



**Proceedings of the
4th International Conference on
Mathematics in Sport**

MathSport International 2013

Leuven, Belgium

5-7 June 2013

Editors: Dries Goossens, Frits Spieksma, Phil Scarf

ISBN: 9789081409964

Contents

Assessing efficiency and setting benchmarks for NBA teams through DEA and DMU clustering	
L. A. Alves, H. H. Kramer, P. B. Tschaffon and J. C. C. B. Soares de Mello	1-7
Estimation of anthropometrical and inertial body parameters	
J. Bastien, Y. Blache and K. Monteil	8-16
Optimization modeling for analyzing fantasy sport games	
J. Beliën, D. Goossens and D. Van Reeth	17-25
A virtual coach for fantasy soccer using mathematical programming	
F. Bonomo, G. Durán and J. Marenco	26-31
2012 UEFA Euro efficiency evaluation based on market expectations	
L. C. Brandão, F. do Valle Silva Andrade and J. C. C. B. Soares de Mello	32-37
Team performance in the Italian Serie A: 2000/01 - 2009/10	
F. Carmichael, G. Rossi and D. Thomas	38-50
Model of joint displacement using sigmoid function: experimental approach for planar pointing task and squat jump	
T. Creveaux, J. Bastien, C. Villars and P. Legreneur	51-61
Quantitative models for retirement risk in professional tennis	
A.C.J. Cutmore and W.J. Knottenbelt	62-79
ODI cricket: characterizing the performance of batsmen using ‘tipping points’	
U. Damodaran	80-86
Inferring the score of a tennis match from in-play betting exchange markets	
A. G. I. Dumitrescu, B. X. Huang, C. W. Knottenbelt, D. Spanias and E. J. Wozniak	87-95
Markov chain volleyball	
M. Ferrante and G. Fonseca	96-107
An Olympic ranking based on sequential use of ordinal multicriteria methods	
S. F. Gomes Júnior, J. C. C. B. Soares de Mello and L. Angulo Meza	108-114
The effect of fatigue from the previous match in Grand Slam tennis	
D. Goossens, J. Kempeneers and F. Spieksma	115-119
Are soccer schedules robust?	
D. Goossens and F. Talla Nobibon	120-134
Probability calculation for tournament format of the 2013 World Baseball Classic	
N. Hirotsu	135-143

Developing an improved tennis ranking system	
D. Irons, S. Buckley and T. Paulden	144-153
Relative importance of the offensive and defensive efficiencies in the FIVB Men's Volleyball World League	
E. Kondo, R. T. Stefani and A. Bedford	154-160
Career lengths and age of retirement of ATP singles players during the Open Era	
S. Kovalchik	161-171
The PEAST algorithm in sports scheduling: evolution of the Major Finnish Hockey League schedules	
N. Kyngäs, D. Goossens, K. Nurmi and J. Kyngäs	172-179
Scheduling a sports league with divisional and round-robin play	
Jeffrey Larson and Mikael Johansson	180-192
Assignment of swimmers to events in a multi-team meeting for team global performance optimization	
S. Mancini	193-205
Fair referee assignment for the Italian soccer Serie A	
S. Mancini and A. Isabello	206-215
Using probabilistic models to simulate tennis matches, with applications to betting strategies	
V. Mogilenko and G. Hunter	216-221
Multivariate analysis of heptathlon results	
A. Murphy and P. Bidgood	222-228
Market making with an inverse Kelly strategy	
E. Noon and W. Knottenbelt	229-232
Elite players' perceptions of football playing surfaces: an ordinal logistic regression model of players' overall opinions	
A. Owen, A. Smith, P. Osei-Owusu, J. Roberts, A. Harland and S. Larman	233-243
A DEA Evaluation and Financial Resources Reallocation for Brazilian Olympic Sports regarding their results in the 2011 Pan-American Games	
R. Pescarini Valério and L. Angulo-Meza	244-249
A closer look at the Independence between points of the top four male players	
G. Pollard	250-257
A comparison of the Masters scoring system and the Knock-out scoring system	
G. Pollard and G. Pollard	258-265

Using forecasting to detect corruption in international football	
J. J. Reade and S. Akie	266-276
Integer programming and sports ranking	
C. Raack, A. Raymond, T. Schlechte and A. Werner	277-286
Dynamic opponent-dependent and position-dependent player ratings in the AFL	
J. Sargent and A. Bedford	287-294
Modelling and optimisation of the sport and exercise training process	
P. Scarf, M. Shrahili, S. Jobson and L. Passfield	295-300
Measuring competitive balance: the case of European premiership rugby union	
P. Scarf, P. Williams and M.M. Yusof	300-309
It is harder, not easier, to predict the winner of the Champions League	
J. Schokkaert and J. Swinnen	310-324
The Macbeth method for ranking Olympic sports: a complementary analysis for the DEA efficiency	
J. C. C. B. Soares de Mello, J. Benício, L. Bragança and V. Guimarães	325-333
Tennis player ranking using quantitative models	
A. D. Spanias and B. W. Knottenbelt	334-341
The London Olympics in perspective: athletics, swimming and home nation medal advantage	
R. Stefani	342-349
Performance inequality at the Olympic Games	
E. Sterken	350-357
Evaluating regional balance in the NCAA men's basketball tournament using the tournament selection ratio	
J. A. Trono	358-368
Metaheuristic Optimisation of Parameterised Betting Exchange Strategies	
P.Tsirimpas and W.J. Knottenbelt	369-375
Sensitivity of court-side in tennis?	
M. Viney, A. Bedford and E. Kondo	376-382
Intelligent computational optimisation of sport skills	
J. Wright and I. Jordanov	383-390

Assessing efficiency and setting benchmarks for NBA teams through DEA and DMU clustering

L. A. Alves* and H. H. Kramer** and P. B. Tschaffon*** and J. C. C. B. S. de Mello****

* Rua Passo da Pátria, 156 – São Domingos – Niterói – RJ- Brazil laura_alves_aa@yahoo.com.br

** Rua Passo da Pátria, 156 – São Domingos – Niterói – RJ- Brazil hugoharry@gmail.com

*** Rua Passo da Pátria, 156 – São Domingos – Niterói – RJ- Brazil pamtschaffon@yahoo.com.br

**** Rua Passo da Pátria, 156 – São Domingos – Niterói – R- Brazil J jcsmello@producao.uff.br

Abstract. In this work, Data Envelopment Analysis (DEA) methodology is applied to assess the efficiency of teams which played the 2011–2012 NBA regular season. To apply such methodology, 30 teams were considered as Decision Making Units (DMU). For each team, the total payroll is considered as input, and the number of victories and average scored points are the outputs. Firstly, Self-Organizing Maps (SOM) were used to group teams into homogeneous clusters. Then, the classic efficiency of each team is computed with the DEA BCC model oriented to outputs and considering weight restrictions. Thus, the composed efficiency score of each DMU is obtained using the inverted frontier method. At the final step, a ranking is determined for each group of DMUs according to their composed efficiency based on the inverted frontier.

1. Introduction

Being one of the most popular sports in the world, mainly in the United States, the basketball is also surrounded by a wide business network which generates great financial operations. A basketball match involves two teams, each one with five players, and the team which scores more points is the winner. The match will never be ended in a draw. If this situation occurs at the end of the regular time, additional times will be played until a team can be declared winner.

The Professional Basketball League of the United States (NBA) is one the most known competitions of the sport, comprising 30 teams spited in two conferences (east and west). In addition, each conference has three divisions which also cluster the teams according to their geographic location.

The competition goes as follows: each team faces all the others at least once along the season. Teams in the same division confront each other exactly four times, and teams in the same conference face each other three times. Between teams from different conferences two matches are played. Thus, a team plays 82 matches, where half of them are held at its own gymnasium (home) and the other half are played away from home. These 82 matches compose the regular season and, at its end, the 8 teams of each conference with more wins are qualified to the playoffs, when the champion of the season will be defined.

Due to its popularity and the large investments that surround this sport, this study proposes an efficiency analysis of 30 teams that played the 2011-2012 regular season using DEA. The methodology proposed in this paper considers the composed efficiency index to assess the efficiency of teams belonging to clusters obtained through Self-Organizing Maps.

In literature, it is possible to find a diversity of works applying the DEA methodology to efficiency assessment in sports, in particular to basketball. Cooper *et al.* (2009) use a procedure in which non-zeros weights are selected to evaluate players' efficiency. Bai (2009) also assesses the efficiency of players, but using cross-efficiency along with environment variables. Aizemberg *et al.* (2011b) present an analysis within a time period of the NBA teams and apply a multi-objective approach in order to determine the benchmarks, and Aizemberg *et al.* (2011a) measure the efficiency of those teams by means of cross-efficiency and the DEA-GAME approach. Besides, a network DEA approach is used by Moreno and Lozano (2012) to assess the efficiency of 30 NBA teams for the regular season 2009-2010, the results suggest that this approach presents more discriminating power than the conventional DEA approach.

It is possible to find a wide application of Self-Organizing Maps (SOM), also known as Kohonen Neural Network, in papers that intend to identify and analyze patterns in sports. Grunz *et al.* (2011) create a neural network to find tactical patterns using position data of typical soccer games. Lees and Barton (2004) use this method to characterize the patterns of kicks in soccer games.

Although studies for grouping sport teams were not found in the literature, various authors use SOM in order to obtain homogeneous clusters (MARKEY *et al.*, 2003; PLETNEY *et al.*, 2002) and, then, apply DEA for assessing efficiency index (YU & LEE, 2013; LI *et al.*, 2009).

This work consists of presenting two steps in sequence: clustering via Kohonen Neural Networks and evaluation of the teams in each cluster using DEA. The first step consists in clustering the DMUs with similar characteristics using inputs that take into account the home advantage. The second step consists in assessing the classic and the inverted efficiencies of teams using the BCC model (Banker *et al.*, 1984) oriented to output and with weight restrictions. Then, the composed efficiencies are calculated in order to improve discrimination and to rank the DMUs in each cluster. The goal is to evaluate and set ranks for DMUs that share similar characteristics.

The paper is organized as follows: Section 2 provides the data used and the methodology proposed describing the SOM method, the DEA BCC model, the inverted efficiency and the composed efficiency index. The result of study is shown in Section 3, as well as its analysis. Finally, Section 4 discusses the conclusions of this work and provides suggestions for future studies.

2. Methodology

2.1. Self-Organizing Map

As noted in the introduction of this paper, the matches take place in different cities with diverse characteristics, which can intensify the home advantage. Some studies (ANDERSON *et al.*, 2012; CARRÉ *et al.*, 2006) suggest that this phenomenon affects significantly the performance of teams in different sports. Hence, for a fair and homogeneous assessment, the DMUs were grouped in clusters defined based on conditions of the cities that hosted their matches.

The literature presents some methods used to group units. This work uses the Self-Organizing Maps for this purpose. This method, as proposed by Kohonen (1982), composes a specific type of neural network, which each neuron is initialized with an identical set of inputs and compete each other in order to be the winning neuron for that specific set of inputs.

Biondi Neto *et al.* (2011) emphasizes that another important feature of SOM is the unsupervised training, which allows the network finds similarities based only on input patterns. Then, the similar input data are grouped into clusters.

So, NBA teams are grouped based in the following set of inputs: average attendance at home matches, total miles traveled by the team during the season and average GDP per capita of host cities.

The choice of such variables to perform clustering by SOM is due to the fact that they reflect the different characteristics of the cities where the teams play their matches. As mentioned in the introduction, half of the games were played in the team's own gym, which results in home advantage effect. The first variable chosen for SOM – “Average home attendance” – expresses the level of public participation in the games that occurred in the team's own gym, and so we have an indicator of home advantage. The second variable – “Total Miles Traveled” – expresses the distance traveled by each team during the competition. This variable can also be interpreted as the effort that fans of a team would make in order to attend it in all its games. Farther a team is playing from its hometown, lower should be the participation of its fans at the games. The last variable chosen for clustering – Average GDP per capita – is the average GDP per capita of all cities that hosted the games that each team participated. This variable reflects the economic condition of the city's public, in other words, if people could pay for tickets for the games occurred in their town.

Data listed in columns 3-5 of Table 1 were normalized and inserted in Matlab, software used to develop the SOM. Section 3 presents the clusters found and the efficiency calculated for each team of clusters. The inputs and outputs used in the assessment of this efficiency are detailed in the next section of this paper.

2.2. Inputs and outputs

The choice of variables to be considered in the DEA model represents the structure of the problem under analysis and reflects the relationship between its inputs and its outputs (Meza, 1998).

In this paper, after clustering the basketball teams, a DEA model is used to calculate the efficiency, considering as DMUs the teams in each cluster. This methodology allows a more appropriated definition of benchmarks for the inefficient DMUs, since the clusters have DMUs with similar realities.

The DMUs that will be analyzed are the 30 teams that played the 2011-2012 NBA regular season and the variables chosen for the studied problem are the total payroll, the number of victories and the average score.

The total payroll represents the investment that each team did in their athletes and employees. This paper considers that the players' expertise and talent are the main factor for achieving a good performance. Much better players and auxiliary staff are, higher will be their wages, since these professionals will be desired by another teams. The number of victories represents the main purpose of a team, that is win the game. But, there is another variable that represents how good a team is, which is the score done. Playing with a team formed by good players, higher the chance of achieving high scores in games and winning the competitions. So, the analyzed problem considers one input – “Total payroll” – and two outputs – “Number of victories” and “Average Score”. The aforementioned data is shown in Table 1.

Table 1: Data used in the study

City	Team	<i>Inputs / SOM</i>		<i>Inputs / DEA</i>		<i>Outputs/ DEA</i>	
		Average Home Attendance	Total Miles Traveled	Average GDP per capita	Total Payroll	Victories	Average Score
Boston	Celtics	18.624	33.617	52.435	\$ 79.503.322,00	39	91,8
Brooklyn	Nets	13.961	33.211	51.533	\$ 56.632.054,00	22	93,1
New York	Knicks	19.763	34.050	51.796	\$ 63.410.254,00	36	97,8
Philadelphia	76ers	17.502	31.962	45.687	\$ 67.030.330,00	35	93,6
Toronto	Raptors	16.835	30.754	46.708	\$ 46.584.937,00	23	90,7
Chicago	Bulls	22.161	30.786	48.324	\$ 69.280.572,00	50	96,3
Cleveland	Cavaliers	15.926	28.567	44.056	\$ 52.838.494,00	21	93,0
Detroit	Pistons	14.413	25.863	42.511	\$ 66.561.901,00	25	90,9
Indiana	Pacers	14.168	31.571	44.239	\$ 51.808.036,00	42	97,7
Milwaukee	Bucks	14.718	33.963	45.809	\$ 60.011.567,00	31	99,0
Atlanta	Hawks	15.199	38.744	44.052	\$ 71.038.778,00	40	96,6
Charlotte	Bobcats	14.757	31.184	44.526	\$ 56.691.609,00	7	87,0
Miami	Heat	19.935	43.280	43.624	\$ 77.363.284,00	46	98,5
Orlando	Magic	18.896	46.221	43.365	\$ 69.377.976,00	37	94,2
Washington	Wizards	16.728	30.158	49.087	\$ 60.066.363,00	20	93,6
Dallas	Mavericks	20.334	33.427	46.142	\$ 75.877.322,00	36	95,8
Houston	Rockets	15.363	39.256	46.490	\$ 57.266.790,00	34	98,1
Memphis	Grizzlies	15.704	39.054	43.427	\$ 70.044.412,00	41	95,0
New Orleans	Hornets	15.109	39.044	47.002	\$ 67.149.408,00	21	89,6
San Antonio	Spurs	18.396	34.515	46.100	\$ 73.194.941,00	50	103,7
Denver	Nuggets	17.029	37.027	49.276	\$ 58.107.300,00	38	104,1
Minnesota	Timberwolves	17.490	42.362	48.295	\$ 56.844.339,00	26	97,9
Portland Trail	Blazers	20.496	43.391	45.871	\$ 67.823.091,00	28	97,2
Oklahoma City	Thunder	18.203	34.584	44.715	\$ 61.331.254,00	47	103,1
Utah	Jazz	19.306	42.836	44.741	\$ 58.288.257,00	36	99,7
Golden State	Warriors	18.857	39.434	49.165	\$ 58.716.176,00	23	97,8
Los Angeles	Clippers	19.219	41.436	49.278	\$ 69.339.317,00	40	97,5
Los Angeles	Lakers	18.997	41.697	49.499	\$ 85.669.424,00	41	97,3
Phoenix	Suns	15.597	36.549	43.982	\$ 63.075.891,00	33	98,4

Sacramento	Kings	14.508	44.335	49.116	\$ 46.514.655,00	22	98,8
------------	-------	--------	--------	--------	------------------	----	------

2.3. BCC model and inverted frontier

Data Envelopment Analysis - DEA (Charnes *et al.*, 1978) is a mathematical tool which main goal is measuring the efficiency of decision making units (DMU) in a given set of observations. This tool allows identifying which units are efficient, besides defining benchmarks between such units. DEA has two classical models, which are the CRS (Charnes *et al.*, 1978) and BCC (Banker *et al.*, 1984). The CRS model works with constant returns to scale, while the BCC assumes that the DMUs can exhibit variable returns to scale. Both models can work with input orientation, where inputs are minimized while outputs remain unchanged, or can work with output orientation, when the objective is to maximize the outputs while inputs do not change.

In this paper, the DEA BCC model was chosen due to the fact that the number of victories is a limited variable, since teams can win at most 82 matches. Therefore, the variables do not vary proportionally. The output orientation was chosen since the main purpose of the basketball teams is having high scores and winning the matches. Also, weight restrictions are imposed in a way that the weights associated to victories are at least twice the weight assigned to points scored.

The Inverted Frontier emerges as a method to distinguish which DMUs are really efficient and consists of a pessimistic assessment of DMUs. This type of frontier, based on the inversion of inputs and outputs, is presented in Yamada *et al.* (1994), Entani *et al.* (2002), Lins *et al.* (2005). This method evaluates the inefficiency of the DMU by constructing a frontier formed by DMUs with the worst management practices, called inefficient border.

Despite the cross-evaluation method is applicable to this work, the inverted frontier was selected for presenting a easier application and because the cross-evaluation method cannot be applied with the DEA BCC model.

In order to solve the problem of low discrimination in DEA and to order the DMUs, it is calculated a composed efficiency index (Soares de Mello *et al.*, 2005), which is the arithmetic average between the efficiency in relation to original frontier and inefficiency in relation to inverted frontier.

Thus, for a DMU be efficient in the composed index, it should have a high degree of relevance to the optimistic border and low degree in the pessimistic border. With this method, all variables are taken into account, and in a far less benevolent way than the efficiency calculation performed on the efficient frontier of classical DEA models.

3. Results and discussion

This section presents the results obtained from the application of the proposed approach described in the previous section of this paper.

The Kohonen network used presents a hexagonal topology with a [2x2] grid. With these parameters, a maximum of four clusters can be found. In fact, exact four clusters were found. Table 2 presents the clusters found and the efficiency scores assessed for teams in each cluster.

Clusters have some similar characteristics, which enables a fair assessment of efficiency, since the teams were compared with other teams that played in similar situations and considering the home advantage.

The teams of cluster 4 played under favorable conditions, since they played in cities with a high average GDP per capita and had a high home advantage. Furthermore, these teams didn't present a high value regarding to total miles traveled.

On the other hand, the cluster 3 presents low values for the average attendance and median values related to the average GDP per capita of cities that hosted the matches.

After clustering step, the efficiency indexes showed interesting results about the performance of the teams. All teams that took first place in the rankings of each cluster have classic efficiency equal to 1. Among six teams that were champions in their divisions, only Thunder (Northwest division) was ranked first in its cluster.

Still analyzing the performance of champions within the divisions, while Thunder and Spurs were the unique efficient according to classic DEA BCC model, Celtics was the worst ranked in its cluster. The study

also found that the champion of the season (Heat) was neither first in its cluster nor efficient in classic point of view.

Table 1: Clusters and efficiency indexes found

Position	City	Team	Cluster	Results / DEA		
			[2x2]	Classic Efficiency	Inverted Efficiency	Composed Efficiency
1st	Miami	Heat	1	0,989	0,982	0,503
2nd	Orlando	Magic	1	0,945	1,000	0,472
3rd	Utah	Jazz	1	1,000	0,945	0,528
4th	Portland Trail	Blazers	1	0,975	0,969	0,503
5th	Minnesota	Timberwolves	1	1,000	0,962	0,519
6th	Golden State	Warriors	1	1,000	0,963	0,519
7th	Los Angeles	Clippers	1	0,978	0,970	0,504
8th	Los Angeles	Lakers	1	0,976	1,000	0,488
1st	Atlanta	Hawks	2	0,927	1,000	0,464
2nd	Houston	Rockets	2	0,950	0,917	0,517
3rd	Memphis	Grizzlies	2	0,912	1,000	0,456
4th	New Orleans	Hornets	2	0,870	1,000	0,435
5th	Denver	Nuggets	2	1,000	0,875	0,563
6th	Phoenix	Suns	2	0,945	0,927	0,509
7th	Sacramento	Kings	2	1,000	0,906	0,547
1st	Brooklyn	Nets	3	0,982	0,955	0,513
2nd	Cleveland	Cavaliers	3	1,000	0,948	0,526
3rd	Detroit	Pistons	3	0,918	1,000	0,459
4th	Indiana	Pacers	3	1,000	0,929	0,536
5th	Milwaukee	Bucks	3	1,000	0,915	0,542
6th	Charlotte	Bobcats	3	1,000	1,000	0,500
7th	Philadelphia	76ers	3	0,945	1,000	0,473
8th	Toronto	Raptors	3	1,000	0,963	0,518
9th	Washington	Wizards	3	0,980	0,953	0,513
1st	New York	Knicks	4	1,000	0,938	0,531
2nd	Boston	Celtics	4	0,885	1,000	0,443
3rd	Chicago	Bulls	4	0,931	0,952	0,489
4th	Dallas	Mavericks	4	0,927	0,958	0,485
5th	San Antonio	Spurs	4	1,000	1,000	0,500
6th	Oklahoma City	Thunder	4	1,000	0,890	0,555

Analyzing the performance of teams in each group, cluster 2 presents the greater difference between maximum and minimum values for each efficiency score, while cluster 1 presents the lowest difference. By the classic BCC point of view, more than a half of teams in cluster 3 are efficient and teams with the highest and lowest composed efficiency were grouped in cluster 2.

4. Concluding remarks

In order to assess the efficiency of 30 teams that played 2011-2012 NBA season, this work developed a hybrid model composed by Self-Organizing Map and the DEA technique. Thus, a rank was proposed for each cluster generated by the neural network used.

The use of SOM to cluster teams became the analysis of efficiency fairer, once considers the different conditions in each match and the existence of home advantage, which can benefit some teams over others. After, using inverted frontier the problem of poor discrimination of teams in classic efficiency method was solved.

The first stage of this work showed that teams of different divisions may share similar characteristics, while the opposite may be observed for teams in the same division. This article also concluded that the champion of 2011-2012 NBA season and most of the division champions are not efficient.

References

- Aizemberg, L., Ramos, T.G., Roboredo, M.C., de Azevedo, G.H.I., Alves, A.M., Caldas, M.A.F. (2011a) Measuring the NBA teams cross-efficiency by DEA game. *Proceedings of the 3rd IMA International Conference on Mathematics in Sport*, Manchester, England.
- Aizemberg, L., Ramos, T.G., Roboredo, M.C., de Azevedo, G.H.I., Alves, A.M., Caldas, M.A.F. (2011b) Medindo a eficiência DEA de times de basquete da NBA: análise temporal da eficiência e enfoque multiobjetivo para obtenção de benchmarks. *Anais do XLIII SBPO 2011*, Ubatuba-SP, Brasil.
- Anderson, M., Wolfson, S., Neave, N., Moss, M. (2012) Perspectives on the home advantage: A comparison of football players, fans and referees. *Psychological of Sport and Exercise* **13**, pp. 311-316.
- Bai, F. (2009) Testing the effects of environmental variables on efficiency and generating multiple weight sets for cross-evaluation with DEA: an application to the National Basketball Association.
- Banker, R.D., Charnes, A., Cooper, W.W. (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science*, pp. 1078-1092.
- Biondi Neto, L., Alves, A.M., Ramos, T.G., Mello, J.C.S. (2011). Estudo da qualidade do ar do Estado de São Paulo por meio de mapas auto-organizáveis de Kohonen. *Relatórios de pesquisa em Engenharia de Produção* **11**, n 13.
- Carré, J., Muir, C., Belanger, J., Putnam, S.K. (2006) Pre-competition hormonal and psychological levels of elite hockey players: Relationship to the “home advantage”. *Psychological & Behavior* **89**, pp. 392-398.
- Charnes, A., Cooper, W. W., Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research* **2**, p. 429-444.
- Cooper, W.W., Ruiz, J.L., Sirvent, I. (2009), Selecting non-zero weights to evaluate effectiveness of basketball players with DEA. *European Journal of Operational Research* **195**, pp. 563-574.
- Doyle, J., Green, R. (1994), Efficiency and cross-efficiency in DEA: derivations, meanings and uses. *Journal of the Operational Research Society* **45**, pp. 567-578.
- Entani, T.; Maea,Y.; Tanaka,H. (2002) Dual Models of Interval DEA and its extensions to interval data. *European Journal of Operational Research* **136**, 32-45.
- Grunz, A., Memmert, D., Perl, J. (2011) Tactical pattern recognition in soccer games by means of special self-organizing maps. *Human Movement Science*.
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics* **43**, n 1, pp. 59-69.
- Lees, A., Barton, G. (2004) A characterization of technique in the soccer kick using a Kohonen neural network analysis. *Journal of Sports Sciences* **22**, pp. 491.
- Li, Z.; Liao, H.; Coit, D. W. (2009) A two-stage approach for multi-objective decision making with applications to system reliability optimization. *Reliability Engineering and System Safety* **94**, pp. 1585-1592.
- Lins, M.P.E.; Novaes, L.F.L.; Legey, L.F.L. (2005) Real estate value assessment: a double perspective data envelopment analysis. *Annals of Operations Research*.
- Markey, M.K., Lo, J.Y., Tourassi, G. D., Floyd Jr, C. E. (2003) Self-organizing map for cluster analysis of a breast cancer database. *Artificial Intelligence in Medicine* **27**, pp. 113-127.
- Meza, L. A. (1998) Data Envelopment Analysis (DEA) na determinação da eficiência dos Programas de Pós-Graduação do COPPE/UFRJ, Tese de Mestrado, COPPE/UFRJ, Rio de Janeiro.
- Moreno, P.; Lozano, S. (2012) A network DEA assessment of team efficiency in the NBA. *Annals of Operations Research*, pp. 1-26.
- Pletnev, I.V., Zernov, V.V. (2002) Classification of metal ions according to their complexing properties: a data-driven approach. *Analytica Chimica Acta* **455**, pp. 131-142.
- Soares de Mello, J.C.C.B.; Gomes, E.G.; Meza, L. A.; Leta, F.R. (2008) DEA advanced models for geometric evaluation of used lathes. *WSEAS Transactions on Systems* **7**, pp. 510-520.

- Yamada, Y.; Matui, T.; Sugiyama, M. (1994). New analysis of efficiency based on DEA. *Journal of the Operations Research Society of Japan* **37**, n. 2, p. 158-167.
- Yu, P.; LEE, J. H. (2013) A hybrid approach using two-level SOM and combined AHP rating and AHP/DEA-AR method for selecting optimal promising emerging technology. *Expert Systems with Applications* **40**, p. 300-314.

Estimation of anthropometrical and inertial body parameters

J. Bastien*, Y. Blache* and K. Monteil*

*EA 647, Centre de Recherche et d'Innovation sur le Sport, Université Claude Bernard-Lyon 1, 27-29, Bd du 11 Novembre 1918, 69622 Villeurbanne Cedex, France; jerome.bastien@univ-lyon1.fr; yoann.blache@univ-lyon1.fr; karine.monteil@univ-lyon1.fr

Abstract. The inertial (IP) and anthropometrical (AP) parameters of human body are mostly estimated from coefficients issue from cadaver measurements. These parameters could involve errors in the calculation of joint torques during explosive movements. The purpose of this study was to optimize the IP and AP in order to minimize the residual torque and force during squat jumping. Three methods of determination have been presented: method A: optimizing AP and IP of each body part; method B: optimizing trunk AP and IP, assuming that the AP and IP of the lower limbs were known; method C: using Winter AP and IP. For each method, the value (degree 0), the integral (degree 1) and the double integral (degree 2) of the residual moment were also used. The method B with degree 2 was the most accurate to determine trunk AP and IP by minimizing the residual force and torque, by providing a linear least squares system. Instead of minimizing the residual force and torque, by classical way, the double integral of the latter provided more accurate results.

1. Introduction

Joint forces and torques are commonly used in motion analysis for orthopedics, ergonomics or sports science [1, 2]. A standard bottom-up inverse dynamic model is often used to calculate joint forces and torques at the lower extremity. Body anthropometric (AP) and inertial parameters (IP) are needed to apply inverse dynamic model. AP and IP can be obtained from many ways. Cadaver measurements have been the first method applied by researchers and is still commonly used [3–7]. Then, predictive linear or non-linear equation [8–12] and imaging resonance magnetic techniques [13–17] have been elaborated. In Hatze study [18], the author has been interested in the accuracy of the different methods. He observed that the methods using γ ray, X ray, tomography and imaging resonance magnetic techniques presented an average accuracy of 5% with a maximal error of 11%. Linear regression yields AP and IP with an average accuracy of about 24% with a maximal error of 40%. The corresponding values for non-linear regression were 16% and 38% respectively. According to Hatze [18], the most accurate technique would be the anthropometrico-computational method with 1.8% of average accuracy and 3% of maximal error.

Some authors tried to evaluate the influence of error in AP and IP on joint torques during gait analyses. In [19], the authors compared joint torques during walking using cadaver AP and IP [7] versus direct measurements [20]. In [21], 6 methods to calculate AP and IP were compared. Even if these methods provided different AP and IP, no effect on torque measurements were pointed out during walking. Nevertheless, in [21], the authors concluded that changes in AP and IP should have a greater influence on the torque measurements for activities involving greater accelerations. This is the case in squat jumping, where the push-off lasts around 350 ms, and the acceleration of the body center mass could reach 20 m.s^{-2} . Researchers mainly use cadaver measurements [22–27] or predictive equations [28–34] during studies about vertical jumping. However these methods give AP and IP which are not specific to the population studied. As a result, some errors in AP and IP could imply inaccuracy in joint torque calculations.

Besides, to the best of our knowledge, no study has focused on the optimization of AP and IP in explosive movements and especially in 2-D squat jumping investigation. Therefore the purpose of this study was to adjust AP and IP of the human segments during squat jumping in order to minimize error in joint torque values. Especially, the optimization will focus on the "head arm trunk" segment (HAT).

2. Methods

2.1 Experimental acquisition

Twelve healthy athletic male adults (mean \pm SD: age, 23.2 ± 3.6 years; height, 1.75 ± 0.06 m; mass, 69.1 ± 8.2 kg) volunteered to participate in the study and provided informed consent. Prior to the experimental protocol, reflective landmarks were located on the right 5-th metatarsophalangeal, lateral malleolus, lateral femoral epicondyle, greater trochanter and acromion. Thereafter, the subjects performed at most ten maximal squat jumps. In order to avoid the contribution of the arms in vertical jump height [29, 31], the subjects were instructed to keep their hands on their hip throughout the jump.

All jumps were performed on an AMTI force plate model OR6-7-2000 sampled at 1000 Hz. Countermovement defined as a decrease of vertical ground reaction force (R_y) before the push-off phase was not allowed.

The beginning of the push-off was considered as the instant when the derivative of the smoothed R_y is different to zero. Simultaneously, the subjects were filmed in the sagittal plane with a 100 Hz camcorder (Ueye, IDS UI-2220SE-M-GL). The optical axis of the camcorder was perpendicular to the plane of the motion and located at 4 meters from the subject.

Jumps recorded were digitalized frame by frame with the Loco® software (Paris, France). A four rigid segments model composed of the foot (left and right feet together), the shank (left and right shanks together), the thigh (left and right thighs together) and the HAT (head, arms and trunk) was used. Squat jump being a symmetrical motion, the lower limb segments were laterally combined together and it was supposed that the left and right sides participate equivalently to the inter-articular efforts. The position of the upper limbs is fixed to limit their influence on I_4 . Moreover the objective is to provide a robust estimate of the I_4 according to a given protocol and to observe that it is different from that of Winter.

Data obtained from the video and the force platform were synchronized and then smoothed in order to be derivated one or twice.

2.2 Calculation of residual error

The dynamics equations applied to each of the segments $[A_j A_{j+1}]$, for $j \in \{1, \dots, q-1\}$, give

$$\vec{R}_j - \vec{R}_{j+1} = -m_j \vec{g} + m_j \frac{d^2 \overrightarrow{OG}_j}{dt^2} \quad (1a)$$

$$-M_j + I_j \ddot{\theta}_j = C_j - C_{j+1} \quad (1b)$$

where

$$M_j = -(x_{j+1} - x_j)(\alpha_j R_{y,j} + (1 - \alpha_j) R_{y,j+1}) + (y_{j+1} - y_j)(\alpha_j R_{x,j} + (1 - \alpha_j) R_{x,j+1}) \quad (2)$$

With boundary condition

$$\vec{R}_1 = \vec{R}, \quad \vec{R}_p = \vec{0} \quad (3)$$

$$C_1 = C, \quad C_q = 0 \quad (4)$$

We obtain classically for all $k \in \{1, \dots, q-1\}$,

$$\vec{R}_k = \vec{R} - \sum_{j=1}^{k-1} m_j \left(\frac{d^2 \overrightarrow{OG}_j}{dt^2} - \vec{g} \right) \quad (5a)$$

$$C_k = C + \sum_{j=1}^{k-1} (M_j - I_j \ddot{\theta}_j) \quad (5b)$$

and

$$C = - \sum_{j=1}^{q-1} M_j - \sum_{j=1}^{q-1} I_j \ddot{\theta}_j \quad (5c)$$

The residual torque is defined by

$$\tilde{C} = C + \sum_{j=1}^{q-1} M_j - \sum_{j=1}^{q-1} I_j \ddot{\theta}_j \quad (6)$$

Where angles θ_j are determined from the smoothed displacements, M_j are defined by (2) and joint forces $R_{x,j}$ and $R_{y,j}$ are calculated by using (5a).

We now explain how to determine I_1, I_2, I_3 and I_4 . The residual torque is defined by (6) or by the following equation:

$$\tilde{C}^{(0)}(t) = C_{exp} - C_{angl} \quad (7a)$$

Where C_{exp} is torque measured experimentally and

$$C_{angl} = - \sum_{j=1}^{q-1} M_j + \sum_{j=1}^{q-1} I_j \ddot{\theta}_j \quad (7b)$$

is defined according to moments M_j and the double derivatives $\ddot{\theta}_j$. $X^{(0)}$ corresponds to the values of function X . The impulsion phase is equal to $[t_0, t_f]$. By integration, between the beginning t_0 and t_i , and since the angular velocities are null at onset of push-off, we obtain

$$\tilde{C}^{(1)}(t_i) = C_{exp}^{(1)}(t_i) - C_{angl}^{(1)}(t_i) \quad (8a)$$

$$C_{exp}^{(1)}(t_i) = \int_{t_0}^{t_i} C_{exp}(s) ds \quad (8b)$$

$$C_{angl}^{(1)}(t_i) = - \sum_{j=1}^{q-1} \int_{t_0}^{t_i} M_j(s) ds + \sum_{j=1}^{q-1} I_j \dot{\theta}_j(t_i) \quad (8c)$$

Where s is the variable of integration. $X^{(1)}$ corresponds to the first order integration of the function X . After a second integration we obtain:

$$\tilde{C}^{(2)}(t_i) = C_{exp}^{(2)}(t_i) - C_{angl}^{(2)}(t_i) \quad (9a)$$

$$C_{exp}^{(2)}(t_i) = \int_{t_0}^{t_i} \int_{t_0}^u C_{exp}(s) ds du \quad (9b)$$

$$C_{angl}^{(2)}(t_i) = - \sum_{j=1}^{q-1} \int_{t_0}^{t_i} \int_0^u M_j(s) ds du + \sum_{j=1}^{q-1} I_j (\theta_j(t_i) - \theta_j(t_0)) \quad (9c)$$

Where u is the second variable of integration. $X^{(2)}$ corresponds to the second order integration of the function X . In order to compare the residual values $\tilde{C}^{(0)}$, $\tilde{C}^{(1)}$ and $\tilde{C}^{(2)}$ obtained with different methods, it is necessary to normalize these values by considering the dimensionless quantity defined by

$$\varepsilon^{(j)} = \frac{\|C_{exp}^{(j)} - C_{angl}^{(j)}\|}{\|C_{exp}^{(j)}\| + \|C_{angl}^{(j)}\|} \in [0,1] \quad (10)$$

Where $\|\cdot\|$ is the l² norm

2.3 Methods of determination IP and AP

2.3.1 Method A: optimization of all inertia I_1, I_2, I_3 and I_4

Considering that the residual is null, (7), (8), and (9) become

$$\sum_{j=1}^{q-1} I_j \ddot{\theta}_j(t_i) = C(t_i) + \sum_{j=1}^{q-1} M_j(t_i) \quad (11a)$$

or

$$\sum_{j=1}^{q-1} I_j \dot{\theta}_j(t_i) = \int_{t_0}^{t_i} \left(C(s) + \sum_{j=1}^{q-1} M_j(s) \right) ds \quad (11b)$$

or

$$\sum_{j=1}^{q-1} I_j (\theta_j(t_i) - \theta_j(t_0)) = \int_{t_0}^{t_i} \int_{t_0}^u \left(C(s) + \sum_{j=1}^{q-1} M_j(s) \right) ds du \quad (11c)$$

As the method used in Section 2.2 to determine α_4 , for Eq. (9), the double derivative of angles is not used for (11c), but only values of these angles.

Each equation (11) is equivalent to determine I_1, I_2, I_3 and I_4 such that

$$\forall i, \sum_{j=1}^{q-1} A_{i,j} I_j = B_i \quad (12)$$

where $A_{i,j}$ and B_i are known. These equations are equivalent to the overdetermined linear system

$$AI = B, \text{ where } I = \begin{pmatrix} I_1 \\ I_2 \\ I_3 \\ I_4 \end{pmatrix} \quad (13)$$

which has no solution in the general case, but has a least square sens solution. In this case, the number $j \in \{0, 1, 2\}$ is called the degree of the method A; the number $\varepsilon^{(j)}$, defined by (10) is denoted $\varepsilon_A^{(j)}$ and the coefficient of multiple determination for the overdetermined system (13) is denoted $R_A^{2(j)}$.

2.3.2 Method B: optimization of inertia I4 only

It can be assumed that I_1, I_2 and I_3 are determined in [11]. Then (7), (8), or (9) can be written under the following form: for all i ,

$$I_q - 1\ddot{\theta}_{q-1}(t_i) = - \sum_{j=1}^{q-2} I_j \ddot{\theta}_j(t_i) + C(t_i) + \sum_{j=1}^{q-1} M_j(t_i) \quad (14a)$$

$$I_q - 1\dot{\theta}_{q-1}(t_i) = - \sum_{j=1}^{q-2} I_j \dot{\theta}_j(t_i) + \int_{t_0}^{t_i} \left(C(s) + \sum_{j=1}^{q-1} M_j(s) \right) ds \quad (14b)$$

or

$$I_{q-1}(\theta(t_i) - \theta_{q-1}(t_0)) = - \sum_{j=1}^{q-2} I_j (\theta_j(t_i) - \theta_j(t_0)) + \int_{t_0}^{t_i} \int_{t_0}^u \left(C(s) + \sum_{j=1}^{q-1} M_j(s) \right) ds du \quad (14c)$$

Here, it is also equivalent to find I_4 such that

$$\forall i, y_i = I_4 x_i. \quad (15)$$

As previously, we consider $\varepsilon_B^{(j)}$ and $R_B^{2(j)}$.

2.2.3 Method C: values of inertia I1, I2, I3 and I4 according to Winter coefficients

The values of I_1, I_2, I_3 and I_4 are estimated from [11]. As previously, we consider $\varepsilon_C^{(j)}$ and $R_C^{2(j)}$. This method is not an optimization method and $R_C^{2(j)}$ is formally defined; this number is not necessarily positive.

To summarize, we have three methods defined by $X \in \{A, B, C\}$ and for each of them the order j belongs to $j \in \{0, 1, 2\}$. The method X with degree j is called method « Xj ». For example « $A2$ » is the method A with degree 2. For each of these three methods and for each degree j are defined $\varepsilon_C^{(j)}$ and $R_C^{2(j)}$. An accurate method corresponds to ε close to 0 and R^2 close to 1.

2.4 Statistics

Main effects of the three methods and the three degrees based on "residual error" were tested to significance with a general linear model one way ANOVA for repeated measures. When a significant F value was found,

post-hoc Tukey tests were applied to establish difference between methods (significant level $p < 0.05$). All analyses were proceeding through the R software [46].

3. Results

The results compared the methods defined by $X \in \{A, B, C\}$ and degree $j \in \{0, 1, 2\}$ and called “ Xj ”. Values of $\text{Log}_{10}(\varepsilon)$ and $\text{Log}_{10}(1-R^2)$ are given in tables 1 and 2.

