the arc belongs to the graph if, and only if, the number of decision makers who preferred u to v is greater or equal to those who preferred v to u.

The representation of the preference relations by means of a graph greatly facilitates determining the dominant and dominated alternatives (when they exist). When only one dominant alternative exists, that is the one chosen. The Condorcet method has the great disadvantage of leading to situations of intransitivity, leading to the well-known Condorcet paradox. This occurs when A is preferable to B, (A P B), B is preferable C (B P C) and C is preferable A (C P A), a situation known as the Condorcet Triplet. This means that the Condorcet method does not always lead to a pre-ranking in the set of alternatives. However, there are situations in which cycles of intransitivity do not occur. In these situations, the Condorcet method can be used without any problems.

When the alternatives present sub-alternatives, the evaluation is made through the components of the alternative. Thus, if a group of people is an alternative, the evaluation of the alternative is performed by the evaluation of each of the individuals in the group, who together form the sub-alternatives. The existence of relations of preference and indifference between alternatives are formalized only considering, for simplicitys sake, the existence of two sub-alternatives for each alternative. Let A and B be two alternatives, each with sub-alternatives A1 and A2 and B1 and B2, respectively, such that A1 P A2 and B1 P B2.

Definition 1: A P B, if and only if A1 P B1 and A2 P B2.

Definition 2: A I B, if and only if A1 I B1 and A2 I B2, or if A1 P B1 and B2 P A2, or if B1 P A1 and A2 P B2; where I represents the indifference relation.

It is easy to see that the relation A P B is transitive. However, this does not occur in the indifference relation. To illustrate this situation, there is the example with three alternatives (A, B and C), with two sub-alternatives each. For a particular decision maker, the ranking occurred as presented in Table 1.

Position	Sub-alternative
1	A1
2	B1
3	B2
4	C1
5	C2
6	A2

Table 1: Ranking of the sub-alternatives.

In this table it can be seen that A1 P B1, and B2 P A2. By Definition 2 it must be that A I B. Similarly comparing A and C we can see that A I C. However, B1 P C1 and B2 P C2, which by Definition 1 indicates that B P C, in other words, $\sim$(B I C), which demonstrates that the indifference was not transitive. This means that, considering the evaluation of the alternatives by means of the sub-alternatives, the decision maker was weakly rational.

This is a situation which occurs in championships in which the ranking of the teams is based on the ranking of the individuals who belong to those teams, as we will see in the next sections.

3 The Formula 1 World Championship

The Formula 1 Drivers World Championship determines that the champion driver of the season is the driver who achieves the highest number of points at the end of all the seasons races. The rank of the other drivers in the championship is determined by the total number of points achieved. In the first championship the 3 worst results of the 7 races held would be discarded. The scoring was as follows: 8 points for first place; 6 for second place; 4 for third; 3 for fourth;

2 for fifth place and one point for the driver who registered the fastest lap in the race. The scoring system over time can be seen in Haigh[2].

The first Constructors World Championship was won by the Vanwall team in 1958. In the majority of the seasons until 1979, only the results of the best driver in the team counted towards the scoring of the championship. In the following year, there was an alteration to the rule and the points were obtained by the sum of the results of both drivers in each team, a change which has lasted until today. Only on ten occasions has the world champion constructors team not had one of their drivers win the title of world champion. Thus, as in the case of the drivers championship, the winning team in the constructors championship is the team which obtains the greatest number of points, adding together the number of points obtained by their two drivers in each race of the championship. It should be noted that the sum of points can cause similar distortions to those pointed out by Soares de Mello [4] for the drivers championship. These distortions can be aggravated by the fact that, currently, a small team (STR-Renault) is, in practice, a branch of one of the larger teams (Red Bull Racing-Renault). In order to mitigate these distortions an adaptation of the Condorcet method for the team championship will be shown.

It should be observed that, as each team (alternative) has two drivers (sub-alternatives), this championship is characterized by the fact that each decision maker (race) can only be weakly rational.

4 Analysis of the 2014 Championship

The aim of this study was to obtain a ranking for the constructors championship independent of irrelevant alternatives. As each team has two drivers in each race, it is characterised as an evaluation of alternatives with sub-alternatives and is therefore weakly rational. Thus, the analysis by the Condorcet method proposed in this article has small differences in relation to the original method.

To construct the Condorcet matrix, the alternatives are compared two by two to establish the preferences of the decision maker. In order to evaluate the preference between the teams which participate in the Formula 1 constructors championship, it is necessary to compare the sub-alternatives of each of them, as each team participates in the championship with two drivers. Table 2, presents the comparison between the Sauber-Ferrari and Williams-Renault teams in the 2013 Formula 1 World Championship. In this table, the first column presents each of the championship races. Columns 2 to 5 present the positions of each of the drivers of these teams at the end of each of the races. For the purposes of the construction of the matrix, when comparing two teams, it was considered that in the case of any driver abandoning a race, the driver who completed most laps would have a better classification than a driver who completed fewer laps. In addition, a driver who abandoned a race would outrank another who did not classify for the starting grid, and this driver in turn was better than one who did not participate in the practice days. It is worth pointing out that that FIA (Fédération Internationale de l'Automobile) already uses the same criteria to present the results of each race.

Basically, the best placed driver in the first team is compared with the best placed driver in the second team, and the worst placed driver in the first team is compared with the worst placed driver in the second team. It is considered to be a victory for one team over another when the first is preferable over the other according to Definition 1. Similarly, the two teams are considered tied when one is indifferent to the other according to Definition 2.

Using a practical example, Table 2 shows the results of the 2013 Malaysia Grand Prix. In

this race Nico Hülkenberg of Sauber-Ferrari finished in eighth place, while the best Williams-Renault driver (Valtteri Bottas) finished in 11th place. Esteban Gutierrez of Sauber-Ferrari also finished ahead of Pastor Maldonado, the Williams-Renault driver. In this case, the first team is preferred in relation to the second. It is important to remember that there is no predetermined first or second driver; this denomination is made according to the position in which the driver finishes in each race separately.

In the 2013 Canada Grand Prix, the inverse occurred. This time the best placed driver from the Sauber-Ferrari team was Esteban Gutierrez, who finished in 20th place, while Valtteri Bottas of the Williams-Renault team finished in 14th place. Both Valtteri Bottas and Pastor Maldonado finished in positions in front of the drivers from the Sauber-Ferrari team, therefore, the Williams-Renault team is preferred.

A third situation occurred in the race in the 2013 British Grand Prix. Nico Hülkenberg, the driver from the Sauber-Ferrari team, finished in front of the drivers from the Williams-Renault team, however, Esteban Gutierrez, also from the Sauber-Ferrari team, finished behind the drivers from the Williams-Renault team. In this case, no team is preferred, characterizing a tie between the alternatives. It is worthwhile pointing out that there will always be a tie between the alternatives when one driver from a team wins the race and the other driver does not leave the grid, in relation to any other team in which the two drivers complete at least one lap of the race: a team will be preferred in one sub-alternative but not preferred in the other.

Team	Sauber-Ferrari		Williams-Renault		Winning Alternative		
Driver	Nico	Esteban	Pastor	Valtteri	Sauber-	Williams-	Tie
Race	Hülkenberg	Gutierrez	Maldonado	Bottas	Ferrari	Renault	
Australia	22	13	21	14			1
Malaysia	8	12	19	11	1		
China	10	22	14	13			1
Bahrain	12	18	11	14		1	
Spain	15	11	14	16	1		
Monaco	11	13	20	12	1		
Canada	21	20	16	14		1	
England	10	14	11	12			1
Germany	10	14	15	16	1		
Hungary	11	21	10	20		1	
Belgium	13	14	17	15	1		
Italy	5	13	14	15	1		
Singapore	9	12	11	13	1		
Korea	4	11	13	12	1		
Japan	6	7	16	17	1		
India	19	15	12	16		1	
Abu Dhabi	14	13	11	15			1
USA	6	13	17	8	1		
Brazil	8	12	16	21	1		
Total					11	4	4

Table 2: Comparison between the Sauber-Ferrari and Williams-Renault teams.

After comparing the sub-alternatives in all the races, the victories of each of the alternatives and the ties are added up. The alternative which has the greater total is the preferred alternative. This same procedure was applied to the 2014 Formula 1 Constructors World Championship, the final adjacency matrix is shown in Table 3. The final rankings of all the races

were obtained from the site http://www.f1.com.

	Mercedes	McLaren	Ferrari	Williams	Force India	Toro Rosso	Sauber	Marussia	Lotus	Caterham	Red Bull
Mercedes		1	1	1	1	1	1	1	1	1	1
McLaren	0		0	0	1	1	1	1	1	1	0
Ferrari	0	1		0	1	1	1	1	1	1	0
Williams	0	1	1		1	1	1	1	1	1	0
Force India	0	0	0	0		1	1	1	1	1	0
Toro Rosso	0	0	0	0	0		1	1	1	1	0
Sauber	0	0	0	0	0	0		1	0	1	0
Marussia	0	0	0	0	0	0	0		0	1	0
Lotus	0	0	0	0	0	0	1	1		1	0
Caterham	0	0	0	0	0	0	0	0	0		0
Red Bull	0	1	1	1	1	1	1	1	1	1	

Table 3: Adjacency Matrix of the Condorcet graph for the 2014 Constructors Championship.

In order to extract a ranking from the matrix one begins by performing a descending distillation [1]. In order to do this, one observes if there is any team which outranks all the others, in other words, if there is any row on which the only zero is on the main diagonal. This team is removed and the procedure is repeated.

The ranking obtained by the modified Condorcet method compared with the official ranking is shown in Table 4.