Table 1. Groups statistics of $\text{Log}_{10}(\varepsilon)$ for the three studied methods with the three degrees: mean \pm standard deviation

degree j	method A	method B	method C
0	-0.46 ± 0.16	-0.35 ± 0.14	-0.28 ± 0.1
1	-1.06 ± 0.29	-0.59 ± 0.3	-0.37 ± 0.25
2	-1.83 ± 0.44	-1.01 ± 0.38	-0.49 ± 0.33

Table 2. Groups statistics of $\text{Log}_{10}(1-R^2)$ for the three studied methods with the three degrees: mean \pm standard deviation

degree j	method A	method B	method C
0	-0.27 ± 0.28	-0.05 ± 0.23	0.11 ± 0.24
1	-1.47 ± 0.48	-0.52 ± 0.31	0.03 ± 0.54
2	-2.99 ± 0.73	-1.37 ± 0.56	-0.2 ± 0.82

First of all, taken into consideration the $\text{Log}_{10}(\varepsilon)$ and the $\text{Log}_{10}(1-R^2)$, the lowest the value, the more accurate the method or the degree of integration.

The general linear model one way ANOVA for repeated measures pointed out significant differences between the three methods (A, B and C) and the three degrees (0, 1 and 2). The post-hoc Tukey tests indicated that for the methods A and B, the values decreased when the degree increased (degree 2 < degree 1 < degree 0). Concerning the method C, the results were lower for degree 2 than degree 1 and no significant difference was observed between the degree 1 and degree 0.

Comparing the methods, for the degrees 1 and 2, for the both values of $\text{Log}_{10}(\varepsilon)$ and $\text{Log}_{10}(1-R^2)$, the values were the lowest for method A, then B, then C (A < B < C). With regard to the degree 0, the method A was significantly lower than method B and no difference was observed between the other methods.

Finally, when all methods and degrees were compared together, the lowest values of $\text{Log}_{10}(\varepsilon)$ and $\text{Log}_{10}(1-R^2)$, were observed for method A2. The later was significantly lower than method A1 and method B2. By comparing A1 and B2, we obtain $p = 0.2719$. The latter were significantly lower than the method B1. Moreover, if we compare B1 to C2 and C2 to C0, we obtain a significant difference. This can be summarized under the following form:

$$A2 < B2 = A1 < B1 < C2 < C0$$

As the results of method A gave unphysical values for inertia, the most accurate method was the method B2. C0 and C2 are the methods with torque value or double integration of torque corresponding to Winter’s data respectively. A0 and B0 are optimization methods on all IP or only trunk IP with Winter’s data.

4. Discussion

The purpose of this study was to adjust AP and IP of the human segments during squat jumping in order to minimize error in joint torque. The results indicated that the method A2 minimized the most the residual torque

(ie. ε and 1-R2) following by the methods B2 and A1 being more accurate than the method B1. Nevertheless, the method A yields unrealistic I_J, therefore the most accurate method retained was the method B2. Consequently, the optimization focused especially on the HAT inertial and anthropometric parameters. It seems to be possible to optimize AP and IP of one segment when the others are known, but the simultaneous optimization of three AP and IP segments seems to be difficult. According to [36] IP and AP optimized for three segments cannot be considered as true, while if two of three segments are known, the IP and AP of the last segment can be calculated.

The IP and AP found with the method B2 are close to the Winter ones but gave better residual joint torque. These differences could be obviously explained by the different position between the subjects performing squat jumps and cadavers. Especially the position of the arms was different, influencing the IP and AP of the HAT segment. [35, 36] used optimization techniques to solve inverse dynamic by determining numerical angles which minimize the difference between the ground reaction force measured and the ground reaction force calculated. They found segmental angles which minimize an objective function under equality and inequality constraints, by taking into account the difference ground reaction force measured and the ground reaction force calculated. From these angles, joint torques are determined. Our approach is different: techniques of [35, 36] to determine angle and torque were not applied. Only experimental displacements smoothed were considered. Then, with a direct inverse method, joint forces and torques were deduced. Finally, an optimization is made on the residual torque and force to determine values of AP and IP. This optimization is very simple and fast, since it is based on the least square linear method. It can be noticed that even if the residual torque or force are minimized, the error at each joint may be increased [35, 36, 37]. However, in our study and for the method B, we only optimized AP and IP of the trunk, consequently the joint torque and forces at the hip, knee and ankle joints remained unchanged. The major difference with classical way is the following point: it is possible to minimize residual torque, or its integral or double integral. The best method corresponds to the minimization of the double integral, which do not use the double derivative of angle.

5. Conclusion

The optimization of inertial and anthropometric parameters seems to be necessary when researchers use inverse dynamic methods. Indeed, cadaver data lead to errors in the calculi of the joint torques, especially in dynamic motions with greater acceleration. These ones could be reduced by optimization methods. The present method of optimization, based on the double integration of residual, has been applied on the HAT segment but could also be applied on more segments.

References

- [1] Pearsall, D. J. and Reid, J. G. (1994) The study of human body segment parameters in biomechanics. An historical review and current status report. *Sports Med* **18**, 126-140.
- [2] Reid, J. G. and Jensen, R. K. (1990) Human body segment inertia parameters: a survey and status report. *Exerc Sport Sci Rev* **18**, 225-241.
- [3] Clauser, C., McConville, I., and Young, I. (1969) Weight, volume and center of mass of segments of the human body. Wright-Paterson A.F.B., Ohio.
- [4] Fujikawa, K. (1963) The Center of Gravity in the parts of Human Body. *Okajimas Folia Anat Jpn* **39**, 117-125.
- [5] Chandler, R., Snow, R., and Young, J. (1978) Computation of mass distribution characteristics of children. B. Soc. Of Photo-Optical Instrumentation Engineers, Washington.
- [6] Hinrichs, R. N. (1990) Adjustements to the segment centre of mass proportions of Clauser et al. 1969. *J Biomech* **23**, 949-951.
- [7] Dempster, W. (1955) Space requirements of the seated operator. Techn. Report WADC-TR-55-159. Ohio (WADC-TR- 55-159): Wright-Paterson AFB.
- [8] Hinrichs, R. N. (1985) Regression equations to predict segmental moments of inertia from anthropometric

- measurements: an extension of the data of Chandler et al. 1975. *Biomechanics* **19**, 621-624.
- [9] Zatsiorsky, V., and Seluyanov, V. (1983). The mass and inertia characteristics of the main segments of the human body. Champaign: Humain Kinetics.
- [10] McConville, J. T., Churchill, T. D., Kaleps, I., Clauser, C. E., and Cuzzi, J. (1980) Anthropometric relationships of body and body segment moments of inertia. Tech. Rep. AFAMRL-TR-80-119, Aerospace Medical Research Laboratory, Dayton, Ohio: Wright-Patterson Air Force Base.
- [11] Winter, D. (2009) Biomechanics and Motor Control of Human Movement, fourth ed. John Wiley and Sons, New York.
- [12] Yeadon, M. R., and Morlock, M. (1989) The appropriate use of regression equations for the estimation of segmental inertia parameters. *J Biomech* **22**, 683–689.
- [13] Durkin, J. (1998) The prediction of body segment parameters using geometric modelling and dual photon absorptiometry. McMaster University: Hamilton, Ontario.
- [14] Huang, H. K. and Suarez, F. R. (1983) Evaluation of cross-sectional geometry and mass density distributions of humans and laboratory animals using computerized tomography. *J Biomech* **16**, 821-832.
- [15] Martin, P. E., Mungiole, M., Marzke, M. W. and Longhill, J. M. (1989) The use of magnetic resonance imaging for measuring segment inertial properties. *J Biomech* **22**, 367-376.
- [16] Pearsall, D. J., Reid, J. G. and Livingston, L. A. (1996) Segmental inertial parameters of the human trunk as determined from computed tomography. *Ann Biomed Eng* **24**, 198-210.
- [17] Cheng, C. K., Chen, H. H., Chen, C. S., Chen, C. L. and Chen, C. Y. (2000) Segment inertial properties of Chinese adults determined from magnetic resonance imaging. *Clin Biomech (Bristol, Avon)* **15**, 559-566.
- [18] Hatze, H. (2002) Anthropomorphic contour approximation for use in inertial limb parameter computation. In 12th international conference on mechanics in medicine and biology.
- [19] Goldberg, E. J., Requejob, P. S. and Fowler, R. G. (2008) The effect of direct measurement versus cadaver estimates of anthropometry in the calculation of joint moments during above-knee prosthetic gait in pediatrics. *J Biomech* **41**, 695–700.
- [20] Fowler, E. G., Hester, D. M., Oppenheim, W. L., Setoguchi, Y. and Zernicke, R. F. (1999) Contrasts in gait mechanics of individuals with proximal femoral focal deficiency: Syme amputation versus Van Nes rotational osteotomy. *J Pediatr Orthop* **19**, 720–731.
- [21] Pearsall, D. J. and Costigan, P. A. (1999) The effect of segment parameter error on gait analysis results. *Gait Posture* **9**, 173-183.
- [22] Lees, A., Rojas, J., Ceperos, M., Soto, V. and Gutierrez, M. (2000) How the free limbs are used by elite high jumpers in generating vertical velocity. *Ergonomics* **43**, 1622-1636.
- [23] Lees, A., Vanrenterghem, J. and De Clercq, D. (2004) The maximal and submaximal vertical jump: implications for strength and conditioning. *J Strength Cond Res* **18**, 787-791.
- [24] Bobbert, M. F., de Graaf, W. W., Jonk, J. N. and Casius, L. J. (2006) Explanation of the bilateral deficit in human vertical squat jumping. *J Appl Physiol* **100**, 493-499.
- [25] Bobbert, M. F., Casius, L. J., Sijpkens, I. W. and Jaspers, R. T. (2008) Humans adjust control to initial squat depth in vertical squat jumping. *J Appl Physiol* **105**, 1428-1440.
- [26] Laffaye, G., Bardy, B. G. and Durey, A. (2005) Leg stiffness and expertise in men jumping. *Med Sci Sports Exerc* **37**, 536-543.
- [27] Laffaye, G., Bardy, B. G. and Durey, A. (2007) Principal component structure and sport-specific differences in the running one-leg vertical jump. *Int J Sports Med* **28**, 420-425.
- [28] Domire, Z. J. and Challis, J. H. (2007) The influence of squat depth on maximal vertical jump performance. *J Sports Sci* **25**, 193-200.
- [29] Domire, Z. J. and Challis, J. H. (2010) An induced energy analysis to determine the mechanism for performance enhancement as a result of arm swing during jumping. *Sports Biomech* **9**, 38-46.
- [30] Cheng, K. B. (2008) The relationship between joint strength and standing vertical jump performance. *J Appl Biomech* **24**, 224-233.

- [31] Hara, M., Shibayama, A., Arakawa, H. and Fukashiro, S. (2008) Effect of arm swing direction on forward and backward jump performance. *J Biomech* **41**, 2806-2815.
- [32] Wilson, C., Yeadon, M. R. and King, M. A. (2007) Considerations that affect optimised simulation in a running jump for height. *J Biomech* **40**, 3155-3161.
- [33] Haguenauer, M., Legreneur, P. and Monteil, K. M. (2006) Influence of figure skating skates on vertical jumping performance. *J Biomech* **39**, 699-707.
- [34] Vanrenterghem, J., Lees, A., Lenoir, M., Aerts, P. and De Clercq, D. (2004) Performing the vertical jump: movement adaptations for submaximal jumping. *Hum Mov Sci* **22**, 713-727.
- [35] Riemer, R. and Hsiao-Wecksler, E. T. (2008) Improving joint torque calculations: Optimization-based inverse dynamics to reduce the effect of motion errors. *J Biomech* **41**, 1503–1509.
- [36] Riemer, R. and Hsiao-Wecksler, E. T. (2009) Improving Net Joint Torque Calculations Through a Two-Step Optimization Method for Estimating Body Segment Parameters. *J Biomech Eng.* **131**, 011007
- [37] Kuo, A.D. (1998) A Least-Squares Estimation Approach to Improving the Precision of Inverse Dynamics Computations. *ASME J Biomech. Eng* **120**, 148–159.

Optimization modeling for analyzing fantasy sport games

J. Beliën***, D. Goossens*** and D. Van Reeth*

*Hogeschool-Universiteit Brussel, Warmoesberg 26, 1000 Brussels, Belgium, jeroen.beliën@hubrussel.be, daam.vanreeth@hubrussel.be

**Center for Operations Management, KU Leuven, Naamsestraat 69, 3000 Leuven, Belgium

*** ORSTAT, KU Leuven, dries.goossens@kuleuven.be

Abstract. In a fantasy sport game, participants act like a team manager building a team of real individual players of a professional sport. The real performances of these players (or their teams) are translated into points for their team managers. The managers' aim is to collect as many points as possible thereby defeating the fantasy teams of opponents. First, we discuss a number of common characteristics of fantasy sport games. Based on these common characteristics, we present a mixed integer programming model that can be used to analyze the ex-post results of fantasy sport games. We discuss how these results can create value for both the game organizer and the participants. Finally, we apply our model in practice on a fantasy cycling game.

1. Introduction

A fantasy sport game allows ordinary people to act like a team manager building a team of real individual professional sport athletes. The real-world performances of these athletes (or their teams) are translated into points for their team managers. The managers' aim is to collect as many points as possible thereby defeating the fantasy teams of opponents. In the remainder of this paper we will use the word 'participant' to describe a virtual team manager while the word 'player' solely refers to a real-world athlete.

Fantasy sports found their origin in the United States in the 1980s. American journalists Glen Waggoner & Daniel Okrent developed a game in which a handful of participants would draft from a pool of active baseball players (Davis & Duncan, 2006). From the 1990s on, the Internet made game results more accessible, and virtual leagues easier to manage. But fantasy sports really took off around the turn of the millennium, when the internet transformed fantasy sports into a mainstream phenomenon. Fantasy sports are now being played by tens of millions of people worldwide. Fantasy American football has become the most popular fantasy sport league, with a market share of 80% in the United States and Canada (<http://www.fantasysportsadnetwork.com>). In Europe, soccer is the most popular fantasy sport subject. The enormous success of fantasy sports has a lot to do with the fact that it allows online participants to assume an active role of a team owner or a team manager in a sport they are heavily interested in, thereby intensifying the way live sport is consumed. As a result, many sports enthusiasts are now obtaining their sports entertainment through fantasy sports rather than solely by watching games on television (Nesbit & King, 2011).

This paper presents a mixed integer programming model for finding an optimal set of decisions in fantasy sport games under the assumption that all data are known beforehand. Thus, the model allows to identify ex post an optimal team selection and player transfers. One could ask why an ex-post analysis for fantasy sports might be useful. After all, participants need to select their team and make transfers before this information becomes available. To describe the value of this ex-post analysis we distinguish between the *information value* and the *commercial value* of the results.

First, our model can be used to create information value. Fantasy sports are often organized by media groups whose core business is to provide information and sell news. On a regular basis they publish the fantasy sport game results and rankings through a website or newspaper. The related articles are often limited to an interview with the week winner and an overview of the top-x best performing teams of the specific week, and the updated general classification. Many readers, however, would be interested in more detailed analyses, providing answers to questions like "what is the optimal team of this week?", "what is the optimal team so far (with respect to different objectives such maximal price money, maximal number of week victories, etc.)?", "what were the optimal team changes for the most recent transfer period?", etc. These kinds of questions can easily be answered by the ex-post optimization model.

Moreover, the model can also be used to provide individual feedback to participants (e.g., through a personal website account) that specifically concerns the participant's team. For instance, given the participant's starting team, which were the optimal transfers with respect to a particular game objective? What was the highest ranking that could be achieved (at this moment)? Our model can thus be used to generate information that goes much further than only providing the participant's current ranking. We are convinced that the availability of this detailed information can be a distinguishing factor for choosing between competing fantasy sports.

Second, the results of applying our model can also have commercial value. It is our experience that an optimal strategy largely outperforms that of the best participant. Having this in mind, the last season's optimal team can be used for the promotion campaign of the new season highlighting the (hopefully large) difference with the score of the best participant, who after all, has won the competition. By comparing these scores, people tend to overestimate their chances of winning the game. Consequently, the game can attract more participants which increases the turnover for the organizer.

A second example of how our model can create business value is by using the results to stimulate people to join the game even after it has already been started. Indeed, most fantasy sport games allow people to submit a (new) team at any time during the game (as long as they pay the participation fee). However, these late participants have to start with a handicap as the other participants have already had opportunities to collect points. That is where our model comes into play. By highlighting the difference between the last season's optimal score (calculated by our model) and the score of the last season's winner, one can possibly convince doubting people of still having a fair chance of winning the general classification. This will certainly be the case if this difference is (substantially) higher than the total score of the current leader. After all, new participants also have an important advantage as they face less uncertainty concerning, for instance, the players' expected performance. Moreover, the ex-post optimal team often looks so familiar that people get the (false) impression that this optimal team could easily be predicted beforehand or that it was, at least, not that difficult to compose a team that would have beaten the current round (e.g., race or matchday) winner. Consequently, by regularly publishing the optimal team for every round and the difference with the round winner, people will be continuously stimulated to submit a new team that aims at winning a specific round.

The remainder of this paper is organized as follows. Section 2 provides a literature review and a number of common characteristics of fantasy sports. Based on these common characteristics, Section 3 presents a mixed integer programming (MIP) model that can be used to analyze the ex-post results of fantasy sport games. We discuss how we applied our model on a real-life fantasy cycling game in Section 4. Finally, Section 5 concludes this paper and lists directions for future research.

2. Fantasy sport games: an overview

2.1 Literature review

The growth of fantasy sports has made it an important part of the sports industry. The booming popularity also stimulated research on fantasy sports and from 2005 on literature on fantasy sports really started to grow. Two directions of research are currently dominant. The first line of research is economically oriented. It sees fantasy sports as a new form of sport consumption and studies how this affects the behavior of sports fans. Randle & Nyland (2008), Drayer et al. (2010) and Karg & McDonald (2011) all take a rather global view and look at the impact of fantasy sport participation on the various media sources fantasy sport participants use. The specific impact of fantasy sport on television ratings is analyzed by Nesbit & King (2010), Dwyer (2011a) and Fortunato (2011), while the live attendance impact is taken up by Nesbit & King (2011). Most of these studies basically conclude that instead of competing with traditional ways of sport consumption, fantasy sport appears to be a complementary and value-adding activity (Dwyer, 2011a). Fan loyalty and how fantasy sports can be used in customer relationship management are studied by Dwyer (2011b) and Smith, Synowka & Smith (2010).

A second line of research analyses sociologically oriented aspects of fantasy sport participation. Participants in fantasy sports are predominantly young, white, male and well-educated. For instance, Davis & Duncan (2006, p. 247) state that 97.9% of the fantasy sports enthusiasts are male, 93.7% are white, and 68.7% are

college graduates or postgraduates. The Fantasy Sports Ad Network (www.fantasysportsadnetwork.com) comes to a similar conclusion. They claim that 92% of fantasy participants are male, 91% are white and 71% have a bachelor's degree or higher. According to them, the average age of a fantasy sports participant is 36.

Some of the studies focus on the reasons why sports fans participate in fantasy sports. Often cited and validated motivational dimensions are fanship, entertainment/escape, competition and social interaction (see, for instance, Farquhar & Meeds, 2007; Dwyer & Kim, 2011; Ruihley & Hardin, 2011; Lee, Seo & Green, 2012, Billings & Ruihley, 2013). Although the conclusions on the importance of some of these dimensions might differ from one study to another, the results generally suggest a pattern of fantasy sport participation that is more purposeful and active than traditional media use (Dwyer & Kim (2011), p. 70). Other studies analyzed gender-related aspects of fantasy sport participation. The findings of these studies are diffuse. While Davis & Duncan (2011, p. 244) conclude that "*fantasy sports reinforce hegemonic ideologies in sport spectatorship, emphasizing authority, sports knowledge, competition, male-bonding, and traditional gender role*", Ruihley & Billings (2012, p. 16) find that "*fantasy sport seems to be filling a specific need for women participants. (...) Motivational research indicates that there are many reasons they are drawn to it - if only they just enjoyed it to the degree that the men do.*"

2.2 Common characteristics

To our knowledge, there are no overview studies that list fantasy sports or classify them according to a certain set of characteristics. Such a work would, in fact, be virtually impossible since the number of fantasy sports is extremely large, even if one would focus on a small country like Belgium or on one single sport. Based on observation, we therefore looked at a number of characteristics of fantasy sports we think are fairly common in fantasy sports (see Table 1). We make a distinction between organizational characteristics and game rules characteristics. Fantasy sports can either be organized by professional organizations or by amateurs. Professional organizations are usually media related, as can be seen from the list of members of the professional Fantasy Sports Trade Organization (<http://fst.org/about/members>). With the organization of fantasy sports, media managers try to build customer loyalty and an increased use of their media platform. For instance in Belgium, the popular Belgian newspapers Het Laatste Nieuws and Het Nieuwsblad were the first to organize fantasy sports in the mid 1990s as a marketing instrument to boost sales. In the case where amateurs organize a fantasy sport, it is usually amongst friends or colleagues. In such cases, participation is often free or at a very low price to cover the costs and there is usually a restriction on the number of participants in the game, especially when such amateur organized games start to create a broader appeal. Fantasy sports organized by media companies normally do not have a limit on the number of participants and a participation fee is often required as part of the commercial exploitation of the game.

More relevant to our analysis are the game rules characteristics. We distinguish between several elements. First, team selection in fantasy sport can either be subject to some sort of budget constraint or can be free from such a constraint. Constraints are based on a value that is given to each player that can be selected. This value is based on past performances of the player and can either be a fictitious monetary value (e.g., based on his salary or transfer price) or a sports score (e.g., ATP points in tennis). Most games usually have a budget constraint to make it impossible to select only the best, (usually high-valued) players. The remaining budget, that is the budget not spent on players, is often used as a tiebreaker in case several participants end up with the same score.

Second, in some fantasy sports players can only be selected once. If one participant picks a given player for his fantasy team, the player becomes unavailable to the other participants. Of course, this reduces significantly the number of people that can participate in the game. Therefore, in most fantasy sports players can be selected infinitely, this means by every participant independent of the decisions of the other participants. In this way, all participants have equal chances of winning and the game organizer can allow a much larger number of participants.

Third, some fantasy sports allow participants to change their fantasy team selection at well-defined moments during the season while in other fantasy sports the team selection is fixed from the start for the whole of the season. Of course, when it is possible to transfer players during the season, the game becomes

much more strategic thus creating an even more intense need and desire to follow the sport closely, which is to the benefit of the media company that is behind the game.

Fourth, while in many fantasy sports all players of a fantasy team earn points for the participant, there are also fantasy sports in which this is not the case. For instance, some soccer fantasy sports allow participants to select a basic team of 11 players and a number of substitutes. Only in case a basic player does not play on a certain matchday, the substitute can earn points for the participant. Another possibility is that only the x players with the highest score of a fantasy team earn points for the participant on each matchday or race.

Fifth, some fantasy sports only have prizes at the end of the competition while others also have day prize winners during the competition. With regular prizes throughout the season, sport loyalty and game affiliation are further increased, again very much to the benefit of the media that are behind the game. Moreover, this offers an extra incentive for participants to enter the fantasy game even after the start of the competition.

Table 1: Characteristics of fantasy sports

Characteristic	Description	Gigabike
Organizational characteristics		
Organizers	Professional (usually media related) OR Amateur (usually friends or colleagues)	Amateur
Participation in the game	Limited OR Unlimited	Limited to 500
Participation cost	A participation fee has to be paid OR Free participation	Free participation
Game rules characteristics		
Budget	Constrained OR Unconstrained	Constrained
	Remaining budget used as a tiebreaker OR Remaining budget irrelevant	Remaining budget used as a tiebreaker
Player distribution	A player can only be part of one fantasy team OR A player can be part of many fantasy teams	A cyclist can be part of many teams
Fantasy team composition	Static throughout the season OR Dynamic with a number of transfer opportunities	Dynamic with 5 transfer opportunities
Calculation of points	All players of a team earn points OR Only a subset of the players of a team earn points	Only the 8 best players of a team earn points
Prizes	Prize(s) at the end of the competition only OR A combination of prizes at the end and daily prizes	Global and daily non-monetary prizes

3. Optimization model

This section presents an MIP model for finding optimal ex-post decisions in fantasy sport games (optimal start selection as well as optimal transfers). In other words, the model allows to identify an optimal strategy if one knows beforehand how many points each player will collect in each game.

We define the following sets and indices:

$p, p' \in P = \{1, 2, \dots, P \}$	Players (athletes, riders, etc.),
$\pi \in \Pi = \{1, 2, \dots, \Pi \}$	Player types (e.g., defenders, attackers, midfielders in football; leaders, domestiques in cycling, etc.),
$P_\pi \subseteq P$	Set of players of type π
$t \in T = \{1, 2, \dots, T \}$	Periods. Between each pair of periods the fantasy team can be modified,
$\tau \in \Theta = \{1, 2, \dots, \Theta \}$	Games, i.e., events where players can gain points (matches, races, etc.),
$G \subseteq \Theta$	Sets of games that are considered as a whole for assigning prices, e.g., matchdays, months, cycling tours, etc.,
$G_t \subseteq \Theta$	Game sets that fall into period t ,
$e \in E_G = \{1, 2, \dots, E_G \}$	Winning ranks that can be obtained in game set G (e.g., first, second, third, etc.).

The model parameters (coefficients, right-hand side constants) are as follows:

v_{pt}	Points obtained by player p in game τ ,
c_{pt}	Cost of player p in period t ,
B	Available budget,
$N_{\pi t}$	Number of players of type π that have to be selected in period t ,
A_t	Maximal number of transfers allowed in period t ,
D_τ	Maximal number of players that can earn points for the team in game τ ,
a_{pt}	= 1 if player p played in game τ ; 0 otherwise,
ω_{eG}	Value (e.g., price money) of obtaining rank e in game set G ,
H_{eG}	Score of the opponent that obtained rank e for game set G ,
R_{eG}	Remaining budget of the opponent that obtained rank e for game set G ,
Q_π	Maximal number of players of player type π in the player pool,
ε	Very small number.

The decision variables are as follows:

x_{pt}	= 1 if player p is in the team in period t ; 0 otherwise,
y_{pt}	= 1 if player p earns points for the team in game τ ; 0 otherwise,
r_t	Remaining budget in period t ,
z_{pt}	= 1 if player p is transferred into the team in period t ; 0 otherwise,
s_{pt}	= 1 if player p is a substitute in period t ; 0 otherwise,
w_{eG}	= 1 if the team obtains rank e for game set G ; 0 otherwise,
q_p	= 1 if player p is included in the player pool.

The optimization model can then be formulated as follows.

$$\text{Maximize} \quad \sum_{p \in P} \sum_{\tau \in \Theta} v_{pt} y_{pt} + \varepsilon r_{|T|} \quad (1)$$

$$\text{st} \quad \sum_{p \in P_\pi} x_{pt} = N_{\pi t} \quad \forall \pi, \forall t \quad (2)$$

$$\sum_{p \in P} c_{p1} x_{p1} + r_1 = B \quad (3)$$

$$\sum_{p \in P} \sum_{t \in T} c_{pt} (x_{pt} - x_{p,t-1}) = r_{t-1} - r_t \quad \forall t > 1 \quad (4a)$$

$$z_{pt} \geq x_{pt} - x_{p,t-1} \quad \forall t > 1 \quad (4b)$$

$$\sum_{p \in P} z_{pt} \leq A_t \quad \forall t > 1 \quad (4c)$$

$$y_{pt} \leq x_{pt} \quad \forall p, \forall t, \forall \tau \in T_t \quad (5)$$

$$\sum_{p \in P} y_{pt} \leq D_\tau \quad \forall \tau \quad (6)$$

$$\sum_{p \in P_\pi} s_{pt} = 1 \quad \forall \pi, \forall t \quad (7a)$$

$$s_{pt} \leq x_{pt} \quad \forall p, \forall t \quad (7b)$$

$$y_{pt} \leq (1 - s_{pt}) + \sum_{p' \in P_\pi \setminus p} (1 - a_{p'\tau}) x_{p'\tau} \quad \forall \pi, \forall p \in P_\pi, \forall t, \forall \tau \in T_t \quad (7c)$$

$$x_{pt} \in \{0,1\} \quad \forall p, \forall t \quad (8)$$

$$y_{pt} \in \{0,1\} \quad \forall p, \forall t \quad (9)$$

$$z_{pt} \in \{0,1\} \quad \forall p, \forall t > 1 \quad (10)$$

$$s_{pt} \in \{0,1\} \quad \forall p, \forall t \quad (11)$$

$$r_t \geq 0 \quad \forall t \quad (12)$$

The objective function (1) maximizes the points collected over all games. In case several solutions exist with maximal points, the solution with the highest remaining budget in the final period is selected, which explains the second term in the objective function. Constraint set (2) ensures that in each period the required number of players of each type is selected. Constraint set (3) models the budget restriction for the first period. If it concerns a static fantasy game, constraint (3) suffices.

For a dynamic fantasy game, the remaining budget of a preceding period can often be used as additional budget in the succeeding period. This is modeled by the constraint sets (4a). The cost of the new players in the team, i.e., players for which $(x_{pt} - x_{p,t-1})$ equals 1, must be smaller than the cost players transferred out, i.e., players for which $(x_{pt} - x_{p,t-1})$ equals -1, plus the remaining budget of the preceding period, r_{t-1} . The difference is the remaining budget r_t of period t . In this scenario, the maximal number of transfers between two periods may be limited. This is handled by the constraint sets (4b) and (4c). Constraint set (4b) ensures that the binary variable z_{pt} is forced to 1 if a player is transferred into the team. Constraint set (4c) guarantees that the maximal number of transfers is not exceeded.

Constraint set (5) models the scenario in which not all selected players automatically earn points for the fantasy team. First, in some fantasy sports, only the points of the D_τ best performing players in each game τ are counted. This is modeled by constraint set (6). Second, in order to avoid that participants lose their interest in case of inactivity (e.g., caused by injury) of a selected player, some fantasy sports require participants to select a substitute player for each player type. This substitute only earns points for the team if at least one of the other players of the substitute's type did not come into action. For instance, in some fantasy football games, participants select four basis midfielders and one substitute midfielder. The substitute does not earn points in a particular game unless (at least) one basis player did not play that game, even in the (unlucky) case that the substitute scored more points than the basis player. The constraint sets (7a)-(7c) model the issue of substitute players in fantasy sports. Constraint set (7a) ensures the selection of one substitute player for each player type. Constraint set (7b) guarantees that the substitute player is one of the selected players. Constraint set (7c) ensures that the substitute player can only earn points if at least one of the basis players did not play. Finally, the constraint sets (8)-(12) define the variable domains.

Model (1)-(12) can be used to find the team that would have scored the maximal number of points for every fantasy sport game that we are aware of. Objective function (1) can be easily adapted for evaluating alternative goals. For instance, if one is interested in the team that scored the most points in particular classes of games (e.g., mountain stages in cycling), only the points scored in those games are retained in the

objective function. Some fantasy game variants, however, require an alternative model. They are listed hereafter.

In a first alternative scenario for a dynamic game, the participants make a start selection of players, referred to as the player pool, out of which they select subsets of players for particular (sets of) games. This should better reflect the decisions made by a real team manager, who has to compose an extensive team at the start of the season and makes a team selection for every game. To model this scenario the constraint sets (4a)-(4c) are replaced by the constraint sets (13)-(15). Constraint set (13) ensures that in future periods only players of the player pool can be selected, while constraint set (14) restricts the number of players of each type in the player pool. Constraint set (15) defines the domain of q_p .

$$x_{pt} \leq q_p \quad \forall p \forall t \quad (13)$$

$$\sum_{p \in P_\pi} q_p \leq Q_\pi \quad \forall \pi \quad (14)$$

$$q_p \in \{0,1\} \quad \forall p \quad (15)$$

Second, it often occurs that one wants to know the team that would have won the most prizes (e.g., to find the team that maximizes the price money). To this aim, the objective function is rewritten as (16) and the constraint sets (17)-(19) are added.

$$\text{Maximize} \quad \sum_G \sum_{e \in E_G} \omega_{eG} w_{eG} \quad (16)$$

$$\text{st} \quad \sum_{p \in P} \sum_{\tau \in G_t} v_{p\tau} y_{p\tau} - H_{eG_t} w_{eG_t} + \varepsilon(r_t - R_{eG_t}) \geq 0 \quad \forall t, \forall G_t, \forall e \in E_{G_t} \quad (17)$$

$$\sum_{e \in E_G} w_{eG} \leq 1 \quad \forall G \quad (18)$$

$$w_{eG} \in \{0,1\} \quad \forall e, \forall G \quad (19)$$

Objective function (16) maximizes the number of weighted game set winning ranks obtained by the team. A game set is defined as a set of games for which a prize can be won. The weight ω_{eG} is typically the prize money associated with rank e in game set G . For instance, in many fantasy games, prizes are assigned on a periodical basis. Constraint set (17) ensures that rank e is obtained for a particular game set G ($w_{eG} = 1$) only if either the team's score is higher than the score of the opponent that obtains rank e or the team obtains a score equal to the opponent's score, but has a larger remaining budget in the corresponding period. Finally, constraint set (18) guarantees that a team can obtain at most one winning rank for each game set. This model assumes that higher ranks entail a higher value ω_{eG} which seems to be a realistic assumption. Finally, constraint set (19) defines the domain of w_{eG} .

As most of the fantasy sports that we are aware of, if not all, entail a simpler game concept than the rules incorporated in model (1)-(12), (13)-(15) and (16)-(19), we are confident that our model can be used for a large variety of fantasy sports. For instance, the ex-post optimization model for a static fantasy sport, in which teams are fixed during the whole length of the game, is less complicated than the model for a dynamic fantasy sport. As a matter of fact, the static model can be seen as a special case of a dynamic model with only one period. Consequently, our model can be used for the ex-post optimization of both static and dynamic fantasy sports. A similar reasoning applies to the characteristic of multiple (daily) prizes. Fantasy sports with only one end prize can be seen as a special case of the situation with multiple prizes. Consequently, model (16)-(19), which involves multiple winning ranks, can be used for both types of fantasy sports. The same holds for the characteristic that only a limited number of players per team yield points. Fantasy sports in which all selected players earn points on every matchday or race can be seen as a special case in which the number of players that yield points equals the team size. Finally, also the use of the remaining budget as a tiebreaker leads to a model extension as compared to the case in which the remaining budget is neglected. The latter can be seen as a special case in which the importance (weight) of the remaining budget is set to 0 in a weighted objective function.

4. Practical application: Gigabike

Gigabike is a fantasy cycling game that is hugely popular amongst Belgian and Dutch professional road cycling fans. The Gigabike population is highly representative for the typical fantasy sports participants. In the 2013 edition of the game, according to the organizers, 96,7% of the participants are male, all are white and they have an average age of 38 years. No information on the educational level of the Gigabike participants was available though. Gigabike differs from many other fantasy sports in the organizational aspect of the game. Participation in Gigabike is free but there is a limit of about 500 participants and entry is difficult and sometimes upon invitation. A further notable difference is the fact that, apart from a challenge cup, there are no prizes nor monetary rewards to be won.

The game basically operates as follows. At the start of the road cycling season, each Gigabike participant selects a team of 30 riders from a world ranking of all professional road cyclists. In this ranking, each cyclist is assigned a so-called *cycling quotient* (CQ) value based on his performance during the last 12 months (see <http://www.cqraking.com>). This value reflects the past quality of the rider and might indicate his future performance. The total CQ value of a team must be less than a given CQ budget. At each of five fixed moments during the season, one can modify the team by substituting at most five cyclists. Hence, the season consists of six periods. However, the sum of the CQ values of the incoming riders cannot exceed the sum of the values of the outgoing riders plus the remaining CQ budget of the preceding period, if any. The winner of the game is the person whose team gains the most points over the course of the whole season. Table 1 also makes clear how the fantasy sports characteristics apply to the Gigabike game.

We illustrate our model on the Gigabike game, because Gigabike represents a typical fantasy sports game when the game rules characteristics are considered. It makes use of a constrained budget, in which the remaining budget is used as a tiebreaker, there are transfer opportunities and a single cyclist can be in many different fantasy teams. Only the eight best scoring riders of each fantasy team in every race yield points. The game also offers a broad range of different daily prizes and overall prizes. These rules make Gigabike by far the most complex fantasy sport game that we are aware of. Moreover, the Gigabike game entails a large number of riders (>3000) to select from and a large number of events (>120 race days) in which points can be earned. Hence, from a computational point of view, Gigabike poses quite a demanding test.

We applied our model to the 2012 edition of Gigabike, for which we computed ex post an optimal starting selection and optimal transfers for each of the 5 transfer periods. Our model was solved with IBM Ilog Cplex 12.3 within a few minutes of computation time. The result learned that the Gigabike winner's score accounted for 74% of the optimal score, leaving a substantial margin for improvement. This also allows to compare the 2012 winner with winners of previous editions. Indeed, comparing the scores in absolute terms is not very meaningful, since the total number of points that can be obtained differs from year to year, because of the races that are included (e.g. Olympics), and rider's performances (e.g. cheap vs. expensive riders collecting the most points). Computing the optimal score, however, provides a meaningful benchmark to evaluate the participant's performance in Gigabike over the years.

We also used our model to compute, for each of the participants, their optimal transfer decisions, given their starting selection. Computationally, this is less of a challenge, since most of the variables are fixed beforehand. Apart from the fact that many participants found the resulting information highly interesting, and tried to induce strategic insights for the next Gigabike edition, it showed that in fact all participants focusing on the end victory could still (easily) have won the game with the right transfers. In other words, selecting a starting team turned out not to be overly important in the sense that a poor choice could ruin the rest of the participant's Gigabike experience. A good balance between the importance of the starting team and the transfers had always been a concern for the organizers, but the results of our model now finally managed to provide well-founded support in favor of the current game rules. Comparing a participant's score with the score that could be obtained given his or her starting team, provided a method to assess the quality of this participant's transfers. Similarly, the quality of the starting selection could be measured. The organizers used this information to give an award to the participants with the best starting team and the best transfers. Upon request, we also computed optimal starting teams and transfers for participants with different goals than the end victory (e.g. winning as many bunch-sprint races as possible).

5. Conclusion

We developed a general ex-post optimization model for fantasy sport games, allowing to compute the best decisions if one knows beforehand how many points each player will collect in each game. We argued that this model is valuable for both organizers and participants, as it produces informational and commercial value.

Based on a survey of fantasy sport games, we listed a number of common characteristics with respect to game rules. Our model is able to deal with each one of them, and hence, is suitable for any fantasy sport game we know.

A practical application to the cycling game Gigabike shows that our model is computationally manageable. Moreover, it illustrates how organizers and participants embraced the new possibilities that this model brings. Indeed, the outcome of our model justified the game rules, provided insight in the fantasy game's tactics, allowed for the creation of extra prizes, and increased the joy of participating in the game.

References

- Billings, A.C. and Ruihley, B.J. (2013) Why we watch, why we play: the relationship between fantasy sport and fanship motivations. *Mass Communication and Society* **16**, 5-25.
- Davis, N.W. and Duncan, M.C. (2006) Sports knowledge is power: reinforcing masculine privilege through fantasy sport league participation. *Journal of Sport and Social Issues* **30**, 244-264.
- Drayer, J., Shapiro, S.L., Dwyer, B., Morse, A.L. and White, J. (2010) The effects of fantasy football participation on NFL consumption: a qualitative analysis. *Sport Management Review* **13**, 129-141.
- Dwyer, B. (2011a) The impact of fantasy football involvement on intentions to watch NFL games on television. *International Journal of Sport Communication* **4**, 375-396.
- Dwyer, B. (2011b) Divided loyalty? An analysis of fantasy football involvement and fan loyalty to individual NFL teams. *Journal of Sport Management* **25**, 445-457.
- Dwyer, B. and Kim, Y. (2011) For love or money: developing and validating a motivational scale for football participation. *Journal of Sport Management* **25**, 70-83.
- Farquhar, L.K. and Meeds, R. (2007) The relationship of fantasy football participation with NFL television ratings. *Journal of computer-mediated communication* **12**, 1208-1228.
- Fortunato, J.A. (2011) The relationship of fantasy football participation with NFL television ratings. *Journal of Sport Administration & Supervision* **3**, 74-90.
- Karg, A.J. and McDonald, H. (2011) Fantasy sport participation as a complement to traditional sport consumption. *Sport Management Review* **14**, 327-346.
- Lee, W.Y., Kwak, D.H., Lim, C., Pedersen, P.M. and Miloch, K.S. (2011) Effects of personality and gender on fantasy sports game participation: the moderating role of perceived knowledge. *Journal of Gambling Studies* **27**, 427-441.
- Lee, S., Seo, W.J. & Green, B.C. (2012) Understanding why people play fantasy sport: development of the Fantasy Sport Motivation Inventory (FanSMI). *European Sport Management Quarterly*, prepublished online.
- Nesbit, T.M. and King, K.A. (2010) The impact of fantasy sports on TV viewership. *Journal of Media Economics* **23**, 24-41.
- Nesbit, T.M. and King, K.A. (2011) Major League Baseball attendance and the role of fantasy baseball. *Journal of Sports Economics* **13**, 494-514.
- Randle, Q. and Nyland, R. (2008) Participation in internet fantasy sports leagues and mass media use. *Journal of website promotion* **3**, 143-152.
- Ruihley, B.J. and Billings, A.C. (2012) Infiltrating the boys' club: motivations for women's fantasy sport participation. *International Review for the Sociology of Sport*; prepublished online.
- Ruihley, B.J. and Hardin, R.L. (2011) Beyond touchdowns, homeruns, and three-pointers: an examination of fantasy sport motivation. *International Journal of Sport Management & Marketing* **10**, 232-256.
- Smith, A.A., Synowka, D.P. and Smith, A.D. (2010) Exploring fantasy sports and its fan base from a CRM perspective. *International Journal of Business Innovation and Research* **4**, 103-142.