Team	Official Ranking	Condorcet Ranking	Difference
Mercedes	1	1	0
Red Bull	2	2	0
Williams	3	3	0
Ferrari	4	4	0
McLaren	5	5	0
Force India	6	6	0
Toro Rosso	7	7	0
Lotus	8	8	0
Sauber	10	9	−1
Marussia	9	10	1
Caterham	11	11	0

Table 4 : Final results for 2014 (official and Condorcet)

5 Conclusions

In this article it has been shown that, due to the formal analogy between the Formula 1 World Championship and a multi-decision maker selection process, there is no regulation which can be considered fair.

The comparison of the official result and that obtained by the Condorcet method shows very similar results, with variations only between the Sauber and Marussia teams.

Using the results of the 2014 championship, it was possible to rank all the teams through the Condorcet method, as there was no cycle of intransitivity among the alternatives. When these intransitive cycles arise, the solution that it supplies is less sensitive to the irrelevant alternatives than the official method.

Acknowledgements

We would like to thank CNPq and FAPERj for their financial support.

References

[1] L. M. C. Dias, L. M. A. T. Almeida, and J. C. N. Climaco. *Apoio multicritério à decisão*. Universidade de Coimbra, Coimbra, 1996.

[2] J. Haigh. Uses and limitations of mathematics in sport. *IMA Journal of Management Mathematics*, 20(2):97–108, 2009.

[3] J. C. Pomerol and S. Barba-Romero. *Multicriterion decision in management: Principles and practice*. Kluwer Academic, Boston, 2000.

[4] J.C.C.B. Soares de Mello, L. F. A. M. Gomes, E. G. Gomes, and M. H. C. Soares de Mello. Use of ordinal multi-criteria methods in the analysis of the formula 1 world championship. *Cadernos Ebape.BR*, 3(2):1–8, 2005.

The Relative Competitive Balance of Male and Female Teams in World Championship Competition and the Relative Predictability of Official International Sports Rating Systems

Raymond Stefani

California State University, Long Beach, USA
Raymond.stefani@csulb.edu

Abstract

There are 13 international team sports having official rating systems which can then be compared match by match with subsequent world championship/world cup results to evaluate predictability and thus competitive balance. Greater predictability implies less competitive balance. In recent consecutive competitions, men changed predictability by +15% in rugby union, +8% in football, +7% in basketball, +6% in rugby league, +3% and +6% in ice hockey, -3% in ODI cricket, -3% in curling and -3% in volleyball. Women changed predictability by +3% in volleyball, -1% in curling, 0% and -3% in netball, -9% and -0% in ice hockey and -13% in football. The predictability of women minus men was +8% in volleyball, +3% in basketball, 0% in handball, 0% in rugby sevens, -2% in ice hockey, -4% in water polo, -4% in field hockey and -7% in curling. Top male athletes appear drawn to (are more predictable for) the major revenue-producing sports while the top female athletes appear drawn to indoor sports. The top three rating systems in terms of predictability were the three self-adjustive systems which adjust ratings by comparing the actual match result with the expected match outcome. The predictability of David Kendix system was 89% in netball, 80% in ODI cricket and 76% in T20 cricket, an average of 84%. The IRB system was 84% accurate in rugby union while the FIFA womens Elo system for football was 82% accurate.

1 Introduction

The term competitive balance invokes both the abstract and the concrete. In the abstract, competitive balance suggests that teams and individuals are reasonably equal in skill (each competitor has a reasonable chance of winning). When applied to a professional sports league, concrete economic considerations come into play. If the less skilled teams can win a few matches and remain competitive through most of the matches lost, then supporters will continue to pay for tickets and maintain team viability financially. If the top teams win most of their matches but are challenged throughout match duration, then supporters of both teams will remain at the venue, enhancing concession sales. TV viewers will continue to watch, so that advertisers and sponsors will be rewarded for their support and the TV networks will receive fees to compensate for the cost of transmission.

A normalized metric is needed to evaluate competitive balance and to evaluate various strategies for increasing balance, making competition less predictable. For a given season of a professional league of sports teams, the standard deviation of actual wins, SD_W can be divided by the ideal standard deviation SD_I that would result if each team had a probability of 0.5 to win each match. From the binomial distribution, SD_I would be $0.5n^{1/2}$ if n games are played by each team. The ratio is one if teams are of equal skill and more than one when the wins are

more widely spread and the matches are more predictable, indicating less-than-ideal balance [1, 2, 3, 4]. Various methods can be applied to try to make teams more balanced, hopefully lowering that ratio, such as salary caps and a reverse drafting of new players [1, 2].

Over a span of seasons, the literature does not seem to have a standard of comparison. The SD of the wins achieved by a team or teams over several seasons is suggested [2]; but there is no obvious value for that statistic, implying balance. The Hirfindahl-Hirchman Index, HHI is also used, the sum of the squares of the number of wins by each champion divided by the number of years [2]. If each champion wins once, the ratio is 1. If some teams win more often, the ratio is larger than one, implying lack of balance. That Index is normalized but does not take into account what the other teams have done. We suggest that the correlation coefficient of two vectors of wins lagged by one season would be a scaled statistic that ought to be used. Values near zero imply balance and value near one imply predictability and lack of balance.

There is very little in the literature regarding competitive balance in international sports, given the lack of an obvious standard of equality. The HHI could be used, given the obvious flaw just noted. We suggest a simple standard to gauge competitive balance in international team sports. When a sport has a recognized rating system, the fraction of matches won by the higher ranked team in the next world championship provides a logically-scaled statistic A value near 50% implies balance while a value nearer to 100% implies predictability. A typical world championship has a group phase wherein a collection of teams plays a round-robin where ties may happen followed by a knockout phase where ties are not allowed. To use a common scale for all sports, only matches resulting in a win are tabulated. Four studies will be covered in this paper.

- Changes in the predictability of men and women are calculated for sports where more than one competition can be evaluated. See Section 3.
- The relative predictability of men and women are calculated for sports where the same rating system if used for men and women. See Section 3.
- The relative predictability of the group phase is compared to that of the knockout phase for each sport. See Section 4.
- The relative predictability of the rating systems are compared. See Section 5.

2 Sports, Data Base and Rating Systems

[6] contains a comprehensive survey of 159 recognized international sports, 99 of which were governed by federations publishing official rating systems (we now have identified 105 such systems). The interest here is for true team sports requiring cooperative behavior, such as football, basketball and rugby. There are 18 officially recognized team sports having official rating systems, five of which lack a world championship. The remaining 13 sports are the focus of this study. Nine sports involve both men and women: basketball, curling, field hockey, football (which has two different rating systems), handball, ice hockey, rugby sevens, volleyball and water polo. Three sports are for men only: cricket (ODI and T20), rugby league and rugby union. One sport is for women only: netball.

A rating is a numerical value given to a team. A ranking is the ordinal position based on the rating. There are three types of rating systems, subjective (not involved here), point accumulation (simply called Accumulative) and self adjustive (simply called Adjustive). An Accumulative system creates a non-decreasing running sum (an accumulation) of points over some window, using non decreasing f_I .

$$\text{Rating} = f_I(\text{result points, importance, ageing})$$

The result points depend mostly on order of finish in a competition and not on opponent or score.

In contrast, the ratings for an Adjustive system increase, decrease or remain the same depending on the difference between the actual result of a match and an expected result, based on rating difference and possibly home advantage. An Adjustive system depends on more information than does an Accumulative system. The value of K below depends on the importance of the match.

$$\text{New Rating} = \text{Old Rating} + K[\text{Actual} - \text{Expected}]$$

The data base includes archived and contemporarily-gathered data for 22 competitions involving 1261 games for the 12 mens sports and for 17 competitions involving 822 games for the 10 womens sports. That is twice as much data for men and three times as much for women than [7]. Nine rating systems are Accumulative and four are Adjustive. Thus, there are 13 sports and 13 rating systems, but not on a one-to-one basis. Football has two different Adjustive systems while one adjustive system is used for both netball and cricket.

3 Changes in the Predictability of Men and Women

Table 1 (for men) and Table 2 (for women) show changes in predictability where more than one competition was evaluated.

Sport	Federation	1st Year	2nd Year	1st Year %	2nd Year %	% Change
Rugby Union	IRB	2007	2011	77	92	+ 15
Football	FIFA	2010	2014	72	80	+ 8
Basketball	FIBA	2010	2014	69	76	+ 7
Rugby League	RLIF	2008	2013	72	78	+ 6
Ice Hockey	IIHF	2010	2011	68	71	+ 3
		2011	2014	71	77	+ 6
ODI Cricket	ICC	2007	2011	81	78	− 3
Curling	WCF	2011	2012	70	67	− 3
Volleyball	FIVB	2010	2014	70	67	− 3

Table 1: Percent Changes in Predictability for Men (Ties are not Counted)

Sport	Federation	1st Year	2nd Year	1st Year %	2nd Year %	% Change
Volleyball	FIVB	2010	2014	75	78	+ 3
Curling	WCF	2012	2014	62	61	− 1
Netball	IFNA	2010	2011	90	90	0
		2011	2014	990	87	− 3
Ice Hockey	IIHF	2011	2012	76	67	− 9
		2012	2013	67	67	0
Football	FIFA	2007	2011	89	76	− 13

Table 2: Percent Changes in Predictability for Women (Ties are not Counted)

Generally, men become more predictable, rising 15% in rugby union to a remarkable 92%. Men increased in rugby union, football, basketball, rugby league and ice hockey, all highly paying sports with significant TV exposure. Women generally decreased in predictability. They

increased only in volleyball. They had a modest decrease in curling (a sport with low predictability) and a modest decrease in netball (a sport that had 90% to 87% predictability). Women had a significant decrease in football by 13%, an encouraging figure, indicating much more competitive balance.

Sport	Federation	Years (M)	Years (W)	P(M) (%)	P(W) %	P(W)−P(M)
Volleyball	FIVB	2010, 14	2010, 14	69	77	+ 8
Basketball	FIBA	2010, 14	2014	72	75	+ 3
Handball	IHF	2015	2013	73	73	0
Rugby Sevens	IRB	2013	2013	80	80	0
Ice Hockey	IIHF	2010, 11, 14	2011, 12, 13	72	70	− 2
Water Polo	FINA	2014	2014	83	79	− 4
Field Hockey	FIH	2014	2014	79	75	− 4
Curling	WCF	2011, 12	2012, 14	69	62	− 7

Table 3: Percent Predictability of Women Minus Men (Ties Not Included, P = Percent Correct)

Table 3 compares men and women in the eight sports contested by both genders and having a common rating system. For Table 3, women were more predictable than men for volleyball and basketball and were even in handball. If we include netball as a predictable sport for women (although not contested by men) women of the top nations appear to be most successful in indoor sports requiring a high degree of team interaction with social crowd interaction. Those characteristics are apparently attractive to top female athletes. Women were equally predictable with men in rugby sevens and somewhat less predictable in the other four sports.

4 Relative Predictability of the Group and Knockout Phases

Football appears twice due to two rating different systems. We would expect the knockout phase to be less predictable than the group phase in that the remaining teams ought to be more balanced in skill, having eliminated the underperforming teams. While that is generally true, the notable counter examples provide insight into the strategy employed by top nations in some sports.

As expected, the group phase in handball (by 35%) and in ice hockey (by 21%) are more predictable. Five sports were 10-16% more predictable in the group phase while three sports were between 8-9% more predictable. The situations for water polo, football and rugby league are remarkable. Water polo teams were only 4% less predictable in the knockout phase. For football, men were 11% more predictable in the knockout phase while women were as predictable. Rugby league teams were 11% more predictable in the knockout phase. It appears that the better teams know they are good enough to advance; hence, they play only well enough to do so in the group phase. They may eschew winning for tying or even losing tactically.

Sport	Games (Comp)	Overall	Group	Knockout	G−K
Handball	165(2)	73%	84%	49%	35%
Ice Hockey	239(6)	72%	76%	55%	21%
Netball	124(3)	89%	94%	78%	16%
Rugby Sevens	98(2)	80%	85%	72%	13%
Field Hockey	76(2)	77%	80%	69%	11%
Cricket	130(3)	79%	80%	69%	11%
Volleyball	386(4)	73%	74%	64%	10%
Rugby Union	96(2)	84%	84%	75%	9%
Basketball	196(3)	73%	75%	67%	8%
Curling	287(4)	65%	66%	58%	8%
Water Polo	48(2)	81%	83%	79%	4%
Football (W)	64(2)	82%	82%	82%	0%
Football (M)	128(2)	76%	73%	84%	−11%
Rugby League	46(2)	76%	72%	85%	−13%

Table 4: Percent Predictability of the Group Phase versus the Knockout Phase

5 Comparison of the Rating Systems

Sport	Games (WCs)	System (Years)	Predictability
Netball	124(3)	Adj	89%
Cricket (ODI)	96(2)	Adj	80%
Cricket (T20)	34(1)	Adj	76%
Kendix (Total)	254(6)		84%
Rugby Union	96(2)	Adj	84%
Football (W)	64(2)	Adj	82%
Water Polo	48(2)	Acc(4)	81%
Rugby Sevens	98(2)	Acc(1) [Current yr]	80%
Field Hockey	76(2)	Acc(4)	77%
Football (M)	128(2)	Adj Points Avg (4) [Includes opponent]	76%
Rugby League	46(2)	Acc(5) [Includes opponent]	76%
Handball	165(2)	Acc(4)	73%
Basketball	196(3)	Acc(8)	73%
Volleyball	386(4)	Acc(4)	73%
Ice Hockey	239(6)	Acc(4)	72%
Curling	287(4)	Acc(6)	65%

Table 5: Relative Accuracy of the 13 Rating Systems

In Table 5 for the 13 rating systems, as for rating systems in sports in general, predictability depends on the information content of each system and on the competitive balance of the sport. The table is divided into three parts: the bottom five systems, the next higher five and top three. The bottom five systems are accumulative, covering 4-8 years and are for sports contested indoors. The curling system is the least predictive. The other four systems are within one percent.

Four of the five sports in the middle panel are contested outdoors. The mens football system

and the rugby league system employ points based on the ranking of the opponent. The football points average system would create the same rank as for point accumulation if teams play about the same number of games [5]. The rugby sevens system employs points gained only for the season leading up to the world cup. Those three systems use more and better information than the bottom five and are more accurate. The field hockey and water polo accumulative systems are likely more predictive then the bottom five due to less competitive balance.

The bottom 10 systems are accumulative or effectively accumulative. The top three systems are adjustive, following Equation (2). These three are summarized in Table 6, David Kendix developed the system used by both netball and cricket. The top three systems are equal best in this survey. All three include importance. Two systems use home advantage and score. Two systems find the expected probability to win based on a linear measure of rating difference while one uses the Elo exponential equation. The netball and cricket federations chose not to include home advantage and score.

Sport (Feder.)	Actual (1-0 scale)	Expected (1-0 scale)	K	Also
Cricket (ICC) Netball (IFNA) (David Kendix)	(1, .5, 0) for (W, T, L)	.5 + .5 (d/50) (limited .1 to.9)	50/n Importance	
Rugby (IRB)	(1, .5, 0) for (W, T, L)	.5 + .5 (d/10) (limited .1 to .9)	Importance Score Diff	Home Adv.
Football (FIFA, W) Elo System	1-0 based on goals scored ad hoc table	$1/(1+10^{-d/[\sqrt{2}\sigma_d]}))$	Importance	Home Adv.

Table 6: Details of the Three Best Systems, all Adjustive

6 Conclusions

In recent consecutive competitions, men increased predictability (especially in rugby union) while women became less predicable (especially in football) exhibiting more competitive balance. Top male athletes appear drawn to (are more predictable for) the major revenue-producing sports while the top female athletes appear drawn to indoor sports. In world group phases, the top-rated nations appear to be playing strategically in water polo, football and rugby league, accepting ties and even losses to advance. The top three rating systems in terms of predictability were the three self-adjustive systems which adjust ratings by comparing the actual match result with the expected match outcome. These top three were David Kendix system (used in netball, ODI cricket and T20 cricket), the IRB system in rugby union and the FIFA womens Elo system used in football.

References

[1] R. Booth. Comparing competitive balance in australian sports leagues: Does a salary cap and player draft measure up? *Sports Management Review*, pages 119–143, 2005.

[2] M. Leeds and P. Allmen. *The Economics of Sport*. Pearson, Boston, 2005.

[3] L. Lenton. Measurement of competitive balance in conference and divisional tournament design. *Journal of Sports Economics*, 16(1):3–25, 2015.

[4] P.D. Owen and N. King. Competitive balance in sports leagues: The effect of variation in season length. NCER Workingpaper Series, 92, http://www.ncer.edu.au/papers/documents/WP92v2.pdf, July 2013.

[5] R. Stefani. How well do fifas ratings predict world cup success? *Significance*, 28, May 2011. http://www.statslife.org.uk/sports/.

[6] R. Stefani. The methodology of officially recognized international sports rating systems. *Journal of Quantitative Analysis in Sports*, 7(4):Article 10, 2011.

[7] R. Stefani. Predictive success of official sports rating systems in international competition. In *Proceedings of the 11th Australasian Conf. on Math. and Computers in Sport*, 2012.

Predicting the Final League Tables of Domestic Football Leagues

Jan Van Haaren and Jesse Davis

KU Leuven, Department of Computer Science
Celestijnenlaan 200A, 3001 Leuven, Belgium
{jan.vanhaaren,jesse.davis}@cs.kuleuven.be

Abstract

In this paper, we investigate how accurately the final league tables of domestic football leagues can be predicted, both before the start of the season and during the course of the season. To this end, we perform an extensive empirical evaluation that compares two flavors of the well-established Elo-ratings and the recently introduced pi-ratings. We validate the different approaches using a large volume of historical match results from several European football leagues. We assess how well each ranking system performs on this task, investigate what is the most natural metric to measure the quality of a predicted final league table, and the minimum number of matches that needs to be played in order to yield useful predictions. We find that the proportion of correctly predicted relative positions is a natural metric to assess the quality of the predicted final league tables and that the traditional Elo-rating system performs well in most settings we considered.

1 Introduction

The prediction of football matches has received significant attention over the past few decades. Predicting the outcomes of individual football matches is a very challenging task due to the low number of goals scored [1]. Despite this fact, the ever-growing interest in football betting makes this an active area of research. Initially, the focus was on purely statistical models (e.g., [3, 6, 7, 8, 9]) but in recent years it has somewhat shifted to predictive ranking systems (e.g., [2, 5]). In contrast to the large body of work in the area of predicting individual match outcomes, the related task of predicting final league tables has remained almost unexplored to date. However, the latter task is of much more interest to managers and directors who want to develop a successful long-term vision for their clubs.

In this paper, we investigate how accurately the final league tables of domestic football leagues can be predicted, both before the start of the season and during the course of the season. We focus our study on different flavors of two popular predictive ranking systems, namely the recently introduced pi-ratings [2] and the well-established Elo-ratings [5]. We validate the different systems on a large volume of historical match results covering over twenty seasons of the most important European football leagues. Besides assessing how well each predictive system performs on this task, we also investigate what is the most natural metric to measure the quality of a predicted final league table, and the minimum number of matches that needs to be played in order to yield useful predictions.