A Virtual Coach for Fantasy Soccer using Mathematical Programming

F. Bonomo^{1,2}, G. Durán^{2,3,4,5} and J. Marenco^{1,6}

1 Departamento de Computación, FCEN, UBA, Argentina. fbonomo@dc.uba.ar

2 CONICET, Argentina

3 Instituto de Cálculo, FCEN, UBA, Argentina.

4 Departamento de Matemática, FCEN, UBA, Argentina. gduran@dm.uba.ar

5 Departamento de Ingeniería Industrial, FCFM, Universidad de Chile, Chile

6 Instituto de Ciencias, Universidad Nacional de General Sarmiento, Argentina. jmarenco@ungs.edu.ar

Abstract. General awareness of the contribution mathematics can make to sports decision-making has been raised by the movie “Moneyball” through its portrayal of a real professional American baseball club that used mathematical tools to improve team performance. This article addresses the same problem using as a test case a fantasy sport game organized by an Argentinian newspaper. Several mathematical programming models are presented that act as virtual soccer coaches which choose a virtual team lineup for each match. One of the models was entered in the fantasy game, achieving results that positioned them among the highest-scoring participants. Further development of such models would provide useful tools for supporting decision-making by coaches or managers of real sports.

1. Introduction

Gran DT is a fantasy soccer game created and run by a major Argentinian newspaper for the First Division of the real Argentinian professional soccer league. The objective of the game is to build the best possible virtual team by combining real players from the Division’s various clubs so as to accumulate the highest team point score. The virtual teams win or lose points depending on the weekly performance of the players, which is measured using both objective statistics (goals scored, shutouts, yellow and red cards) and subjective ones as defined by the newspaper (individual player performance in each match, Man of the Match awards). The game rules require that each team satisfy a series of restrictions (budget, number of players per position, number of players per club). The virtual teams are dynamic in the sense that participants can make changes after each round of matches to improve their lineups.

First played in the 1990s during a few tournaments, Gran DT was relaunched in August 2008 and has continued without interruption through all 9 First Division tournaments held since that date. Participation in the game has been massive, never falling below 1 million competitors and reaching a peak of almost 2 million (close to 5% of the Argentinian population) in the first tournament of 2009.

The inspiration for Gran DT was “Fantacalcio” [10], the fantasy game of Italy’s Serie A professional soccer league. Currently there are various other fantasy sport games around the world such as English soccer’s Fantasy Premier League [9] and the NBA fantasy basketball game in the United States [13]. Similar games for virtual soccer have also become very popular [12,14]. Particularly interesting, however, is the growing use of fantasy games in recent years to improve the teaching of mathematics and stimulate student motivation in the subject at all education levels in the U.S. [18].

This article presents three mathematical programming models. Two of them were designed *a posteriori* and are referred to as descriptive because they address two variants on the problem of what would have been the optimal team round by round over the course of the tournament if the results had been known in advance.

They are thus able to identify the ideal team that would have had to be devised in order to obtain the highest possible point total while satisfying the game constraints.

The third model was designed *a priori* and is called prescriptive because it proposes changes to the team lineup round by round that are intended to improve team robustness as the tournament progresses without knowledge of future results. It does this by using existing information on player performance in earlier tournaments and past matches in the current one as well as data on key characteristics of the upcoming round.

The prescriptive model was tested by entering it in the fantasy game as a virtual participant. It routinely finished the tournament in the top 3% of all competitors, on one occasion ending up among the top one-tenth

of 1 percent. If the six tournaments our model has played in so far are considered as a single tournament, the model ranks among the best two-tenths of 1 percent.

The literature in the field known as “sports analytics” (SA), which covers mathematical and computational developments for solving sports-related problems, has grown considerably in recent years. Prominent journals in operations research, applied mathematics, statistics, management and economics have published numerous articles on SA [1]. In the case of mathematical programming for sports, the area of greatest development is the definition of league season schedules for various sports and competition formats, a subfield of SA known as “sports scheduling” (SS). Excellent surveys of the state of the art in SS, with analyses of a number of open problems, are found in [2,7]. A review of the main instances for various sports that have been studied in the literature appears in [6].

Although software packages for assisting coaches of different sports in compiling, storing and consulting data have been commercially available for some years, to the best of our knowledge there are no optimization or mathematical programming algorithm applications of the sort proposed in the present work that provide support for real or virtual coaches. The one that comes closest is perhaps Fantarobot [11], an optional functionality on the Fantacalcio webpage that reorders the squad of starting and substitute players chosen by the participant into a “best” team, but as far as we aware, its recommended selection is based only on the current average performance of each player. As regards the use of mathematical techniques for supporting decision-making in real sports, the best-known example is in American baseball [3], where the applications employed are primarily statistical tools.

2. Description of the game

The Gran DT fantasy game begins in the fourth round of the Argentinian First Division’s Closing Tournament, which despite its name is played in the first half of the year, and again in the fifth round of the Opening Tournament, played in the second half of the year. Since in each half-year tournament the 20 First Division clubs play 19 fixtures in a round-robin format, the game consists of 15 or 16 rounds.

After signing up for the game under his or her Argentinian National Identity number, each competitor then creates a virtual team (hereafter simply “team”) by choosing 11 starting and 4 substitute players from the First Division. The 11 can play in any one of 3 different formations: 1 goalkeeper, 4 defenders, 4 midfielders and 2 forwards; 1 goalkeeper, 4 defenders, 3 midfielders and 3 forwards; or 1 goalkeeper, 3 defenders, 4 midfielders and 3 forwards. The 4 substitutes consist of a goalkeeper, a defender, a midfielder and a forward, and can play only if one of the starting players in the same position either does not play or plays less than 20 minutes and is therefore awarded no points by the newspaper for his performance.

Each player chosen for a team is assigned a symbolic monetary value of anywhere from 300,000 pesos for those who are just starting in the First Division to more than 10 million for the top players in previous tournaments. The total value of the team must not exceed the budget limit of 65 million pesos (this figure has varied over the years between 60 and 70 million pesos). No more than 3 players can be chosen from any one First Division club.

Each starting player wins or loses points for each round according to subjective criteria (the score awarded to him by the newspaper, whether he was Man of the Match) as well as objective criteria (goals scored, whether or not he was cautioned, whether or not he was expelled). The point score awarded by the newspaper on each criterion is an integer ranging from 1 to 10. Scoring a goal in open play wins 10 extra points for a goalkeeper, 9 points for a defender, 6 points for a midfielder and 4 points for a forward. A penalty kick goal is worth 3 points regardless of the scorer’s position. The Man of the Match as decided by the newspaper is awarded 4 extra points. For a clean sheet (shutout), the goalkeeper is awarded 3 extra points and the defenders 2 extra points. The goalkeeper loses 1 point for each goal given up and wins 4 extra points for stopping a penalty kick (4 points are also deducted from the player who took the unsuccessful kick). A yellow card costs a player 2 points and a red card costs him 4. A player who plays less than 20 minutes is deemed not to have played, and if he is a starter in the fantasy team, his place is taken by the substitute in the same position. If more than one starting player in the same position does not participate in a given round, their team plays with less than 11 players due to the rule limiting substitutes to one per position.

Starting players and substitutes can be switched from one round to the next as often as desired. Up to 4 transfers, in which a current player on the team is replaced with a new one, are also allowed per round as long as the basic restrictions regarding formation and budget are met (this rule has changed since the first editions of the game, when only three such transfers were permitted).

3. The descriptive models

The *a posteriori* or descriptive models are run once the tournament is over. They determine what the optimal team configurations would have been once the results of the tournament are known. In other words, they use as data the points obtained by each player in each round (or a given set of rounds in a partial analysis).

The first of the two descriptive models chooses a fixed team of 11 starting players satisfying the restrictions imposed by the game and makes no changes over the course of the tournament. Substitute players are not used, though 1,200,000 pesos are allocated for four of them at 300,000 pesos each to conform with the game rule requiring that four be chosen. The model attempts to maximize the team score over the entire tournament with only the 11 starters.

The second descriptive model builds what we call the “perfect team”. Beginning with an initial lineup in the first round of the game, it indicates round by round what changes should be made between starters and substitutes and which new players should be incorporated in order to obtain the highest possible final team point total.

Both models were formulated as integer linear programs in which the objective function maximizes the team’s point total while the constraints ensure the solution meets the game restrictions on team selection, permitted transfers and the budget.

The results obtained by the two descriptive models in the four tournaments of the 2009 and 2010 seasons are set forth in Table 1 together with their respective solution times. The points obtained by the actual Gran DT game winner is also shown.

As can be seen, the Fixed Team model with the original starting players was more or less tied with the human winners of the game, the model winning by a small number of points in three of the tournaments and losing by a few in the remaining one. These scores demonstrate that the winners played very well considering they could not have advance knowledge of the results, which are known only *a posteriori*. Clearly, good gamers are able to take effective advantage of the First Division clubs’ performance dynamic as the tournament progresses to improve their teams from round to round and thus compete at the same level as the model with a fixed lineup over the entire tournament but full knowledge of the future results.

The Perfect Team model, on the other hand, did much better than any human competitor, with point totals that were 50% to 70% higher than the game winner depending on the tournament. The big difference was due fundamentally to the fact that this model captures players who perform well sporadically, something not even the most expert Gran DT gamers are normally able to achieve.

The point total differences between the Closing and Opening tournaments are due to the fact that in the latter the game starts one round earlier.

The results obtained by the two models were published in the newspaper running the game on various occasions [15,16,17].

As regards solution times, the Fixed Team model executed very quickly but the Perfect Team model took up to 2 hours. The Fixed Team formulation had about 500 variables (the total number of players in the First Division) and slightly more than 20 constraints whereas the Perfect Team version had some 2,000 variables and 3,500 constraints. In every case, the models solved the problem to optimality in the times indicated in Table 1. The experiments were run using the CPLEX 12.2 optimizer on a PC with 2 GB of RAM and a 1.6 GHz processor.

Table 1. Results of descriptive models for the 2009 and 2010 Tournaments.

Tourney	GDT Winner	Fixed Team	Solution Time	Perfect Team	Solution Time
Cl. 09	1279	1318	2 sec	1990	10 min
Op. 09	1375	1336	1 sec	2173	120 min
Cl. 10	1227	1232	1 sec	2027	50 min
Op. 10	1394	1412	2 sec	2289	116 min

4. The prescriptive model

The *a priori* or prescriptive model poses a greater challenge since in this case it must find good teams without knowledge of the players' future performance. An index is therefore constructed for each player that generates a prediction of the point score he will obtain in the next round. The model itself is applied in two very similar versions. The first version identifies the initial team by maximizing the global team index (i.e., the sum of the individual players' indexes) for the first round of the game while satisfying the game restrictions. The second version consists in determining the round-to-round changes and transfers to be made, starting with the second round of the game, that again maximize the global team index while satisfying the restrictions. The entire process is similar to the descriptive model except that here, instead of the players' actual point score, the model uses prediction indexes.

The challenge is therefore to build an index that will produce a reasonable representation of what will actually happen. After some initial testing we concluded that a player's point average in recent rounds was not by itself a good predictor of the points he would earn in the next round because it took no account of key match characteristics such as the rival club to be played, the match's home or away status, the current performance or situation of his club, etc.

We therefore decided to construct each player's index as his point average for the rounds already played in the current tournament, weighted by three factors: the home or away status of his club's next round match (the weights applied were 1.05 for home games, 0.95 for away games), the league table position of his club's next round rival, (1 to 1.05 if in the bottom five of the table, 0.95 to 1 if in the top five), and the current performance or situation of the player or his club (up to 5% more if on a scoring or winning streak, respectively; up to 5% less if on a scoreless or winless streak, or tired after, for example, a recent league or international match). The averages obtained by a player in the previous two tournaments are incorporated as though they were two additional rounds in the current one.

One last consideration included in the indexes is the "starting lineup" factor, which is 1 for those players who were announced either in the press or by the club coach as starters in the next round, and 0 for all the other players. This factor is incorporated as an attempt to ensure that the team's 11 starting players will indeed be playing in the coming round. Though it is a crucial piece of information, it cannot always be known *a priori* due to the game rule stipulating that all lineup changes for a given round must be made no later than one hour before kickoff in that round's first scheduled match. Since First Division rounds typically spread the various matches across the "weekend" (Friday to Monday), there will be Fridays when the definitive starting lineup of some clubs playing later in the weekend will not have been announced yet. In such cases the substitute players take on considerable importance and having "good" substitutes is thus advisable, though it may still be prudent not to spend much of the budget on them since in most cases they will not be used.

Both of the prescriptive models have about 1,000 variables and 500 constraints, and were solved to optimality in a couple of seconds.

We analyzed the performance of the prescriptive model for the 2010 Closing and Opening tournaments. As was to be expected, since the model chooses the players who have performed best as of the last round and the individual performance of the players is quite variable, the predictions tend to be higher than the actual values. To check this behaviour we created a team of randomly chosen players for the 2010 Closing tournament, the only additional condition being that in each round all 11 starting players must play, that is, that the starting lineup factor for each of them is equal to 1. The random team won 899 points whereas the predicted total was 934 points. In other words, the prediction was much closer to the true figure, the

difference being less than 4%, as compared to the differences of 21% and 14% for the teams generated by the model for the 2010 Closing and Opening tournaments, respectively.

In the Gran DT game for the 2010 Closing tournament there were 1,442,682 competitors and the winner obtained 1,227 points. Our model came in 13,547th place with 1,070 points, positioning it within the top 1% of all participants. The random team ended up in 498,726th place with 899 points. Since a randomly chosen team would finish around the middle of the league table, this latter result suggests the number of active teams (by “active” we mean teams that are updated from round to round) would have been approximately 1 million. A non-active team will typically finish the tournament participating with less than 11 players in each round given that it does not replace those who are injured or dropped from the starting lineup as the tournament progresses. The estimate of 1 million active participants agrees with that of the game organizers, who have observed that 2/3 of the teams update from round to round. If we consider only active teams, our model’s performance would place it among the top 1.5% of competitors in the game.

In the 2010 Closing tournament there were 1,445,531 competitors and the winner obtained 1,394 points. Our model came in 643rd place with 1,322 points, positioning it in the top 0.1% of all participants even if only active teams are considered. This was easily the model’s best performance of the 6 tournaments it was entered in.

A final result: if we consider all 6 tournaments in which our model participated as one single tournament, the model came in 530th place out of 343,017 competitors, placing it among the top 0.2%.

5. Conclusions and future research

The purpose of this study was to investigate the contribution mathematical programming can make to the design of a virtual sports coach or providing support to a real coach. The analysis took the form of a case study of an established fantasy sport game based on the two annual tournaments of Argentina’s First Division soccer league.

The article presented one *a priori*-designed mathematical programming models referred to as “prescriptive” and two others designed *a posteriori* denoted “descriptive”, all of which were used to identify optimal or good teams for the fantasy game. Each edition of the game, organized by a major local newspaper, has attracted more than a million participants.

The descriptive models were able to identify the ideal teams that would have obtained the highest possible point total while satisfying all of the constraints imposed by the game rules.

The prescriptive model used historical data and the characteristics of the next match round to create a competitive team which was then tested by entering it into the competition. The results obtained by the model systematically positioned it among the top 3% of the game participants, in one case reaching the top 0.1%. Indeed, if the 6 tournaments the model participated in are considered as one, our virtual competitor placed within the top 0.2%. We can safely expect that the longer is the tournament, the better our statistical and optimization tools will function.

In our case, the indications generated by the prescriptive model were followed 100% of the time, but it could also be used as a complementary support tool by an expert game competitor.

As regards future research, various lines of inquiry could be pursued to improve our virtual gamer model. One of them is a more global and not so greedy optimization. Changes made to the team at any point would take into account not only the upcoming round but also the one or two rounds following it. Another idea is to form a “high risk” team with players whose point scores from round to round display high variance but good, although not the best, indexes as defined by the model.

Yet another alternative would be to find the best players for each round without attempting to predict their point scores given their high variability. This could be done by implementing sophisticated statistical models using historical data to determine which indication of the players that should be chosen for the team.

It is also interesting to note that the problem analyzed in this study has some similarities to the problem of selecting a stock portfolio that maximizes investor income. It may be, therefore, that certain models used in finance to predict share behaviour, such as the capital asset pricing model (CAPM) [4,8] based on Markowitz’s portfolio theory [5], could be usefully applied to the determination of robust virtual soccer coach models.

Finally, tools such as those developed in this paper could provide valuable support for coaches in real sports. The combination of sports and mathematics for this purpose was well portrayed in the Bennett Miller film “Moneyball”, based on a book by Michael Lewis [3] and starring Brad Pitt and Jonah Hill. The film tells the true story of the manager of an American baseball club who radically changed the team’s strategy after incorporating mathematical techniques into his decision-making, with excellent results.

Acknowledgements: The authors would like to thank Javier Romero, Carlos Prieto and Jorge Blanco, organizers of the Gran DT game at the *Clarín* newspaper in Buenos Aires, for their constant help in bringing this article to fruition. They are also grateful to Andrés Farall, Leonardo Faigenbom, Leonel Spett and Pablo Groisman for the many discussions that contributed to this project; to Mario Guajardo, who carefully reviewed this paper and made numerous suggestions for its improvement; and to Sebastián Ceria, Kenneth Rivkin and Gustavo Braier for their comments and suggestions. This study was partially funded by project nos. ANPCyT PICT-2007-00518 (Argentina), CONICET PIP 112-200901-00178 (Argentina) and UBACyT 20020090300094 (Argentina), and by the Complex Engineering Systems Institute (ISCI, Chile). The second author was partly financed by project no. FONDECyT 1110797 (Chile).

References

- [1] Coleman B.J. (2012). Identifying the “Players” in Sports Analytics Research. *Interfaces* **42** (2), 109—118.
- [2] Kendall G., Knust S., Ribeiro C.C., Urrutia S. (2010). Scheduling in sports: An annotated bibliography. *Computers & Operations Research* **37** (1), 1—19.
- [3] Lewis M. (2003). *Moneyball: The Art of Winning an Unfair Game*. Norton & Company.
- [4] Lintner J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *Review of Economics and Statistics* **47** (1), 13—37.
- [5] Markowitz H.M. (1952). Portfolio Selection. *The Journal of Finance* **7** (1), 77—91.
- [6] Nurmi, K., Goossens, D., Bartsch, T., Bonomo, F., Briskorn, D., Durán, G., Kyngas, J., Marenco, J., Ribeiro, C.C., Spieksma, F., Urrutia, S., Wolf-Yadlin, R. (2010). A framework for scheduling professional sports leagues. In *IAENG Transactions on Engineering Technologies, American Institute of Physics* (Ao, S-I., Katagir, H., Xu, L., Chan, A.H-S., Eds). Vol. **5**, pp. 14—28.
- [7] Ribeiro C.C. (2012). Sports scheduling: Problems and applications. *International Transactions in Operational Research* **19**, 201—226.
- [8] Sharpe W.F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance* **19** (3), 425—442.
- [9] <http://fantasy.premierleague.com/>
- [10] <http://www.fantacalcio.kataweb.it/>
- [11] http://fantacalcio.repubblica.it/index.php?page=faq&ck_fantacalcio#16
- [12] <http://www.hattrick.org/>
- [13] <http://www.nba.com/fantasy/>
- [14] <http://www.xperteleven.com/>
- [15] <http://edant.clarin.com/diario/2009/07/09/deportes/d-01955551.htm>
- [16] <http://edant.clarin.com/diario/2009/12/20/deportes/d-02104838.htm>
- [17] <http://edant.clarin.com/diario/2010/05/20/deportes/d-02197713.htm>
- [18] <http://www.fantasysportsmath.com/>

2012 UEFA Euro efficiency evaluation based on market expectations

Luana Carneiro Brandão*, Fernando do Valle Silva Andrade** and João Carlos C. B. Soares de Mello***

* Rua Passo da Pátria, 156 – São Domingos – Niterói – RJ + email address:luanabrandao@id.uff.br

** Rua Passo da Pátria, 156 – São Domingos – Niterói – RJ + email address:fernando_do_valle@hotmail.com

*** Rua Passo da Pátria, 156 – São Domingos – Niterói – RJ + email address:jccbsmello@id.uff.br

Abstract. The present work uses Data Envelopment Analysis to evaluate performance of the 2012 UEFA Euro national teams, considering the market's expectations and favouritism. In this model, efficiencies represent the relationship between results and expectations. Before applying DEA, we use MACBETH to transform the tournament ranking into cardinal numbers. We use two different traditional DEA models, as well as a Smooth DEA model. The latter has been developed to avoid some problems of traditional DEA, and consists on replacing the original, piecewise linear frontier by a smooth frontier. This smooth DEA frontier is used to deal with the default efficiency problem.

1. Introduction

The UEFA European Football Championship (UEFA Euro) is a European tournament organized by the Union of European Football Associations (UEFA). It is considered to be the second most prestigious football tournament contested by senior men's national teams in the world, after the FIFA (Fédération Internationale de Football Association) World Cup, and to gather the most powerful football teams, except for among others Brazil and Argentine.

All 53 teams affiliated to UEFA compete in a qualifying process, with the exception of the host nations, who qualify automatically. However, only 16 compete in what is called the Final Phase, which is the object of this study. Starting on the next edition to be held in France in 2016, 24 teams will compete in the Final Phase for the first time (UEFA, 2012).

The idea for the tournament started in the mid 50s, after the creation of the European entity, in 1954, and it was proposed by Henri Delaunay, the then UEFA general secretary. Other similar tournaments were already settled, and may have inspired UEFA, such as the British Home Championship that started in 1883 and lasted 100 years, the Nordic Football Championship, first played in 1924, and the Central European International Cup, in 1927 (ROBINSON, 1996). Others were being created at that time, such as the Africa Cup of Nations, in 1957, and the AFC Asian Cup, in 1956.

Tournaments between national teams are very different from tournaments between clubs, in which investments actually indicate the expectation on the team's performance. The expectation and favoritism towards the national teams are generally based on historical aspects, such as the team's tradition, and very hard to measure. This is why the object of this study is to analyze the team's performance based on its favoritism.

2. Bibliographic Review

Data Envelopment Analysis (DEA) seeks a comparative relation between inputs and outputs of Decision Making Units (DMUs), which are in this case the teams in the UEFA Euro 2012. The method calculates the DMUs' efficiencies, which is the ratio between the weighted sum of the outputs and the weighted sum of the inputs. These weights are individually calculated so the efficiency of the DMU being analyzed is maximized. The efficient DMUs form what is called an efficient frontier.

There are two basic models in DEA: CCR (Charnes, Cooper e Rhodes, 1978) and BCC (Banker, Charnes e Cooper, 1984). The first assumes constant returns to scale and proportionality between inputs and outputs, while the latter assumes variable returns to scale and convexity of the efficient frontier. The DEA models may also be divided in output and input oriented. The first assumes that inefficient DMUs must increase their production until they reach the efficient frontier, while the latter assumes that they should reduce their inputs until the frontier is reached.

The Smooth DEA Frontier Theory (SOARES DE MELLO et al., 2002; SOARES DE MELLO et al., 2004; NACIF et al., 2009) has been developed to avoid some problems of traditional DEA. One of them is

the multiple solutions associated with the efficient DMUs, which makes it impossible to calculate the weights of inputs and outputs used by the efficient DMUs. These weights may be interpreted as trade-offs (COOPER et al., 2000a) or shadow-prices (COELLI et al., 1998), and therefore are very important. Another problem of traditional DEA is the Pareto inefficient regions, where Pareto inefficient units are considered to be efficient in the model. For this, the original, piecewise linear frontier is replaced with a smooth frontier that has derivatives at all points. This new frontier also maintains the essential properties of traditional DEA.

3. Proposed Improvements

Many works use the CCR model even when it's not as proper as BCC because of the latter's low discrimination capacity. This is partially due to default efficiency, which means that in BCC any DMU with the smallest of one of the inputs or the greatest of one of the outputs will be considered efficient. This considerably affects evaluation, also because DMUs nearby these default-efficient units suffer different evaluation as well.

Therefore, we propose a slight change in the smooth models to reduce this problem, which is to relax the equality restrictions for these default-efficient DMUs. In other words, instead of guaranteeing that the new frontier includes all efficient DMUs, we allow it not to include these default-efficient units, as long as it passes over them and includes the others.

However, since the DMUs that are close to these default-efficient units are also affected by this problem, we may broaden this concept and consider that every DMU that is BCC efficient, but CCR inefficient may be considered default efficient. Therefore, we may allow the new frontier to pass over all of these DMUs. If it does pass over them, they will be considered inefficient, but if it passes by them, they will be considered efficient.

We can propose a default efficiency index, as in equation (1). This index calculates how much of the traditional BCC efficiency is due to the default efficiency problem. Therefore, the greater the index, the less efficient the DMU really is.

$$\% \text{Efficiency}_{\text{default}} = \frac{\% \text{Efficiency}_{\text{BCC}} - \% \text{Efficiency}_{\text{Smooth}}}{\% \text{Efficiency}_{\text{BCC}}} \quad (1)$$

On the other hand, there are differences in the efficiency values of these two models that are not due to the default efficiency problem. Thus, we propose that this index is calculated only where there might be default efficiency, which is where at least one of the inputs is smaller than the smallest CCR efficient DMU and where at least one of the outputs is greater than the greatest CCR efficient DMU.

4. OR in Football

Operational Research is extensively used in Football. García-Sánchez (2007) analyze the 2004/2005 Spanish Championship on and off the field with a 3 stage DEA model. The study first calculates the operational efficiency, by relating offense and defense statistics to goals made, then the operational effectiveness, which relates the operational efficiency and the points made, and finally the social effectiveness, which relates the operational effectiveness with total audience.

5. Case Study

The target of this work is to analyze performance of the UEFA Euro 2012 teams, based on their favoritism. For that we use a DEA model where these teams are the DMUs analyzed, the sum of the players' market value is the 1st input and the total points in the FIFA ranking is the 2nd input. The main outcome is the final tournament ranking, but in order to use it as output, we must transform it into cardinal numbers, using MACBETH (BANA E COSTA and VANSNICK, 1994). This software helps the decision maker to grade different options, by comparing them with each other in terms of attractiveness.

We transform the tournament ranking into groups of ranking to be compared with each other (according to only one criterion): 1st place, 2nd place, eliminated in semi-finals, eliminated in quarterfinals, eliminated at group stage, nonparticipant. Then, we consider that the difference between the 1st place and 2nd place is greater than the one between the 2nd place and the eliminated in semi-finals, and so on until the last judgment to be made. This results in the following final scale: 100 points for 1st place, 54 points for 2nd

place, 30 points for the eliminated in semi-finals, 17 points for the eliminated in quarterfinals, 10 points for the eliminated at group stage, and 0 points for the nonparticipants.

Moreover, we also model another DEA problem with the number of points won in the tournament as a second output. The purpose of this is just to draw other conclusions that might be interesting for the teams.

Due to the lack of proportionality between inputs and outputs and also to the limited radius of the Macbeth variable, we use a traditional DEA BCC model, instead of the CCR. In order to eliminate the default efficiencies, we also use the smooth DEA BCC model.

5.1 Preliminary Analysis

In the first place, we use SIAD (ANGULO MEZA et al., 2005) to calculate the traditional DEA BCC models (with 1 output and with 2 outputs), as shown in table 1. We consider output orientation because the target of the teams is to improve their performance, instead of reducing their market value and FIFA scores.

We can see in table 1 that Germany has the greatest difference between the two models, meaning that she was poorly ranked considering her points in the tournament. She had 3 more points than Italy (in 2nd place) and yet she was eliminated in semi-finals. It's as if she won the "wrong games".

Though it doesn't matter the number of points won in the tournament, the model with both outputs allows us to draw other interesting conclusions about the teams. For instance, Germany might not need to work on her technical skills, although she is considered to be inefficient in the first model. She might only need to work on the psychological side of her players, because they seem to work very well during group stage, but not in the knock-out stages. On the other hand, Italy might need to improve her technical skills, although she is considered efficient in the first model. She might do well when it comes to knock-outs, but not as well when there are no pressures involved.

Table 1 – Traditional DEA BCC results

DMU	Team	1 Output Model	2 Outputs Model	Difference
1	SPAIN	1	1	0
2	GERMANY	0,384	1	0,616
3	ENGLAND	0,245	0,725	0,480
4	PORTUGAL	0,537	1	0,463
5	FRANCE	0,322	0,503	0,181
6	NETHERLANDS	0,180	0,180	0
7	ITALY	1	1	0
8	RUSSIA	0,332	0,573	0,241
9	CROATIA	0,351	0,587	0,236
10	SWEDEN	0,425	0,471	0,046
11	UKRAINE	0,739	1	0,261
12	CHECK REPUBLIC	1	1	0
13	POLAND	1	1	0
14	DENMARK	0,561	0,667	0,106
15	GREECE	1	1	0
16	IRELAND	1	1	0

In fact, Germany traditionally has a strong team (3 times world champion and also 3 times champion of the Euro tournament), but in the UEFA Euro 2012 tournament, she was the youngest one. She also had a similar performance in the last FIFA world cup, when she was also eliminated in semi-finals, after a very well played tournament. On the other hand, it is well known that Italy performs very well at important games and she also had more experienced players, such as the 2006 world champions Andrea Pirlo and Gianluigi Buffon.

The same problem with pressure might be valid for other teams, such as Portugal and England. On the other hand, Greece and Check Republic might perform well under pressure, and they must improve their technical skills, despite being efficient.

5.2 Smoothing DEA

We finally use smoothing DEA with the single output model to calculate the default efficiencies, using standardized data to simplify calculations for Excel's Solver. This standardized data and the Smooth DEA results are shown in table 2.

The frontier's polynomial equation is $Z=F(x,y)=-0,13012+1,07772x+0,23530y-0,18290x^3$. Here, x represents input 1 (team's market value), y represents input 2 (FIFA points) and Z represents output 1 (MACBETH's transformation of the tournament ranking).

Since y is only associated with a linear term, this indicates that there's a linear relationship between the FIFA points and the tournament ranking. In other words, if all teams were equally efficient in the tournament, the FIFA ranking would equal the tournament ranking. This may be considered a validation for the FIFA ranking.

Analyzing the standardized values in table 2, we can see that there's a significant gap between Spain and the other teams. Considering this and the fact that she won 2 European and 1 world tournament in the past few years, we can say that she is in fact an outlier.

Table 2 – Smoothed DEA BCC Results

DMUs	x	y	z	Z_{smooth}	Eff BCC Output Or.	Eff CCR	Eff smooth Output Or.	Eff default
SPAIN	1,00	1,00	1,00	1,00	100%	100%	100%	0%
GERMANY	0,76	0,88	0,30	0,82	38%	38%	37%	
ENGLAND	0,66	0,79	0,17	0,72	25%	24,5%	23,7%	
PORTUGAL	0,56	0,68	0,30	0,60	54%	51%	50%	
FRANCE	0,55	0,66	0,17	0,59	32%	29%	29%	
NETHERLANDS	0,51	0,85	0,10	0,60	18%	18%	17%	
ITALY	0,50	0,67	0,54	0,54	100%	100%	100%	
RUSSIA	0,26	0,67	0,10	0,31	33%	33%	32%	
CROATIA	0,25	0,72	0,10	0,30	35%	35%	33%	
SWEDEN	0,21	0,63	0,10	0,24	42%	41%	42%	2%
UKRAINE	0,18	0,39	0,10	0,15	74%	50%	66%	10%
CHECK REPUBLIC	0,17	0,53	0,17	0,17	100%	86%	97%	3%
POLAND	0,15	0,36	0,10	0,12	100%	58%	86%	14%
DENMARK	0,144	0,70	0,10	0,19	56%	56%	53%	6%
GREECE	0,136	0,65	0,17	0,17	100%	100%	100%	0%
IRELAND	0,11	0,62	0,10	0,14	100%	71%	73%	27%
Averages								
62,97% 55,67% 58,62%								

The smoothed efficiencies output oriented are calculated as the targets divided by the original values for output 1 ($[\text{Eff}]_{smooth}=Z_{smooth}/z$). In average, they were smaller than the CCR average, but greater than the BCC average. This is an interesting result because it means that the smoothed frontier maintains the benevolence property inherent to DEA.

Some DMUs have higher efficiencies in the CCR model, which is acceptable, since the two models have different assumptions: CCR assumes that there should be a linear frontier between two efficient DMUs, while smooth DEA assumes that there should be slight changes throughout the whole frontier. Other DMUs have their smooth efficiencies closer to the BCC efficiencies. This means that, despite the relaxed restrictions, the smooth model still considers sufficiently variable returns to scale. It is also important to point out that the BCC efficiencies are never higher than the smooth efficiencies, as expected.

Concerning default efficiencies, Ireland, Poland and Check Republic are not really efficient, as considered by the BCC model. Ukraine also isn't as efficient as she seems, meaning that she benefits from default efficiencies.

5.3 Result Analysis

The basic result shown in table 2 is that Italy and Spain achieved their target. Italy started off slowly, but was very efficient by winning the “right games” to achieve her goals. Spain had extremely consistent results and dominated the competition, as she had done 4 years before.

Greece also achieved her target, while Check Republic almost did so, by passing the group stage while having very low market expectation. But we must also consider the fact that they were both in a group of teams with also low expectations, making it easier to win the games.

Poland and Ireland almost achieved their target, which was considerably low – between just participating and passing the group stage. The fact that Ireland was in Spain's and Italy's group, might have impaired her performance.

Poland and Ukraine were the home teams. Therefore, despite having a home advantage that wasn't considered in this model, they were also automatically qualified to participate in the tournament. Since their only achievement was their participation, they might not have “deserved” any output points at all. However, they did make a few points in the tournament.

Germany, Portugal and England lost important games and were eliminated sooner than the market's expectation, which was based on their performance in the past few years (represented by their FIFA points) and on the quality of their players. France was also far behind her expectations, and she might have done even worse if she hadn't won the “right games”. Russia, Croatia, Sweden and Denmark are medium sized teams, according to the FIFA ranking, but were using players with relatively low market value, which might have affected their performance. Russia had the best chance of passing the group stage, since she was in a relatively weak group.

The major disappointment in the tournament was Netherlands: the FIFA 2010 World Cup runner-up lost every game. This might be partially justified by the fact that her team aged since then – it wasn't adequately renewed. Moreover, important players, such as Robben and Sneijder, have been going through a terrible phase ever since the World Cup.

6. Conclusions

The main target of this study was to analyze the 2012 UEFA Euro outcome, based on the market's expectations. The efficiencies indicate how much of the team's target is achieved, considering the other teams' results and expectations. According to this model, not many teams were efficient. In fact, it would be impossible for all of them to be efficient. In order to obtain more interesting results, other studies may apply this method to a longer tournament, such as the FIFA World Cup or the 2016 UEFA Euro, in which 24 teams will participate.

The present study indicates that there should actually be more teams participating in the tournament, so more phases could be played, allowing teams in general to have higher efficiencies. If the 2016 tournament does not adopt new phases, there will also be an increase in average efficiency, since it is natural for the added teams to have lower expectations than the ones that usually qualify.

The DEA model proposed in this study provides important results that aid the teams' managerial decisions. If these results are repeated throughout several championships, some teams might decide to invest more in different players or more in their players' skills, which would affect the FIFA ranking. These results might also be important for the market: it may indicate that some players are overpriced or that FIFA should review her scores. However, the smooth DEA was able to indicate that FIFA has an adequate punctuation system, at least for the 2012 UEFA Euro tournament.

In general, the smooth DEA proposed in this study allows a more exact interpretation of the situation, since it accounts for default efficiency in a variable return to scale model, meaning it does not impose proportionality between inputs and outputs.

References

- Angulo Meza, L., Biondi Neto, L., Soares de Mello, J.C.C.B. and Gomes, E.G. (2005) ISYDS – Integrated System for Decision Support (SIAD - Sistema Integrado de Apoio à Decisão): A software package for Data Envelopment Analysis model. *Pesquisa Operacional*, **25**(3), 493-503.
- Bana e Costa, C.A. and Vansnick, J.C. (1994) MACBETH: an interactive path towards the construction of cardinal value functions. *International Transactions on Operations Research* **1**, 489–500.
- Banker, R.D., Charnes, A. and Cooper, W.W. (1984) Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis. *Management Science* **30**(9), 1078-1092.
- Charnes, A., Cooper, W.W. and Rhodes, E. (1978) Measuring the Efficiency of Decision Making Units. *European Journal of Operational Research* **2**, 429-444.
- Coelli, T., Rao, D.S.P. and Battese, G.E. (1998) *An Introduction to Efficiency and Productivity Analysis*. Boston-Kluwer Academic Publishers.
- Cooper, W.W., Park, K.S. and Pastor, J.T. (2000) Marginal Rates and Elasticities of Substitution with Additive Models in DEA. *The Journal of Productivity Analysis* **13**(2), 105-123.
- García-Sánchez, I.M. (2007) Efficiency and effectiveness of Spanish football teams: a three-stage-DEA approach. *Central European Journal of Operations Research* **15**(1), 21-45.
- Nacif, F.B., Soares De Mello, J.C.C.B. and Angulo Meza, L. (2009) Choosing Weights in Optimal Solutions for DEA-BCC Models by means of a N-dimensional Smooth Frontier. *Pesquisa Operacional* **29**(3), 623-642.
- Robinson, J. (1996) *Soccer: The European Championships 1958-1996*. Cleethorpes-Soccer Book Publishing.
- Soares de Mello, J.C.C.B., Gomes, E.G., Biondi Neto, L. and Lins, M. P. (2004) Suavização da Fronteira DEA: O Caso BCC Tridimensional. *Investigação Operacional* **24**(1), 89-107.
- Soares de Mello, J.C.C.B., Lins, M.P.E and Gomes, E.G. (2002) Construction of a Smoothed DEA Frontier. *Pesquisa Operacional* **22**(2), 183-201.
- UEFA/EURO, UEFA (2012) Focus turns to France in 2016. Available at:
<http://www.uefa.com/uefaeuro/abouteuro/uefaeuro2016/index.html?mobile=true>, Access on 11/27/2012 at 3 p.m.

Team Performance in the Italian Serie A: 2000/01 - 2009/10

F. Carmichael*, G. Rossi** and D. Thomas***

*University of Birmingham, Edgbaston, Birmingham, BT15 2TT, UK: f.carmichael@bham.ac.uk

**University of East London, Docklands Campus, University Way, London, E16 2RD, UK: g.rossi@uel.ac.uk

***Aberystwyth University, Room F5, Cledwyn Building, Aberystwyth, SY23 3DD, UK: det@aber.ac.uk

Abstract. This paper uses data on the Italian Serie A to estimate a production function for the league and estimate the relative efficiency of the clubs playing in it. While there has been considerable research on production and efficiency in La Liga and the EPL, corresponding evidence relating to Serie A is limited. This paper addresses this imbalance utilising a panel data set comprising season aggregated match statistics for 36 clubs that played in Serie A over ten seasons (2000/01 - 2009/10). The seasons covered by the data include those affected by the Calciopoli scandal (2005/06 and 2006/07) and we incorporate indicators for these events in the statistical model. Factor analysis is used to construct composite measures of direct inputs reflecting playing performance into the team production function. This procedure uses factor loadings from principal components factor analysis. The use of composite measures reduces the number of performance measures and increases the available degrees of freedom. This allows for the inclusion of additional team specific measures as well as indicators for managerial change and Calciopoli effects in the estimating model. The results highlight the importance of attacking play in Serie A, the role played by historic success or lack of it and, more tentatively, the potential gains and also costs from fraudulent behaviour.

1. Introduction

Following Scully's (1974) study there has been a substantial body of work analysing production function in the economics of sports. These models assume that teams, like other enterprises, are assumed to be involved in a production process with 'output', in terms of sporting success, arising from the combination of various player and non-player inputs. Team output is conventionally viewed in terms of outcomes of individual matches or over a complete season or tournament, and variously measured by such indicators as league position, win rates, points achieved or, as in the particular case of association football, goals or goal difference. The inputs into team production are player performances in terms of various 'plays' or 'tasks' during matches; the quality of which are themselves dependent on such factors as inherent ability and skill or talent, physical characteristics, age and experience, form and fitness. To varying extents some of these can also be influenced by team coaching and management styles and decisions, which also affect teamwork and tactics.

This paper follows the tradition in this literature of explicitly focusing on the direct relationships between team success on the field of play and the contribution of different player skills and abilities captured by a range of different aspects of match play performance. The contribution of the paper is twofold. First, the analysis utilises panel data set comprising season aggregated match statistics for 36 clubs in the Italian Serie A over ten seasons. Hitherto, production function analysis has not been undertaken for this league. Second the method of analysis is innovative in that composite measures of attacking, other constructive and defensive performance are constructed in a systematic way using the factor loadings from principal components factor analysis. This is in contrast to most previous research which has tended to construct composite measures of playing performance in an arguably more ad hoc way. The use of composite measures reduces the number of performance measures and in so doing reduces scope for type 1 errors and increases the available degrees of freedom. This allows for the inclusion of additional team specific measures, including indicators for the impact of the Calciopoli scandal and managerial change.