An experimental evaluation on four seasons in seven different leagues shows that the proportion of correctly predicted relative positions and the mean squared error in terms of positions are natural metrics to assess the quality of the predicted final league tables. Furthermore, the evaluation shows that the traditional Elo-rating system performs well in most settings. The pi-rating model performs reasonably well in many settings but excels when only match results from previous seasons are available.

2 Predictive Models

We consider multiple variants of both the widely used Elo-ratings [4, 5] and the more recent pi-ratings [2]. While these ranking systems are very similar in spirit, they differ in several important aspects. We now discuss the specifics of both ranking systems.

2.1 Elo Ratings

The Elo system uses a single number or *rating* to represent the strength of a team at a particular point in time. This rating increases or decreases based on the outcomes of matches. After each match, rating points are transferred from the losing team to the winning team. The number of transferred rating points depends on the difference between the ratings of the teams prior to the match. More rating points are transferred when a low-ranked team wins against a high-ranked team than when a high-ranked team beats a low-ranked team. Hence, the Elo system is self-correcting in the sense that underrated teams will gain rating points until their rating reflects their true strength, while overrated teams will lose rating points.

More formally, the rating of a team is computed as follows after each match:

$$R_{new} = R_{cur} + I \times G \times (R_{act} - R_{exp}) \tag{1}$$

In this equation, R_{cur} and R_{new} denote the rating of a team before and after a match, respectively. The parameter I is a positive number that denotes the importance of the match, and higher numbers correspond to more important matches. The parameter G is a positive number that denotes the importance of the goal difference and a bigger goal difference corresponds to an higher number. This parameter is typically defined as a function of the goal difference:

$$G = \begin{cases} 1 & \text{if GD } \leq 1 \\ 1.5 & \text{if GD } = 2 \\ \frac{(GD+11)}{8} & \text{if GD } \geq 3 \end{cases} \tag{2}$$

The parameters R_{act} and R_{exp} represent the actual and expected outcome of the match, respectively. The parameter R_{act} takes one of three possible values: 1 for a win, 0.5 for a draw, and 0 for a loss. The parameter R_{exp} takes a value between 0 and 1 and is computed as follows:

$$R_{exp} = \frac{1}{1 + 10^{\left(\frac{R_{away} - R_{home}}{400}\right)}} \tag{3}$$

In this equation, R_{home} and R_{away} denote the ratings of the home and away team.

2.2 Probabilistic Intelligence Ratings

The Probabilistic Intelligence system uses two numbers or *pi-ratings* to represent the strength of a team at a particular point in time. These numbers represent the expected goal difference against an average opponent in a league both at home and on the road. Similar to the Elo system, the ratings increase or decrease based on the outcomes of matches. After each match, the ratings of the teams involved are adjusted based on the goal difference. The key idea is that the system updates the ratings to reduce the discrepancy between the predicted and observed goal differences. The convergence speed depends on two learning rates, which denote how important recent results are for assessing the current ability of a team.

More formally, the ratings of a team are computed as follows after each match:

$$R_{\alpha,H} = C_{\alpha,H} + \psi_H(e) \times \lambda \tag{4}$$

$$R_{\alpha,A} = C_{\alpha,A} + (R_{\alpha,H} - C_{\alpha,H}) \times \gamma \tag{5}$$

$$R_{\beta,A} = C_{\beta,A} + \psi_A(e) \times \lambda \tag{6}$$

$$R_{\beta,H} = C_{\beta,H} + (R_{\beta,A} - C_{\beta,A}) \times \gamma \tag{7}$$

In these equations, $C_{\alpha,H}$ and $C_{\alpha,A}$ are the current home and away ratings for team α, $C_{\beta,H}$ and $C_{\beta,A}$ are the current home and away ratings for team β. $R_{\alpha,H}$, $R_{\alpha,A}$, $R_{\beta,H}$, and $R_{\beta,A}$ are the respective revised ratings. Furthermore, λ and γ are learning rates, and $\psi(e)$ is a function of the difference between the observed and predicted goal difference, which is computed as follows:

$$\psi(e) = 3 \times \log_{10}(1 + e) \tag{8}$$

Due to space limitations, we refer to [2] for the technical details and a concrete example.

3 Experimental Evaluation

This section presents the experimental evaluation. We first present the dataset, the methodology and the predictive models before discussing the experimental results.

3.1 Dataset

In our experimental evaluation, we use a large dataset of historical football match results.[1] The dataset contains match results for eleven European football leagues including the English, German, Spanish, Italian and French top divisions. The dataset dates back to the 1993/1994 season for most leagues and contains nearly 150,000 match results. Due to space limitations, we restrict the evaluation to the four seasons from 2010 through 2014 and the following seven leagues: Belgium, England, France, Germany, Italy, The Netherlands, and Spain.

3.2 Methodology

To evaluate each of the models, we split the match results for each league in two sets: a training set and a test set. We use the training set to learn the parameters of the model and the test set to assess the performance of the model. We predict the outcomes of the matches in the test set in an iterative way. We use the model learned on the training set to predict the outcomes of the matches on the first match day. We then update the parameters of the model using the expected outcomes and predict the outcomes of the matches on the second match day. We repeat this procedure until all matches have been predicted.

To account for the fact that performances tend to vary throughout a season, we employ a probabilistic approach to predict match outcomes. For the Elo-rating models, we predict a match outcome by first computing the probability distribution over the possible outcomes and then sampling an outcome from this distribution. For the pi-rating model, we sample a match

[1] http://www.football-data.co.uk/data.php

outcome from a normal distribution whose mean is the expected outcome. We repeat each experiment 10,000 times and report the average results across all runs.

In each league, we predict the outcomes for the 2013/2014 season and use the match results from the earlier seasons to learn the parameters of the models.

3.3 Models

We compare the following four models in our experimental evaluation:

- **Random model (RM):** This model predicts a random outcome for each match.
- **Elo-rating model (EM):** This is a traditional implementation of the Elo-rating system (see Section 2.1), which represents each team by a single rating. We account for the home-field advantage by increasing the rating of the home team with an empirically determined number of rating points.
- **Split Elo-rating model (SEM):** This is a modified implementation of the Elo-rating system, which represents each team by two ratings. The first rating represents the strength at home, while the second rating represents the strength on the road.
- **Pi-rating model (PM):** This is a traditional implementation of the pi-rating system (see Section 2.2), which represents each team by two ratings.

We have not rigorously tuned the parameters. For each model, we used reasonable parameter values that work well on our large volume of training matches.

3.4 Results

In our evaluation, we investigate what is the most natural metric to assess a predicted league table (Q1), how well each model performs on predicting the final league table (Q2), and what proportion of the matches needs to be played in order to yield useful predictions (Q3).

Q1: What is the most natural metric to assess a predicted league table?

We investigated the following four metrics to assess the models:

- **Proportion of correct absolute positions:** This is the ratio between the number of correctly predicted absolute positions and the total number of positions in the league table.
- **Proportion of correct relative positions:** This is the ratio between the number of correctly predicted relative positions and the total number of relative positions in the league table. This metric compares the relative positions between any two teams.
- **Mean squared error in terms of positions:** This is the mean squared error between the actual positions and the predicted positions for each team.
- **Mean squared error in terms of points:** This is the mean squared error between the actual number of points and the predicted number of points for each team.

We found that the proportion of correctly predicted relative positions and the mean squared error in terms of positions are the two most natural metrics to assess a predicted league table. Predicting the absolute position for a team is hard because it depends on the positions of all other teams in the league. Furthermore, both the Elo-rating models and the Pi-rating model have difficulties to predict draws, which often leads to an higher variance on the predicted number of points than on the actual number of points for each team.

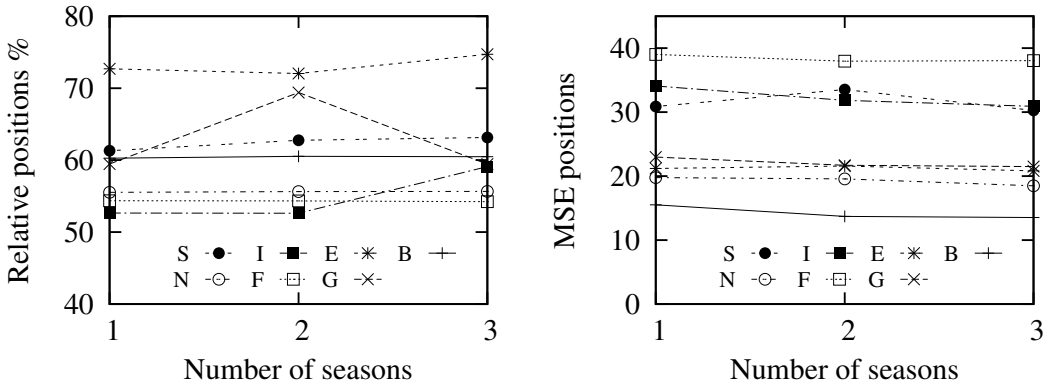


Figure 1: Learning curves for the proportion of correctly predicted relative positions and the mean squared error in terms of positions, where we vary the size of the training set.

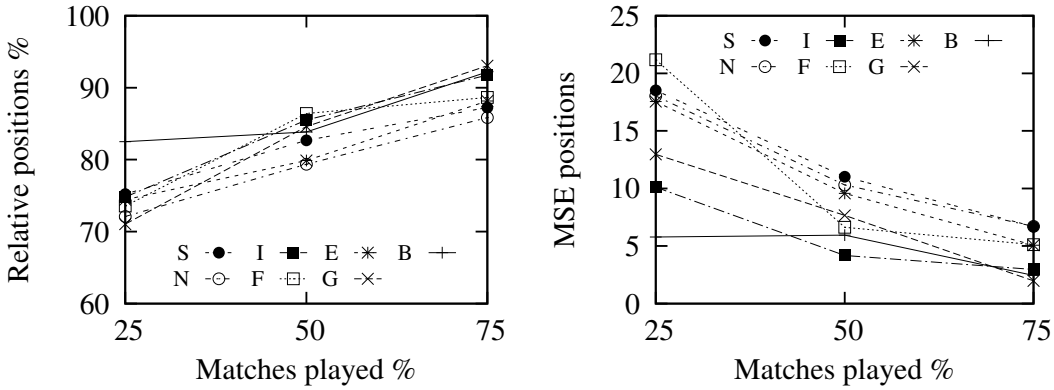


Figure 2: Learning curves for the proportion of correctly predicted relative positions and the mean squared error in terms of positions, where we vary the amount of played matches.

Q2: How well does each model perform on predicting the final league table?

Figure 1 shows learning curves for the proportion of correctly predicted relative positions and the mean squared error in terms of positions, where we vary the size of the training set from one season to three seasons. For the former metric, the PM performs best in 18 of the 21 settings and the SEM in the remaining three settings. For the latter metric, the EM performs best in 17 of the 21 settings, the PM in three settings, and the SEM in one setting.

Q3: What proportion of the matches needs to be played to yield useful predictions?

Figure 2 shows learning curves for the proportion of correctly predicted relative positions and the mean squared error in terms of positions, where we vary the number of matches played in the final season. For example, 25% means that the model is trained on the results for the 2012/2013 season as well as the results for the first quarter of the 2013/2014 season. Unsurprisingly, the

predicted final league tables become more accurate as more matches have been played. For the former metric, the EM performs best in 15 of the 21 settings and both the SEM and PM in three settings. For the latter metric, the EM performs best in 18 of the 21 settings and the SEM in the remaining three settings.

4 Conclusions

This paper investigates whether popular predictive rankings are capable of accurately predicting the final league tables in football leagues. The experimental evaluation shows that the proportion of correctly predicted relative positions and the mean squared error in terms of positions are the most natural metrics to assess predicted league tables. Furthermore, the evaluation shows that the traditional Elo-rating system performs particularly well in most settings we considered. While the Pi-rating model performs reasonably well in many settings as well, it excels when only match results from previous seasons are available.

The suggested directions for future research include devising more sophisticated metrics to assess the predicted final league tables (e.g., assigning weights to positions such that positions qualifying for European football become more important) and comparing to more advanced predictive ranking systems (e.g., systems that also consider performance data).

Acknowledgments

Jan Van Haaren is supported by the Agency for Innovation by Science and Technology in Flanders (IWT). Jesse Davis is partially supported by the Research Fund KU Leuven (OT/11/051), EU FP7 Marie Curie Career Integration Grant (#294068) and FWO-Vlaanderen (G.0356.12).

References

[1] C. Anderson and D. Sally. *The numbers game: Why everything you know about football is wrong.* Penguin UK, 2013.

[2] A. C. Constantinou and N. E. Fenton. Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *Journal of Quantitative Analysis in Sports*, 9(1):37–50, 2013.

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

[4] A. E. Elo. *The rating of chessplayers, past and present.* Batsford London, 1978.

[5] L. M. Hvattum and H. Arntzen. Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, 26(3):460–470, 2010.

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

[7] A. J. Lee. Modeling scores in the Premier League: Is Manchester United really the best? *Chance*, 10(1):15–19, 1997.

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

[9] H. Rue and O. Salvesen. Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(3):399–418, 2000.

A Low-cost Spatiotemporal Data Collection System for Tennis

Silvia Vinyes Mora, George Barnett, Casper O. da Costa-Luis, Jose Garcia Maegli, Martin Ingram, James Neale and William J. Knottenbelt

Imperial College London, London, U.K.
silvia.vinyes-mora12@imperial.ac.uk, george.barnett10@imperial.ac.uk, casper.dcl@physics.org, jose.garcia-maegli14@imperial.ac.uk, martin.ingram14@imperial.ac.uk, james.neale14@imperial.ac.uk, wjk@imperial.ac.uk

Abstract

Technology has revolutionised tennis, especially the Hawk-Eye system which is now an integral part of professional tennis tournaments with the line-call challenge system. While such technologies are highly accurate, they are also equipment-intensive, costly and require substantial expertise to install on a court. Our aim is to investigate the feasibility of constructing a low-cost alternative from commodity hardware components and to investigate what compromise can be achieved between cost, accuracy and speed. The system that we implement is composed of four high frame rate, high definition grey scale cameras and a GPU-enabled desktop computer - a highly portable system that can be installed on court in minutes. The current system provides accurate spatio-temporal information about the players, the ball and their position relative to the court lines, and is able to replicate results in a 3D virtual environment.

1 Introduction

The integration of technology in sports has represented a new era in how matches are broadcast; the data that is collected from them and therefore the analysis that can be achieved. The state-of-the-art technology involved in officiating and collecting data in sports is the Hawk-Eye system [5]. Hawk-Eye is used in many sports and has become an integral part in professional tennis matches by the introduction of new rules regulating the line-call challenges. This system is extremely accurate but also highly sophisticated, comprising between 8 to 10 high speed cameras (more than 1000fps) and an extremely powerful computer. The fact that this technology is equipment-intensive, costly and requires expertise to install on a court restricts its availability to high profile venues of major tournaments. Alternatives exist but are similarly costly and equipment-intensive. Our aim is to develop a low-cost alternative from commodity hardware components able to replicate the events that occur on the court in a 3D virtual environment in real-time. Therefore, the specific hardware components have been selected based on our careful examination about the compromise that can be achieved between cost, accuracy and speed. The system that we propose is composed of four high frame rate, high definition greyscale cameras and a GPU-enabled desktop computer.

2 System Architecture

2.1 Hardware

The principal constraints when choosing the hardware equipment for this project are cost minimisation and portability. The system configuration is shown in Figure 1 and is described

in detail in this section. A clear image of the ball from at least two cameras is required at all times so that its 3D position can be inferred. To guarantee this, it was decided that the minimum number of cameras required is four, one at each corner of the court. The camera model chosen is the Basler Ace acA 1300-60gm, these are high resolution (1280 x 1024 pixels) and high frame rate (60 fps) monochrome cameras, therefore minimising motion blur. We have complemented them with low distortion Kowa LMVZ4411 lenses, which have dimensions matching the area to be captured (the standard tennis court dimensions). The cameras are connected to a four-port Gigabit Ethernet card through Gigabit Ethernet cables that have a dual function: deliver the camera capture at high speed to the computer and provide Power over Ethernet to the cameras, reducing cabling requirements. The computer used for processing is an HP Z620 Desktop Workstation with a fast 400GB PCI-E SSD and a GTX 980 nVidia graphics card. The system is unobtrusive and highly portable.

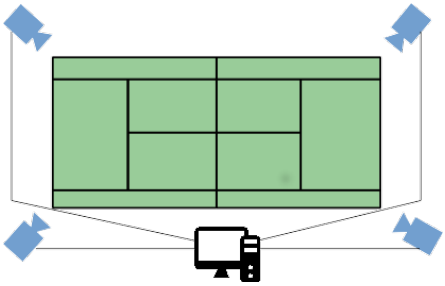


Figure 1: Configuration of the Detection System Hardware

2.2 Software

The software architecture is composed of three main layers as shown in Figure 2:

1. Frame capture: The Pylon Camera Software Suite was used to interface with the cameras and set properties such as exposure. An external 60 Hz signal generator was built to produce synchronized frame capture signals for the four cameras.
2. Frame processing: The computer vision part of the project is implemented in C++ using the OpenCV library [1]. It is organized as a multithreaded application in which player and ball are detected independently in each frame before the information is combined to obtain a 3D position.
3. Display: The data obtained from the frame processing module is fed into the data display module. This presents a 3D virtual environment in which the player and ball position are presented. It is a user-friendly display designed in UNITY, facilitating cross-platform portability.

3 Detection and tracking

3.1 Player Detection

The player detection process comprises two main tasks: background reconstruction and player detection itself. Figure 3 shows the intermediate images in detecting the player. For the

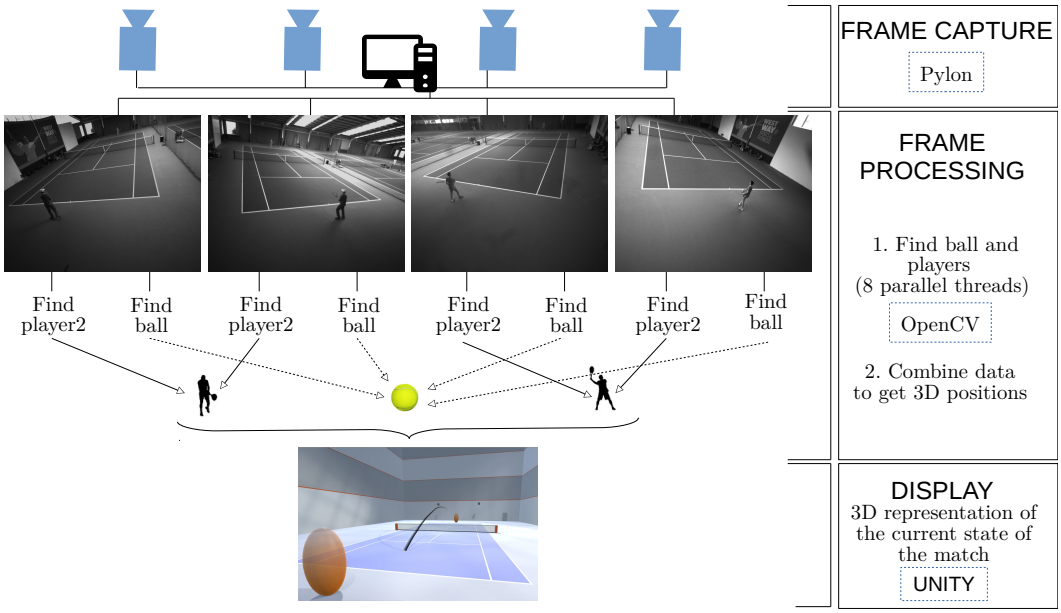
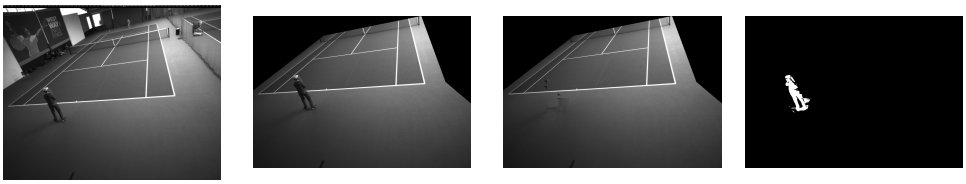