2. The Italian Serie A and the Calciopoli scandal

Established in 1898, Serie A is the top football division in the pyramid structure of four professional leagues in Italy. Nowadays, Serie A is composed of 20 teams and it is separately run by Lega Calcio Serie A under the supervision of the FIGC, the Italian football association, which provides the guidelines for the

operation of the Italian League Championships. The other professional divisions, the second tier Serie B (22 teams) and the third and forth tiers Lega Pro Division 1 and Division 2¹ are respectively managed by Lega Serie B and Lega Pro. Through the traditional system of promotion and relegation, at the end of each season the bottom three Serie A clubs are demoted and replaced by the top three Serie B clubs. This transfer is determined by accumulated match points over regular season fixtures (3 for a win, 1 for a draw and 0 for a loss) with teams equal on total points ranked by the so called *classifica a vulsa*². The Serie A league winner is awarded with the “Scudetto” (the small shield) and directly qualifies for the next UEFA Champions League together with the clubs finishing second and third in the final Serie A ranking³. The winner of the League Cup, “Coppa Italia”, together with the other clubs positioned from the 4th to the 6th places in the final Serie A ranking also qualify for the next UEFA Europa League (formerly UEFA Cup).

During the study period covering the seasons from 2000/01 to 2009/10, Serie A changed its league structure (Hamil et al., 2010). In 2003, Catania Calcio, a Serie B club, was at the centre of a controversy that led to the enlargement of Serie B from 20 to 24. Known as the *Catania case*, the club claimed that Siena, another Serie B club, fielded an ineligible player in a 1-1 draw. This match result saw Catania relegated, but the two extra points from a victory would have kept them safe. Catania was initially awarded a 2-0 win but the win was reverted. In August 2003, the FIGC decided to let Catania, along with Genoa and Salernitana stay in Serie B, and the newly re-instated Fiorentina were also included in the 2003-04 season. These decisions took effect in the following season 2004-05 and Serie B was reduced to 22 teams, while at the same time Serie A expanded from 18 to 20 teams. Since then, Italian Serie A has comprised 20 clubs playing each other on a home and away basis and composition varies from season to season due to the system of promotion and relegation. 33 clubs have competed in Serie A during the study period. 12 teams achieved a top six position at least once⁴, with 4 teams that featured in all 10 seasons sharing in the championship honours⁵, and three other clubs appearing in the top six on all but one occasion⁶.

Italian football has been involved in several major scandals in its modern history because of poor practice in corporate governance and administration (Agnew, 2007; Foot, 2007; Jones, 2007; Di Meo and Ferraris, 2012). Amongst the scandals linked variously to doping, false passports, bribery and match-fixing, arguably the most damaging was the *Calciopoli* scandal which emerged shortly before the World Cup in Germany in 2006. Investigations carried out by the Italian police discovered a network of close relationships between team managers, referees, agents and club executives with the intention of affecting the final results of a number of league matches (Hamil et al., 2010). The system was settled by choosing “favourable” referees to officiate specific matches in order to improve the chances of positive results in favour of certain clubs (Boeri and Severgnini, 2011). At the heart of the system were FC Juventus’ General Director Luciano Moggi in collaboration with Pierluigi Pairetto, vice chairman of UEFA’s referees’ commission and head of FIGC’s refereeing selection, and Paolo Bergamo, co-head of the FIGC’s refereeing selection. In addition to FC Juventus, the Serie A clubs, AC Milan, ACF Fiorentina, SS Lazio and Reggina Calcio, and the Serie B club, Arezzo, were also involved in the scandal through some of their club officials. The final punishments awarded after an appeal are set out in Table 1. In addition to these club level punishments, several club officials were also banned from taking part at any level of Italian football for specific periods (Hamil et al., 2010).

¹ While Lega Pro Division 1 is made of 36 clubs divided in two subdivisions of 18 teams each, Lega Pro Division 2 is composed of 41 clubs separated by two subdivisions of 20 teams and 21 teams.

² From the 2005/06 season, if two or more teams end the season with the same number of points, the ordering is determined by their head-to-head records. If two or more teams have same total points and same head-to-head records, goal difference is decisive.

³ Before season 2011-12, four teams from the Italian Serie A had the right to compete in the UEFA Champions’ League. In 2011, Serie A was ranked 5th in the UEFA ranking that determined the loss of one place available for Serie A teams to compete in the major European club competition.

⁴Roma, Juventus, Lazio, Parma, Inter, Milan, Chievo, Udinese, Sampdoria, Palermo, Livorno, Fiorentina, Genoa, Napoli.

⁵Inter, Milan, Roma, Juventus.

⁶Genoa, Napoli, Livorno.

Table 1: Calciopoli scandal punishments

Team	Final Punishment
Ac Milan	Deduction of 8 points for the 2006/07 season Ex post deduction of 30 points 2005/06 1 home game to be played behind closed doors
Fiorentina	Deduction of 15 points for the 2006/07 season Excluded from UEFA Champions' League for the 2006/07 season 2 home games to be played behind closed doors
Juventus	Removal of 2004/05 and 2005/06 Serie A League titles Relegated to Serie B Deduction of 9 points for the 2006/07 season 3 home games to be played behind closed doors
Lazio	Deduction of 3 points for the 2006/07 season 2 home games to be played behind closed doors
Reggina	Deduction of 11 points for the 2006/07 season £68,000 fine
Arezzo	Deduction of 6 points for the 2006/07 season

In particular, FC Juventus was stripped of their 2004/05 and 2005/06 Serie A titles and was the only club to be relegated to Serie B with a 9 point deduction. The club won the Serie B Championship in 2006/07 giving them promotion back to Serie A at the first opportunity. AC Milan, ACF Fiorentina, SS Lazio and Reggina Calcio suffered respectively 8, 15, 13 and 11 point deductions in the 2006/07 Serie A Season. After having avoided a ban from the 2006/07 Champions' League, AC Milan won the European tournament in 2007. AFC Fiorentina finished sixth in Serie A in 2006/07 leading to qualification for the 2007/08 UEFA Europe Cup. Reggina Calcio was able to maintain its position in Serie A into the following season.

The Calciopoli scandal and the sums of money that circulate within Italian football are linked. During the 2011/12 season, professional football in Italy is estimated to have generated an aggregate turnover equal to approximately €2.5bn for all teams participating in professional football tournaments; i.e. 0.15 per cent of GDP (Arel et al., 2012). Together the professional soccer teams contributed a total of €1.3bn in tax revenues. Nevertheless, Italian professional football has been facing a constant and slow decline since the last decade due to clubs' rising payrolls, needed to attract worldwide football stars in a highly competitive transfer market, combined with sluggish growth in revenues compared to the other four major European football leagues (Baroncelli and Lago, 2006; Bof, et al. 2008; Boeri and Severgnini, 2012); players' salaries have accounted for 90 per cent of total costs since 2004/05. Nevertheless, the wage cost per point in Serie A varies considerably across clubs according to the presence of superstars, with top teams spending significantly more per point than the other clubs (Boeri and Severgnini, 2012). In particular, AC Milan, FC Inter Milan, AS Roma and FC Juventus on average spent €1.3m and €3.3m per point in the last five years, while the second layer teams (Fiorentina, Napoli, Lazio and Udinese) spent between €0.5m and €1.5m per point. As Boeri and Severgnini (2012) report, Italian clubs are perhaps unsurprisingly subject to a high mortality rate; 9 out of 37 teams participating in Serie A, 25 per cent of the total, declared bankruptcy in

the period from 2001 to 2011⁷. In the same period, accounting data reveals an upward trend in net losses, with an average of €250m in losses per year⁸, and a compound annual growth rate in operational losses of 7 percent⁹. The total debt of Serie A clubs has been increasing at a compound growth rate of about 9 percent per year; i.e. by over 60 per cent since 2006/07. On the revenue side, while TV and new media rights revenues are stable and represented about 56 percent of total revenues in 2010/11, gate revenues are declining; from 2001-2011, Italian stadia had an average of less than 25,000 spectators per Serie A match, the lowest among the top European leagues. As Caruso and Di Domizio (2012) argue, the declining attendance trend cannot be explained by the high price of tickets or by excess exposure of football on TV. Several episodes of violence and hooliganism occurring in the proximity, if not within Italian stadiums have negatively characterised the image of Italian football. As Babatunde et al. (2012) show, the falls in stadium attendance may also be related to the corruption highlighted by the Calciopoli investigation in 2006. The scandal exacerbated the declining trends in gate attendances and revenues leading to deteriorating balance sheets for many clubs. Revenues from match attendance declined strongly for all the teams involved and there was a negative spill over on attendance for the other teams. This was only partially offset by rising income from sales of television broadcast rights.

3. Production functions in sport: Previous research

The production function for any firm represents the technical relationship between inputs used in production and their relative contribution to output and it is also a critical aspect of management in any organisation. Rottenberg (1956) was the first to recognise that sport teams as other enterprises provide a product, victory or success, by employing and combining various inputs, the skills and other characteristics of the team. The first study to estimate a production function of this kind for team performance was conducted by Scully (1974). He compared salaries and players' marginal revenue product in the MLB in order to evaluate the extent of monopsonistic exploitation. The methodology explicitly related sports' team output to team input measures. Since this seminal work there has been a tradition of production function analysis in the economics of sports.

The subsequent research on team production functions has estimated production functions for baseball, soccer, cricket, American professional and college football, rugby league, and basketball (e.g. Zech, 1981; Carmichael et al., 2000; Schofield, 1988; Atkinson et al., 1988; Carmichael and Thomas, 1995; Scott et al., 1985). Studies have used estimates of sporting team production functions to examine whether managers' turnover is consistent with labour market theories of matching (Borland and Lye, 1996) incentive effects on player performance (Krautmann, 1990) and to assess managerial efficiency (Dawson et al., 2000). Apart from the context of the sport selected and the objective of the study, estimated production functions can be also differentiated by functional form, estimation method, selection of output and input measures and the time frame.

For sports such as European football, points won rather than win percentage, is usually considered a more appropriate output measure because of drawn matches and season-long tournaments (Schofield, 1998; Dawson et al., 2000). However, a range of measures of output have been used including in addition to points won: league position, win rates, and goals or goals difference (Espitier-Escuer and Garcia-Cebrian, 2004; Barros and Leach, 2006 a, b; Gerrard, 2006; Hofler and Payne, 2006).

Input measures have traditionally been selected to reflect factors such as player ability, skills and personal and physical characteristics. Most of the early research is concentrated on North American sports as these provide accurate measures of inputs and outputs through detailed and specific performance statistics for the NBA (Zak et al. 1979; McCormick and Clement 1992; Chatterjee et al. 1994) and for the NFL (Atkinson et al. 1988). The first two studies to investigate production functions for European team sports were Schofield (1988) and Carmichael and Thomas (1995) respectively for English country cricket and

⁷ These teams are, in alphabetical order: Ancona, Como, Fiorentina, Messina, Perugia, Piacenza, Torino, Treviso and Messina.

⁸ Loss peaks were above €300m in 2002/03, and 2010/11.

⁹ A trend at which losses double in about 10 years.

rugby league football. Carmichael *et al.* (2000) were the first to estimate a production function for association football.

In contrast to North American sports, cricket and rugby, Football's interactive nature and the limited number of set plays do not facilitate decomposition, record, and measurement. In the production function studies of football, the inclusion of match-play inputs has varied considerably including various measures of attacking and constructive plays, aggressive and defensive plays and non-playing inputs including managerial inputs. More recently, Carmichael *et al.* (2010) investigate the production function as part of a wider empirical investigation of the relationship between playing success and commercial success in English Premier League.

To summarise, the literature on sporting production functions is wide but, to date, the Italian Serie A has not been analysed in depth. This study addresses this imbalance and also builds on previous studies by incorporating a rich set of direct performance measures. This avoids having to use proxy measures to represent particular aspects such as defensive performance. The analysis also has the advantage of covering a relatively long time period, 10 seasons. This increases the number of observations as well as the number of clubs involved.

4. Data and empirical specification

Team success, whether match specific or cumulative, ultimately depends on winning performances, which by the nature of association football is reflected by points won and depends on positive goal difference. In any match the number of goals scored is basically determined by the number of effective attacking moves, also involving passing play and associated ball possession, culminating in shots on goal in combination with the opposition's defensive plays manifested in several ways. In a parallel way, the number of goals conceded is determined by a combination of defensive skills and opponent attacking performance. Some plays such as shot-taking are unambiguously attacking in nature and others are more specifically defensive, some also lend themselves to more ambiguous interpretation, as discussed in Carmichael *et al.* (2000, 2001).

The estimating model is based on a system of two behavioural equations:

$$\text{League success}(S)=S(\text{Goals For and Against, Team Specific Attributes, Manager Change}) \quad (1)$$

and

$$\text{Goal Difference } (G)=G(\text{Shot Making and Assists, Constructive Play, Defensive Play}) \quad (2)$$

Equation 1 models league success as a simple function of goals plus team specific influences that give a team its individuality and managerial change. Equation 2 states that goals depend on a combination of attacking, other constructive and defensive performance. In addition we estimate the reduced form of (1);

$$\text{League Success } (S) = S(\text{Attacking Play, Constructive Play, Defensive Play, Team Specific Attributes, Manager Change}) \quad (3)$$

League success in the reduced form estimation measured by the number of points won as a percentage of the maximum winnable over the season, POINTS%. To capture the impact of the Calciopoli scandal we estimate additional versions of equation (3). In the first of these the dependent variable incorporates the points deducted in 2006/07 as a result of the scandal (POINTS%_deduc_06/07). Our hypothesis is that the relationship between underlying performance and success should be less well captured in this estimation as this measure of success is in some sense artificial. In the second alternative estimation the dependent variable (POINTS%_deduc_05/06) is calculated with the points deduction subtracted in 2005/06 rather than 2006/07. This variable uses the points deducted in 2006/07 as measure of the gains from fraudulent behaviour in 2005/06 and subtracts these to construct a 'truer' measure of success. The hypothesis here is that the relationship between underlying performance and success will be better captured in this estimation.

In the remaining estimations the dependent variable is unchanged but either a dummy variable for the teams implicated in Calciopoli or a variable interacting implicated clubs with actual points deducted is included as an independent variable. We consider timing in relation to these impacts by interacting with either the 2005/06 season (when the fraudulent behaviour took place and implicated teams were by assumption advantaged) or the 2006/07 season (when the implicated teams were punished). Our aim in these estimations is to examine whether (i) implicated teams gained any ‘unfair’ advantage in 2005/06 and; (ii) whether their performance was in any way affected by their punishment in 2006/07.

The data set contains aggregated club level data over 10 seasons for a varied mix of performance indicators as summarised in Table 2. Although there are 10 seasons of data, the number of observations is still quite limited because the number of observations in any one season is at most 20. In order to utilise the dataset more effectively it is common practice to reduce the number of independent variables by constructing composite variables to reflect implicitly latent and unobserved aspects of overall playing performance. This method increases degrees of freedom and reduces problems linked to multicollinearity. Traditionally, researchers have used their knowledge of the sport in question to construct composite measures of performance. This process inevitably involves an element of subjective judgement in the weighting of the components. In this paper we remove this element of subjectivity by using principal components factoring to construct the composite measures (Kim and Mueller, 1978; StatSoft, 2012; Torres-Reyna, 2012). This process uses the factor loadings following the default varimax rotation in STATA 10 to create the composite variables. The factors were extracted from the available performance measures and reflect all attacking, shot taking, other constructive and defensive play (as per Table 3). For each set of variables, only the factor that accounted for the largest percentage of the total variance was extracted (in each case the selected factor or principal component satisfied the Kaiser (1958) criterion, having an eigenvalue greater than 1). The four new composite variables are: ATTACK, SHOTS, CONSTRUCTIVE and DEFENCE (Table 3).

The team-specific characteristics that provide a team with its individual qualities are by assumption difficult to capture. In this analysis we attempt to measure some of these features by the measures listed in the lower section of Table 1. The team specific playing quality of a club is likely to be reflected to some extent by their achievements in the previous season and this is measured by points won and qualification for participation in European club competitions. The degree of reliance on home performance is an alternative negative indicator of playing quality. The wage bill at a club is another indicator of team playing ability but wages are potentially endogenous given that success in the league is the source of funding for players’ wages. Managerial change is controlled for by including a dummy variable recording whether or not there was such a change in the season concerned. However, there are also potential endogeneity issues linked to the inclusion of managerial changes in independent variable given that poor performance is likely to lead to the replacement of the manager. Definitions of all the variables used in the analysis are provided in Tables 2-3. That the two rankings are strongly correlated is not surprising. However, there are also some interesting differences between the ranks in the two columns. For example, Roma appears to have made more efficient use of its resources than its average league rank, as have Palermo and Parma. Interestingly Milan appears to have had the potential to achieve a higher overall ranking than third while Inter has achieved second place overall in an equally efficient way.

Table 2: Definitions and summary statistics for variables used in the estimations

Variable	Definition	Obs.	Mean	St. Dev.	Max.	Min.
Measures of output/league success						
GOAL_DIF	Goal difference	192	0	20.409	-49	54
POINTS%	Points total as % share of maximum possible	192	45.205	14.105	85.08	12.74
POINTS%_deduc_06/07	Points total less deductions in 2006-7 as % share of maximum possible	192	45.031	14.025	85.08	12.74
POINTS%_deduc_05/06	Points total less 2006-7 deductions subtracted 2005-6	192	45.032	14.013	85.08	12.74

	as % share of maximum possible					
Measures of playing performance						
Attacking play						
GOALS	Goals scored	192	47.697	12.595	85	21
SHOTS	Total shots attempted	192	476.989	78.697	755	326
SHSTARG	Total shots on target	192	180.25	36.505	393	99
ASSIST	Passes leading directly to goals scored	192	30.322	9.484	56	10
CONSATTRat	Ratio of constructive attacking play (sum of goals scored, assists and shots on target) to total shots	192	.535	.068	.688	.219
OFSDE	Offsides	192	114.099	24.789	64	180
Other constructive play						
CRCOMP	Total crosses completed received by own team player	192	136.0781	26.93154	212	73
CROSS	Total crosses made	192	612.031	97.95	405	923
CROSSrat	Ratio of crosses completed to crosses made	192	.222	.024	.15	.298
PSCOMP	Total passes completed received by own team player	192	11366.15	1943.095	18095	3265
TOTCH	Total balls touched	192	19562.14	2052.64	27007	13577
TOUCHN	Net total balls touched	192	7893.2	984.6	15084	3474
DРИBuse	Total useful dribbles	192	302.791	68.961	507	151
Defensive and aggressive play						
TACKLES	Total tackles made	192	719.145	81.525	989	493
CLEARS	Total clearances made	192	147.531	33.194	291	69
OPOFSD	Opponents offsides	192	114.718	38.141	244	49
INTERCEP	Total interceptions made	192	3679.495	274.973	4613	3082
ANTICIPA	Total anticipations made	192	625.45	104.822	378	1057
RECOVER	Recovered balls	192	5839.016	323.47	5024	6859
EF_REC	Effective recovered balls	174	3750.345	219.97	3243	4347
TEF_REC	Team effective recovered balls	174	2050.506	162.609	1638	2512
RUNS_REC	Runs after recovering ball	177	1069.458	105.671	836	1388
GC/GKSV	Ratio of goals conceded to keeper saves	192	.4008	.089	.656	.206
GKSV	Total saves made by goalkeeper	192	119.854	19.776	174	55
GKCT	Goalkeeper catches	192	263.11	38.78	167	350
GKDIS	Goalkeeper distribution	177	493.475	79.225	279	737
GOALSCon	Goals conceded	192	47.364	10.685	73	19
YC	Yellow cards	192	78.906	17.709	118	10
RC	Red cards	192	5.755	2.665	17	1
FOULS	Total fouls committed conceded	192	713.416	80.859	924	511
Team specific non-playing inputs/measures						
POINTS_Prev	Points won previous season as a percentage of total league points	192	5.824	1.406	3.14	9.903
HOME	Ratio of home points to away points	192	1.951	1.1406	12	0.739
CHMAN	Change of manager during season	192	.3906	.489	1	0
PROM	Promoted club	192	.192	.395	1	0
CHAML	Club competing in UEFA Champions League	192	.213	.4108	1	0
UEFACUP	Club competing in UEFA Cup	192	.156	.364	1	0
CPI_2006/7	Dummy variable interacting club receiving points deduction in 2006-7 with 2005-6 season (season previous to that in which	192	0.026	0.160	1	0

	points deducted) Interaction variable: season 2006-7 with points deducted in season 2006-7	192	0.198	1.470	15	0
CPI_pnts2006/7	Interaction variable: season 2005-6 with points deducted in season 2006-7	192	0.198	1.470	15	0

Table 3: Composite variable created using factor analysis

ATTACK	Composite variable created from STATA generated factor loadings for the factor explaining most variance among attacking performance variables (eigenvalue=4.85; percentage of variance explained = 69.26; max value=3.15 (Roma) (a); min value=-2.09 (Piacenza)). Cronbach's alpha =0.71
SHOTS	Composite variable created from STATA generated factor loadings for the factor explaining most variance among measures of shot taking/making performance variables (eigenvalue=3.27; percentage of variance explained = 65.30; max value=3.203 (Roma) (a); min value=-2.069 (Piacenza)). Cronbach's alpha =0.60
CONSTRUCTIVE	Composite variable created from STATA generated factor loadings for the factor explaining most variance among variables measuring constructive passes, crosses and dribbles (eigenvalue = 5.26; percentage of variance explained = 62.22; max value=3.40 (Milan); min value= -2.64 (Ancona)). Cronbach's alpha = 0.64
DEFENCE	Composite variable created from STATA generated factor loadings for the factor explaining most variance among defensive performance variables (eigenvalue=3.92; percentage of total variance explained=21.78; max value=3.15 (Atalanta) min value= -2.42 (Bari). Cronbach's alpha = 0.69

Stochastic frontier models were used to estimate the production functions represented by equations 1-3. The *xfrontier* command in Stata was used to model the longitudinal features of the data. For comparison, we also estimated fixed and random effects models. However the results did not vary substantially and therefore for ease of interpretation we only present the results from the frontier estimations. These are summarised in tables 4 - 6.

5. Results

In Table 4 estimation 1 (corresponding to equation 1) the dependent variable is POINTS% and the independent variables in the measure of goal difference. As expected this relationship is positively significant. We also estimated an alternative version of equation 1 including separate measures of goals scored and goals conceded. In this estimation the absolute marginal effects of goals scored (positive) and conceded (negative) were equal suggesting that overall success in the league was determined as much by attacking as defensive performance. However, the positive significance of goal difference indicates that winning by a larger margin is conducive to league success overall. The effect of previous season's points is positive which in itself is indicative of both historic club specific effects and competitive imbalance. A change of manager is linked negatively to league success, however, as noted; the direction of causality between playing success and a managerial change is unlikely to be uni-directional. Home reliance is insignificant.

In Estimation 2 the dependent variable is GOAL_DIF. The included composite variables measuring shot making performance (SHOTS) and other constructive (CONSTRUCTIVE) performance are both positively significant. The composite measure of defensive performance (DEFENSIVE) has no direct impact on goal difference. As in estimation 1, previous season's points are positively significant and CHMAN is significantly and negatively related to goal difference. In contrast to estimation 1, HOME is negatively significant in estimation 2. As an alternative to the stochastic frontier model we additionally estimated

equation 2 using the Tobit estimator since GOAL_Diff is truncated, however, the results were similar and not presented.

Table 4: Stochastic frontier estimates of equations (1)-(3)

Dependent variable	POINTS%	GOAL_DIF	POINTS%	POINTS% _deduc _06/07	POINTS% _deduc _05/06
Independent variable	(1)	(2)	(3)	(4)	(5)
GOAL_DIF	0.642*** (0.011)				
ATTACK			3.294*** (0.900)	3.172*** (0.902)	3.152*** (0.875)
SHOTS		3.825*** (1.331)			
CONSTRUCTIVE			4.576*** (1.754)	2.949*** (1.081)	2.642** (1.108)
DEFENCE			-0.800 (0.954)	-1.070* (0.614)	-1.047* (0.619)
POINTS_Prev	0.323** (0.124)	1.901** (0.895)	1.351** (0.573)	1.174** (0.569)	1.466** (0.548)
CHMAN	-0.874*** (0.229)	-8.367*** (1.214)	-6.001*** (0.811)	-5.870*** (0.811)	-5.795*** (0.792)
HOME	0.060 (0.42)	-3.511*** (1.154)	-2.236*** (0.768)	-2.151*** (0.773)	-2.198*** (0.752)
Constant	43.675*** (0.826)	19.836 (31.492)	58.220 (36.148)	60.202** (29.107)	57.392 (28.666)
Number of observations	192	159	159	159	159
Number of groups	36	31	31	31	31
Log likelihood		-609.131	-544.545	-545.813	-541.388
Wald χ^2	8288.64***	92.73***	119.43***	105.81***	122.63***

Notes: Reported figures are coefficients (marginal effects). Figures in parenthesis are standard errors. ***, **, *: significant at 1%, 5%, 10% levels.

Estimation 3 is the reduced form of equation 1. The dependent variable measuring league success is POINTS% and the independent variables include the composite measures of performance, ATTACK, CONSTRUCTIVE and DEFENSIVE. The variables capturing attacking performance and other constructive performance are both positively and significantly related to POINTS%. The composite variable measuring defensive performance is negatively and (weakly) significantly related to POINTS%. The larger absolute size of the coefficient on the attacking measure relative to the measure of defensive performance suggests that attacking play is a more important determinate of league success overall. Previous season's points are negatively significant but weakly so and as in estimation 2, both a change of manager and reliance on home performance are negatively related to league success.

Estimations 4 and 5 are the alternative specifications in which we explore the impacts of Calciopoli through a transformed dependent variable. In estimation 4 the dependent variable is POINTS%_deduc_06/07 and the results are very similar to those in estimation (3). However, the overall significance of the estimation is higher as reflected in the Wald and Log likelihood statistics, the smaller standard errors of some of the independent variables but the increased significance of the constant. This is consistent with the hypothesis that the transformed dependent variable, incorporating as it does the artificiality of the points deduction, provides a less accurate representation of performance. In contrast, in estimation (5) where the dependent variable is POINTS%_deduc_05/06, the overall significance of the estimation is somewhat increased as is the significance of some of the included variables. This is consistent with the hypothesis that this measure of success is a better reflection of underlying performance and that the implicated clubs gained points unfairly in 2005-6.

Table 5 shows the alternative estimations in which we explore the impacts of *Calciopoli* by including additional independent variables. Estimation 6 includes the variable CPI_2005/6 and equation 7

includes CPI_pnts2005/6. Both variables are significant and positive, although the former only weakly so, indicating that clubs implicated in the scandal and subsequently punished in 2006/07 gained from their behaviour in the previous season. The stronger significance of CPI_pnts2005/6 suggests that these gains were proportionate to the punishment inflicted. Interestingly CPI_pnts2006/7 is not significant in estimations 8, suggesting that the subsequent punishment had no impact on actual performance (excluding CPI_pnts2005/6 does not affect this result).

Table 5: Stochastic frontier estimates of reduced form production function (equ. (3))

Dependent variable	POINTS% (6)	POINTS% (7)	POINTS% (8)
Independent variable			
ATTACK	3.297*** (0.886)	3.150*** (0.876)	3.070*** (0.872)
CONSTRUCTIVE	2.745** (1.070)	2.863*** (1.044)	2.688** (1.059)
DEFENCE	-1.166* (0.609)	-1.202** (0.601)	-1.192** (0.598)
POINTS_Prev	1.458** (0.564)	1.470*** (0.552)	1.381** (0.548)
CHMAN	-5.835*** (0.807)	-5.790*** (0.796)	-5.718*** (0.791)
HOME	-2.163*** (0.761)	-2.198*** (0.752)	-2.154*** (0.749)
CPI_2005/6	5.724* (3.285)		
CPI_pnts2005/6		0.896** (0.351)	0.944*** (0.354)
CPI_pnts2006/7			0.389 (0.354)
Constant	57.394* (29.651)	57.355** (28.473)	58.347** (23.966)
No. of observations	159	159	159
No. of groups	31	31	31
Log likelihood	-543.051	-541.380	-540.788
Wald χ^2	126.19	133.42	133.74

Notes: Reported figures are coefficients (marginal effects).

Figures in parenthesis are standard errors. ***: significant at 1%, **: significant at 5%, *: significant at 10% levels.

Lastly we use the results to examine efficiency in the Serie A. Measurement of efficiency in professional team sports has been explored widely in the sport economic literature using a range of techniques: for the MLB (Horowitz, 1994a, 1994b; Porter and Scully, 1982; Ruggiero et al. 1996), for the NFL (Hadley et al. 2000), the NBA (Zak, 1979) the MLS (Haas, 2003); and for European football (Dowson et al., 2000; Carmichael et al. 2000; Carmichael et. al., 2010). Here we use the inefficiency terms (u_i) in the frontier estimations of the reduced form equation 3 to rank the Serie A clubs in terms of their performance in the league over the 10 seasons. These rankings are shown beside overall league rankings in Table 6. That the two rankings are strongly correlated is not surprising. However, there are also some interesting differences between the ranks in the two columns. For example, Roma appears to have made more efficient use of its resources than its average league rank, as have Palermo and Parma. Interestingly Milan appears to have had the potential to achieve a higher overall ranking than third while Inter has achieved second place overall in an equally efficient way.

Table 6: League efficiency ranks

Club	Overall League Ranking by mean POINTS%	Efficiency Ranking from equations (3) Table 3
Roma	4	1
Inter	2	2
Palermo	7	3
Parma	11	4
Atalanta	17	5
Sampdoria	9	6
Udinese	10	7
Lazio	5	8
Bologna	16	9
Napoli	13	10
Brescia	14	11
Milan	3	11
Chievo	12	13
Torino	28	14
Juventus	1	15
Empoli	20	16
Cagliari	18	17
Lecce	29	18
Reggina	22	19
Siena	23	20
Livorno	26	21
Catania	21	22
Fiorentina	8	23
Perugia	15	24
Genoa	6	25
Messina	31	26
Bari	30	27
Verona	19	28
Modena	27	29
Ascoli	32	30
Piacenza	25	31
Vicenza	24	32
Ancona	36	33
Como	33	34
Treviso	34	35
Venezia	35	36
Spearman coef.		0.608***
Kendall's coef. tau-a-tau-b		0.4291-0.443***

*Rankings are the same for estimations 3 and 5

6. Summary and conclusions

This article has analysed the production function and the technical efficiency of Italian professional football in the last decade focusing on the on-the-field performance of Serie A football clubs. The analysis was performed using factor analysis to construct composite measures of playing performance into the team

production function. Specifically, defensive, offensive and constructive playing performance factors were considered in the two behavioural equations measuring team success. We also include additional team specific measures as well as indicators for managerial change and the *Calciopoli* scandal effects which deeply scarred Italian football over the last decade. Using statistical methods of optimisation, specifically stochastic frontier methods, we estimated the efficient production frontier in order to determine the most efficient Serie A clubs.

The results highlight the importance of attacking and constructive play in Serie A. In contrast to the study conducted by Boscà et al. (2009) that analysed Italian football over a much shorter period of time, our study suggests that to obtain a high ranking in Serie A, it is much more important to be offensively, rather than defensively, efficient. Additionally, the results show how important constructive playing performance is for final league position. Finally, both home reliance and changing a manager are less relevant for league success in Serie A.

Looking at the *Calciopoli* scandal, the punishments imposed on the implicated clubs do not appear to have had a significant impact on actual performance. However, the implicated clubs outperformed relative to expectations in the previous season, suggesting that they made short-term gains from their behaviour. In the longer-term, the league as a whole appears to have been negatively impacted by the *Calciopoli* scandal through lower attendances particularly at the grounds of the implicated clubs (Babatunde et al. 2012).

References

- Agnew, P. (2007) *Forza Italia*, Random House: London.
- Arel, PricewaterhouseCoopers and Federcalcio (2012) *ReportCalcio*, AREL: Roma.
- Atkinson, S. E. and Stanley, L. R. and Tschirhart, J. (1988) "Revenue Sharing as an Incentive in an Agency Problem: An Example from the National Football League", *The RAND Journal of Economics*, Vol. 19, N. 1: pp. 27-43.
- Babatunde, B., Migali, S. and Simmons, R. (2012) "Corruption does not pay: An analysis of consumer response to Italy's Calciopoli Scandal", *Working Paper*, Lancaster University, Management School.
- Baroncelli, A. and Lago, U. (2006) "Italian Football", *Journal of Sports Economics*, Vol. 7, N. 1: pp. 13-28.
- Barros, C. P., and Leach, S. (2006a) "Performance Evolution of the English Premier Football League with Data Development Analysis", *Applied Economics*, Vol. 38, N. 12: pp. 1449-1458.
- Barros, C. P., and Leach, S. (2006b) "Analyzing the Performance of the F. A. English Premier League with an Econometric Frontier Model", *Journal of Sports Economics*, Vol. 7, N. 4: pp. 391-407.
- Boeri, T. and Severgini, B. (2011) "Match rigging and the career concerns of referees", *Labour Economics*, Vol. 18, N. 3: pp.349-359.
- Boeri, T. and Severgini, B. (2012) The decline of professional football in Italy, *Discussion Paper Series*, Forschungsinstitut zur Zukunft der Arbeit, No. 7018, <http://hdl.handle.net/10419/67318>.
- Bof, F., Montanari, F. and Silvestri, G. (2008) *Il Management del Calcio*, Franco Angeli Editore: Milano.
- Boscà, J. E., Liern, V., Martínez, A. and Sala, R. (2009) "Increasing offensive or defensive efficiency? An analysis of Italian and Spanish football", *Omega: The International Journal of Management Science*, Vol. 37, N. 1: pp. 63-78.
- Borland, J. and Lye, J. (1996) "Matching and Mobility in the Market for Australian Rules Football Coaches?", *Industrial and Labour Relations Review*, Vol. 50: pp. 143-158.
- Carmichael, F., Thomas, D. (1995) "Production and Efficiency in Team Sports: An Investigation of Rugby League Football", *Applied Economics*, Vol. 27, pp. 859-869.
- Carmichael, F., Thomas, D. and Ward, R. (2000) "Team performance: The case of English Premiership football", *Managerial and Decision Economics*, Vol. 21: pp. 31-45.
- Carmichael, F., Thomas, D. and Ward, R. (2000) "Production and efficiency in Association football", *Journal of Sports Economics*, Vol. 2, N. 3: pp. 228-243.
- Caruso, R. and Di Domizio, M. (2012) "Hooliganism and football demand in Italy. Evidence for the period 1962-2011", *DISCE - Quaderni dell'Istituto di Politica Economica*, Ispe 062, Università Cattolica del Sacro Cuore, Dipartimenti ed Istituti di Scienze Economiche (DISCE).

- Chatterjee, S., Campbell, M. R. and Wiseman, F. (1994) "Take that jam! An Analysis of Winning Percentage for NBA Teams", *Managerial and Decisions Economics*, Vol. 15: pp. 521-555.
- Dawson, P., Dobson, S. and Gerrard, B. (2000) "Estimating Coaching Efficiency in Professional Team Sports: Evidence from English Association Football", *Scottish Journal of Political Economy*, Vol. 47, N. 4: pp. 399-421.
- Di Meo, S. and Ferraris, G. (2012) *Il Pallone Criminale*, Ponte alle Grazie: Milano.
- Espitier-Escuer, M. and Garcia-Cebrian, L. (2004) "Measuring the Efficiency of Spanish First-Division Soccer Teams", *Journal of Sports Economics*, Vol. 5: pp. 329-346.
- Foot, J. (2007) *Calcio: A History of Italian Football*, updated edition, Harper Perennial: London.
- Gerrard, B. (2006) "Analysing the Win-Wage Relationship in Pro Sports Leagues: Evidence from the FA Premier League 1997/98 – 2001/02?", in P. Rodriguez, S. Kesenne and J. Garcia (ed.), *Sports Economics After Fifty Years: Essays in Honour of Simon Rottenberg*, Ediciones de la Universidad de Oviedo, Oviedo.
- Haas, D. J. (2003) "Technical Efficiency in the Major League Soccer", *Journal of Sports Economics*, Vol. 4, N. 3: pp. 205-215.
- Hadley, L., Poitras, M., Ruggiero, J. and Knowles, S. (2000) "Performance Evaluation of National Football League clubs", *Managerial and Decision Economics*, Vol. 21: pp. 63-70.
- Hamil, S., Morrow, S., Idle, C., Rossi, G. and Faccendini, S. (2010) "The governance and regulation of Italian football", *Soccer and Society*, Vol. 11, N. 4: pp. 373-413.
- Horowitz, I. (1994a) "Pythagoras, Tommy Lasorda, and Me: An Evaluating Baseball Managers", *Social Science Quarterly*, Vol. 75: pp. 187-194.
- Horowitz, I. (1994b) "On the Manager as Principal Clerk", *Managerial and Decision Economics*, Vol. 15: pp. 413-419.
- Jones, T. (2007) *The Dark Heart of Italy*, Faber & Faber: London.
- Kaiser, H. F. (1958) "The varimax criterion for analytic rotation in factor analysis", *Psychometrika*, Vol. 23: pp. 187-200.
- Kim, J. and Mueller, C. (1978) *Factor Analysis: Statistical methods and practical issues*, Sage Publications: Beverly Hills, Calif.
- Krautmann, A. C. (1990) "Shirking or Stochastic Productivity in Major League Baseball?", *Southern Economic Journal*, Vol. 56, pp. 567-579.
- McCormick, R. E. and Clement, R. C. (1992) "Intra-firm Profit Opportunities and Managerial Slack Evidence from Professional Basketball" in Scully, G. W. (ed.), *Advances in the Economics of Sports*, JAI Press, Greenwich.
- Porter, P. K. and Scully, G. W. (1982) "Measuring Managerial Efficiency: The Case of Baseball", *Southern Economic Journal*, Vol. 48: pp. 642-650.
- Rottenberg, S. (1956) "The Baseball Players' Labour Market", *Journal of Political Economy*, Vol. 64, pp. 242-258.
- Ruggiero, J., Hadley, L. and Gustafson, E. (1996) "Technical Efficiency in Major League Baseball", in Fizel, J., Gustafson, E. and Hadley, L. (ed.), *Sports Economics: Current Research*, Praeger, Westport.
- Schofield, J. A. (1988) "Production Function in the Sports Industry: An Empirical Analysis of Professional Cricket", *Applied Economics*, Vol. 15, pp. 283-296.
- Scott, F. A. J., Long, J. E. and Somppi, K. (1985) "Salary Vs. Marginal Revenue Product under Monopsony and Competition: The Case of Professional Basketball", *Atlantic Economic Journal*, Vol. 13, N. 3: pp. 50-59.
- Scully, G. W. (1974) "Pay and Performance in Major League Baseball", *American Economic Review*, Vol. 64, N. 6: pp. 915-930.
- Zak, T. A., Huang, C. J. and Sigfried, J. J. (1979) "Production Efficiency: The Case of Professional Basketball", *Journal of Business*, Vol. 53: pp. 379-302.
- Zech, C. E. (1981) "An Empirical Estimation of a Production Function: The Case of Major League Baseball", *American Economist*, Vol. 25: pp. 19-23.

Model of joint displacement using sigmoid function. Experimental approach for planar pointing task and squat jump

T. Creveaux*, J. Bastien, C. Villars and P. Legreneur

*Université de Lyon, CRIS, 27-29 Bd du 11 Novembre 1918, 69622 Villeurbanne Cedex, France. thomas.creveaux@univ-lyon1.fr

Abstract. Using an experimental optimization approach, this study investigated whether two human movements, pointing tasks and squat-jumps, could be modeled with a reduced set of kinematic parameters. Three sigmoid models were proposed to model the evolution of joint angles. The models parameters were optimized to fit the 2D position of the joints obtained from pointing tasks and 120 squat-jumps. The models were accurate for both movements. This study provides a new framework to model planar movements with a small number of meaningful kinematic parameters, allowing a continuous description of both kinematics and kinetics. Further researches should investigate the implication of the control parameters in relation to motor control and validate this approach for three-dimensional movements.

1. Introduction

Quantitative analysis of human movement usually relies on the time history of reflective markers fixed to anatomical landmarks obtained from optical systems. These raw data are further used to compute relevant parameters such as velocities, accelerations, moments or powers. During the recent years, the performance of acquisition systems greatly increased, especially considering acquisition rate and accuracy. However, raw data still remain noisy, due to the movement of the skin with regard to the bones and finite accuracy of such systems. Furthermore, the effect of noise increases as the data is derived with respect to time, which is a very common task in movement analysis.

To overcome the above issues, raw data are quite always smoothed or filtered, resulting in well-known decrease of movement amplitude. Specific filtering methods accounting for properties of the skeletal system such as constant length of the limbs have been used but such approaches still suffer from the motion of the markers relatively to the skeletal system. An interesting feature of human motion is the necessity for decelerating the joint displacement before its maximal amplitude (anatomical constraint) in order to protect this joint from any damage (van Ingen Schenau, 1989). Regarding to kinematics, the anatomical constraint implies that joint angular time history should match an asymmetric sigmoid shape (Zelaznik et al., 1986) and thus an asymmetric bell-shaped velocity profile (Soechting and Lacquaniti, 1981), which accounts for synergistic actuators' activations at a joint, i.e. agonist and antagonist muscle-tendon systems. In the field of human movement analysis, Plamondon proposed a model of asymmetric sigmoid (Plamondon, 1995, 1998; Plamondon et al., 2003). However, the velocity is not null at the end of the movement so that the anatomical constraint is not satisfied.

Therefore, this study aimed at modeling two different movements, i.e. a pointing task and an explosive movement, the squat-jump, using a generic model of sigmoidal joint displacement based on meaningful kinematic parameters which accounts for the anatomical constraint. Three sub models were used to achieve best fitting of experimental data obtained from both movements (Creveaux et al., 2012).