Figure 2: Software design

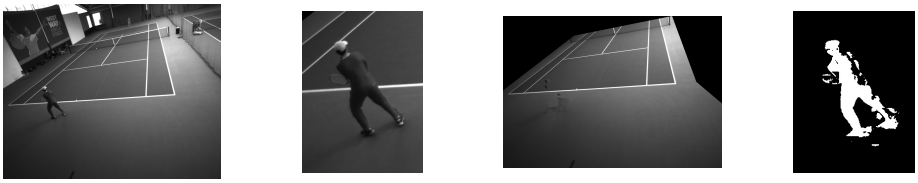
background reconstruction, the frame is divided in blocks. A block is incorporated to the background if the percentage of pixels that do not change over two frames is above a certain threshold. Blocks that do not satisfy this condition are discarded. After applying this process to a few frames, a complete background image is obtained for each of the cameras. Once this is complete, player detection can take place by selecting the foreground pixels within a region of interest (ROI), corresponding to the area in which the player can be located. For the frames in which the prior player position is unknown, the ROI for each player corresponds to their half court; detection is performed on these two areas separately. For the frames in which the prior positions are known, detection is performed only in areas within a given distance d of the player, d is defined according to the scale of the image and maximum speed of a player's movement. Once the collection of foreground pixels within the appropriate ROI is obtained, the center of mass is calculated using the moments of the shape and this is stored as the player position.

3.2 Ball Detection

The detection of the ball in a given frame n starts with the isolation of the pixels representing motion. These are stored in a matrix that will be referenced here as a motion matrix. The motion matrix is filled by computing the absolute pixel intensity difference between frames $n-1$ and frame n (backward difference). This is further refined by identifying bright points from the absolute difference of frames $n+1$ and $n-1$ (central difference) and setting these pixels to zero in the backward difference, thus eliminating the ball ghost from frame $n-1$. Next, contours of ball candidates are detected on the motion matrix using the OpenCV implementation [4]. To select the best candidate, a score is assigned to each potential ball according to its semblance



(a) Frame (b) Region of Interest (c) Background (d) Foreground



(e) Frame (f) Region of Interest (g) Background (h) Foreground

Figure 3: Player detection at the first iteration (a to d) and at a later stage (e to h)

to an ellipse [3]. Sequences of ball candidates which form possible trajectories are identified based on the method described in [2]. From these trajectory candidates, good trajectories are then selected using the following criteria: a quadratic fit of the points must have a negative first coefficient, the trajectory must comprise a large number of ball candidates (at least 8), and its ball candidates must fit an ellipse well on average. These good trajectories are then used to compute a best guess of the ball position in each frame. Ball detection is performed in parallel for the four cameras and the results are integrated by the next module to obtain the 3D position of the tennis ball.

3.3 3D Ball Position

Finding the 3D coordinates of a point, given its known position in images taken from two different views is a process known as triangulation. As shown in Figure 4, a given point y in a 2D image corresponds to a line R_y in 3D space. The 3D coordinates of that point are given by the intersection of the lines (R_y and $R_{y\prime}$ in Figure 4) corresponding to the two 2D points (y and $y\prime$) in the images taken from different cameras. However, noise in 2D coordinates (if we detect x and $x\prime$ instead of y and $y\prime$) generally means these lines (R_x and $R_{x\prime}$) do not intersect. Finding the 3D coordinate thus becomes a challenging problem for which different techniques can be applied. We have employed the mid-point method, in which the 3D point a is the value that minimizes the equation:

$$d(R_x, a)^2 + d(R_{x'}, a)^2 \text{ with } d(R, a) \text{ the euclidean distance between } R \text{ and } a$$

4 Results

As described, the ball and player are detected in each camera independently before the 3D coordinates are obtained via triangulation. Visual inspection of the 2D detection of the player and ball showed good accuracy across a large number of videos. Figure 5 shows the detailed analysis of the 2D results comparing our data to the ground truth for circa 5000 frames.

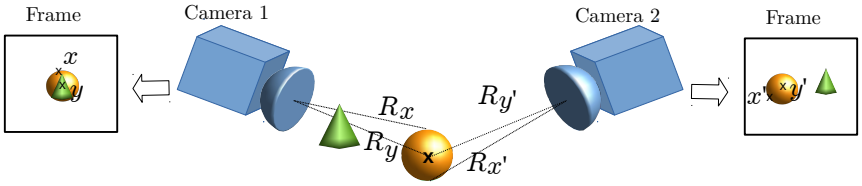
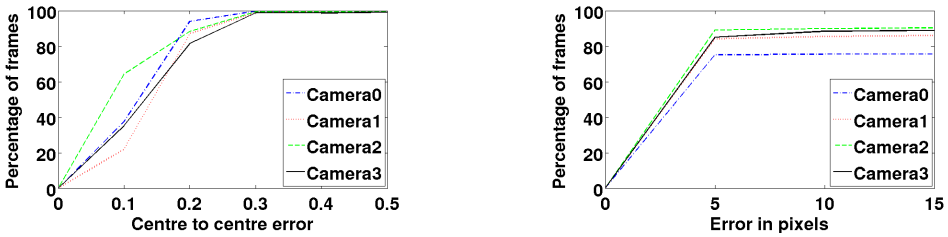


Figure 4: Triangulation



(a) Player detection results

(b) Ball detection results

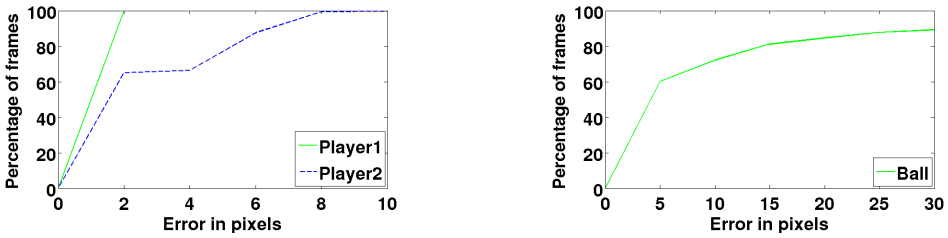
Figure 5: Analysis of the results for the player and ball detection in 2D

4.1 Player Detection

The player detection results in a 2D coordinate that represents the estimated center of mass of the player. The ground truth player data is stored as a rectangle surrounding the player. Note that each camera detects only the closest player, therefore there is a two to one camera to player ratio. The detected center of mass was within the ground truth bounding box for 100% of the analyzed frames. Figure 5a shows a fine grained analysis for each camera, depicting a curve representing the cumulative percentage of frames with a center-to-center error smaller than the x coordinate. This error represents the euclidean distance between the center of the ground truth bounding box and the detected center of mass difference. This distance is then represented as the percentage of the bounding box dimension to reflect that a given error distance of x pixels represents a larger error if the bounding box of the player is smaller. In general, errors are of the order of 15 pixels. This is a good result since the center of the bounding box does not correspond exactly to the center of mass of the player for many positions.

4.2 Ball Detection

The ball detection also produces a 2D coordinate, however the ball itself has a diameter of up to 10 pixels depending on the proximity to the camera. The objective is to determine the ball's center. Figure 5b shows the error in ball detection as the euclidean distance between the detected point and the ground truth for every frame in which a ball candidate was obtained. The ground truth data was obtained by manually annotating the same sequence of videos. Every curve in Figure 5b represents frames captured by a different camera, and it can be observed the error is less than 5 pixels for more than 75% of the frames.



(a) 3D Player detection results

(b) 3D Ball detection results

Figure 6: Analysis of the results for the player and ball detection in 3D

4.3 3D data

The results for the 2D can be accurately analyzed because ground truth data exists. However this is not the case for 3D data. As a first step, the 3D results have been visually inspected ensuring that the representation of the data in the 3D environment and the videos are consistent. Then we have also compared the ball 3D location obtained from the 2D ground-truth data to our 3D ball location estimates. It is important to note that the 3D data is produced for every single frame, while 2D data isn't. In fact, 3D data integrates data from all cameras and estimates the values for frames in which the ball was not detected. The same analysis was run for the player 3D location. Figure 6b shows the accuracy of our 3D results.

4.4 Running time analysis

The running time of the algorithm is 59.404 ± 0.595 fps, this is almost three times faster than when the code is run without using parallel threads (20.328 ± 0.083 fps) and does not include the I/O time since we run the experiments after buffering the frames.

5 Conclusion

We have designed a system able to collect 3D spatio-temporal data from a tennis match in real-time from commodity hardware. It is a low-cost alternative that has proven accurate results. In addition to this, its design is modular, with the capability of incorporating improvements both in terms of speed and accuracy. From the previously described results it would be interesting to investigate other triangulation methods and further increase the accuracy of the system.

References

[1] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[2] Ciarán Ó Conaire, Philip Kelly, Damien Connaghan, and Noel E O'Connor. Tennissense: A platform for extracting semantic information from multi-camera tennis data. In *Digital Signal Processing, 2009 16th International Conference on*, pages 1–6. IEEE, 2009.

[3] Andrew W Fitzgibbon, Robert B Fisher, et al. A buyer's guide to conic fitting. *DAI Research paper*, 1996.

[4] Satoshi Suzuki et al. Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing*, 30(1):32–46, 1985.

[5] Baodong Yan. Hawkeye technology using tennis match. *Computer Modelling and New Technologies*, 18(12):400–402, 2014.

A Competing Risks Model for the Time to First Goal

Petr Volf

Institute of Information Theory and Automation, Czech Ac. Sci., Prague
volf@utia.cas.cz

Abstract

In the contribution the time to the first goal in a football (soccer) match is analyzed, in the framework of competing risks scheme. Potential random times to the first goals scored by both teams are modelled by exponential distributions with parameters depending on attack and defence strengths of teams. Mutual dependence of these two times is described with the aid of a conveniently chosen copula ensuring the model identifiability. As a real example the data from the 2014 World Championship are analyzed. It is shown that the correlation is, as a rule, negative, and is absolutely larger in more competitive matches. Possible extensions of the approach are discussed, too.

1 Introduction

A basic probability model for final score of a football (soccer) match, presented for instance already in [3], consist of two conditionally independent Poisson random variables. It means that they are dependent just through shared parameters or covariates. More flexible models are obtained by generalizations, for instance the distribution of number of scored goals can be inflated to cover certain more frequent results. Another generalization can consist in considering a time development of model parameters as well as covariates during the match (see for instance [6]), in such a way a model based on counting process scheme is obtained. The present contribution concerns yet another direction of basic model improvement, namely to models considering an explicit form of dependence of both teams scoring distributions. Thus, in [2] a special case of bivariate Poisson distribution was employed. In the same context, McHale and Scarf [4] have described the dependence with the aid of a copula model. Interesting is the comparison of conclusions of both approaches. While the correlation in the former model is non-negative (by definition), the latter concludes that the correlation is negative and is absolutely larger in more competitive matches. It has to be also said that the use of copula in the discrete distribution models is not easy technically (and then computationally), because marginal distribution functions are as a rule expressed by sums of point probabilities, not having a reasonably closed form.

In the present contribution we analyze continuous distribution of time to the first scored goal in a match. Consequently, we deal with the scheme of competing risks. On the one hand the use of copula for two-dimensional continuous distribution can lead to a 'nice' closed form of the model, on the other hand it is well known that in the competing risks setting the model may be non-identifiable. A proof and an example of this phenomenon is given in [5], some instances of identifiable (or not) models are treated in [1] – in these classical studies the notion of copula has not been used yet. Therefore we are facing the problem of reasonable copula selection. Fortunately, it is known (cf. [7]), that the selection of copula type is not so crucial, that the finding proper value of its parameter (connected with correlation) is much more important.

Potential times to the first goal scored by each team are modelled by exponential distribution following from the basic Poisson model proposed in [3]. The parameters again consist of attack

and defence parameters of both teams. Further, for joint survival function we use a copula derived from Tsiatis' [5] example, its form is convenient for work with exponential distribution. Such a combination of marginal and simultaneous distributions has already been analyzed in [1] and proved to be identifiable. From this fact the consistency of estimates follows.

The outline of the paper is the following: The next section recalls the scheme of competing risks and the problem of possible non-identifiability. Then the copula model and corresponding likelihood is formulated. Section 4 then contains a real example, namely the analysis of data from the 2014 Football World Championship (in Brazil). The results are quite comparable with conclusions in [4], namely that estimated correlation is, as a rule, negative, and is absolutely large in more competitive matches, i.e. the matches of teams with good defence and comparable attack abilities.

2 Competing Risks Scheme

Let us assume that certain event (e. g. a failure of a device) can be caused by K reasons. Therefore we consider K (possibly dependent) random variables - survival times $T_j, j = 1, ..., K$, sometimes plus variable C of random right censoring (C is then independent of all T_j). Let $\overline{F}_K(t_1, ..., t_K) = P(T_1 > t_1, ..., T_K > t_K)$ be the joint survival function of $\{T_j\}$. However, instead the 'net' survivals T_j we observe just 'crude' data (sometimes called also 'the identified minimum') $Z = \min(T_1, ..., T_K, C)$ and the indicator $\delta = j$ if $Z = T_j$, $\delta = 0$ if $Z = C$.

Such data lead to direct estimation of the distribution of $Z = \min(T_1, ..., T_K)$, for instance its survival function $S(t) = P(Z > t) = \overline{F}_K(t, ..., t)$. Further, we can estimate **cause–specific hazard functions** for events $j = 1, 2, \ldots, K$:

$$h_j^*(t) = \lim_{d \to 0} \frac{P(t \le Z < t + d,\, \delta = j \mid Z \ge t)}{d},$$

and the **cumulative incidence functions**

$$F_j^*(t) = P(Z \le t, \delta = j) = \int_0^t S(s) \cdot h_j^*(s)\, \mathrm{d}s.$$

As both components, i.e. S and h_j^*, are estimable consistently by standard survival analysis methods, there also exist consistent estimates of F_j^*.

2.1 Problem of Non-Identifiability

However, in general, from data (Z_i, δ_i), $i = 1, \ldots, N$ it is not possible to identify neither marginal nor joint distribution of $\{T_j\}$. A. Tsiatis [5] has shown that for arbitrary joint model we can find a model with independent components having the same incidences, i.e. we cannot distinguish the models. Namely, this 'independent' model is given by cause-specific hazard functions $h_j^*(t)$. In other words, even if the model is parametrized and the MLE yields consistent estimates, in general we do not know parameters of which model are estimated. The situation can be better in the case of a regression model, because the covariates provide an additional information, especially when their structure is rich enough. On the other hand, as a consequence of the Tsiatis [5] result, in competing risks models without regressors it is necessary to make certain functional form assumptions about both marginal and joint distribution in order to identify them. Several such cases are studied in [1] and in some other papers.

3 Competing Risks and Copula

In the sequel we shall consider just 2 random variables S, T and data $Z_i = min(S_i, T_i, C_i), \delta_i = 1, 2, 0$. The notion of copula offers a way how to model multivariate distributions, we prefer here to use it for modelling the joint survival function $\overline{F_2}(s,t)$ of S, T:

$$\overline{F_2}(s,t) = C(\overline{F_S}(s), \overline{F_T}(t), \theta), \tag{1}$$

$\overline{F_S}$, $\overline{F_T}$ are marginal survival functions of S, T, $C(u, v, \theta)$ is a copula, i.e. a two-dimensional distribution function on $[0,1]^2$, with uniformly on $[0,1]$ distributed marginals U, V. θ is a copula parameter, which is, as a rule, uniquely connected with correlation of U, V, hence also with correlation of S, T. It is seen that the use of copula allows to model the dependence structure separately from the analysis of marginal distributions. From another point of view, the identifiability of the copula (and its parameter) and marginals can be considered as two separate steps.

Zheng and Klein [7] proved that when the copula is known, the marginal distributions are estimable consistently (and then the joint distribution, too, from (1)), even in non-parametric (so that quite general) setting. However, in general, also the knowledge of θ is needed. They also discussed importance of proper selection of copula form. As it has already been said, the knowledge (or a good estimate) of parameter θ is much more crucial for correct model of joint distribution. As a consequence, because the knowledge of copula type is still an unrealistic supposition, we can try to use certain sufficiently flexible class of copulas, as approximation, and concentrate to reliable estimation of its parameter.

3.1 Copula Based on Tsiatis' Example

Let us return to the example of Tsiatis [5], considering just $K = 2$ random variables S, T with exponential distribution and the following marginal and joint survival functions,

$$\overline{F}_S(s) = e^{-\lambda s}, \ \overline{F}_T(t) = e^{-\mu t}, \ \overline{F}_2(s,t) = e^{-\lambda s - \mu t - \gamma s t}.$$

Hence, $S(t) = \overline{F}_2(t,t) = exp(-\lambda t - \mu t - \gamma t^2)$. Corresponding cause-specific hazard rates and their integrals are

$$h_S^*(t) = (\lambda + \gamma t), \ h_T^*(t) = (\mu + \gamma t), \ H_S^*(t) = (\lambda t + \frac{\gamma}{2} t^2), \ H_T^*(t) = (\mu t + \frac{\gamma}{2} t^2).$$

It follows that $S^*(t) = exp(-H_S^*(t) + H_T^*(t))$ is the same as $S(t)$ above, which means that independent random variables with marginal survival functions

$$\overline{G}_S(s) = e^{-\lambda s - \frac{\gamma}{2} s^2}, \ \overline{G}_T(t) = e^{-\mu t - \frac{\gamma}{2} t^2}$$

yield the same competing risk scheme. Notice, however, that 'true' marginal distributions are exponential while derived independent distributions are not. It gives a chance that, when the type of marginals is assumed, they (and parameter γ, too) can be estimated, uniquely. Tsiatis' example actually uses the following copula:

$$C(u,v) = u \cdot v \cdot \exp(-\theta \cdot \ln u \cdot \ln v) \tag{2}$$

with $\theta \geq 0$, corresponding correlation $\rho(U,V) \leq 0$, $\theta = 0$ means independence of U, V. The parameters are connected in the following way: $\gamma = \theta \cdot \lambda \cdot \mu$. Figure 1 shows the dependence of correlation on parameters.