2. Methods

2.1 General model of joint displacement

Accounting for a monotone evolution of a given angle and considering the anatomical constraint requirements, it is assumed that each angle θ is characterized by the following properties (figure 1):

- at the beginning and at the end of the movement, the velocity and the acceleration are equal to zero;
- the angle increases (respectively decreases) throughout the whole movement;
- during the movement, the velocity increases (respectively decreases) until it reaches its maximum (respectively minimum), then decreases (respectively increases).

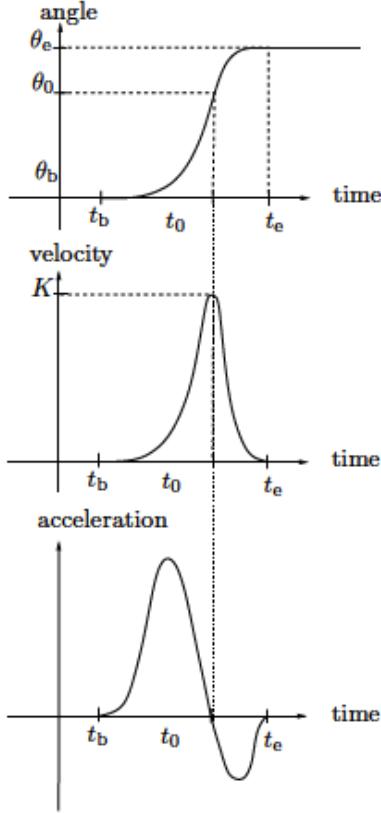


Figure 1. Shape of used sigmoid: angle, velocity and acceleration versus time (for the increasing case).

More precisely, we try to determine a function θ from $[0, T]$ to \mathbb{R} of class C^2 . Let t_b, t_0, t_e be three instants such that

$$0 \leq t_b < t_0 < t_e \leq T \quad (2.1)$$

Let $\theta_b, \theta_0, \theta_e$ be three real numbers such that

$$\theta_b < \theta_0 < \theta_e \text{ or } \theta_e < \theta_0 < \theta_b \quad (2.2)$$

We assume that

- θ is constant and equals to θ_b on $[0, t_b]$;
- θ is constant and equals to θ_e on $[t_e, T]$;
- there exists $\varepsilon \in \{-1, 1\}$ such that $\varepsilon\theta$ is strictly increasing on $[t_b, t_e]$;
- $\varepsilon\theta$ is strictly convex on (t_b, t_0) ;
- $\varepsilon\theta$ is strictly concave on (t_0, t_e) ;

We set

$$\varepsilon = \text{Sign}(\theta_e - \theta_b) \epsilon \{-1, 1\} \quad (2.3)$$

Let K be the number defined by

$$K = \begin{cases} \max \theta'(t), t \in [t_b, t_e] & \text{if } \varepsilon = 1 \\ \min \theta'(t), t \in [t_b, t_e] & \text{if } \varepsilon = -1 \end{cases} \quad (2.4)$$

Since θ is of class C^2 , we have

$$\theta(t_b) = \theta_b, \quad \theta'(t_b) = 0, \quad \theta''(t_b) = 0 \quad (2.5a)$$

$$\theta(t_e) = \theta_e, \quad \theta'(t_e) = 0, \quad \theta''(t_e) = 0 \quad (2.5b)$$

$$\theta(t_0) = \theta_0, \quad \theta'(t_0) = 0, \quad \theta''(t_0) = 0 \quad (2.5c)$$

$$\forall t \in (t_b, t_0), \quad \varepsilon\theta''(t) > 0 \quad (2.5d)$$

$$\forall t \in (t_0, t_e), \quad \varepsilon\theta''(t) < 0 \quad (2.5e)$$

We consider $\alpha, \beta \in (0, 1)$ and $k \in \mathbb{R}$ defined by

$$\alpha = \frac{t_0 - t_b}{t_e - t_b}, \quad \beta = \frac{\theta_0 - \theta_b}{\theta_e - \theta_b}, \quad k = K \frac{t_e - t_b}{\theta_e - \theta_b} \quad (2.6)$$

Applying the following change of scale,

$$\forall t \in [t_b, t_e], \quad u = \frac{t - t_b}{t_e - t_b} \in [0, 1] \quad (2.7a)$$

$$\forall u \in [0,1], \quad g(u) = \frac{\theta((t_e - t_b)u + t_b) - \theta_b}{\theta_e - \theta_b} \quad (2.7b)$$

The problem can be reformulated as follow: we search a function g of class C^2 defined on $[0,1]$ satisfying:

$$g(0) = 0, \quad g'(0) = 0, \quad g''(0) = 0 \quad (2.8a)$$

$$g(1) = 1, \quad g'(1) = 0, \quad g''(1) = 0 \quad (2.8b)$$

$$g(\alpha) = \beta, \quad g'(\alpha) = k, \quad g''(\alpha) = 0 \quad (2.8c)$$

$$\forall u \in (0, \alpha), \quad g''(t) > 0 \quad (2.8d)$$

$$\forall u \in (\alpha, 1), \quad g''(t) < 0 \quad (2.8e)$$

Remark 2.1. Under the assumptions given in (2.8), we have necessarily

$$k \geq \max \left(\frac{\beta}{\alpha}, \frac{1-\beta}{1-\alpha} \right) > 1 \quad (2.9)$$

Finally, the function θ is defined for all $t \in [0, T]$ by

$$\theta(t) = \begin{cases} \theta_b, & \text{if } t \leq t_b \\ (\theta_e - \theta_b)g\left(\frac{t-t_b}{t_e-t_b}\right) + \theta_b, & \text{if } t_b < t < t_e \\ \theta_e, & \text{if } t \geq t_e \end{cases} \quad (2.10)$$

This function is defined by 7 independent parameters:

- 2 time scale parameters (t_b and t_e);
- 2 angle scale parameters (θ_b and θ_e);
- 3 shape parameters (α, β, k).

Thus, θ can be written under the form $\theta_{t_b, t_e, \theta_b, \theta_e, \alpha, \beta, k}$. Aimed to solve the system (2.8), we have to include into the model 3 control parameters, which have to be related with α, β and k . In the next section, 3 sigmoid models, *i.e.* SYM, NORM and INVEXP are presented.

2.2 The SYM model

The SYM model was built using a pseudo-symmetry approach. Its function g is defined by $\alpha, \beta \in (0,1)$ and $\kappa > 1$. Let $g_{\alpha, \beta, k}$ be a function of class C^2 from $[0, \alpha]$ to \mathbb{R} satisfying (2.8a), (2.8c) and (2.8d). If the function g is defined from $[0,1]$ to \mathbb{R} by,

$$g(u) = \begin{cases} g_{\alpha, \beta, k}(u), & \text{if } u \leq \alpha \\ 1 - g_{1-\alpha, 1-\beta, k}(1-u), & \text{if } u > \alpha \end{cases} \quad (2.11)$$

Then, g is of class C^2 on $[0,1]$ and (2.8) holds. Considering the function $H_{a, b, \kappa}$ defined on $[0, \alpha]$ for all $a, b > 0$ and $\kappa > 2$ as

$$H_{a, b, \kappa}(u) = a(1 - e^{-bu^\kappa}) \quad (2.12)$$

a, b and κ have to be determined so that (2.8a), (2.8c) and (2.8d) hold. We set

$$r_0 = \frac{1}{e^{0.5} - 1} \approx 1.54 \quad (2.13)$$

For all $(\alpha, \beta) \in (0,1)^2$, for all k such that $k > r_0 \beta/\alpha$, there exists $(a, b, \kappa) \in \mathbb{R}_+^{*2} \times (2, \infty)$ such that (2.8a), (2.8c) and (2.8d) hold for function $H_{a, b, \kappa}$. a, b and κ still need to be defined. We set

$$\gamma = \frac{\beta}{k\alpha} \in (0, e^{0.5} - 1) \quad (2.14a)$$

It exists an unique $X \in (0.5, 1)$ such that

$$(e^X - 1) \frac{1-X}{X} = \gamma \quad (2.14b)$$

It follows

$$a = \frac{\beta}{1 - e^{-X}} \quad b = \frac{X}{\alpha^\kappa} \quad \kappa = \frac{1}{1-X} \quad (2.14c)$$

By setting $(a, b, \kappa) = G(\alpha, \beta, k)$, the function g is defined for all $u \in [0,1]$ by

$$g(u) = \begin{cases} H_{G(\alpha, \beta, k)}(u), & \text{if } u \leq \alpha \\ 1 - H_{G(1-\alpha, 1-\beta, k)}(1-u), & \text{if } u > \alpha \end{cases} \quad (2.15)$$

2.3 The NORM model

The NORM model (named from its relation to the normal law) function g is defined by 3 parameters $a \in (0,1)$, $p > 0$ and $s > 0$. As a reminder, the density function of the normal, or Gaussian distribution with mean m and variance s^2 is given by:

$$\forall x \in \mathbb{R}, f(x) = \frac{1}{s\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-m}{s}\right)^2\right) \quad (2.16)$$

Considering the erf function defined by

$$\forall x \in \mathbb{R}, \text{erf}(t) = \frac{2}{\pi} \int_0^x e^{-t^2} dt \quad (2.17)$$

The cumulative distribution function of the normal law is given by

$$\forall x \in \mathbb{R}, \Phi(x) = \frac{1}{2} \text{erf}\left(\frac{x-m}{\sqrt{2}s}\right) + \frac{1}{2} \quad (2.18)$$

For all $p > 0$

$$\forall u \in (0,1), G(u) = \ln\left(\frac{u^p}{1-u^p}\right) \quad (2.19)$$

The function g is defined by

$$\forall t \in (0,1), g(t) = \Phi(G(t)) \quad (2.20a)$$

$$g(0) = 0 \quad (2.20b)$$

$$g(1) = 1 \quad (2.20c)$$

2.4 The INVEXP model

The INVEXP model (derived from the inverse exponential) function g is defined by 3 parameters $\lambda, \mu > 0$ and $a \in \mathbb{R}$. For all a , for all λ and μ we set

$$\alpha = \frac{\lambda}{\lambda + \mu} \in (0,1) \quad (2.21)$$

and we consider the function $g_{a,\alpha}$ as:

$$g_{a,\alpha} = 1, \quad \text{if } a = 0 \quad (2.22a)$$

$$\begin{cases} \forall y \in [0, \alpha] & g_{a,\alpha}(y) = 1 - \exp\left(\frac{t}{a(t-\alpha)}\right) \\ \forall y \in [\alpha, 1] & g_{a,\alpha}(y) = 1 \end{cases} \quad \text{if } a > 0 \quad (2.22b)$$

For all $a \in \mathbb{R}$ and for all $\alpha \in (0,1)$, we consider the function $G_{a,\alpha}$ defined by

$$\begin{cases} \text{if } a \geq 0, & G_{a,\alpha} = g_{a,\alpha} \\ \text{if } a < 0, & G_{a,\alpha} = g_{-a,1-\alpha} \end{cases} \quad (2.23)$$

For all $\lambda, \mu > 0$, $f_{\lambda,\mu}$ is defined by

$$\forall t \in (0,1), f_{\lambda,\mu}(t) = \exp\left(-\frac{1}{t^\lambda(1-t)^\mu}\right) \quad (2.24a)$$

$$f_{\lambda,\mu}(0) = 0 \quad (2.24b)$$

$$f_{\lambda,\mu}(1) = 1 \quad (2.24c)$$

For all $a \in \mathbb{R}$, $\lambda, \mu > 0$, $h_{\lambda,\mu,a}$ is defined by:

$$h_{\lambda,\mu,a} = f_{\lambda,\mu} G_{a,\lambda/(\lambda+\mu)} \quad (2.25)$$

The function g is defined by

$$\forall t \in (0,1), g(t) = \frac{\int_0^t h_{\lambda,\mu,a}(u) du}{\int_0^1 h_{\lambda,\mu,a}(u) du} \quad (2.26)$$

2.5 Definition domains of the sigmoid models

Each of the 3 functions is defined by 3 parameters. For all $k > 1$, there exist a part S_k of $(0,1)^2$ such that for all $(\alpha, \beta) \in S_k$, there exist at least one sigmoid of kind g satisfying (2.8) whose parameters can be determined by splitting (2.8) in 3 non-linear equations which can be solved with a numerical solver. This part S_k differs for the 3 sigmoid models. The bigger is obtained with the INVEXP model (Figure 2).

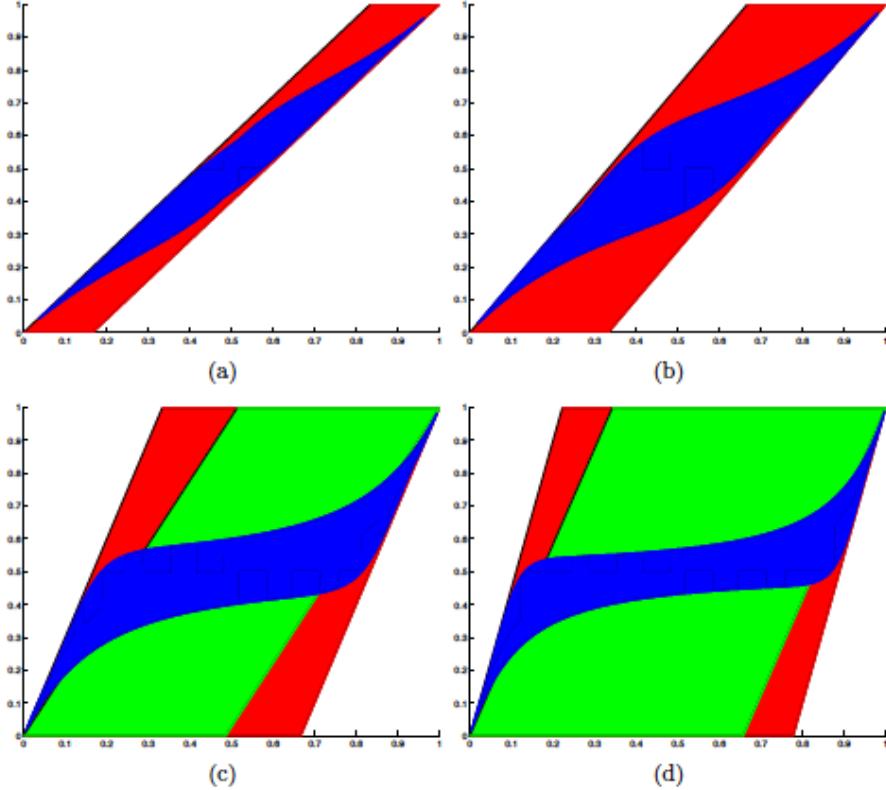


Figure 2. S_k domains of the three sigmoids for $k=1.2$ (a), $k=1.5$ (b), $k=3$ (c), and $k=4.5$ (d). INVEXP, NORM and SYM domains are plotted in red, blue and green respectively. According to (2.28), SYM domain is empty for $k=1.2$ and $k=1.5$.

Examples of position and velocity curves obtained from the three models are provided in figure 3.

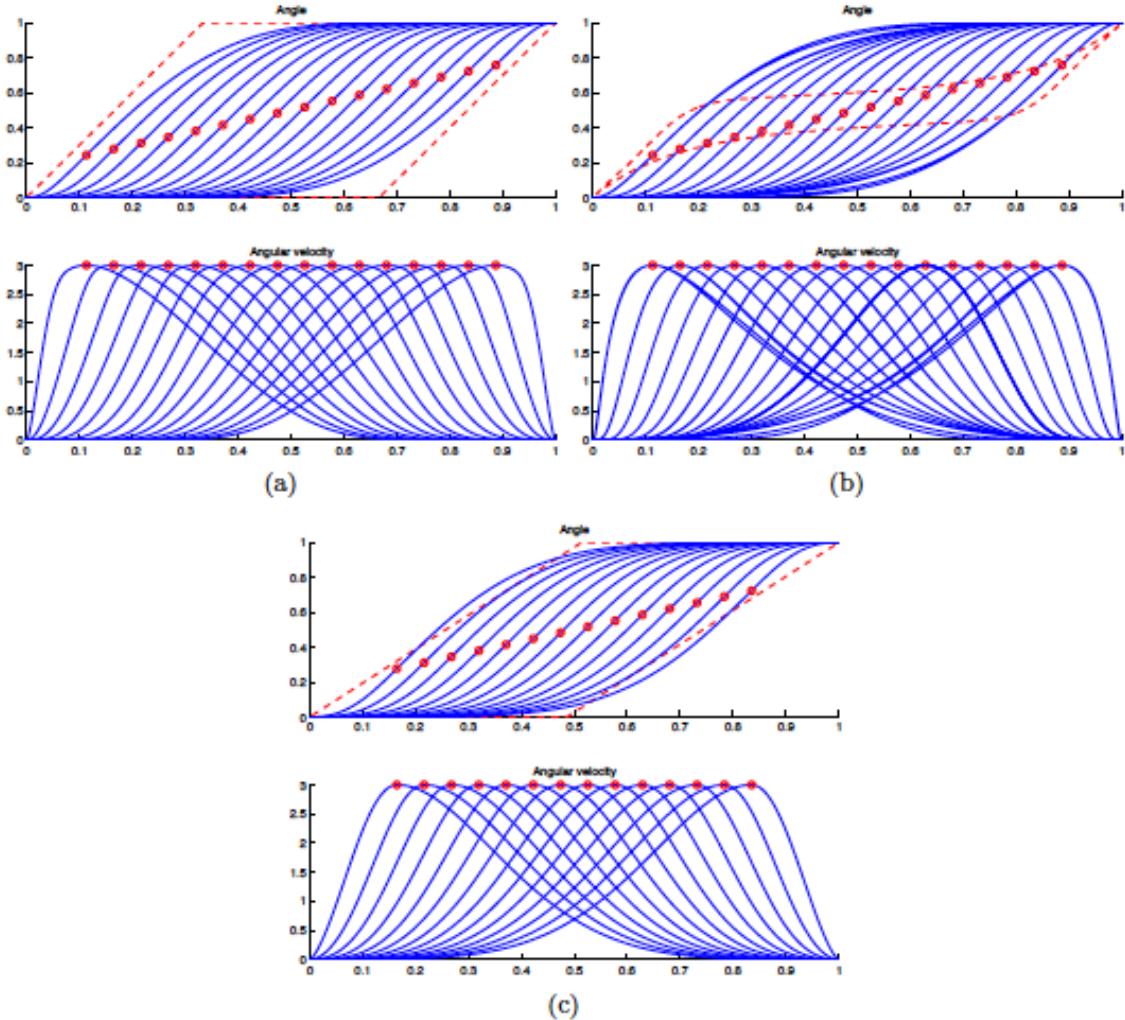


Figure 3. Examples of curves for angle and angular velocity for INVEXP (a), NORM (b) and SYM (c) models. The boundaries of domains are plotted in red dashed lines.

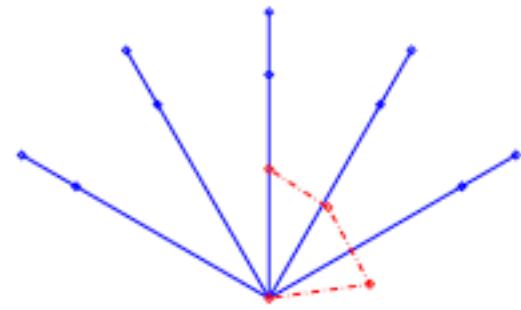
3. Experimental procedures

3.1 Pointing task

9 right-handed male subjects (age = 24.9 ± 2.42 years, height = 177.6 ± 5.83 cm and mass = 68.8 ± 8.18 kg) were asked to perform pointing tasks in the horizontal plane. The total number of pointing tasks was 304. Movements were performed for five directions and two distances (Figure 4). For each direction, two spherical targets were placed on a table at 60 and 80 cm from the shoulder. Directions of pointing task ranged regularly from 30 to 150 degrees including pointing along the antero-posterior axis. The described position of the targets ensures that each of them is located inside the subjects workspace (Figure 5) when considering a 80 cm upper limb length and the corresponding anthropometric dataset (Bastien et al., 2010). At the beginning of the movement, subjects had to position their arm so that the forefinger was located at 40 cm of the shoulder in the antero-posterior direction. During the experiment, subjects sat on a chair whose height was adjusted so that the upper limb remained in the horizontal plane while moving over the table from starting point to targets and the trunk was immobilized by using straps. In order to ensure that the upper limb remained in the horizontal plane, the subjects were instructed to keep the upper limb lying on the table during the movements. Video reflective markers were placed on the subjects at the shoulder (acromion), elbow (olecrane), wrist (middle of radial and ulnar styloid processes) and forefinger extremity to allow further modeling of the upper limb. For each target, subjects performed three movements which were filmed at 25 Hz with a numeric camera JVC © Everio placed above the subjects and oriented vertically. Raw experimental data, *i.e.* the position of the joints throughout the movement, were extracted from videographic recordings.



(a) Upper view of the task environment



(b) Targets (continuous lines) and initial arm position (dashed line)

Figure 4. Pointing task experimental procedure.

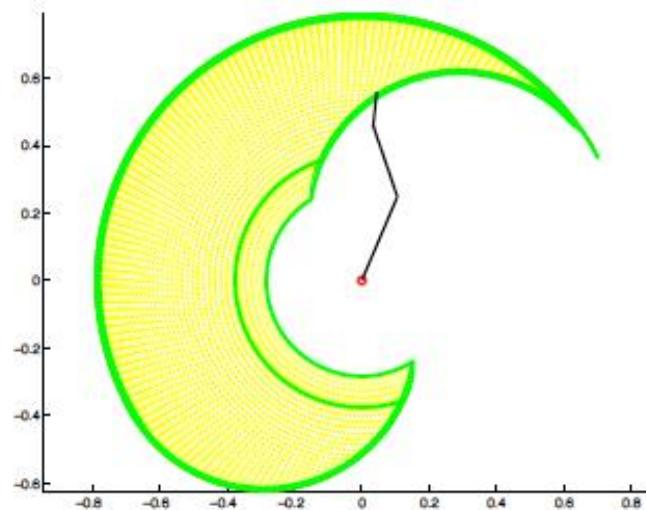


Figure 5. Upper limb workspace (adapted from (Bastien et al., 2010)).

3.2 Squat jumping

13 subjects performed 10 vertical jumps. Instructions were given for keeping the hands on the hips during the movement to limit the contribution of the upper limbs to the performance. Furthermore, subjects were asked to do no countermovement. The jumps that did not meet both of these requirements were excluded from the study. In order to model the skeleton in a 4 rigid segments system, landmarks were placed on the left fifth metatarsophalangeal, lateral malleolus, lateral femoral epicondyle, greater trochanter and acromion. These landmarks define the foot, the shank, the thigh and the upper body (Head, Arms and Trunk: HAT). The subjects were filmed orthogonally to the sagittal plane at 100 Hz and the ground reaction force was recorded at 1000 Hz from an OR6-7-2000 AMTI force plate. The center of mass (CoM) position of limbs was computed using anthropometric data (Winter, 2009). The whole body CoM (Center of Mass) position was determined on the one hand from kinematic data and on the other hand from force plate measurements using a double numerical integration procedure. For the latter, subject mass, initial body CoM position and velocity had to be set. These values were computed so that the difference between CoM path obtained from kinetic and kinematic data was minimized in a least square sense. This optimization step was also used to synchronize both recording sources.

4. Data processing

4.1 Skeletal model

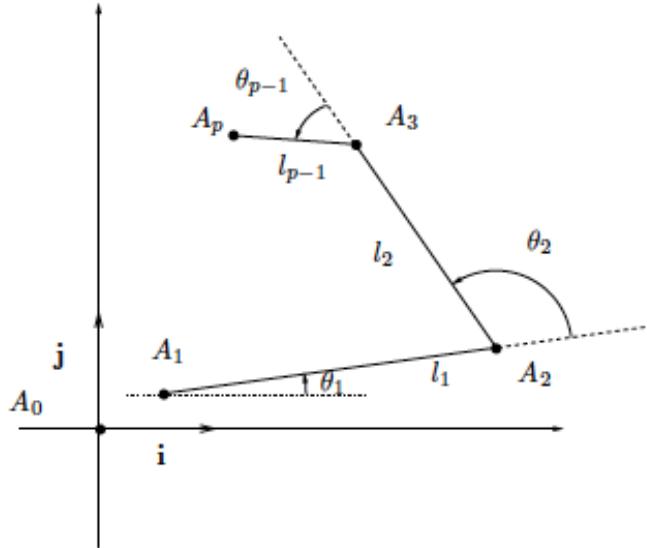


Figure 6. General geometrical representation of the skeletal model

For both tasks, the limbs were modeled as rigid bodies rotating around frictionless hinge joints. Given p limbs, the joint positions are defined by the points $A_j(x_j, y_j)$ with $j \in \{1 \dots p\}$ and $p = 3$ and $p = 4$ for pointing task and squat jump respectively. Thus, the position of any joint is given by

$$zA_j = zA_1 + \sum_{n=1}^{j-1} l_n \exp \left(i \sum_{k=1}^n \theta_k \right) \quad (4.1)$$

Where I is the imaginary unit and zA_1 is the affix of A_1 .

4.2 Determination of sigmoid parameters

Sigmoid parameters were obtained from a multi-stage optimization procedure. First, t_b , t_e , θ_b , θ_e , α , β and κ were estimated from experimental data. The scale parameters were defined so that the absolute angular velocity peak occurs between t_b and t_e and its sign changes at the endpoint of this interval. Thus, the shape parameters α , β and κ were determined tanks to (2.6).

First optimization consisted in minimizing the sum of square of differences between experimental angles and those obtained from the sigmoid models at each instant. The optimization was achieved for each sigmoid with the lsqcurvefit function provided in Matlab software. Initial values of parameters were set from estimations of experimental data described previously. This optimization stage will be further referred to as local optimization.

Secondly, differences between experimental and model reconstructed joint positions were minimized in a least square sense. Compared to the previous stage, this optimization can be considered as global since for the latter, the parameters of the sigmoids were determined simultaneously. Computation of model-based joint positions implies the lengths of the limbs to be provided. For the pointing tasks, the optimization was performed using (i) mean experimental limb lengths (semi-global optimization) and (ii) limb lengths as model parameters (global optimization).

4.3 Squat jump specific procedure

The modeling of the jump focused on the position of the joints in a reference frame located at the distal extremity of the foot. Thus, the optimization consisted in fitting the experimental joint positions of ankle, knee, hip and shoulder with the model parameters in this reference frame. Since joints do not remain fully extended after the takeoff, differences were not taken into account during the whole movement. This prevented the model from underestimating the necessary amplitude of joint extensions. Therefore, differences between experimental and model-based data were considered during the intervals corresponding to increase of vertical joint coordinates in the given reference frame (e.g. the error at the ankle joint was only taken into account while the vertical distance between the knee and the foot extremity increased).

Second stage of optimization included non-linear constraints on position, velocity and acceleration of the body CoM computed from sigmoid model. It was imposed that the body CoM position computed from both the sigmoid

model and the force plate data were similar at the instant t_1 for which the marker located on the distal extremity of the foot started to move upward. At this instant, equality for the coordinates of both velocity and acceleration of body CoM obtained from kinetic and kinematic data was also required. Finally, body CoM vertical acceleration was constrained to be greater than -9.81 m.s^{-2} before t_1 ensuring that takeoff occurs necessarily after t_1 . From t_1 to the end of the jump, the movement of A_0 was set so that kinetic and kinematic-based movement of the CoM were similar. This results in a continuous characterization of the movement position, velocity and acceleration. It should be noticed that using similar constraints for jerk and further derivatives could have led to description of class C_3 and higher.

4.4 Modeling accuracy

At each instant i of the joint j , the optimization accuracy can be quantified by the difference between experimental data (x_j^i and y_j^i) and sigmoid-modeled data (X_j^i and Y_j^i):

$$\varepsilon_{i,j} = \sqrt{(X_j^i - x_j^i)^2 + (Y_j^i - y_j^i)^2} \quad (4.2)$$

In further analysis, maximal ε_{max} and mean ε_{mean} values of these differences were used to account for the fitting accuracy of the modeling procedures.

4.5 Statistical analysis

For both maximal and mean accuracies, the Shapiro-Wilk test reported unnormal distributions. Thus, statistical tests were realized on normally distributed \log_{10} of observations (*i.e.* errors and computation time). Firstly, anovas for repeated measures were performed for errors and computation time. When anovas reported significant results, post-hoc tests were performed to check for differences between the sigmoid models and the optimization procedures. All the tests were realized with R and statistical significance was set at 95% confidence level, *i.e.* $p<0.05$.

5. Results

The results were obtained from 304 pointing tasks and 120 squat-jumps. These results are very accurate from numerical viewpoint (Figure 7). Indeed, in pointing task, in 95% of cases, for the three models, ε_{max} , was smaller than 2.894 cm. ε_{mean} was smaller than 0.824 cm. For squat jumps, ε_{max} and ε_{mean} were respectively equal to 8.492 cm and 2.882 cm.

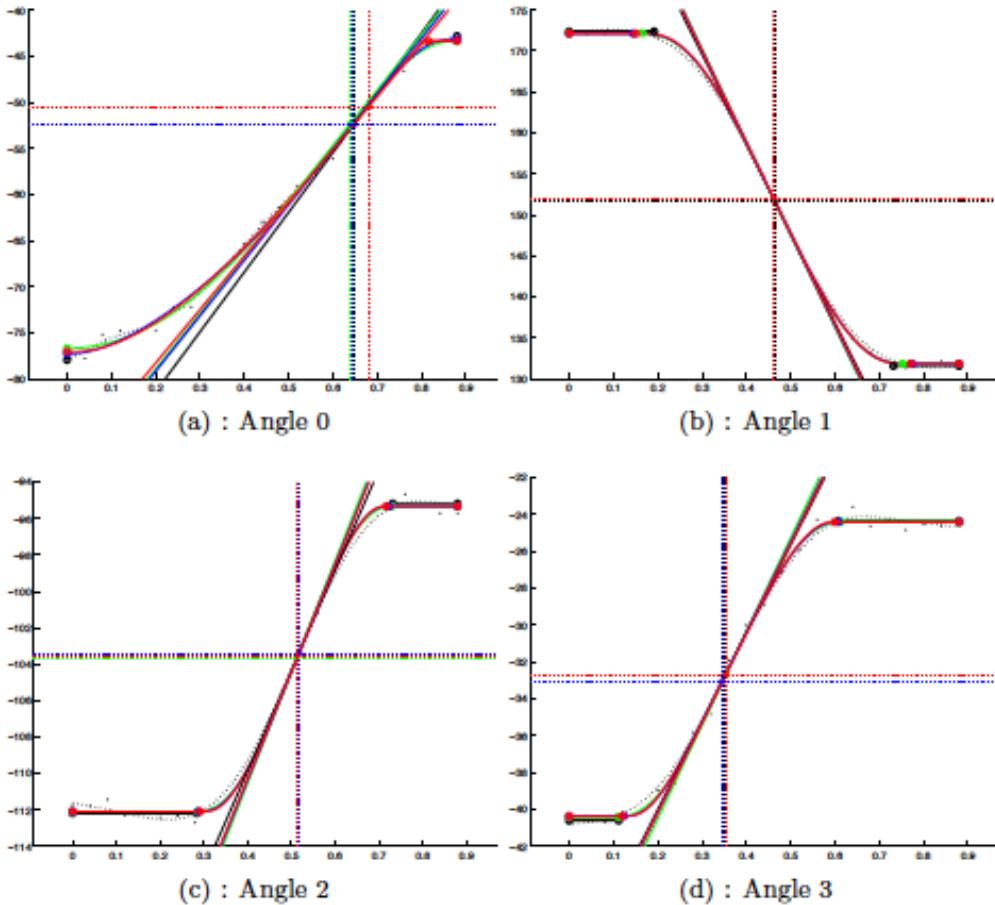


Figure 7. Time histories of joint angles in pointing task. Experimental data are plotted with black points, INVEXP model with red line, NORM model with blue line and SYM model with green line.

6. Discussion

This study evaluated different optimization methods to fit joint trajectories produced during pointing tasks and squat jumps. The evolution of joint angles during the movements was modeled using three sigmoid shaped functions. Assuming a constant length of the limbs, the whole movement was reconstructed from the sigmoid models parameters. For each movement type (i.e. pointing tasks and squat jumps) and sigmoid model, different optimization methods were investigated. In the literature, only Plamondon used a similar approach. However among the published articles, experimental data were presented only in (Plamondon, 1998). Furthermore, no quantitative results were provided and the data was presented for a single subject. This does not allow to compare the present models with Plamondon's one. However, as mentioned earlier, the models used in the present study are defined on a bounded time interval contrarily to the log-normal models for which the end of the movement is not clearly defined.

Differences between original and reconstructed data were lower for pointing tasks than for squat-jumps. The relatively greater amplitude of the joint trajectories could explain this result during the jumping movement. Moreover, the modeling of the skeleton assumes rigid bodies between the joints. Considering the pointing tasks, it can be supposed that the length of the modeled limbs is quite constant. This assumption is supported by the similarity of the errors observed for global and semi-global methods. The rigid bodies assumption would be less true for squat-jump, especially for the trunk limb. Indeed, the spine is composed of many joints which allow bending of the trunk and thus, the trunk may be divided into two segments to ensure that the rigid bodies model is close enough to the reality of the movement.

The modeling methods proposed in this study deal with planar movements. The higher errors obtained with modeling of squat jumps may be explained by the movement of the joints along the transverse axis, especially for the knee. In comparison, pointing tasks would be closer to a real planar movement since the movement is performed on a planar surface.

Concerning optimization methods computing velocity, computation lasted longer for semi-global method than for local one in pointing task. Global optimization executed with similar velocity compared to semi-global method. Thus, global optimization should be used unless specific purposes are researched. For the squat-jumps, the present results show that unsurprisingly, using the constrained method is much more longer than the unconstrained optimization.

For both pointing tasks and squat jumps, similar accuracy was obtained with the three models of sigmoids. Among the two movements and the optimization methods, it appears that the NORM model allows fastest computation. Considering SYM and INVEXP models, the non-linear equation solving and the numerical integration can explain their relatively slower execution respectively. NORM model formulation takes advantage of the native implementation of the erf function in Matlab software thus ensuring fast computation.

7. Conclusion

The present results show that joint trajectories during planar movements such as pointing tasks or squat-jumps can be modeled using meaningful kinematic parameters. Among the three sigmoid models tested in this study, it appears that the NORM model is computed faster and allows better data fitting of the pointing tasks than other models. On the contrary, for squat-jumps, INVEXP and SYM models fitted better original data. From these results, it can be suggested that INVEXP and NORM models should be used preferentially. Indeed, the INVEXP model did not lead to better results and needs substantial computation time compared to other models. Despite the important computation time, INVEXP model may be useful for modeling specific movements, especially fast movements, which may not allow a good fitting with NORM model due to the relatively small definition domain of this model. For relatively slow and smooth movements, NORM model should be primarily used. Considering the class of the three models, INVEXP or NORM models should be used when the jerk has to be computed, since it can be analytically determined from the models formulation. If the jerk is not considered as a relevant parameter, both velocities and accelerations can be obtained analytically whatever the used model. Furthermore, slow data acquisition rates should not affect much the quality of the fits since only three points are needed to compute the shape parameters of the three models.

References

- Bastien, J., Legreneur, P. and Monteil, K. (2010). A geometrical alternative to Jacobian rank deficiency method for planar workspace characterisation. *Mechanism and Machine Theory* **45**, 335-348.
- Creveaux, T., Bastien, J., Villars, C. and Legreneur, P. (2012). Model of joint displacement using sigmoid function. Experimental approach for planar pointing task and squat jump. *arXiv preprint arXiv:1207.2627*, 1-28.
- Plamondon, R. (1995). A kinematic theory of rapid human movements. *Biological Cybernetics* **72**, 309-320.
- Plamondon, R. (1998). A kinematic theory of rapid human movements: Part III. Kinetic outcomes. *Biol Cybern* **78**, 133-145.
- Plamondon, R., Feng, C. and Woch, A. (2003). A kinematic theory of rapid human movement. Part IV: a formal mathematical proof and new insights. *Biol Cybern* **89**, 126-138.
- Soechting, J. F. and Lacquaniti, F. (1981). Invariant characteristics of a pointing movement in man. *J Neurosci* **1**, 710-720.
- van Ingen Schenau, G. J. (1989). From rotation to translation: constraints on multi-joint movements and the unique action of bi-articular muscles. *Human Movement Science* **8**, 301-337.
- Winter, D. A. (2009). Biomechanics and motor control of human movement - Fourth Edition. Hoboken, New Jersey: John Wiley & Sons, INC.
- Zelaznik, H. N., Schmidt, R. A. and Gielen, S. (1986). Kinematic properties of rapid aimed hand movements. *Journal of motor behavior* **18**, 353-372.

Quantitative Models for Retirement Risk in Professional Tennis

A.C.J. Cutmore* and W.J. Knottenbelt**

*7 Larks Grove, Barking, Essex, IG11 9UB, UK; adam.cutmore@imperial.ac.uk

** Department of Computing, Imperial College London, South Kensington Campus, London, SW7 2AZ, UK;
w.knottenbelt@imperial.ac.uk

Abstract. Tennis attracts both spectators and speculators. Indeed, every on-court development impacts on in-play financial markets. For singles matches, sophisticated traders make use of quantitative models. These yield the probability of each player winning the match given the current score and the probability of each player winning a point on serve. Significantly, these models ignore retirement risk, which can drastically and rapidly alter in-play odds. The precise impact on a market depends on that market's payout policy in the case of player retirement. This paper proposes novel quantitative models for the in-play evolution of retirement risk due to injury and the corresponding impacts on in-play match outcome markets under different retirement payout policies. Injury occurrence is assumed to be a Bernoulli process, with a magnitude sampled from a truncated exponential distribution. Retirement risk is propagated from one point to another subject to a damping parameter that reflects a player's ability to "run off" an injury. An important practical challenge is to parameterise our models given a score feed and odds data for two match outcome markets with different retirement policies. We demonstrate the feasibility of not only inferring model parameters from synthetic data effectively but also of mimicking real in-play markets.

1. Introduction

1.1. The Tennis Betting Industry

Tennis is one of the world's most popular individual sports and thus is also one of the most heavily traded on betting exchanges. The market continues to grow at a remarkable pace; for example, during the Wimbledon 2006 final between Roger Federer and Rafael Nadal, Betfair processed approximately £25 million worth of matched trades. In comparison, the Wimbledon 2012 men's final between Andy Murray and Roger Federer matched almost £56 million worth of bets. Such interest is not limited to Grand Slam finals either; during the women's semi-final of the Sony Ericsson Open 2011 between Maria Sharapova and Andrea Petkovic, £10 million was traded on Betfair¹.

Tennis is well suited to in-play trading on exchanges since points are played at a steady rate, are clearly separated, and are consistently won and lost by both participants, leading to frequent (and often dramatic) changes in fortune for players but at relatively predictable intervals. Consequently, the odds can very quickly swing back and forth as traders react to on-court events, generating potential money-making opportunities. This, in combination with the fact that tennis markets have the ability to offer only a few outcomes (e.g. in a Match Odds market, there are only two outcomes, either one player wins or the other does), ensures its popularity. Around 80% of money wagered on tennis matches is bet while the match is in progress².

Tennis also happens to be a relatively simple game to model in comparison with other highly complex sports such as football or cricket. It is just a series of discrete repeated contests, i.e. points, and calculations essentially boil down to the probability each player has of winning a given point. The scoring system has a fixed number of hierarchical states; points are nested within games, games within sets, and sets make up a match.

1.2 Player Retirement in Professional Tennis

In professional tennis, a player may 'retire hurt' from a match at any time should they feel they are unable to complete the match due to injury or illness, or that it is unwise to continue in case they aggravate their condition. The match is consequently awarded to the opponent regardless of the current match state. A *walkover* occurs when a player withdraws from a match before it has begun. In-play injuries are common occurrences in tennis as

¹ <http://www.fracsoft.com>

² http://www.sportspromedia.com/guest_blog/peter_webb_why_tennis_is_big_business_for_bookmakers

a whole. Between 2000 and 2009, there was a retirement during approximately 3.9% of Grand Slam men's singles matches³. Betting companies take different approaches when dealing with the issue of player retirement. Typically, they fall into one of four categories (with regards to Match Odds markets)⁴:

Category 1: Ball-Served Rule For a bet to stand, at least one ball must have been served, e.g. Ladbrokes.

Category 2: One-Set Rule For a bet to stand, at least one set must have been completed in the match. However, if a player retires from the match before the first set is over, all bets are cancelled and stakes are refunded, e.g. Betfair.

Category 3: Two-Sets Rule For a bet to stand, at least two sets must have been completed in the match, e.g. TheGreek.

Category 4: Match-Completed Rule The entire match must be completed for a bet to stand, e.g. Paddy Power.

Markets offered on the outcomes of individual games or sets are always rendered void unless the result has been unconditionally determined. UK-based Betfair, the world's first betting exchange, falls into *Category 2* of the tennis betting retirement payout policies with respect to its in-play Match Odds market. Betfair's Set Betting market is always voided in the event of a retirement since if the match is not finished, we do not have a final score.

1.3 Premise

Figure 1 displays the evolution of implied match-winning probabilities for Novak Djokovic when he played Rafael Nadal in the US Open 2011 Men's Final. The blue line shows implied probabilities extracted from the Betfair Set Betting market, the red line shows implied probabilities extracted from the Betfair Match Odds market, and the green line shows the positive difference of the Set Betting probability minus the Match Odds probability. As you can see, both markets are closely matched. This is intuitive since the probability of Djokovic winning the match should be the same as the sum of the probabilities of the final score being 3-0, 3-1, or 3-2 in Djokovic's favour. This is especially true in high profile matches such as this one where the markets are very liquid and millions of pounds are being traded in both, leading them to produce very accurate odds.