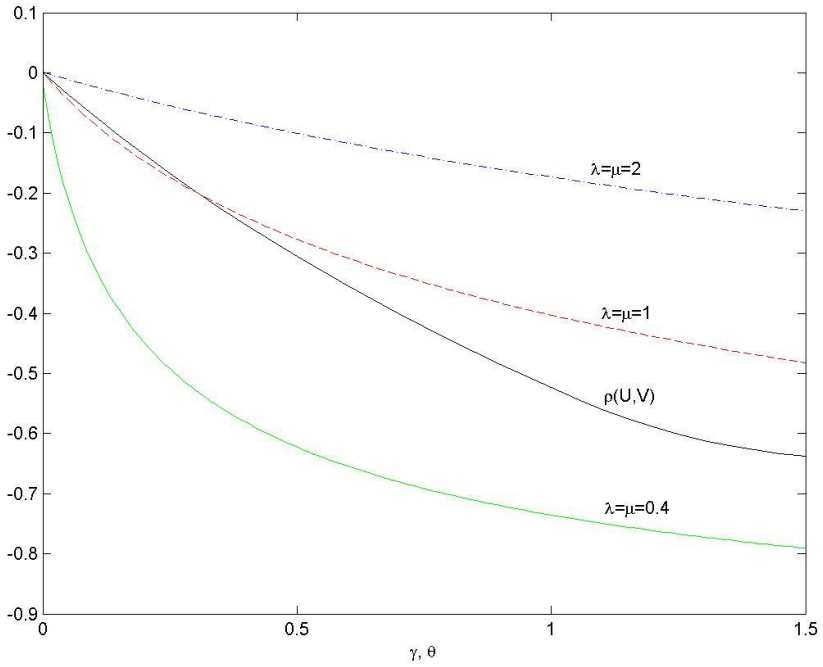


Figure 1: Dependence of $\rho(U,V)$ on parameter θ and $\rho(S,T)$ on γ, when $S \sim Exp(\lambda)$, $T \sim Exp(\mu)$.

It is easy to show that the case of competing risks with two exponential marginal distributions tied together by copula (2) is identifiable. It actually has been proven already by Basu and Ghosh [1] – though authors did not use a notion of copula yet. It is also easy to verify that the case fulfils the regularity conditions and therefore yields unique ML estimates of parameters. The likelihood function has the following form:

$$L = \prod_{i=1}^{N} (\lambda + \gamma Z_i)^{[\delta_i=1]} \cdot (\mu + \gamma Z_i)^{[\delta_i=2]} \cdot S(Z_i),$$

where again $Z_i = min(S_i, T_i, C_i), \delta_i = 1, 2, 0$.

3.2 Other Two-Dimensional Exponential Distributions

The identifiability results obtained above need not hold for other selection of copula type. For instance, let us consider the Gumbel copula

$$C(u,v) = \exp\{-[(-\ln u)^\theta + (-\ln v)^\theta]^{1/\theta}\},$$

with $\theta \geq 1$. Here $\rho(U,V) \geq 0$, $\theta = 1$ corresponds to independence. Let again $S \sim Exp(\lambda)$, $T \sim Exp(\mu)$, then

$$\overline{F_2}(s,t) = \exp\{-[(\lambda s)^\theta + (\mu t)^\theta]^{1/\theta}\}, \;\; i.e. \;\; S(z) = \exp\{-[\lambda^\theta + \mu^\theta]^{1/\theta} \cdot z\},$$

It is easy to check that the corresponding competing risks model is 'over-parametrized', determined by any couple of parameters only, i.e. we cannot estimate λ, μ, θ uniquely.

Another often used model for bivariate exponential distribution is the Marshall–Olkin model: Let X_1, X_2, X_3 be independent exponential random variables with parameters $\lambda_1, \lambda_2, \lambda_3$, respectively, set $S = \min(X_1, X_3)$ and $T = \min(X_2, X_3)$. Then marginal distributions of S, T are also exponential, with parameters $\lambda_1 + \lambda_3$, $\lambda_2 + \lambda_3$, resp., their correlation equals $\lambda_3/(\lambda_1 + \lambda_2 + \lambda_3)$. However, as $P(S = T) = \lambda_3/(\lambda_1 + \lambda_2 + \lambda_3)$, too, the joint distribution of S, T is not of continuous type and, therefore, is not convenient for our purposes. Let us note that this distribution is closely connected with bivariate Poisson model used for instance in [2].

4 Application to Time of the First Goal

We shall now use the competing risk model derived in Part 3.1 to modelling the time to first scored goal during a football (soccer) match. Marginal variables are the 'latent' times of 1-st goal of each time, however only the incidence of one of them is observed. Or, in the case of draw 0:0, we have censoring by a fixed value 90 minutes (or 120 minutes in the case of prolonged match). Except statistical estimation of model parameters, we are interested in the main question: How dependent are these 'latent' times to 1-st goal of both teams?

In our rather small study we shall use the data from the Football World Championship 2014 in Brazil. 32 participating teams played together 64 matches, some of them just 3 matches in a group. In order to improve this proportion (matches to team), we considered just 11 'teams': 8 teams passing to quarterfinal, then 'team' No 9 - aggregated data of 8 teams losing in eight-final matches, No 10 - teams taking 3-rd places in groups, No 11 - teams ending 4-th in groups. Let us recall also the final order of the championship: 1. Germany, 2. Argentina, 3. The Netherlands, 4. Brazil.

As regards marginal models, the source was the standard model of Maher [3]. More specifically, each team (i) was characterized by its attack parameter a_i and defense parameter b_i. The sequence of scoring in a match between teams i and j is then described by two Poisson processes with intensities $a_i \cdot b_j$, $a_j \cdot b_i$, respectively. Consequently, the time to the 1-st goal followed from two competing exponential random variables

$$S_{ij} \sim Exp(a_i \cdot b_j)\,,\ T_{ij} \sim Exp(a_j \cdot b_i).$$

Further, it was assumed that their mutual dependence can be expressed via 'Tsiatis' model described in Part 3.1.

Thus, we were facing the problem of the maximum likelihood estimation (MLE) of 23 parameters, a_i, b_i of 11 teams and γ characterizing the dependence. It was assumed that γ was the same for all couples of teams, i.e. in all matches. The results of the MLE of teams parameters are displayed in Table 1. For computational convenience, we estimated $\alpha_i = \ln a_i$, $\beta_i = \ln b_i$. Finally, the MLE of parameter γ was 0.605, with half-width of approximate 95% confidence interval 0.143.

It is possible to say that our result is comparable with the findings of McHale and Scarf [4]. Inspection of the graph on Figure 1 indicates that in the match of teams with very good defense (as for instance Germany and Argentina) the first goal really matters, correlation is large (absolutely), while in a opposite case of weaker defense and sufficiently good attack ability (here for instance Columbia and - rather surprisingly - Brazil) the correlation is smaller (but still negative).

Team	alpha		beta		a	b
Brazil	0.8408	(1.1756)	0.6683	(1.1519)	2.3181	1.9509
Netherlands	0.3580	(1.2784)	-0.9680	(2.1790)	1.4305	0.3799
Columbia	0.8542	(1.0408)	-0.2337	(2.0061)	2.3496	0.7916
Costa Rica	-0.9342	(2.3540)	-1.6044	(3.6409)	0.3929	0.2010
France	-0.2146	(1.5123)	-0.8994	(2.1512)	0.8068	0.4068
Argentina	0.4830	(0.9885)	-4.6885	(13.4755)	1.6209	0.0092
Germany	0.6888	(0.9692)	-5.0360	(17.2193)	1.9913	0.0065*
Belgium	-0.7982	(2.1896)	-0.2884	(1.5620)	0.4501	0.7495
No 9	-0.1707	(0.6335)	-0.3715	(0.6572)	0.8431	0.6897
No 10	0.1572	(0.6805)	0.4176	(0.6815)	1.1702	1.5183
No 11	-1.3428	(1.6872)	0.4444	(0.5148)	0.2611	1.5596

Table 1: **Results:** Estimated parameters $\alpha_i = \ln a_i$, $\beta_i = \ln b_i$ (with half-widths of approximate 95% conf. intervals in brackets), then a_i, b_i

5 Conclusion

We have studied the dependence of random variables – latent times of scoring the first goal in a football matches, with the aid of the competing risk model. Achieved results lead to conclusion that the correlation is, as a rule, negative, and is absolutely larger in more competitive matches. The approach can be extended to the analysis of times to next goals, further generalization can consider different copula parameters for certain groups of matches and/or teams. Further, in a more general models the intensities can also depend on other factors and on match development (see also [6] for an overview of models).

Acknowledgement. The research has been supported by the project No 13-14445S of the Czech Scientific Foundation (GA ČR).

References

[1] A.P. Basu and J.K. Ghosh. Identifiability of the multinormal and other distributions under competing risks model. *Journal of Multivariate Analysis*, 8:413–429, 1978.

[2] D. Karlis and I. Ntzoufras. Analysis of sports data by using bivariate poisson models. *J. R. Stat. Soc. Ser. D*, 52:381–394, 2003.

[3] M.J. Maher. Modelling association football scores. *Stat. Neerl.*, 36:109–118, 1982.

[4] I. McHale and P.A. Scarf. Modelling the dependence of goals scored by opposing teams in international soccer matches. *Statistical Modelling*, 11:219–236, 2011.

[5] A. Tsiatis. A nonidentifiability aspects of the problem of competing risks. *Proc. Nat. Acad. Sci. USA*, 72:20–22, 1975.

[6] P. Volf. A random point process model for the score in sport matches. *IMA Journal of Management Mathematics*, 20:121–131, 2009.

[7] M. Zheng and J.P. Klein. Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika*, 82:127–138, 1995.