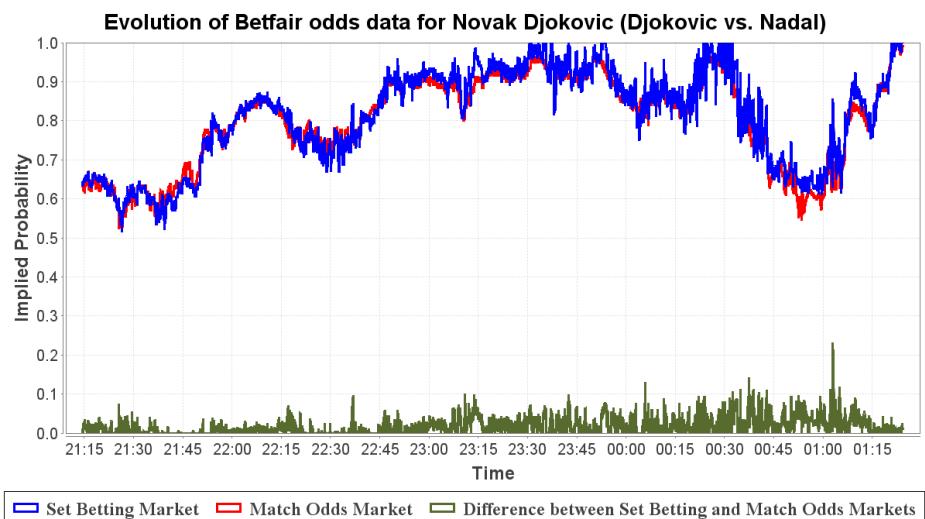


Figure 1: Evolution of implied match-winning probabilities extracted from the Betfair Match and Set Betting markets as well as the gap between them for Novak Djokovic - Djokovic vs. Nadal (*US Open 2011 Men's Final*)

³ <http://www.tennis.ukf.net/stats15.htm>

⁴ <http://rebelbetting.com/faq/tennis-rules>

Compare with Figure 2 which displays the evolution of implied match-winning probabilities for Andy Murray when he played Michael Berrer in the French Open 2011 Men's Third Round. In particular, observe where the Match Odds probability of Murray winning the match suddenly drops *almost 60%*. This phenomenon in the odds data occurred during a *single point in the match*. Although there are a few anomalous events which could have led to this huge and rapid swing in the market such as a disqualification, the evidence here weighs heavily towards the *injury* that was suffered by Andy Murray. We quote from the *BBC Sport*⁵ live text commentary of the match during this point:

- "Big, big trouble for Andy Murray, who has gone over on his right ankle and looks in real pain. Not sure whether he will be able to continue. Unbelievable."
- "We are going to have a medical time-out while Murray has treatment. He slipped as he ran in to put away a forehand, and the replays are not very pleasant to watch."

In this case, the injury did not cause Murray to retire from the match and he went on to win in straight sets, as reflected in the subsequent recovery of his odds to win. Nevertheless, it is clear the market reacted to this event and its opinion on Murray's chances of winning the match was severely affected.

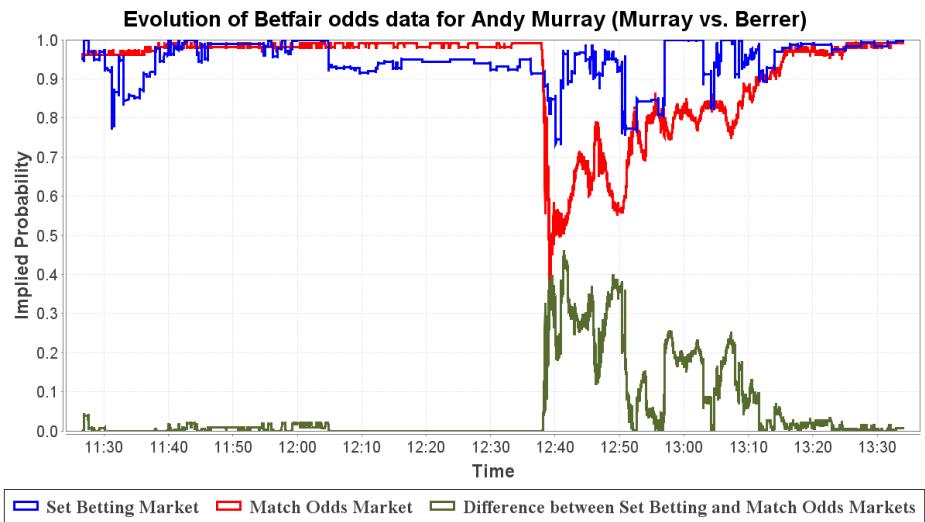


Figure 2: Evolution of implied match-winning probabilities extracted from the Betfair Match and Set Betting markets as well as the gap between them for Andy Murray - *Murray vs. Berrer (French Open 2011 Men's Third Round)*

We can observe the evolution of the Betfair Set Betting and Match Odds markets in order to help us *quantify* a player's risk of retirement. For example, we can see from the Murray vs Berrer match when the injury occurred that the Match Odds implied probability dropped sharply but the Set Betting implied probability remained relatively stable. If a player becomes injured and concedes the match at any point, traditional bookmakers will usually cancel bets in all markets (they fall into Category 4). Betfair will refund money on final score bets (Set Betting) but will still pay out on bets to win (Match Odds) as long as at least one set has been played (since the match has a winner but no final score). In this way, the Betfair Set Betting market *simulates* a Match Odds market that **ignores risk of retirement** whereas the actual Betfair Match Odds market **takes into account retirement risk**, given one set has been played. Consequently, we see discrepancies between the match and final score odds in such scenarios (possibly more so beyond the first set) as we do in the Murray match. The theory is that the probability of a player retiring *at some point during the remainder of the match* should be somehow encapsulated within the difference between the two markets, i.e. the combined opinions of all the traders betting

⁵ <http://www.bbc.co.uk/sport>

in those markets (the wisdom of the crowd). All the complex factors which lead to a player deciding to retire such as the state of the match, the seriousness of the injury, the importance of the match, even their past history of retirements, should be *some function* of this gap in the odds.

2 Background

2.1 The Physical Demands of Professional Tennis

Various papers have been written on tennis-related injuries from a physiological perspective. In particular, Johnson and McHugh (2006) attempted to quantify the demands in professional male tennis by analysing the number and type of strokes played per game for 22 players from three Grand Slams. They found that the serve was the predominant stroke played in service games (up to 60%) whereas topspin forehands and backhands were more frequent when receiving. The 2003 US Open winner hit over 1000 serves during his seven matches at the tournament. More strokes are played at the French Open than Wimbledon due to the relative speeds of the clay and grass court surfaces leading to longer rallies at Roland Garros. Importantly, Johnson and McHugh discuss the strain that playing a point inflicts on the body. They report that over 50% of world-class tennis players experience shoulder discomfort during their career and 80% of these cases stem from overuse. Stroke production in tennis involves generating repetitive forces and motions that are of high intensity and short duration. For example, the serve is the most strenuous stroke on the upper extremity with internal rotation velocities of the humerus reaching 2420 degrees per second for elite players during the acceleration phase. This, coupled with the relentless demands the ATP (Association of Tennis Professionals) and WTA (Women's Tennis Association) tours place on players, suggest that it is no surprise injuries are an issue in the world of professional tennis.

2.2 Existing Tennis Models

There have been many past works on the topic of modelling tennis. O'Malley (2008) presents what are considered the *tennis formulae*. The tennis formulae are a hierarchical series of equations that compute the probability a given player will win a tennis match given the probabilities that the given player and his/her opponent will win any of their service points. Combined together are individual formulae for the probabilities of winning games, sets, and tiebreaks.

Newton and Keller (2005) more comprehensively explore the use of recurrence relations to model tennis, utilising them to calculate probabilities of winning tournaments and also proving explicitly that the probability of winning a set or match does not depend on which player serves first. Barnett and Clarke (2002) experiment with the same idea in Microsoft Excel. They investigate using six parameters rather than just the two point-winning probabilities taking into account service faults. Consequently, for each player they input the probability of a successful first serve, the probability of winning a point on first serve, and the probability of winning a point on second serve. Spanias and Knottenbelt (2012) also designed a low-level model of tennis points, taking into account such events as service faults and service aces, in order to predict the outcome of matches. They tailor their input probabilities to the specific returning capabilities of the given opponent (not an average player), estimated using historical statistics. They find that their model was able to successfully predict 67% of outcomes from a sample of 1839 ATP matches played during 2011. Since we are already adding further complexity with the probability of retirement, we shall concentrate on extending a simpler, two parameter version of a tennis model. Barnett, Brown, and Clarke (2006) continued with an investigation into player momentum in tennis matches by slightly perturbing point-winning probability depending on how much the given player is leading or trailing the match by (essentially introducing a dependency between points).

The vital thing to note is that no previous tennis model incorporates retirement risk as a factor. They are therefore unable to account for the evolution of odds in in-play betting markets that utilise different retirement payout policies.

3 A Model for Retirement Risk

3.1 Definition

The granularity of the model is point-level as this is compatible with the structure of standard tennis models. When a tennis player plays a point, he or she potentially puts great strain on their body. This strain naturally leads to a chance of an injury occurring. For the vast majority of points played, the strain is perfectly manageable and does not lead to injury. For example, Grand Slam winners can hit over 1000 serves during the course of the tournament without issue. Occasionally however, the body fails to cope with the strain, or the strain is for some reason much more acute than normal (e.g. remember Andy Murray twisting his ankle), and an injury occurs. When a player does get injured during a match, it does not always end in retirement. Players often soldier on at least for a few points and may even recover from the injury as the match progresses. Using this intuition, we present an elegant model for the probability a player will retire *on a given point* in a tennis match:

$$\begin{aligned} r_0 &= 0 \\ r_{n+1} &= \min(\rho r_n + XY, 1) \end{aligned}$$

where r_n is the given player's risk of retirement on point n of the match, $0 \leq \rho \leq 1$, and X and Y are random variables. Do not confuse r_n with what we are hoping to eventually calculate which is R_n , the risk of retiring at some point during the remainder of the match from point n . We say that players start a match with zero probability of retiring ($r_0 = 0$), although in principle it is possible players might start a match with a niggling injury, in which case the magnitude could be inferred from observations of two or more match markets with differing retirement policies. When setting r_{n+1} , we make sure to take the minimum of the calculated r_{n+1} and 1, since retirement risk is a probability and cannot be greater than 1. The Bernoulli random variable, X , with success parameter c , models the chance of an injury occurring on a given point. The random variable, Y , models the magnitude of an injury should it occur and is truncated exponentially distributed with rate parameter λ and upper bound 1, i.e. minor bumps and bruises are more common than serious, match-ending injuries. The decay parameter, ρ , models the idea that players recover from injuries as matches progress. For the purpose of avoiding too complex a model, we make the simplifying assumption that ρ and the distributions of X and Y are the same for both players.

We incorporate our model of retirement risk into a *tennis match simulator* capable of approximating match-winning probabilities. We input the point-winning probability for each player as well as the current score and current point-level retirement risks, and simulate a large number of matches using the same parameters for each match. The simulator is probabilistic but we expect convergence towards exact match-winning probabilities. The greater the number of *runs* (i.e. matches played), the greater the accuracy of the approximation generated by the simulator. The proportion of matches the modelled player wins out of the total number of runs is an estimation of the chance of winning. Similarly, the proportion of matches the modelled player retires from out of the total number of runs is an estimation of the risk of retiring.

A given point in a simulated match has four possible outcomes. Either player can win the point or either player can retire from the match. To resolve a point, we generate a random number between 0 and 1 and observe which of the bins (as shown in Figure 1) the number falls in. The match ends if it falls in one of the retirement bins (which is why r_A and r_B are usually 0 else you will rarely be able to get through even a single match without retiring), otherwise one of the players wins the point (P_A is the point-winning probability on serve of Player A) and the match continues (unless it was their match point).

Table 1: The four possible outcomes of a point in our modified tennis match simulator where Player A is serving

0			1
P_A	$1 - P_A$	r_A	r_B

To imitate the Betfair Set Betting market (or equivalently, a Match Odds market using a Paddy Power-style *match-completed* payout policy), we can use O’Malley’s tennis formulae which ignore retirement risk (the *No Retirement Risk* column in Figure 2). We use our tennis match simulator to imitate a Match Odds market as it might behave using the different retirement betting payout policies by calculating the remaining probabilities in the right-most three columns.

Table 2: Probabilities that can be closely approximated by our modified tennis match simulator

Player	No Ret. Risk	Normal Win With Ret. Risk	Retirement in 1st set	Retirement after 1st Set
A	W_A	W'_A	R'_A	R''_A
B	W_B	W'_B	R'_B	R''_B

W'_A , for example, is the probability of Player A winning the match normally by achieving 3 sets and not via Player B retiring. Note that we can (re-)calculate the match-winning probability for Player A ignoring retirement risk (W_A) as a sanity check by computing:

$$\frac{W'_A}{W'_A + W'_B}$$

In addition, the ratio of W_A to W_B is the same as the ratio of W'_A to W'_B . This is because we assume that the point-winning probabilities of each player are unaffected by injury. We can imitate a Betfair-style *after one set* payout policy market for Player A by computing:

$$\frac{(W'_A + R''_B)}{(W'_A + R''_B) + (W'_B + R''_A)}$$

which is the probability that Player A wins normally plus the probability that Player B retires after the first set, normalised by the sum of the probabilities that either player wins the match normally and either player retires after the first set. Similarly, we can imitate a Ladbrokes-style *after one ball* payout policy market for Player A by computing:

$$\frac{(W'_A + R_B)}{(W'_A + R_B) + (W'_B + R_A)}$$

where $R_A = R'_A + R''_A$ and $R_B = R'_B + R''_B$. This is essentially a re-calculation of the *Normal Win With Retirement Risk* column.

3.2 Simulated Results

We have introduced three unknowns into our simulator; c is the success parameter of the Bernoulli distribution (per point injury probability) corresponding to random variable X , λ is the rate parameter of the injury magnitude truncated exponential distribution corresponding to random variable Y , and ρ is the injury recovery factor.

Figure 3 displays a table describing the majority of the variables we have introduced thus far. An appropriate parameterisation of the model would be values for the unknown input variables, c , λ , and ρ , that generate $R_A \approx R_B \approx 1.95\%$ when input into our modified simulator at the start of a 5-set match. We believe these match-level retirement risks are justified as it corresponds to the 3.9% chance of such a match ending in retirement as seen in men’s Grand Slam singles matches between 2000 and 2009. Given Klaassen and Magnus (2000) find that the average point-winning probability on serve for a top-level professional tennis player is 0.645 for men and 0.560 for women, we choose 0.6 as a reasonable value for P_A and P_B which denotes an even

match. The goal now is to *fit* the model by adjusting the parameter values to accurately reflect real-world events.

It is possible to use a multivariate direct-search optimisation algorithm such as the Nelder-Mead simplex method (an algorithm primarily designed for statistical parameter estimation problems such as ours) to approximate a set of values for these parameters. Specifically, we used a constrained version of the Nelder-Mead algorithm in order to ensure logical bounds $0 < c, \rho < 1$ and $\lambda > 0$.

Table 3: Table describing the variables used in our system

Variable	Known	Input / Output	Description
P_A	Yes	Input	The probability Player A wins a point on serve (assumed for the moment)
P_B	Yes	Input	The probability Player B wins a point on serve (assumed for the moment)
W_A	Yes	Input	The probability Player A wins the match using a standard tennis model
W_B	Yes	Input	The probability Player B wins the match using a standard tennis model
G_A	Yes	Input	The Betfair Set Betting implied probability minus the Betfair Match Odds implied probability for Player A for $G_A \geq 0$
G_B	Yes	Input	The Betfair Set Betting implied probability minus the Betfair Match Odds implied probability for Player B for $G_B \geq 0$
-	Yes	Input	The current score in the match
W'_A	No	Output	The probability Player A wins the match normally given the possibility of retirement
W'_B	No	Output	The probability Player B wins the match normally given the possibility of retirement
R_A	No	Output	The probability Player A retires at some point during the remainder of the match (can be categorised by set)
R_B	No	Output	The probability Player B retires at some point during the remainder of the match (can be categorised by set)
c	No	Input	The Bernoulli success probability parameter representing the chance a player suffers an injury on any given point
λ	No	Input	The rate parameter for the truncated hyper-exponential distribution dictating the magnitude of the injury suffered by a player should such an event occur
ρ	No	Input	The decay constant representing recovery from injuries in our retirement risk equation

Vitally important to the success of the Nelder-Mead method is the choice of objective function to minimise. In our case, our modified simulator is essentially the function, but we must still define what it means to minimise it. This is the main way we mould the algorithm to solve our problem. We want the difference between the probability Player A wins the match normally given no risk of retirement and the probability Player A wins the match normally with risk of retirement to initially be 0.0195, which is the probability at the beginning of the match that Player A wins the match via Player B retiring (and similarly for Player B). Consequently, we aim to minimise the below expression:

$$|W_A - (W'_A + R_B)| + |W_B - (W'_B + R_A)|$$

More specifically, at the minimum and since the contest is evenly matched, we would have:

$$|0.5 - (0.4805 + 0.0195)| + |0.5 - (0.4805 + 0.0195)| = 0$$

We executed the Nelder-Mead method a number of times under the same conditions and took the means of the approximations found for our three input parameters each time as our chosen values. Remember, c is the per point injury probability, λ is the point-level retirement risk magnitude exponential distribution rate parameter, and ρ is the injury recovery factor. We find:

- $c = 0.000115$ (implying an injury approximately every 8500 points)
- $\lambda = 10.0$ (implying injuries cause a point-level retirement risk of 0.1 on average when they occur)
- $\rho = 0.95$ (point-level retirement risk is multiplied by 0.95 on each point)

In practice, there are likely to be many combinations of values for these parameters that would generate the retirement rates we require. For example, the effect of decreasing ρ (quicker recovery) could be countered by increasing λ (injuries are more severe) or increasing c (injuries are more common). We note that not enough top-level matches end in retirement to tailor these parameters to individual players.

In order to assess the accuracy of our model, we try it out in a totally artificial environment where we control all the variables. We run a single simulated match many times with our chosen set of parameters. We assume that P_A and P_B remain unchanged throughout the match. We are looking for 'ideal' scenarios, e.g. where Player A receives an injury during the match but does not retire and still goes on to achieve victory. When we generate such a match, we record into a CSV file the score, the server, and the point-level retirement risks, r_A and r_B , at each point during the match. We then read this information back in, calculating match-winning probabilities at each point in the match.

Figure 3 shows the effect on simulated markets with varying retirement payout policies of Player A suffering an injury in the second set but still managing to go on to win the match. We see a sharp drop in the red and orange lines (Match Odds markets with *after one set* and *after one ball* payout policies, respectively) when the injury occurs. Traders now fear a retirement and are less willing to back Player A to win the match. This is followed by recovery, similar to the Murray vs. Berrer example, as traders' faith in a Player A victory is slowly restored. The blue line (Match Odds market with *match-completed* policy) ignores retirement risk and therefore does not react to the injury. Figure 4 shows the evolution of both the point-level (magenta line) and match-level (green line) retirement risks for Player A in this match. They are clearly inversely related to the behaviour of the markets; where match-winning probability falls, retirement risk rises, and where match-winning probability recovers, retirement risk decays.

Figure 5 shows a similar situation but where the injury occurs in the first set. This time the red line does not fall as drastically as the orange and the two markets merge as we approach the end of the first set. This models the idea that if an injury occurs in the first set of a match, traders in an *after first set* market will only display increased reluctance to back the player in question as it becomes more likely he or she will finish the set (in case they decide to default afterwards and payouts happen).

Figure 6 shows a match where Player A decided not to continue. Player A does attempt to play one or two more points after the injury and the market anticipates recovery but those hopes are soon dashed. Note that any discrepancies you might see between the three markets are due to the fact that the modified simulator provides only approximations whereas the standard model provides exact solutions (particularly noticeable at deuce or 6-6 in a tiebreak).

The model appears to produce markets that behave in the theoretically correct manner. We have sudden, sharp drops in match-winning probability corresponding to an injury on a point, followed by gradual recovery as the traders realised retirement might not happen.

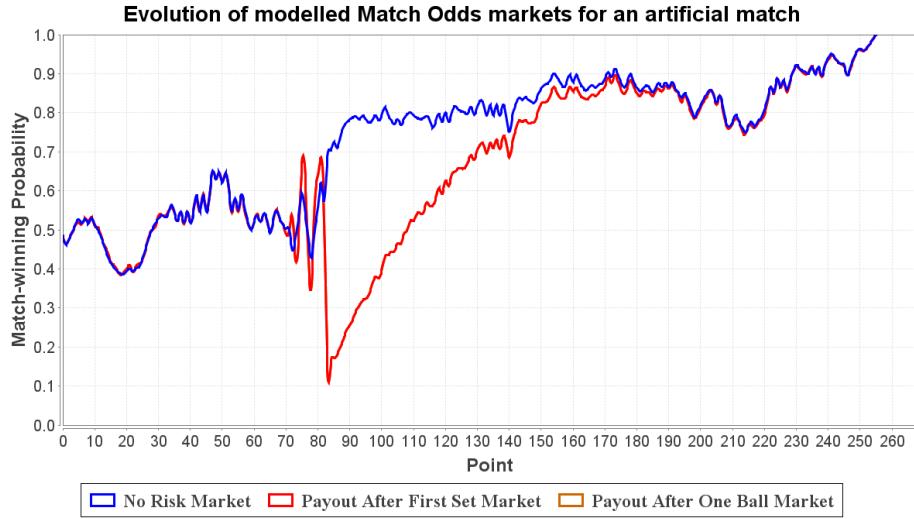


Figure 3: Evolution of Match Odds markets with a variety of retirement payout policies for an artificial match ($P_A = P_B = 0.6$, $c = 0.000115$, $\lambda = 10.0$, $\rho = 0.95$) with respect to Player A. In this match, Player A receives an injury in the second set but rallies and continues on to victory

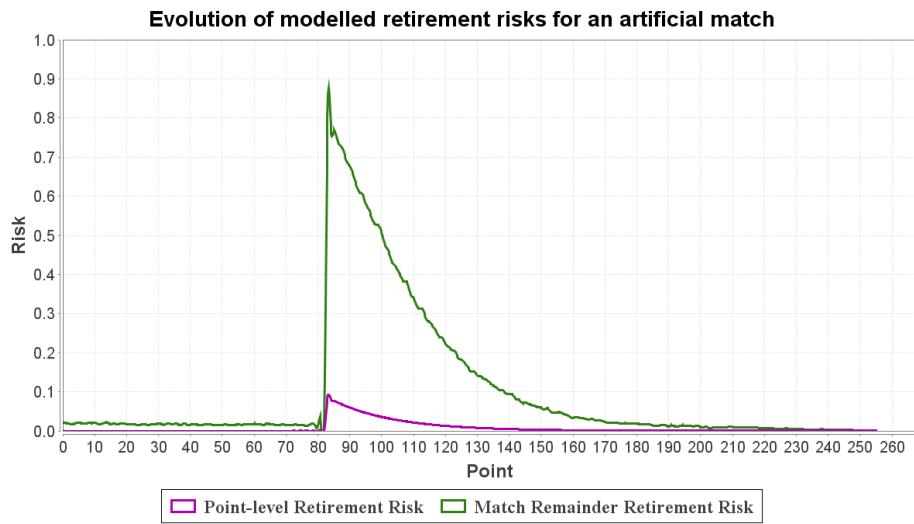


Figure 4: Evolution of point-level and match-level retirement risks for an artificial match ($P_A = P_B = 0.6$, $c = 0.000115$, $\lambda = 10.0$, $\rho = 0.95$) with respect to Player A. In this match, Player A receives an injury in the second set but rallies and continues on to victory

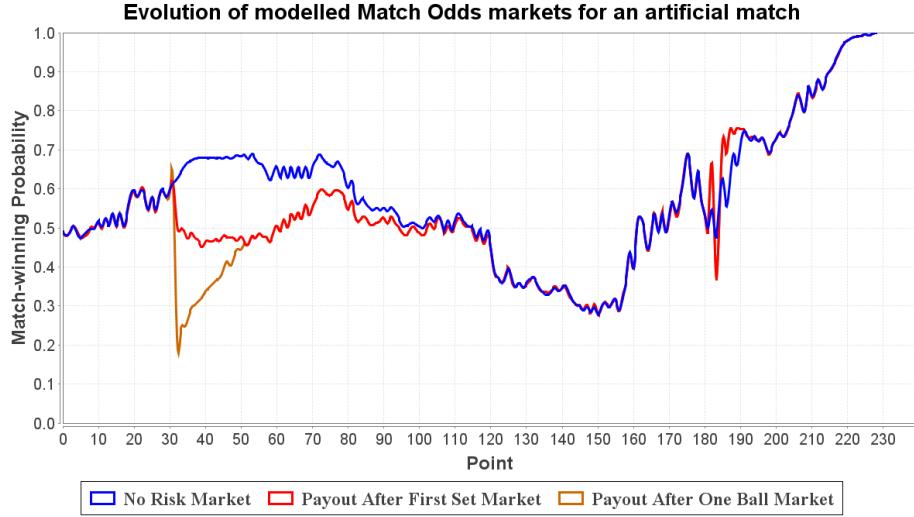


Figure 5: Evolution of Match Odds markets with a variety of retirement payout policies for an artificial match ($P_A = P_B = 0.6$, $c = 0.000115$, $\lambda = 10.0$, $\rho = 0.95$) with respect to Player A. In this match, Player A receives an injury in the first set but rallies and continues on to victory

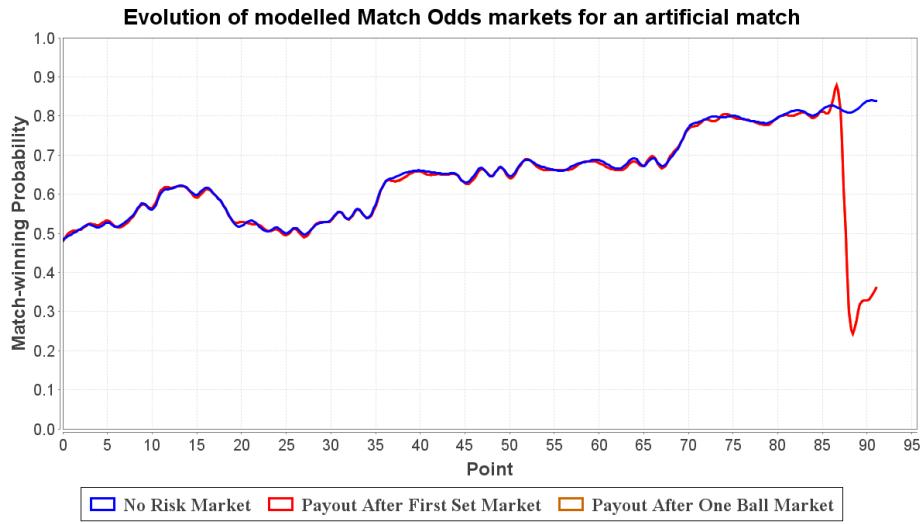


Figure 6: Evolution of Match Odds markets with a variety of retirement payout policies for an artificial match ($P_A = P_B = 0.6$, $c = 0.000115$, $\lambda = 10.0$, $\rho = 0.95$) with respect to Player A. In this match, Player A receives an injury and has to retire

3.3 Application to Real Data

Our model appears to produce reasonable results from the beginning of a match given our assumption that players begin matches with no risk of retiring ($r_0^A = r_0^B = 0$). Our model must also be applicable for any given match state, including the situation where an injury has recently occurred but the affected player is continuing to play. At this stage, we know the gap in the odds for each player (G_A and G_B), but how can we go *backwards* from this and calculate the retirement risk for the *current point*, r_n^A and r_n^B (which will naturally be a lot smaller)? Once again, we can use the Nelder-Mead method to approximate these values for each point in the match. Given this information, our simulator will be able to calculate the match-level retirement risk for each player at each point (R_n^A and R_n^B).

Remember, we can imitate a Betfair-style *after one set* payout policy market for Player A (and similarly for Player B) by computing:

$$W_A'' = \frac{(W'_A + R_B'')}{(W'_A + R_B'') + (W'_B + R_A'')}$$

Now we are dealing with the gaps in the odds rather than the retirement risks to be output so we redefine our objective function as:

$$|W_A - (W_A'' + G_A)| + |W_B - (W_B'' + G_B)|$$

We now also have all the information we need to be able to predict the evolution of a market with an *after one ball* payout policy for a real match using just the two Betfair markets!

3.3.1 Preparing the Match Data

Obtaining the Betfair odds data for a match was only the first step. We also needed to acquire other match-specific information. We needed to be able to read the current score into the model at any stage in the match. We entered point-by-point data manually into CSV files using the archived point-level live scoring provided by website *TennisEarth.com*⁶ and tennis statistics software OnCourt⁷.

Up until this point, we have assumed that we will be able to manually input the point-winning probabilities on serve of both players in the match. Due to Klaassen and Magnus (2000), we know that the average point-winning probability on serve for a top-level professional tennis player, γ , is 0.645 for men and 0.560 for women. We also have the insight, arising from O'Malley (2008) and described by Marek (2011), that the important thing in determining the winner of a match is the *difference*, δ , between the point-winning probabilities of each player and not the absolute values. Consequently, if we can find what δ *should* be, we can choose appropriate values for P_A and P_B by using the constraint that their average equals γ . We must find both; we cannot, for example, fix P_A to γ and linearly search for P_B , since the model would cease to be symmetric for both players. We know the current implied match-winning probability of the Set Betting (no retirement risk) market for both players and so we can do a simple binary search to find a value for δ (and therefore values for P_A and P_B) for which O'Malley's tennis formulae generate the same results as the *no risk* market. Marek also states that for one to be able to express match-winning probability as a function of δ , one must constrain $-0.1 \leq \delta \leq 0.1$. Since the market's opinion of each player's point-winning probability changes throughout the match, we recalculate P_A and P_B on every point. Note that this does not invalidate the assumption that points are iid as no dependency between points is introduced.

Another issue was how to match the odds data to the score at each point. For instance, we may have many thousands of odds data lines for a match but only a couple of hundred points. Although we have a timestamp for each line of odds data, we do not know the exact time that each point was played. To overcome this challenge, we make the reasonable assumption that points happen at regular intervals. For example, if we have 9000 lines of odds data for a match and 300 points were played, we would sample every $9000/300 = 30^{\text{th}}$ line. Consequently, our modelled markets will appear more sparse than the original odds data as we only have as many points as there were points played in the match.

We now test our system on a number of real-world case studies. We consider only the evolution of the in-play markets from the point of view of the player that was injured in the match (it is extremely rare that both players in a match receive significant injuries).

⁶ <http://www.tennisearth.com>

⁷ <http://www.oncourt.info>

4 Case Studies

4.1 Andy Murray vs Michael Berrer

Figure 7 displays processed Betfair odds data for the French Open 2011 third round match between Andy Murray and Michael Berrer that we discussed earlier. In Figure 8, we use our system to model Match Odds markets under three different retirement payout policies using the current score, estimated point-winning probabilities, and the positive gap created by subtracting the Betfair Match Odds implied probabilities on each point from the Betfair Set Betting probabilities, as input. As you can see from the graph, our system reproduces the Betfair Match Odds market (red line) quite accurately with the *after one set* modelled market. This shows that we were successful in finding point-level retirement risks that would recreate the gap between the Betfair Set Betting and Match Odds markets. The orange line shows our predicted *after one ball* modelled market. In this case, it closely follows the red *after one set* market since the injury to Andy Murray occurs after the first set has been completed.

Bear in mind that there will always be a certain amount of variation due to the inexact nature of the simulator and our method of aligning our point-by-point data with the odds data.

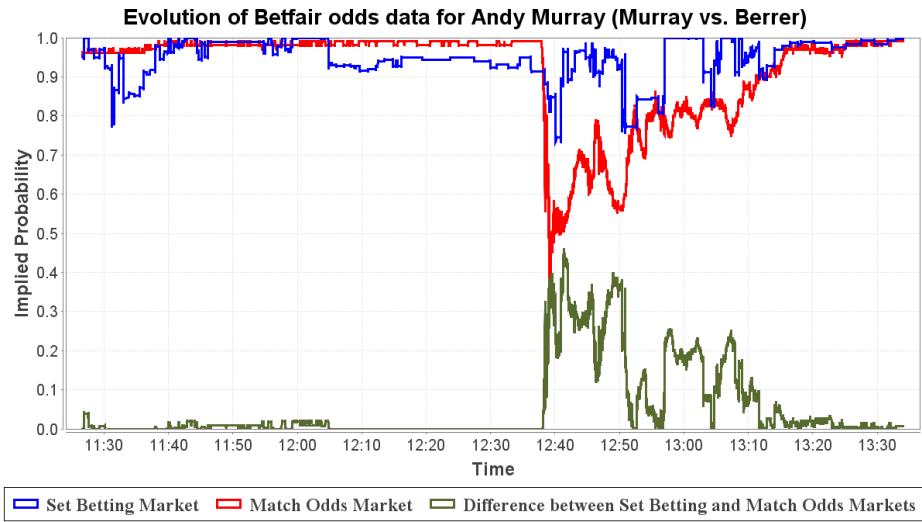


Figure 7: Evolution of implied match-winning probabilities extracted from the Betfair Match and Set Betting markets as well as the gap between them for Andy Murray - *Murray vs. Berrer (French Open 2011 Men's Third Round)*

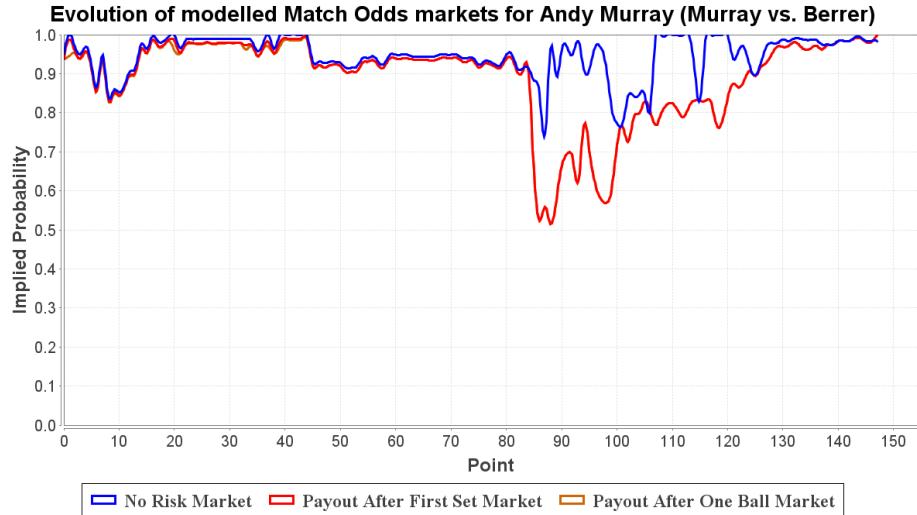


Figure 8: Evolution of modelled Match Odds markets under three different retirement payout policies for Andy Murray - *Murray vs. Berrer (French Open 2011 Men's Third Round)*

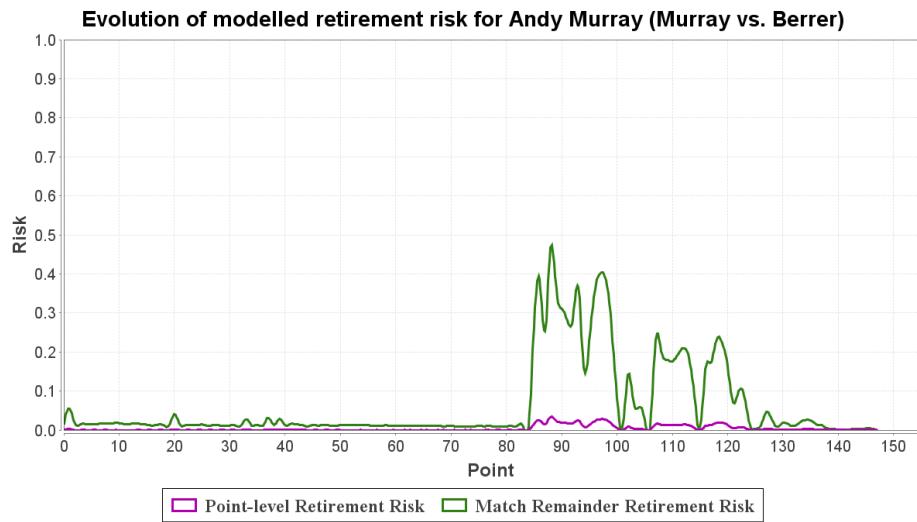


Figure 9: Evolution of modelled point-level and remainder of match retirement risks for Andy Murray - *Murray vs. Berrer (French Open 2011 Men's Third Round)*

Figure 9 shows the evolution of Andy Murray's risk of retirement throughout the match. The green line follows the probability of retirement during the remainder of the match from the given point (R_n), whereas the magenta line gives the probability of retiring on a particular point itself (r_n). Our model predicts a peak match-level retirement risk of 45% during the injury period. This coincides with a peak point-level retirement risk of 3%. Despite the severity of the injury, Murray's position of dominance in the match may be one explanation for why his risk of retirement is not greater. Berrer is still not expected to win the match irrespective of his opponent's misfortune.

4.2 Rafael Nadal vs David Ferrer

An all-Spanish Australian Open 2011 Men's Quarter Final saw Rafael Nadal battle veteran and fellow clay court specialist David Ferrer. A mystery injury early on in the first set meant Nadal struggled throughout the match but he refused to retire and allowed his opponent the three-love win.

Figure 11 illustrates a scenario where an *after one ball* modelled market can react differently than an *after one set* market. As we would expect (since the injury occurred in the first set), our predicted *after one ball* market (orange line) anticipates an even greater drop in Nadal's match-winning probability around the time of his injury than the Betfair Match Odds *after one set* market. In the event of a Nadal retirement, such a market would pay out for a David Ferrer win as long as one ball has been played so at this point, traders would be very reluctant to back Nadal if he showed signs he might retire. As the match moves past the first set, the *after one ball* and *after first set* lines merge since potential injuries now contribute to both markets equally.

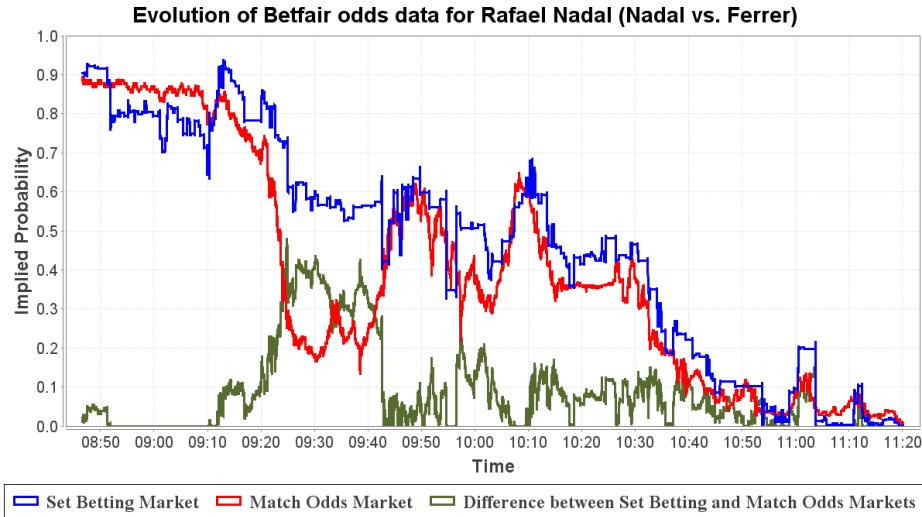


Figure 10: Evolution of implied match-winning probabilities extracted from the Betfair Match and Set Betting markets as well as the gap between them for Rafael Nadal - *Nadal vs. Ferrer (Australian Open 2011 Men's Quarter Final)*

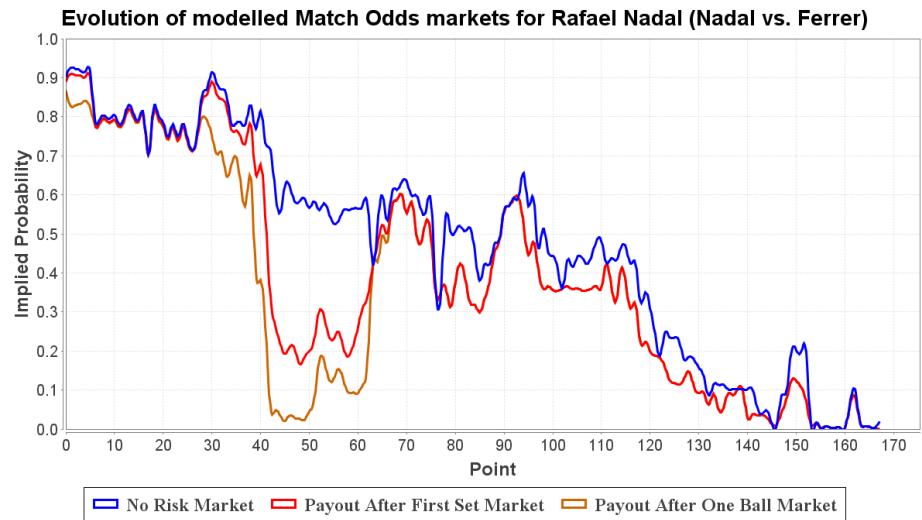


Figure 11: Evolution of modelled Match Odds markets under three different retirement payout policies for Rafael Nadal - *Nadal vs. Ferrer (Australian Open 2011 Men's Quarter Final)*

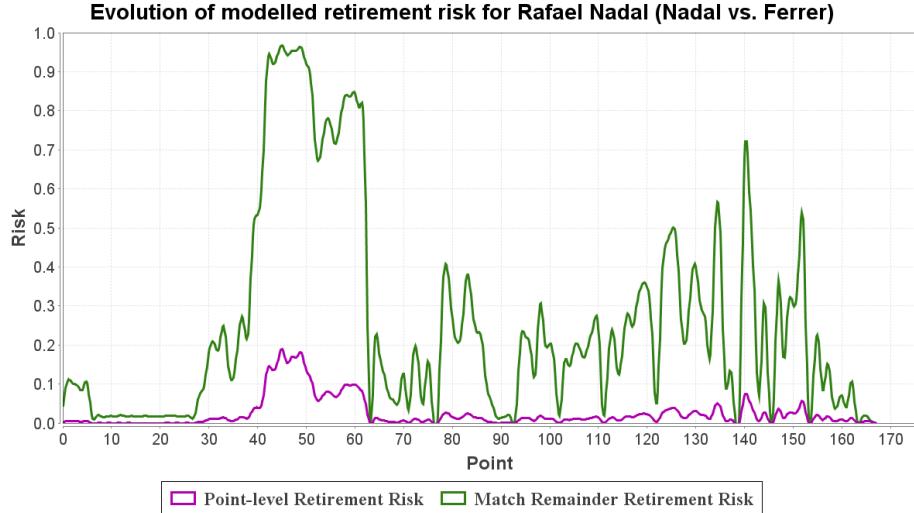


Figure 12: Evolution of modelled point-level and remainder of match retirement risks for Rafael Nadal - Nadal vs. Ferrer (Australian Open 2011 Men's Quarter Final)

Figure 12 shows the evolution of Rafael Nadal's risk of retirement throughout the match. Our model predicts a high peak match-level retirement risk of 96% during the injury period. This coincides with a peak point-level retirement risk of 18%.

As the match progresses and Nadal falls further behind, we see spikes in his retirement risk despite there being no large gaps between the Betfair markets. This happens when the Match Odds market gives a player little chance of winning the match. We illustrate how this occurs with an example. Say that on a particular point, the Set Betting market tells us that the implied probability of Nadal winning the match is 0.5. This means that we can assign $P_A = P_B = 0.645$. We also happen to know that the difference between the Betfair Set Betting and Match Odds markets for Nadal is 0.4. We now look for a value of r_A (retirement risk of Nadal on this point) such that W''_A (probability of winning the match normally with retirement risk after the first set only) is only 0.1. Since we have $P_A = P_B$ and therefore both players have an equal chance of winning the match if it does not end in retirement, we have $W''_B = 0.1$ and so the sum of the probabilities of either player winning the match normally is only 0.2. Making the assumption that r_n^B is negligible, we must have that $R_A \approx 0.8$ (retirement risk of Nadal in the match). This can happen in any situation where W'' is small and there is a risk of retirement, like towards the end of this quarter final.

Such scenarios, as well as our choice of matches where only one player was injured, help to explain why our predicted retirement risk is generally much larger than the gap in the Betfair markets.

4.3 Victoria Azarenka vs. Maria Sharapova

Victoria Azarenka faced off against Maria Sharapova in the Rome Masters 2011 Quarter Final. Unfortunately, Azarenka suffered a hand injury early in the second set while leading one set to love and was unable to continue the match.

As you can see in Figure 15, we have spikes of retirement risk during the first set which correspond with sporadic drops in the implied probability of our predicted *after one ball* market. These anomalies appear to coincide correctly with small gaps created where the *after one set* market model is beneath the Set Betting market. However, the tiniest of these gaps can correspond to a significant drop in the implied probability of *after one ball* market. Whenever we have such a gap, we attempt to find a point-level probability (r_n) that will recreate this gap. Since the Betfair Match Odds market only takes into account retirement after the first set, only retirements after the first set can affect our model of this market. However, the system has no direct control over retirements *after* the first set if we are still in the first set. The more you try to increase the immediate point-level retirement risk, the more likely the given player will retire straight away (still in the first set). Lower it so the player will survive past the first set and he or she probably will not retire. The consequence of this is that our *after one ball* market model (and therefore our predicted retirement risk) is very sensitive to the Betfair odds

source data, particularly towards the beginning of the first set.

Nonetheless, this match is still a good example of how unpredictable injury occurrence is. Azarenka's match-level retirement risk still stays relatively low until the injury occurs on point 75. The risk shoots up to 54%, peaking at 89% at retirement on point 93 after a brief dalliance with the possibility of recovery.

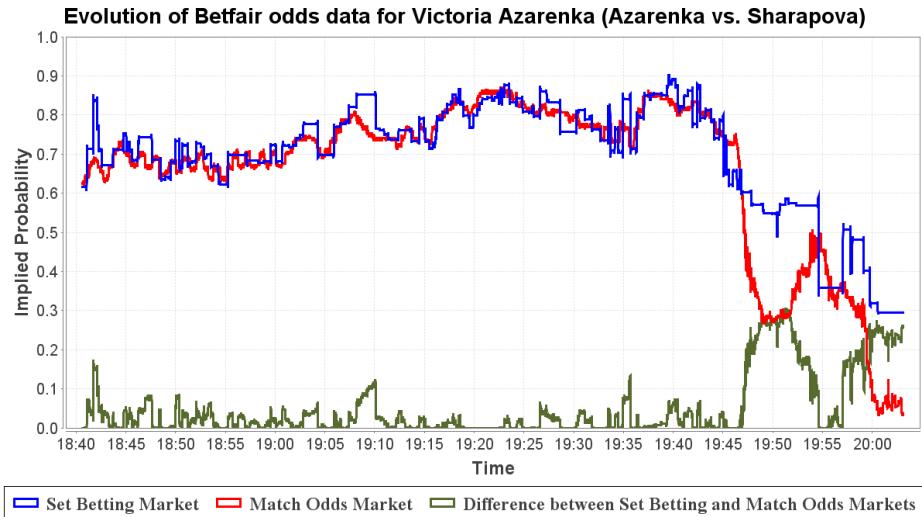


Figure 13: Evolution of implied match-winning probabilities extracted from the Betfair Match and Set Betting markets as well as the gap between them for Victoria Azarenka - Azarenka vs. Sharapova (*Rome Masters 2011 Women's Quarter Final*) [ended in retirement]

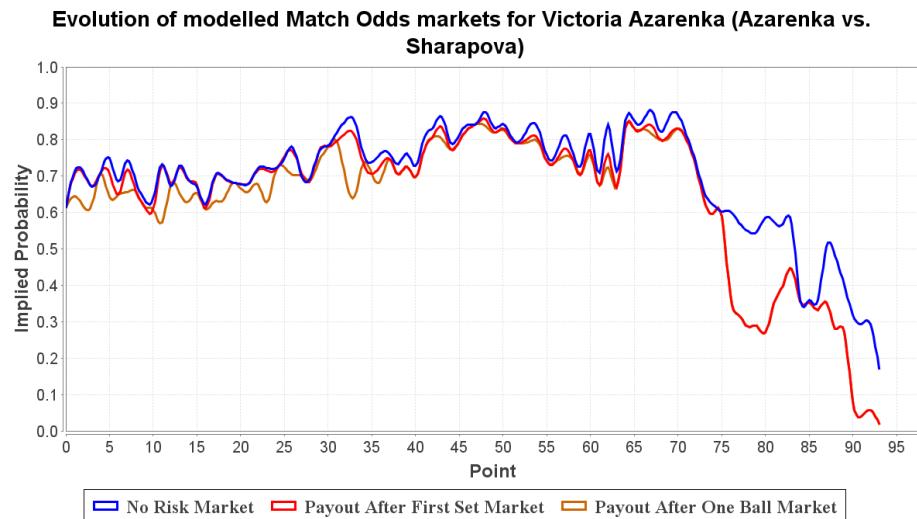


Figure 14: Evolution of modelled Match Odds markets under three different retirement payout policies for Victoria Azarenka - Azarenka vs. Sharapova (*Rome Masters 2011 Women's Quarter Final*) [ended in retirement]

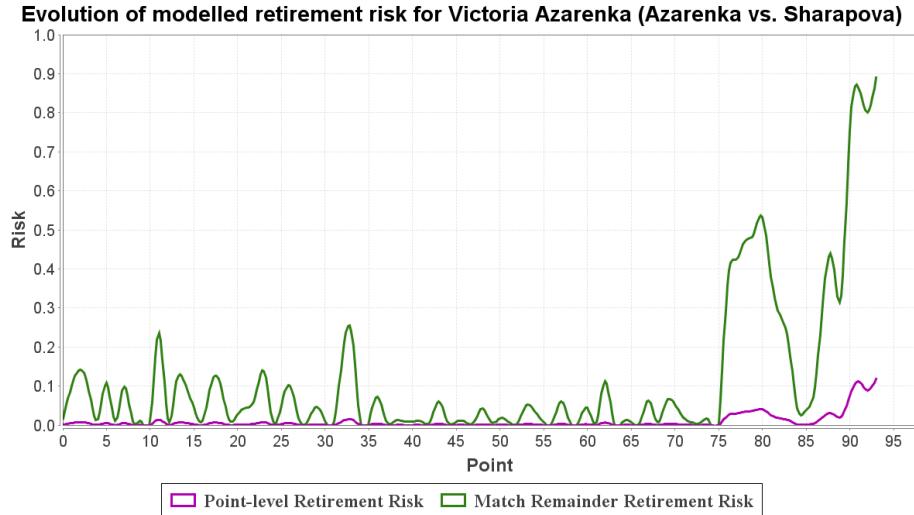


Figure 15: Evolution of modelled point-level and remainder of match retirement risks for Victoria Azarenka - Azarenka vs. Sharapova (Rome Masters 2011 Women's Quarter Final) [ended in retirement]

5 Conclusions

We confirm that it is possible to observe the occurrence of an injury in a given tennis match by observing the evolution of the in-play betting odds. We examine historical Betfair odds for a number of real-life matches and find that a gap between markets that use different player retirement payout policies is created in correspondence with the occurrence of an injury in the match. We note that injuries have a drastic effect on the odds data, in many occasions turning the overwhelming favourite into the underdog in an instant.

We have created a new model for retirement risk in tennis matches and implemented it into a tennis match simulator by incorporating extra parameters approximated using real-world averages and betting odds data as well as additional match outcomes compared to standard tennis models. To the best of our knowledge, this is the world's first published attempt to create such a model. Testing the model in a completely artificial environment, we found it was able to mimic the expected patterns of in-play betting markets with different retirement payout policies for matches with varying injury scenarios.

We conclude that a given player's risk of retirement at some point during the remainder of a match is a function of the difference between the odds of a Match Odds market for that player that ignores risk of retirement and the odds of a Match Odds market for that player that takes into account risk of retirement, for any given point in the match. Our system provides a value for the retirement risk of a given player at any point in a match.

We applied our model to a number of real-world matches (from the point of view of an injured player) using Betfair odds data and produced imitations of the in-play betting markets following different player retirement payout rules. We find that we can mimic to a good degree of accuracy the progress of the Betfair Set Betting and Match Odds in-play markets throughout matches, i.e. a *match-completed* market and an *after first set* market, respectively. We attempt to use the Betfair odds data to predict the evolution of an *after one ball* market. We find that such a market generally correctly produces slightly lower match-winning implied probabilities than the Betfair Match Odds market when the given match is still in the first set, although it can be somewhat erratic. The retirement risk values are very sensitive to any gap between the Betfair markets due to the lack of control our system has over retirement risk *after* the first set when the match is still in the first set. Furthermore, we notice that the system generates retirement risk spikes when the player in question has a low match-winning probability. The problem underlying these fragilities in our predictions is that the Betfair in-play tennis betting markets are not perfect. Since this odds data is vital input for our system, it is no surprise that fluctuating, anomalous, and sparse data heavily influences the output of our system.

We note that a potentially useful application of our model is in the real-time estimates of player retirement risk during professional tennis matches. Such information could be displayed on-screen for the viewer during a live broadcast or used as a commentary aid.

References

- Barnett, T., Brown, A. and Clarke, S. (2006) Developing a model that reflects outcomes of tennis matches. *Proceedings of the 8th Australasian Conference on Mathematics and Computers in Sport*, 178–188.
- Barnett, T. and Clarke, S. (2002) Using Microsoft Excel to model a tennis match. URL: <http://www.strategicgames.com.au/excel.pdf>
- Johnson, C. and McHugh, M. (2006) Performance demands of professional male tennis players. *British Journal of Sports Medicine* 40 (8), 696–699.
- Klaassen, F., Magnus, J. (2000) How to reduce the service dominance in tennis? Empirical results from four years at Wimbledon. *Open Access publications from Tilburg University*.
- Marek, A. (2011). Applying a hierarchical markov model to tennis matches. (final year project), Imperial College London, South Kensington Campus, London, SW7 2AZ.
- Newton, P. and Keller, J. (2005) Probability of winning at tennis I: Theory and data. *Studies in Applied Mathematics* 114 (3), 241–269.
- O'Malley, J. (2008) Probability formulas and statistical analysis in tennis. *Journal of Quantitative Analysis in Sports* 4 (2), Article 15.
- Spanias, D., Knottenbelt, W.J. (2012) Predicting the outcomes of tennis matches using a low-level point model. *IMA Journal of Management Mathematics*.

ODI cricket: characterizing the performance of batsmen using ‘tipping points’

Uday Damodaran*

*XLRI Xavier School of Management, Jamshedpur, Jharkhand, India. Email: uday@xlri.ac.in

Abstract: As a game, cricket has a large fan following in the countries that constituted the erstwhile British Empire. In spite of this huge fan following and the resultant TV viewership, there are surprisingly few measures available to describe and represent player performance in all its richness. The focus in this paper is to develop an approach to help arrive at a richer description of the performance of batsmen. There is a widely held belief amongst commentators and experts that some batsmen become more ‘set’ and confident as they progress in their innings and therefore become dangerous – from the opponents’ perspective- as they settle down. Is this really true for some batsmen? How does a batsman’s performance evolve as he progresses in his innings? This paper uses data for members of the Indian One Day International (ODI) cricket team between 1989 and 2012 to answer these, and similar questions. Earlier work on cricket has suggested that cricket scores follow an approximate geometric distribution and therefore exhibit the memory-less property. Building on the extant literature, this paper identifies “tipping points” using departures from the memory-less property, both in terms of runs scored as well as balls faced.

1 Introduction

As a game, cricket commands a very large fan following in the countries that constituted the erstwhile British Empire. Compared to other games, cricket is relatively slow paced. And in terms of length of play, no other game comes anywhere close to cricket: while the classical ‘test’ version of cricket can stretch to five complete days of play, even the modern and shorter versions of the game (like twenty-twenty or T20 cricket) last for more than three hours. Cricket is also a statistician’s delight because every event in cricket throws up one or more statistic; in fact it is in the very nature of the game that every ball bowled and every run scored has to be formally recorded.

Given its popularity, its slow pace, its length of play time and the fact that it is a game that naturally throws up a huge amount of data, it is surprising that cricket has not attracted more attention from statisticians and mathematicians than it already has. For example, even today, in describing and representing a batsman’s performance, what is primarily used is just an ordinary measure of ‘average’, biased upwards because of an idiosyncratic treatment of the ‘not out’ scores (Damodaran, 2006). This bias is caused because the numerator is the total runs scored over all innings while the denominator excludes the innings in which the player has remained ‘not out’.

Describing or representing a batsman’s performance using just the average disregards and ignores all the nuances and richness that resides in the natural process that constitutes a batsman’s innings. For example, using only the average, we cannot provide answers to questions that are posed in Kimber and Hansford (1993). Or more importantly, we cannot understand the phenomenon underlying cricketing lore that has led to these questions being posed. The questions that are posed relate to the nature of the process of playing and run making that constitutes a batsman’s innings; questions such as: is a batsman more vulnerable (that is, is his probability of getting out or ‘failing’ greater) at the beginning of the innings (when he is yet to ‘settle down’), at the later stages of the innings (when he is tired) and at certain ‘unlucky’ scores (because of psychological reasons)?

In this paper an attempt is made to describe the batsman’s performance not in terms of an average but in terms of the process that unfolds as he plays out his innings. Section 2 of the paper reviews the earlier literature in the area and describes the motivation behind the paper. Section 3 describes the data and the method and Section 4 discusses the results. Section 5 is a conclusion.

2. Literature Review and Motivation

The sparse academic literature in cricket has largely looked at the game either from the point of view of captains and coaches (focusing on batting and bowling strategies) or from the point of view of administrators (focusing on, for example, arriving at a fair result when the game is interrupted and left incomplete due to weather or other reasons; a distinct possibility given the long time spans over which the game stretches).

Examples of work that has focused on devising optimal playing strategies are: Clarke (1988), Clarke and Norman (1999), Preston and Thomas (2000), Swartz et al. (2006) and Rajadhyaksha and Arapostathis (2000). Examples of work that has focused on arriving at a fair result when a game has to be prematurely terminated due to weather conditions or other disturbances are: Duckworth and Lewis (1998), Preston and Thomas (2002) and Carter and Guthrie (2004).

Understanding the underlying playing process that generates the data, and then going on to the development of player-specific performance statistics that capture this process, has, however, received scant attention barring some exceptions (Kimber and Hansford, 1993; Lemmer, 2004; Lewis, 2005; Lewis, 2008).

Chronologically, however, interest in the statistical analysis of cricketers' performance surfaced very early in literature. One of the earliest attempts was by Sir William Elderton (1945). Elderton (1945) observed that the frequency distributions of the individual scores of the players that he studied were 'approximately in Geometric Progression'. Wood (1945) examined the question of the nature of the distribution in greater detail and concluded that the evidence (that the individual scores follow a geometric distribution) was 'definite and unmistakable' in the case of one of the cricketers that he studied and 'is sufficient to lend support to the theory that it *should* apply in every case'. He also observed that scores of zero (a 'duck' in cricket parlance), scores ranging from 1 to 4, and scores of 100 (a 'century') were 'too many'.

This very early interesting and intuitively appealing idea that batsmen's scores follow a geometric distribution has surprisingly not been developed on very much in subsequent literature, though Wood's (1945) work is very often cited. In his paper, in conclusion, Wood (1945) had commented on Elderton's (1945) paper: 'At last a great statistician has discovered what is, I believe, the richest field of statistical material left untilled'. Going by the paucity of literature that has developed on this very early idea, the field still seems left untilled.

The Geometric Distribution is intuitively appealing because of the memoryless property that it possesses. The memoryless property would imply that a batsman's probability of getting out at any score 'n' is constant and independent of the value of 'n' itself. Moreover, in the cricketing context the geometric distribution has certain desirable statistical properties, linked to the presence of 'not out' scores in cricket (Kimber and Hansford, 1993). Kimber and Hansford (1993) also observed that a bathtub hazard model with varying hazard rates (probability of getting out, in the cricketing context) might be a better descriptor of cricketing scores than a simple geometric distribution that assumes a constant hazard rate; they borrowed the term 'hazard rate' from the fields of reliability and survival analysis. Reflecting the surprising lack of work in the area, Kimber and Hansford's (1993) paper, coming after almost fifty years after Wood's (1945) and Elderton's (1993) paper, has only the two 1945 papers as references from the cricketing context; all the other references are from the fields of statistics and reliability. Brewer (2013) uses a Bayesian Survival Analysis method to study the departures from the geometric distribution at scores close to zero. Coming twenty years after Kimber and Hansford's (1993) paper, Brewer's paper lists only Kimber and Hansford's (1993) paper from a cricketing context.

From the limited work done in the area of the analysis of batsmen's scores in cricket it appears that the geometric distribution is a fair descriptor of the distribution of individual scores. There is much scope to develop more on this thinking along various dimensions: are there systematic departures from the geometric distribution and can these be used either to arrive at a 'signature' description of the individual batsman or to arrive at bowling and fielding strategies for the opposing team? Can the concept be built on to arrive at measures that can better describe a batsman and be used by television and other media? Can the idea be transported from the domain of 'runs scored' by a batsman to 'balls faced' by a batsman and if so, can this have implications for strategy formulation or for arriving at better descriptive measures? The existence of this vast, untilled territory provided the motivation for this paper.

This paper builds on the idea of the memoryless property of the geometric distribution and attempts to develop better 'signature' description measures for batsmen. The methods proposed are demonstrated for both runs scored and balls faced. There is no attempt at using the analysis for strategy formulation or for predictive purposes. The methods are demonstrated using statistics for the some of the key members of the Indian One Day International (ODI) cricket team between 1989 and 2012.

3. Data and Method

Since 1989, the Indian One Day International Cricket (ODI) team has seen some very prolific players who have continued playing for long years. Data for seven cricketers who have played ODI for India at some points of time during the period 1989-2012 have been used in the analysis. For each cricketer every match played by him during the period has been included in the analysis. The ‘not out’ scores in the data set have been replaced by the conditional average of the batsman at that score, following the procedure suggested in Damodaran (2006).

Wood (1945) starts off his paper by remarking that Elderton (1945) in his paper ‘makes a rapid progress from “these frequency distributions are approximately in Geometrical Progression” through “If a geometrical progression holds” to “Having noticed that cricket scores were in geometrical progression”. Though Wood (1945) himself and later Kimber and Hansford (1993), Brewer (2013) and others have shown that distributions of individual scores depart from a simple constant hazard rate geometrical distribution, we will, in this paper- with a fair degree of audacity- go ahead and assume that individual cricket scores do indeed follow the constant hazard rate geometrical distribution; in other words, we will in one single leap make the move that Wood (1945) observed Elderton (1945) to have rapidly progressed to. This audacious assumption provides us with a theoretical benchmark the departure from which can be used to arrive at ‘signature’ descriptions of batsmen.

In a constant hazard rate model, for a given cricketer the probability of getting out on a particular score (this probability is the hazard rate) is the same at every score. Formally, a hazard function $H(x) \in [0,1]$ is the probability of the cricketer being dismissed on score x (i.e. $P(X = x)$) given that the batsman is currently on score x (i.e. given $X \geq x$) where $X \in \{0,1,2,3,\dots\}$ is the score that the batsman will make in an innings (Brewer, 2013). In a constant hazard rate model for a particular batsman, $H(x) = \text{a constant } h$ for all levels of x ; and this is the simple geometric distribution.

While the game is on, when the batsman is at a particular score, commentators and spectators of cricket rarely think in terms of the probabilities of the batsman getting out at that score; it is much easier and more natural (and thus more common) to think in terms of how many more runs will the batsman score from thereon. In a constant rate hazard model (i.e. the geometric distribution) the score that the batsman can score from then on is given by the expectation of the geometric distribution $\mu = (1/h) - 1$. Now since h is constant, μ too is invariant with the current score of the batsman. For example, if the value of the hazard rate, h , for a particular batsman is 2% or .02, then μ for this batsman is $(1/.02) - 1 = 49$. This means at whatever score that the batsman is, he can be expected to go on and score 49 runs more. This is the memoryless property of the geometric function; a property that is very well understood in the fields of failure and survivability analysis under reliability engineering.

The procedure that is adopted in the paper is thus to first compute the average for each player adopting the approach suggested in Damodaran (2006) to deal with ‘not out’ scores. This average is then taken to be the estimate of the geometric distribution expectation μ for the player. Under the assumption that the geometric distribution holds, the conditional average for this player at every score x should then be $= x + \mu$; call this E_x . The actual observed empirical value of the conditional average at the score x is then found out; call this O_x . Define δ_x as $O_x - E_x$. If the memoryless property holds exactly then δ_x should be equal to zero for all x . Taking the score x on the X axis and δ_x on the Y axis would help us identify possible signature tipping points for the batsman.

Almost all the literature in cricket has focused on the analysis of runs scored and the nature of distribution of these scores. This analysis of scores is then used to explain phenomena observed in the real world. For example, Brewer (2013) says “It is well known to cricketers of all skill levels that the longer a batsman is in for, the easier batting tends to become”. He then goes on to an analysis of runs scored to develop on this theme. However, drawing on ideas from survivability analysis, it is more intuitive to think of the batsman surviving balls bowled at him rather than runs scored by him; balls bowled at him is what he survives and runs scored are what he makes out of the extra life that he has gained.

So, in this paper, for the selected players, we also try out the procedure by replacing runs scored by balls faced. Again, assuming a geometric distribution for the number of balls faced by an individual batsman, and using the memoryless property of the geometric distribution, the remaining life of the batsman in terms of

number of balls that he will go on to face, can be taken to be a constant and invariant with the number of balls that he has already faced. Since the nature of the distribution of number of balls faced by a batsman has not been much analyzed in the literature, rudimentary goodness of fit tests were done; the Anderson-Darling (AD) values not being very large (and in some cases being very small) the assumption was not rejected. Under this assumption, the δ_x s are again computed (for balls faced) and plotted against balls faced on the X axis to identify ‘signature’ patterns for players.

4. Results and Discussion

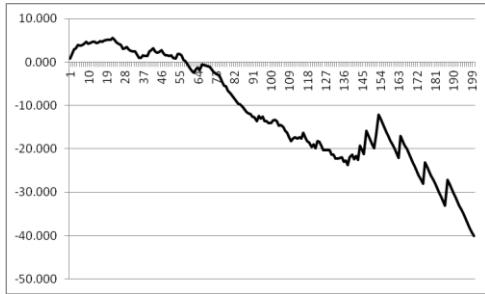
Figures 1 and 2 show the results of the analysis. Running down the left hand side of the panels, the differences between the seven batsmen is apparent. Tendulkar’s signature pattern is distinctive. His most frequently occurring score is zero; a duck. However, once he starts scoring, Tendulkar’s confidence grows as reflected in the series of positive deltas. In fact his deltas are positive right up to the score of fifty-eight; in other words he does better than the theoretical benchmark provided by the geometric distribution right up to the score of fifty-eight. This point, the point at which a player’s delta curve, after entering the positive territory early on, crosses over from the positive territory to the negative territory and remains in negative territory-irreversibly- thereafter, is in a sense the ‘tipping point’; the point at which the player’s performance finally dips irreversibly below the benchmark geometric distribution.

Singh’s and Ganguly’s signatures look similar to that of Tendulkar’s, but with tipping points lower than Tendulkar’s (forty-three for both). Gambhir’s and Kohli’s signature patterns look similar with less pronounced up-climbs from zero and tipping points around a score level of forty five. Sehwag and Dhoni, both known for their hard hitting swashbuckling styles, have very similar profiles. Their profiles differ significantly from those of the other five batsmen: while the other five batsmen seem to draw comfort from runs scored, Sehwag and Dhoni do not seem to become psychologically stronger as they score runs in the initial phase of their innings (or do not seem to need runs to be scored to improve their psychological strength; they do not seem nervous to start with); their graphs of deltas never cross over to the positive territory. So, for Sehwag and Dhoni it is not possible to define tipping points as defined above.

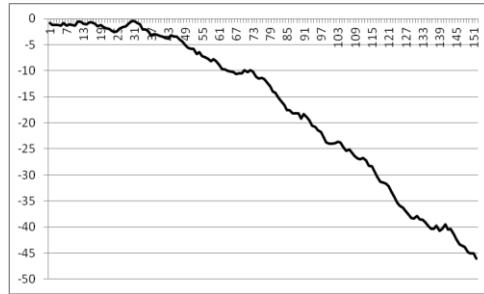
Moving to the right hand side of the panel and running down the profiles of the seven batsmen indicates that an effect similar to the confidence boosting effect of scoring runs is not present in the case of balls faced. In other words, just by successfully facing some balls without getting out, batsmen do not get a boost in terms of increased probabilities of facing more balls successfully. So the author’s initial hunch that ‘playing in’ or ‘settling down’ is more about surviving balls bowled to the batsmen rather than scoring runs does not seem to be vindicated. What seems to be more important for a batsman’s confidence is to score runs rather than occupy the crease. This is consistent with the oft repeated remarks by commentators of the importance of getting ‘runs on the board’. And after all, the game is won not in terms of balls faced, but in terms of runs scored.

Though we have defined the tipping points as points at which the delta curves cross over from the positive to the negative territory irreversibly, it is apparent that there are other sub-patterns in the delta curves. Both in the positive and negative territory the delta curve might rise or drop. Taking a second order difference of deltas and defining these as $\delta_{x''} = \delta_x - \delta_x$ helps us capture these points. A positive $\delta_{x''}$ indicates that at the score of x, while the score itself has incremented by one run from the previous score, the conditional average has incremented by more than one run; at the score x the batsman has done better than the benchmark geometric distribution. To separate the significant values of $\delta_{x''}$ from the insignificant, we compute the mean and the standard deviation of the $\delta_{x''}$ s for each player. Values of $\delta_{x''}$ more than one standard deviation above the mean are then identified as ‘josh’ (Hindi word for ‘energy’) points. This analysis has only been done for the deltas based on scores and not for the deltas based on balls. Moreover, it has only been done for two batsmen, Tendulkar and Sehwag. Figure 3 shows the results in terms of josh points for Tendulkar (the lower line) and Sehwag (the upper line). It is clear that Tendulkar has more josh points at the beginning of the innings. That means he gets energy on crossing these points; it also simultaneously means that before crossing these points, these points were points of stress for him. This is further reinforced by the fact that both players have josh points in the ‘nervous nineties’

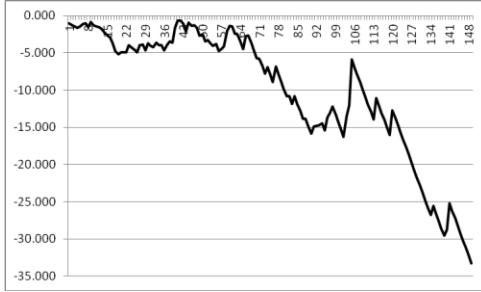
S. Tendulkar: Plot of Deltas, Runs Scored



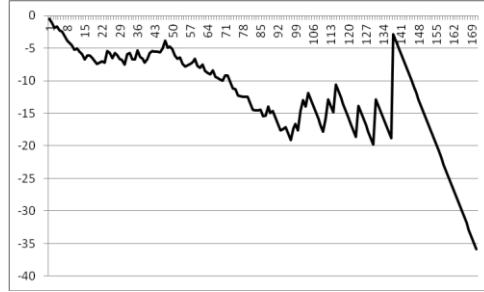
S. Tendulkar: Plot of Deltas, Balls Faced



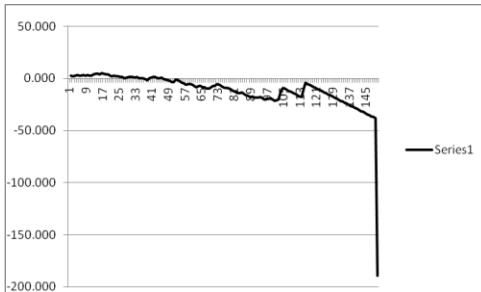
V. Sehwag: Plot of Deltas, Runs Scored



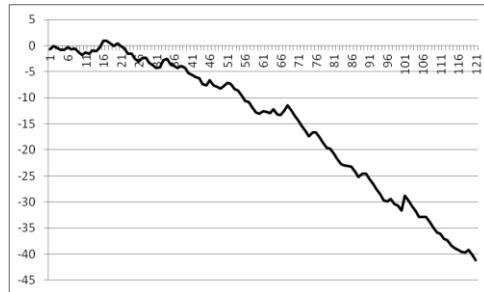
V. Sehwag: Plot of Deltas, Balls Faced



G. Gambhir: Plot of Deltas, Runs Scored



G. Gambhir: Plot of Deltas, Balls Faced



Y. Singh: Plot of Deltas, Runs Scored



Y. Singh: Plot of Deltas, Balls Faced

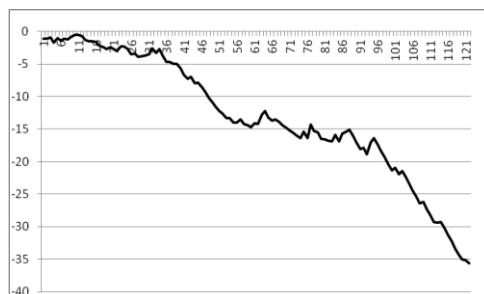
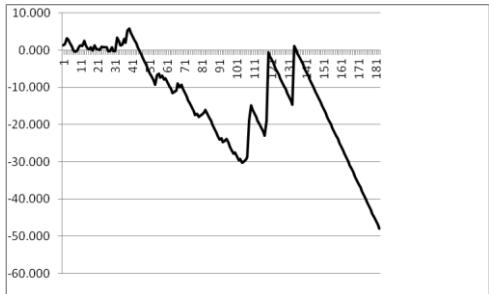
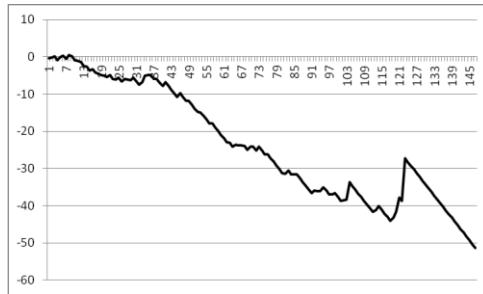


Figure 1: Delta Plots for Tendulkar, Sehwag, Gambhir, Singh

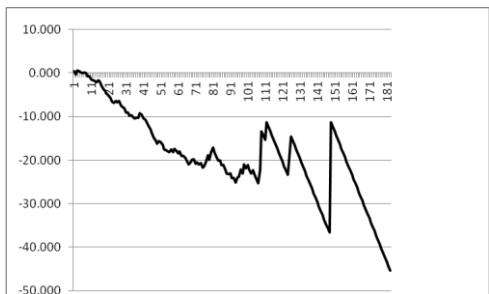
V. Kohli: Plot of Deltas, Runs Scored



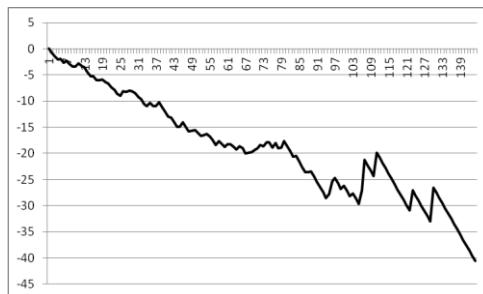
V. Kohli: Plot of Deltas, Balls Faced



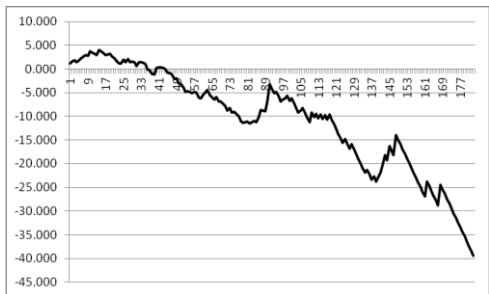
M.S. Dhoni: Plot of Deltas, Runs Scored



M.S. Dhoni: Plot of Deltas, Balls Faced



S. Ganguly: Plot of Deltas, Runs Scored



S. Ganguly: Plot of Deltas, Balls Faced

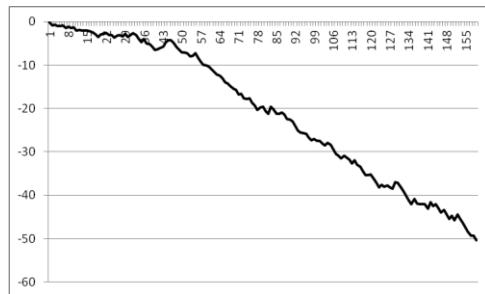


Figure 2: Delta Plots for Kohli, Dhoni, Ganguly

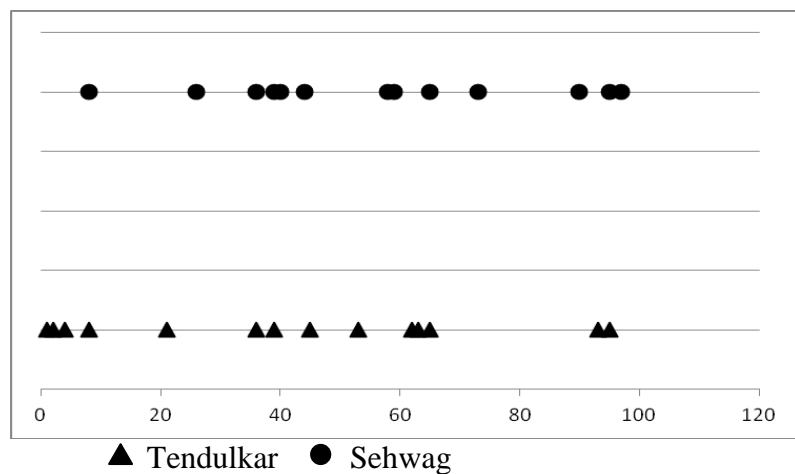


Figure 3: Plot of josh points for Tendulkar, Sehwag

5. Conclusion

The paper demonstrated a simple procedure of using the memoryless property of the geometric distribution to arrive at performance descriptors for batsmen in cricket. The process was studied using both runs scored and balls faced. Used on a larger sample, the procedure can be used to identify systematic patterns in player behavior, both on scores and balls faced.

References

- Brewer, B.J. (2013): Getting Your Eye In: A Bayesian Analysis of Early Dismissals in Cricket. arXiv:0801.4408v2 26 May 2008 [stat.AP]. Retrieved on March 22, 2013
- Carter, M. and Guthrie, G. (2004) Cricket interruptions: fairness and incentive in limited overs cricket matches. *Journal of the Operational Research Society* **55**(8), 822-829.
- Clarke S.R. (1988) Dynamic Programming in one-day cricket- optimal scoring rates. *Journal of the Operational Research Society* **39**(4), 331-337
- Clarke S.R. and Norman J.M. (1999) To run or not?: Some dynamic programming models in cricket. *Journal of the Operational Research Society* **50** (5), 536-545.
- Damodaran, U. (2006) Stochastic Dominance and Analysis of ODI Batting Performance: The Indian Cricket Team, 1989-2005. *Journal of Sports Science and Medicine* **5**, 503-508
- Duckworth, F.C. and Lewis, A.J. (1998) A fair method of resetting the target in interrupted one-day cricket matches. *Journal of the Operational Research Society* **49**, 220-227.
- Elderton, W. (1945) Cricket Scores and Some Skew Correlation Distributions (An Arithmetical Study) *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **108**, 1-11.
- Kimber, A.C. and Hansford, A.R. (1993) A Statistical Analysis of Batting in Cricket. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **156**, 443-455
- Lemmer, H.H. (2004) A measure for the batting performance of cricket players. *South African Journal for Research in Sport, Physical Education and Recreation* **26**(1), 55-64
- Lewis, A.J. (2005) Towards fairer measures of player performance in one-day cricket. *The Journal of the Operational Research Society* **56**(7), 804-815.
- Lewis, A.J. (2008) Extending the Range of Player-Performance Measures in One-Day Cricket. *The Journal of the Operational Research Society* **59**(6), 729-742.
- Preston, I. and Thomas, J. (2000) Batting Strategy in Limited Overs Cricket. *Journal of the Royal Statistical Society: Series D (The Statistician)* **49**, 95-106. Preston, I. and Thomas, J. (2002) Batting Strategy in Limited Overs Cricket. *Journal of the Royal Statistical Society: Series D (The Statistician)* **51**, 189-202.
- Rajadhyaksha, G. and Arapostathis, A. (2000) Using a Bayesian network to recommend the best bowling action. Available from [URL:<http://webspace.utexas.edu/gaureshr/pubs/cricket.pdf>](http://webspace.utexas.edu/gaureshr/pubs/cricket.pdf)
- Swartz, T.B., Gill, P.S., Beaudoin, D. and de Silva, B.M. (2006) Optimal batting orders in one-day cricket. *Computers and Operations Research* **33** 1939-1950
- Wood, G.H. (1945) Cricket scores and geometric progressions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **108**, 12-22.

Inferring the score of a tennis match from in-play betting exchange markets

A. Gabriela Irina Dumitrescu*, B. Xinzhuo Huang**, C. William Knottenbelt***, D. Demetris Spanias****,
E. James Wozniak*****

Department of Computing, Imperial College London, South Kensington Campus, London, SW7 2AZ

* irinadumi@gmail.com

** xh208@ic.ac.uk

*** wk@doc.ic.ac.uk

**** d.spanias10@imperial.ac.uk

***** jhwozniak@gmail.com

Abstract. Over the past few years, betting exchanges have attracted a large variety of customers ranging from casual human participants to sophisticated automated trading agents. Professional tennis matches give rise to a number of betting exchange markets, which vary in liquidity. A problem faced by participants in tennis-related betting exchange markets is the lack of a reliable, low-latency and low-cost point-to-point score feed. Using existing quantitative tennis models, this paper demonstrates that it is possible to infer the score of a tennis match solely from live Match Odds betting exchange market data, assuming it has sufficient liquidity. By comparing the implied odds generated by our quantitative model during play with market data retrieved from the betting exchange, we devise an algorithm that detects when a point is scored by either player. This algorithm is further refined by identifying scenarios where false positives or misses occur and heuristically correcting for them. Testing this algorithm using live matches, we demonstrate that this idea is not only feasible but in fact is also capable of deducing the score of entire sets with few errors. While errors are still common and can lead to a derailment of the detection algorithm, with more work as well as improved data collection, the system has the potential of becoming a precise tool for score inference.

1. Introduction

Tennis is one of the more popular sports in the world, enjoyed by millions of spectators. The popularity of the sport makes it attractive to betting exchanges which offer a range of betting markets related for nearly all professional tennis matches. One of the largest betting exchanges is Betfair which offers its customers the opportunity to take up (back) or offer (lay) bets in a peer-to-peer manner. These bets can take place before the start of the match or during play. Being a large betting exchange, Betfair attracts a large volume of bets from customers which in turn results in very liquid markets – especially for the more popular professional tennis matches. High liquidity in an exchange market translates to exchange odds which change at a high speed and which react very efficiently to real life on-court events (such as a point being scored in a match).

This high liquidity in the exchange market is provided by bets placed by both humans and automated trading systems alike. Like in finance, the majority of liquidity is provided by automated systems which trade algorithmically. Any algorithmic trading software has some underlying quantitative model which drives the decisions made by the software. Speculating on how these models work is out of the scope of this paper, but certainly there must be an input for the current state of the match (like the score).

A good in-play automated trading system must monitor and adapt to real life on-court events. It should immediately change the target odds it trades at according to the current state of the game. As most systems do that in an efficient way, the market odds offered on the exchange appear to adapt very quickly to represent the state of the match.

A problem that new systems may face is the automated retrieval of data which can be used to identify the current state of the match. Manual entry of the current score of the match has been a common method used by current professional systems. The problem with this is that it is not cost effective and it is prone to data entry errors. A more reliable source of the current score of a tennis match is a score feed connected directly to the umpires' chair of every match but this is expensive to acquire. There are some other free sources available online, like the ATP World Tour official live scoreboard but most of these are susceptible to large delays and do not provide the data in an easily retrievable way. This paper aims to introduce a system which takes advantage of the variable odds of bets matched on a betting exchange to infer the current state of the

match. The target result is to be able to provide a reliable score feed of any tennis match with sufficient liquidity in its betting exchange market.

2. Hierarchical Markov Model

The hierarchical scoring structure of tennis allows for the modelling of games, sets and matches through Markov Chains which are linked together recursively. The method, described in Liu (2001) is further adapted to account for the difference in the probability of a player winning a point when (s)he plays the role of the server or the receiver. This section details how Hierarchical Markov Chains can be employed to determine a player's odds in a best-of-three sets match – a best-of-five match can be modelled analogously.

As one of the key characteristics of Markov Chains is memorylessness – the next state that the system transitions to depends only on the current state; previous states have no impact on the transition. Consequently, this technique assumes that points in tennis are independent and identically distributed (i.i.d.). Klaassen & Magnus (2001) show that while the i.i.d. hypothesis is not strictly a realistic assumption, the divergence from this in reality is small enough that making the i.i.d. assumption can produce good results.

The most straightforward way to illustrate this technique is by starting at the highest level – that of the match – and work our way down to the lowest level; a single point. The match level is the simplest one, as it is comprised of a finite set of states, or set scores. A best of 3 match, starts with the 0-0 score, and ends when a player is first to win 2 sets. Let p be the probability that player A wins a set, and $(1-p)$ the probability of A losing a set. Using recursion, one can calculate the probability of A winning the match from any level in the chain, knowing that in the final two states, the probability of winning the match is either 0 (player A already lost) or 1 (player A already won). For example, given a 1-1 set score, to compute A's probability of winning the match, we calculate p multiplied by probability of winning the match from the Win final state (i.e. 1) plus $(1-p)$ multiplied by the probability of winning the match from the Lose final state (i.e. 0).

$$P(A \text{ wins } | 1 - 1) = p \times 1 + (1 - p) \times 0 = p$$

We then can use the result obtained above to compute the probability of A winning given a 1-0 score:

$$P(A \text{ wins } | 1 - 0) = p \times 1 + (1 - p) \times P(A \text{ wins } | 1 - 1) = 2p - p^2$$

The Markov Chain for the set level is similarly constructed, with a few variations. There is a differentiation between the probability, p_A , of player A winning a game when A is the server and the probability, p_B , of Player B winning a game when B is the server.

Assuming the set starts with A serving, then p_A shall be used to compute the probability of A winning at every even game, and $1-p_B$ is the probability of Player A winning at every odd game. For a tie-break set, the Markov Chain has one special case, as the probability of moving from 6-6 to either final state is computed using p' , the probability that A wins the tie-break.

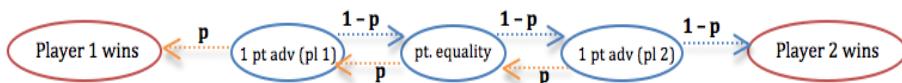


Figure 1 – Random walk with two absorbing states.

The advantage set model, the tie-break model and the game model all share one common characteristic – if there is a tie before winning, one of the two players needs to win two consecutive games/points to win the set/tiebreak/game. These two steps to victory are defined by a random walk with two absorbing states as shown in Figure 1.

From the point equality state in a game, the only way to reach the winning state is by winning two consecutive points (p^2), and analogously for losing the game $(1 - p)^2$. However, there is an indefinite number of transitions that can occur between the deuce point and either of the advantage points.

To return from deuce point back to itself, one of two things can occur: either player A loses a point and then wins the next one or player A wins the point and loses the next. The corresponding probabilities are:

$$p \times (1 - p) + (1 - p) \times p = 2p \times (1 - p)$$

Since this can happen an infinite number of times, the probability of winning the point from deuce is:

$$\sum_{k=0}^{\infty} [2p \times (1 - p)]^k = \frac{1}{1 - [2p \times (1 - p)]}$$

More detail on recursively calculating a player's probability of winning the match from any score given the two players' probabilities of winning on serve (PWOS) can be found in Barnett & Clarke (2002). Figure 2 demonstrates the end-result of this hierarchical Markov chain showing how the probability of Player 1 winning the match alters as the PWOS parameters of the two players are varied. On the left graph of the figure we can see the probabilities at the beginning of the match and on the right graph we can see how the odds have changed in favour of Player2 when the score-line is 0-1 Sets and 0-5 Games.

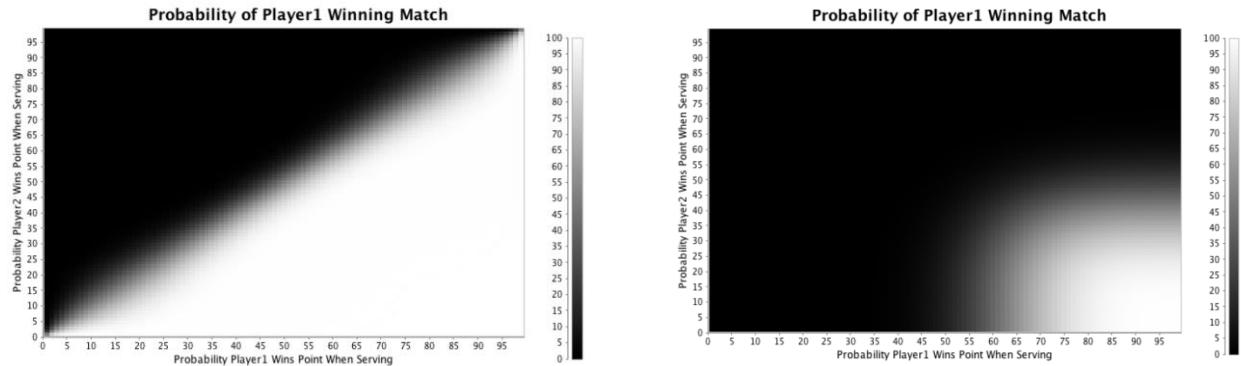


Figure 2 - A representation of the Probability of Player1 winning the match varying the probabilities of the players winning points while serving. (Left: from a score-line of 0-0 Sets, 0-0 Games, Right: from a score-line of 0-1 Sets, 0-5 Games and Player1 Serving)

Having discussed the higher levels of the hierarchical Markov model, all that is required to complete the model is the estimation of the parameter of the PWOS of the players involved in a match. These are the parameters which are used in the models to determine the probabilities of each player winning a game while serving and winning a tie-breaker.

There have been a few proposed methods for estimating the value of the PWOS parameters. A simplistic way would be to use the average number of points won while serving by each player over a period of matches and divide by the total points played while serving in the same period. Barnett & Clarke (2005) propose a technique of estimating PWOS by combining on a high level the statistics of the players compared and Spanias & Knottenbelt (2012) further refine this idea by modeling the point itself as a Markov chain to take into account individual strengths and weaknesses of players. Newton & Aslam (2009) also build upon Barnett & Clarke's idea by incorporating player instability in the estimation of the PWOS parameters. In Section 3.2 of this paper we will discuss a new concept for calculating adaptable PWOS parameters which take into consideration match odds from betting exchanges.

3. Tennis Score Inference

In order to do any sort of score inference from betting exchange odds, one first needs to estimate the implied-PWOS of both players from the live match odds, and then given the current known score, detect when a new point is scored. In this section we describe how one can calculate the implied probability of a player winning the match from the exchange matched odds. We then proceed to show a basic method of detecting when points are scored and then make suggestions on how to heuristically improve on this basic method.

3.1. Implied probability of winning the match from exchange odds

Live betting odds can be obtained from the world's largest betting exchange, Betfair. In our paper we used the Betfair API to retrieve information about the Match Odds betting market for a particular match at regular intervals (every second or so), for its entire duration.

The odds information supplied by Betfair is given in a decimal format, for example the odds of player A winning might be 2.36. These can be easily transformed into the probability of player A winning the match by simply reciprocating the decimal odds.

$$P(\text{player wins match}) = 1 / \text{decimal odds for player winning match} = 1/2.36 = 42.4\%$$

Hence, when the market displays decimal odds of value 2.36 for a player, this implies that the market roughly perceives the player to have 42.4% chance of winning the match. It is this implied probability which we will later use to infer the score of a match.

Another issue to consider is the potential of overrounds within the data collected from Betfair. Betfair's cross matching algorithm will sometimes round the odds offered. As a result, the sum of both players' implied probabilities, as taken from Betfair, sometimes exceeds 100%. The following example is taken from the Federer – Del Potro match (2012 Olympics), roughly two hours into the match. The odds are 1.46, and 3.0 for Federer and Del Potro, respectively. Converting these to probabilities results in a 68.49% probability in favour of Federer, and a 33.33% probability in favour of Del Potro. The resulting sum is 101.8%. To eliminate this overround, the values are normalized so that they sum to 100%.

3.2 Market Adaptive PWOS

In this paper we introduce a new method to calculate the PWOS of players involved which also takes into account the implied match winning probability provided by the exchange odds. The PWOS calculated in this way serves the purpose of inferring the score from the exchange odds in a much more efficient way as it accounts for offsets and adapts to the PWOS introduced by the market.

In their research, Klaassen & Magnus (2001) figured out the average sum of the PWOS for men is of 1.29, and 1.12 for women, with little variance. O'Malley (2008) shows that the probability of a player winning the match is determined by the relative difference between the two PWOS, rather than the absolute PWOS values. This is also visible in Figure 2 as it is the difference in the two PWOS values that cause the change in colour of the graph and not the absolute value. Therefore, given two equally strong male players, each of their PWOS values is approximated as $1.29/2 = 0.645$. For players with different strengths, let 0.645 be the pivot value, and have the two players' PWOS values lay symmetrically on either side of the pivot:

$$\begin{aligned}\text{PWOS(Player A)} &= 0.645 + \text{difference}/2, \\ \text{PWOS(Player B)} &= 0.645 - \text{difference}/2.\end{aligned}$$

The PWOS difference can then be calculated by solving for the difference whose match probability would yield the same normalised implied odds that can be found in the Betfair Match Odds market. The PWOS difference can therefore be estimated using the following binary search type algorithm.

```

calculatePwosDifference (impliedProb, score ){
    double pwosDifference = 0
    double increment = 0.5
    double calculatedProb = Calculator.calculateMatchWinningProb(score, pwosDifference)
    while (calculatedProb != impliedProb){
        if(calculatedProb < impliedProb){
            pwosDifference += increment
        }
        else if (calculatedProb > impliedProb) {
            pwosDifference -= increment
        }
        increment = increment/2
        calculatedProb = Calculator.calculateMatchWinningProb(score, pwosDifference)
    }
}

```

3.3. Score Inference

Before trying to infer the score, it is important to be able to extract an accurate probability for a player to win the match based on the activity on the betting market. We found that the best results came from using the ‘last price matched’ (LPM) values from Betfair, which gave the odds over which the last bet was made on the exchange.

To begin the score inference process, the current score-line (e.g. at the start of a match) must be known. The PWOS values for each player can then be calculated as outlined in Section 3.2, using the latest odds information from Betfair. By initializing our hierarchical Markov model with these PWOS values, we could therefore calculate the probability of either player winning the match from any given score-line, and hence get an idea of the odds we would expect to see from Betfair whilst the match is in play.

To track the score on a point by point basis, we chose one of the players and used our model to calculate the odds we would expect to see if the next point was won or lost. The actual live odds were then examined to see if there were any such changes. To allow for a slight variation from our model, we calculated a pair of threshold values that gave a fixed degree of leniency when testing the current Betfair odds against our predictions. Figure 3 gives a visual representation of the threshold values created for each point.

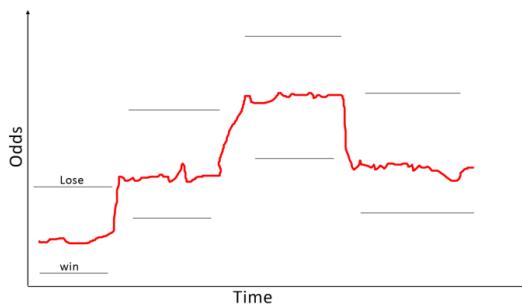


Figure 3 - Thresholds set while tracking live odds to detect the outcome of a point.

If the actual market odds breached either of the threshold values, we assumed a point had been resolved and the assumed score-line was appropriately updated. The new market odds were then used to recalibrate the PWOS estimations in attempt to capture player momentum, and the threshold process was repeated.

The success of this method was limited as the algorithm often either missed a point or detected a false positive. As this method is based on having a correct prediction of the score when the threshold values are generated, errors are carried over and rapidly derail the predicted score-line. From analysing the score inference errors alongside the correct score feed, there are occasions where the odds altered significantly

without any change in the actual score-line, leading to a false positive. Figure 4 (Bartoli/Schiavone) shows a circular region where a point was predicted incorrectly.

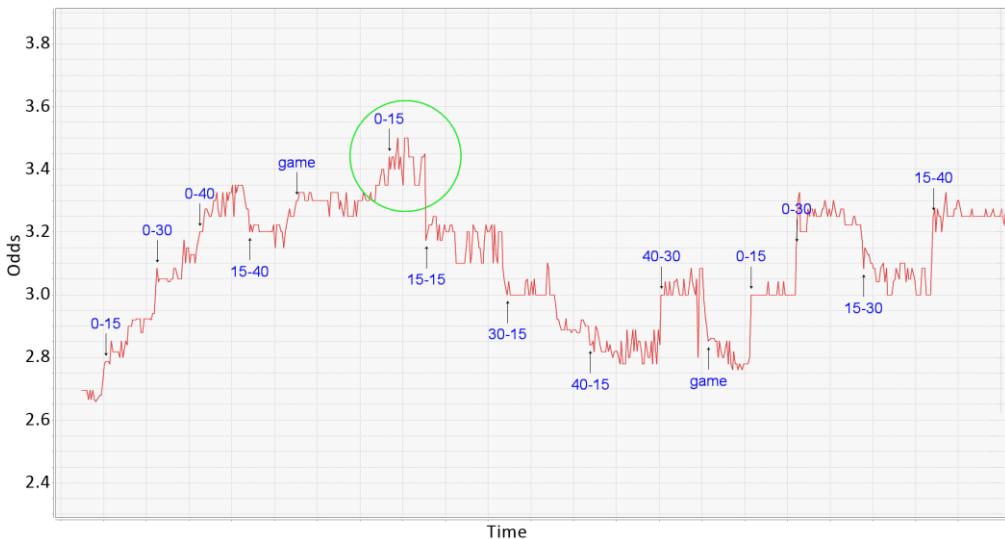


Figure 4 - Schiavone vs. Bartoli - Women's singles Semi final. Roland Garros 2011. Graph of Bartoli's odds during first/second game

In this case, looking at the match itself, the change in odds may be attributed to Schiavone missing an easy smash shot that in reality should have almost guaranteed her the point. For the financial incentive, bets may have been placed preemptively on Schiavone winning the point, causing a temporary shift in the market odds. Such activity highlights the need for any algorithm to be robust and able to handle errors should they occur.

3.4 Improvements

We can significantly improve inference accuracy by using additional heuristics. These techniques reduce false-positives and improve reliability by recognising and accounting for various market behaviour patterns.

3.4.1 Accounting for Volatility

False-positives often arise due to odds fluctuation between points. Therefore, we dynamically calculate odds volatility throughout a match and adjust point-scoring thresholds accordingly. Thresholds are widened when volatility is greater and vice versa. Knowledge of relative volatility is also useful in other heuristics, such as that described in section 3.4.2.

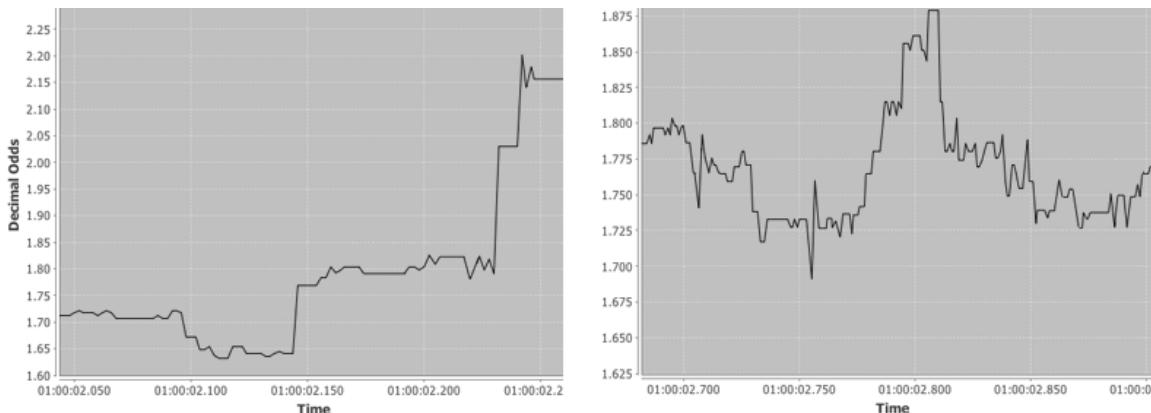


Figure 5 – LHS diagram shows match where volatility is low, score changes can easily be detected with narrow thresholds. RHS diagram shows match with high volatility. (Wozniacki vs. Pennetta - Qatar Ladies Open 2011 & Troicki vs. Djokovic - Sony Ericsson Open 2011)

We take volatility as the interquartile range of odds over the most recent 10 second period, updated continuously. Any statistical measure of spread may be used but we found interquartile range to be most suitable due to its robustness – it is less affected by small numbers of outliers than standard deviation, range, etc.

3.4.2 Recognising Large Odds Change as Scoring

We not only want to suppress false-positives but also ensure all scoring is captured. A change in odds is determined to be a scoring event if it is sufficiently large, sharp and sustained, even if the threshold is not crossed. *Sufficiently large* is defined relative to the difference between thresholds as well as adjusted for volatility, as described in section 3.4.1. When volatility is low, magnitude of odds change required for inference of scoring is lower than when volatility is high.

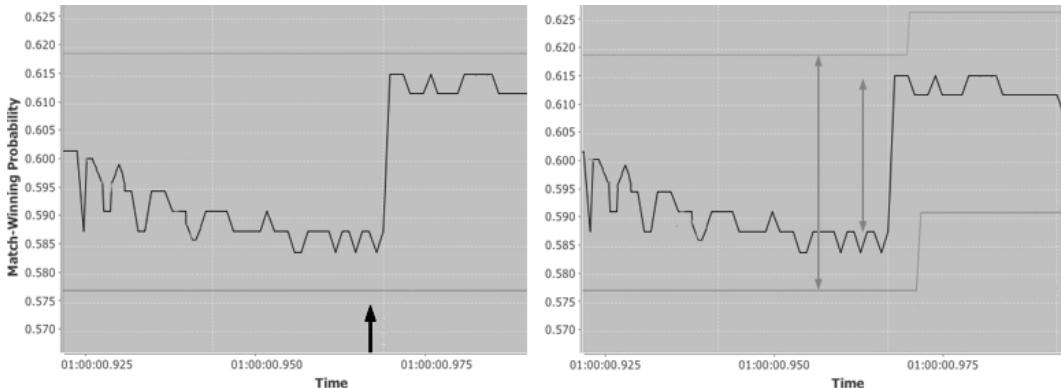


Figure 6 - Identical situation before and after implementing large odds change detection. Arrow on LHS indicates time of point scoring and subsequent odds change. Arrows in RHS denote threshold difference and odds change to which it is compared. In the RHS scenario, the program correctly detects scoring. (Wozniacki vs. Pennetta - Qatar Ladies Open 2011)

This is an important heuristic as we estimate that, out of all inferred points, two thirds are attributed to threshold crossing and the rest to odds-change detection.

3.4.3 Periodic Point Synchronisation

An incorrect inference of a single point may offset the score for the remaining match. Therefore, we periodically synchronise the score throughout the match. During changes of serve and end of sets, the match odds remain relatively constant for long periods. On detection of such patterns, we may recognise that our current score inference is incorrect and adjust it to the nearest game or set accordingly.

4. Case Study - Cilic vs Baghdatis, Roger's Cup 2012

Time-stamped point-by-point data is not freely available in bulk in tennis. Consequently, the results we present in this paper are based on manually collected points and synchronized in time with live market information from Betfair.

Figure 7 shows Cilic's match-winning probability as the match progresses, in his 2012 Roger's cup match against Baghdatis played in Toronto, and compares the expected match-winning probability (calculated using a pair of static PWOS values and the actual scoreline) and the probability implied by the Betfair odds. It can be seen that, while the trends of the two curves are very similar, the two values can diverge, emphasising the importance of dynamically updating the PWOS values to agree with the market odds.



Figure 7- A comparison of modelled match odds with Betfair implied odds of Cilic winning the match against Baghdatis in Roger's Cup 2012.

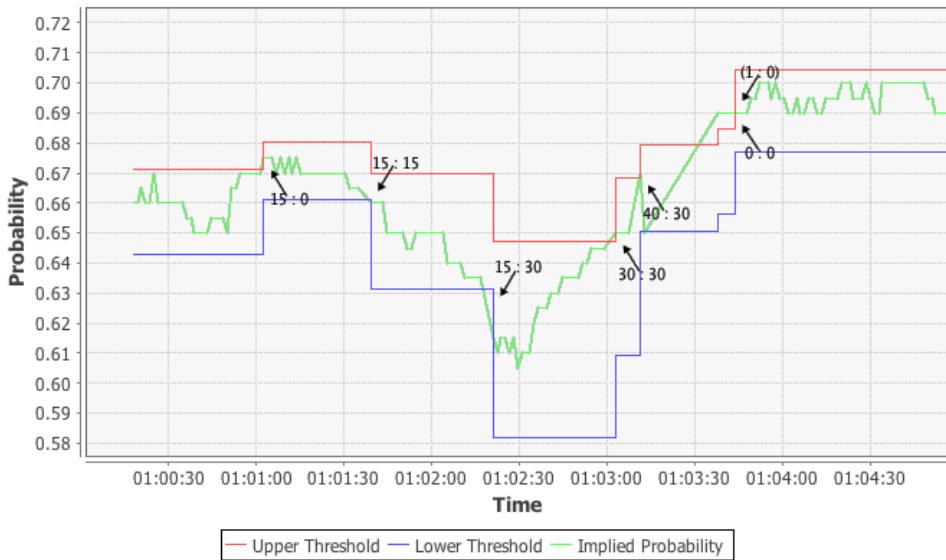


Figure 8 – First game of the Cilic vs. Baghdatis match in Roger's Cup 2012

Figure 8 shows the first game of the Cilic vs. Baghdatis match, and the corresponding thresholds used to infer scores. All of the inferred scores are correct. We note that the incorrect inference of a point being scored shortly after 40-30 is avoided by the use of a timeout mechanism that enforces a delay of at least 8 seconds between points.

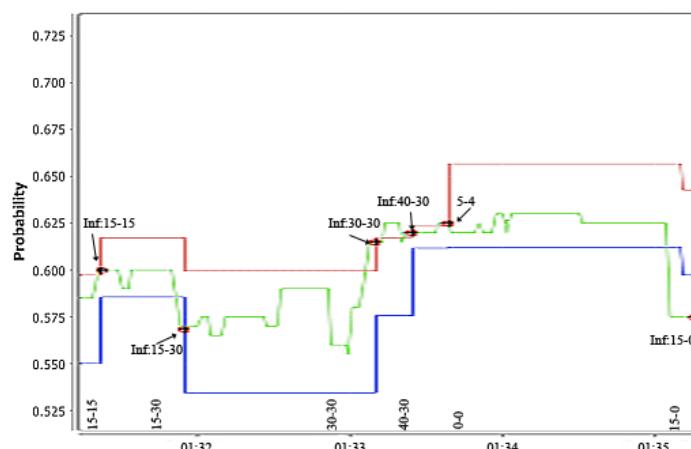


Figure 9 - Live capture of the system's point thresholds and the Betfair implied odds in the match Cilic vs. Baghdatis Roger's Cup 2012. The dots represent the different points detected.

Points continue to be detected correctly up to 5-4 in the first set, the points around which are shown in Figure 9. We note the delayed detection of the 30-30 point, which we speculate might have been caused by market uncertainty about the point outcome (e.g. because of a challenge). In the next game, the inference of a key point is missed because of an insufficient movement in market odds. This rapidly leads to a decoupling of the model and market's implied match odds (even when the former uses dynamically updated PWOS values, since these are now based on different score lines), resulting in cascading point errors. Indeed, while the final score is 7-5, 6-3 in favour of Cilic, the score-detection algorithm infers the final score to be 6-4, 5-3.

5. Conclusion

Using a hierarchical Markov scoring model, the parameters of which automatically adapt to live betting exchange odds, we have managed to build a system which can infer when points are scored in a match from live betting exchange market data. This is done by detecting when the live odds cross some pre-estimated upper or lower bound and depending on which bound was crossed the system interprets it as a point scored by a particular player. The system is not perfect and any errors made in the form of false positives or missed points carry over to future deductions can quickly destabilize the system. Nevertheless heuristic improvements have been added to the system and are effective at improving detection of points and reducing the effect of errors on future deductions.

References

- Barnett, T. J. & Clarke, S. R., (2002) Using Microsoft Excel to model a tennis match. *6th Australian Conference on Mathematics and Computers in Sport* (G. Cohen ed.), pp. 63-68.
- Barnett, T. J. & Clarke, S. R., (2005) Combining player statistics to predict outcomes of tennis matches. *IMA Journal of Management Mathematics*, pp. 113-120.
- Klaassen, F. J. G. M. & Magnus, J. R., (2001) Are Points in Tennis Independent and Identically Distributed? Evidence from a Dynamic Binary Panel Data Model. *Journal of the American Statistical Association*, 96(454), pp. 500-509.
- Knottenbelt, W. J., Spanias, D. & Madurska, A. M., (2012) A common-opponent stochastic model for predicting the outcome of professional tennis matches. *Computers & Mathematics with Applications*, 64(12), pp. 3820 - 3827.
- Liu, Y., (2001) Random walks in tennis.. *Missouri Journal of Mathematical Sciences*, 13(3).
- Newton, P. K. & Aslam, K., (2009) Monte Carlo Tennis: A Stochastic Markov Chain Model. *Journal of Quantitative Analysis in Sports*, 4(3), pp. 1-42.
- O'Malley, J. A., (2008) Probability Formulas and Statistical Analysis in Tennis. *Journal of Quantitative Analysis in Sports*, 4(2), p. 15.
- Spanias, D. & Knottenbelt, W., (2012) Predicting the outcomes of tennis matches using a low-level point model.. *IMA Journal of Management Mathematics*.

Markov Chain Volleyball

M. Ferrante* and G. Fonseca**

*Dip. di Matematica, Università di Padova, via Trieste 63, 35121 Padova – Italy. Email address:
ferrante@math.unipd.it

**Dip. Di Scienze Economiche e Statistiche, Università di Udine, via Tomadini 30/A, 33100 Udine – Italy. Email address: giovanni.fonseca@uniud.it

Abstract. In this paper we consider the volleyball under the Markovian assumption that the probability of winning a single rally is independent of the other rallies and constant during the game. Fixing two parameters which indicate the probabilities of winning a rally for the serving team, we derive the exact expression of the probability of winning a set and a match in the present rally point- and in the former side-out scoring systems. We observe that the present point system reduces the winning probability of the stronger team, adding interest/randomness to the game. Furthermore we study the mean duration of the games in both the scoring systems, obtaining, as well known in the practice, that this change reduced the (expected) length of the matches.

1. Introduction

Volleyball, as well as Tennis, Badminton and Squash, is a game that can be nicely described with a simple mathematical model. Assuming that the probability that a team wins any point is independent of the previous rallies and constant during the game, it is immediate to prove that the score of a single set and of the whole match can be described by homogeneous Markov chains. Even though the model is simple to be described, the rules of this sport (as opposite to other games, e.g. the tennis, the serving team is the one that scored the last point) makes the model huge and difficult to be handled theoretically. For example, the stochastic process that describes the volleyball set has 1254 states and the direct use of the powerful results of the Markov chain theory is in this case of little interest. On the contrary, using a recursive approach, as done by Lee and Chin (2004) for a set of volleyball and by Newton and Keller (2005) for the whole tennis match, we derive in Section 2. a simple and usable expression for the conditional probabilities that the team that starts serving, wins the set and then wins the match. We then apply the same computation to the former scoring system (partially considered by Simmons (1989)), known as side-out scoring system, while the present one is known as rally point scoring system. Comparing the probabilities in the two cases, we show that the change in the scoring system facilitated the weaker teams and introduced a source of randomness in the outcomes of the sets (and therefore of the matches).

Since the scoring system has changed thirteen years ago in order to make this sport more spectator- and television-friendly, being now the matches of a more predictable length, we consider the expected duration of a given set. In the rally point scoring system it is possible to compute again explicitly this expectation, while in the side-out scoring system we approximate the value using a suitable simulation on a large number of independent sets. Results of this simulation, and a similar one carried out for the present scoring system, show that both the mean duration (in terms of number of rallies) and the standard deviation of the duration decreased consistently, in accord to the goal of making the duration of the match shorter and more predictable.

In the last section we use data from the Italian Volley League in order to check if the model is close to reality, at least for top teams.

To conclude, let us spend just few words on our model. At first sight, this model appear too simple to catch the variety of a single set and the sequence of psychological constraints that different scoring situations during a match can provide to the players. Nevertheless, the training that it is carried out on the player, since they are very young, is that they have to try to forget the previous rally and play any single rally as the first one. This is clearly the spirit of the present

model and in this sense the present results can be thought as a theoretical benchmark. In conclusion, the difference between a normal player and a champion is that the latter is able to play the decisive point of a match as this were any other point. As a joke, we could say that a champion is at the end nothing else than a Markovian player.

2. Winning probabilities: Set

Let us consider the following model for a set of volleyball. The probability that a team wins each point is constant during the set, independent from the other points played and depends just on the fact that the team serves or returns the serve. So, calling the two teams 0 and 1, we will define two parameters p_0 and p_1 which represents, respectively, the probabilities of winning a point when the team 0 or 1 serves. To avoid trivial cases, we will always assume that $0 < p_0 < 1$ and $0 < p_1 < 1$. Furthermore and in contrast to the similar model for the tennis given e.g. in Newton and Keller (2005), it will be here reasonable to consider both these numbers less than 0.5 (see the final section for the estimated values in the Italian Volley League).

In order to analyze the probability of winning a set under these assumptions, one recognizes that the score of a set can be thought as the realization of a discrete-time Markov chain, whose transition matrix will be specified in the sequel. Since the scoring system has recently changed, we will consider separately the two cases, starting from the present system.

2.1 Rally point scoring system

Let us start by defining the set S of the states of the Markov chain that describes the evolution of a volleyball set under the rally point scoring system. We shall define

$$S := \{(i, j, s) : i \in \{0, 1, \dots, 24, Ad, W\}, j \in \{0, 1, \dots, 24\}, s \in \{0, 1\}\}$$

where the first number represents the score of the serving team, the states Ad and W in the first position stand for Advantage and Winning of the serving team, and similarly for the numbers in second position relative to the returning team, while the last number indicates which team serves. State Ad includes all the scores larger than 24 with the team serving for a set-point. It is simple to see that the transition probabilities are defined as follows: when $\max\{i, j\} < 24$ then

$$(i, j, s) \rightarrow (i+1, j, s) \text{ with probability } p_s$$

$$(i, j, s) \rightarrow (j+1, i, 1 - s) \text{ with probability } 1 - p_s$$

while

$$(23, 24, s) \rightarrow (24, 24, s) \text{ with probability } p_s$$

$$(23, 24, s) \rightarrow (W, 23, 1 - s) \text{ with probability } 1 - p_s$$

$$(24, 23, s) \rightarrow (W, 23, s) \text{ with probability } p_s$$

$$(24, 23, s) \rightarrow (24, 24, 1 - s) \text{ with probability } 1 - p_s$$

$$(24, 24, s) \rightarrow (Ad, 24, s) \text{ with probability } p_s$$

$$(24, 24, s) \rightarrow (Ad, 24, 1 - s) \text{ with probability } 1 - p_s$$

$(Ad, 24, s) \rightarrow (W, 24, s)$ with probability p_s

$(Ad, 24, s) \rightarrow (24, 24, 1 - s)$ with probability $1 - p_s$

Let us now compute the conditional probability that the team who starts serving, wins the set. In order to obtain the above probability, we have to compute the probabilities that the Markov chain starting from the state $(0, 0, s)$ reaches one of the states $(W, 0, s), (W, 1, s), \dots, (W, 23, s), (W, 24, s)$. One possible approach would be to consider the whole Markov chain and to compute the absorbing probabilities of these states starting from $(0, 0, s)$. Although this is theoretically correct, it is not viable in practice, since the Markov chain representing a volleyball set can be described by a huge 1265×1265 transition matrix, not suitable for any, at least simple, computation.

As an alternative (see e.g. Lee and Chin (2004)), we can consider directly the computation of this probability, obtaining that

$$\begin{aligned} P(s/s) &= P[s \text{ wins a set serving first}] = \\ &= \sum_{i=0}^{23} p(W, i, s) + p(24, 24, s)p(Adv, s) + p(24, 24, 1-s)(1 - p(Adv, 1-s)) \end{aligned}$$

where $p(W, l, s)$ denotes the probability that team s wins the set while team $1-s$ scores exactly l points, while $p(24, 24, s)$ is the probability that team s reaches the score $(24, 24)$ serving next and $p(Adv, s)$ that team s wins the tie break at the end of the set. Recall that the Markov property allow us to consider the contribution of the past rallies independent to the future rallies, known the present score. A simple computation gives first that

$$p(W, 0, s) = p_s^{25}$$

since the team s has to win all the played rallies. The situation becomes slightly more complicated once the loosing team scores points itself. In this case to evaluate the value of $p(W, l, s)$ we have to take into account all the possible breaks (changes in the serving team) that happened during the set and their relative position in the set. This computation leads to this formula

$$p(W, l, s) = \sum_{k=1}^l A(k, 25, l) p_s^{25-k} p_{1-s}^{l-k} (1 - p_s)^k (1 - p_{1-s})^k$$

for $l \geq 1$, where for positive integers k, m, l , with $k \leq l$,

$$A(k, m, l) = C((k, l-k))C((k+1, m-k)),$$

and $C(n, k)$ denotes the number of combinations with repetitions of k objects from a set of cardinality n , which is equal to

$$C((n, k)) = \binom{n+k-1}{k}.$$

Note that the term $A(k, m, l)$ counts all the possible sequences of consecutive points won by the serving team, between two breaks. If the set arrives to the score $(24, 24)$ we have to consider the probability of winning the final tie break. By the Markov property, we can first compute the probability to reach the score $(24, 24, s)$ or $(24, 24, 1 - s)$ and multiply it by the probability,

respectively, that team s wins the tie break serving first, that team $1 - s$ loose the tie break serving first. Proceeding as before, we obtain that

$$p(24,24,s) = \sum_{k=1}^{24} A(k,24,24)p_s^{24-k} p_{1-s}^{24-k} (1-p_s)^k (1-p_{1-s})^k;$$

$$p(24,24,1-s) = \sum_{k=1}^{23} B(k+1,25,24)p_s^{24-k} p_{1-s}^{23-k} (1-p_s)^{k+1} (1-p_{1-s})^k,$$

where

$$B(k,m,l) = C((k,m-k))C((k,l-k)).$$

In order to compute the probability $p(\text{Adv},s)$, let us consider the sub Markov chain consisting only of the states $\{(24, 24, 0), (24, 24, 1), (\text{Ad}, 24, 0), (\text{Ad}, 24, 1), (\text{W}, 24, 0), (\text{W}, 24, 1)\}$. The computation of the absorbing probability, starting from $(24, 24, s)$ of the state $(\text{W}, 24, s)$, gives

$$p(\text{Adv},s) = \frac{p_s^2}{p_s^2 + p_{1-s}^2 + p_s p_{1-s} - p_s^2 p_{1-s} - p_s p_{1-s}^2}.$$

In Table 1. we present the values of $P(0|0)$ for different couples of probabilities p_0 and p_1

Table 1: Rally point scoring system. Probability of winning a set by team 0 when it serves first.

p_0	p_1						
	0.1	0.2	0.3	0.4	0.5	0.6	0.7
0.1	0.42650	0.10414	0.01844	0.00233	0.00020	0.00001	0.00000
0.2	0.83438	0.45763	0.17102	0.04462	0.00789	0.00086	0.00005
0.3	0.96807	0.78324	0.47516	0.20804	0.06349	0.01258	0.00140
0.4	0.99581	0.94163	0.76498	0.48834	0.22934	0.07361	0.01424
0.5	0.99965	0.98970	0.92862	0.76175	0.50000	0.24006	0.07476
0.6	0.99998	0.99891	0.98637	0.92639	0.76890	0.51172	0.24125
0.7	1.00000	0.99994	0.99860	0.98685	0.93345	0.78633	0.52511

2.2 Side-out scoring system

The side-out scoring system can be modeled in the Markov chain framework as follow: let us define the set S of the states as

$$S := \{(i, j, s) : i \in \{0, 1, \dots, 14, \text{Ad}, \text{W}\}, j \in \{0, 1, \dots, 14\}, s \in \{0, 1\}\}$$

where the first number represents the score of the serving team, the states Ad and W in the first position stand for Advantage and Winning of the serving team, and similarly for the numbers in second position relative to the returning team, while the last number indicates which team serves next. As before, p_s denotes the probability that the team s wins a rally when it is serving. In this scoring system, we have to compute also the probability p_{ps} , that denotes the probability that team s starts serving and scores a point. This can be easily preformed by defining a four states Markov chain, with state space $\{\text{A0}, \text{A1}, \text{W0}, \text{W1}\}$, where A0 , respectively A1 , stands for team

0, resp. 1, serves, while W_0 , resp. W_1 , stands for team 0, resp. team 1, marks the point, and transition probability matrix

$$\begin{bmatrix} 0 & 1-p_0 & p_0 & 0 \\ 1-p_1 & 0 & 0 & p_1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

It is clear that the probability pp_s is equal to the absorbing probability of state W_s starting from A_s , which is equal to

$$pp_s = \frac{p_s}{p_s + p_{1-s} - p_s p_{1-s}}. \quad (1)$$

Remark 1.

It is worth noting that if the probabilities $p_0 = p_1 = 1/2$, in the side-out scoring system it is not anymore true that the probability of scoring a point is independent from the event of who is serving first. Indeed, from (1) we get that in this case the above probability is equal to $pp_0 = pp_1 = 2/3$. It is easy to prove that in general $pp_s \geq p_s$ and that $pp_0 = 1/2$ if

$$p_0 = \frac{p_1}{1 + p_1}.$$

Proceeding as before, it is easy to see that if the first serving team is s , then the transition probabilities are defined as follows: when $0 \leq i, j \leq 13$, then

$$(i, j, s) \rightarrow (i+1, j, s) \text{ with probability } pp_s$$

$$(i, j, s) \rightarrow (j+1, i, 1 - s) \text{ with probability } 1 - pp_s$$

while

$$(14, 13, s) \rightarrow (W, 13, s) \text{ with probability } pp_s$$

$$(14, 13, s) \rightarrow (14, 14, 1 - s) \text{ with probability } 1 - pp_s$$

$$(13, 14, s) \rightarrow (W, 13, 1 - s) \text{ with probability } 1 - pp_s$$

$$(13, 14, s) \rightarrow (14, 14, s) \text{ with probability } pp_s$$

$$(14, 14, s) \rightarrow (Ad, 14, s) \text{ with probability } pp_s$$

$$(14, 14, s) \rightarrow (Ad, 14, 1 - s) \text{ with probability } 1 - pp_s$$

$$(Ad, 14, s) \rightarrow (W, 14, s) \text{ with probability } pp_s$$

$$(Ad, 14, s) \rightarrow (14, 14, 1 - s) \text{ with probability } 1 - pp_s$$

In order to compute the probability of winning a set of team s , when it starts serving, we have to compute the probabilities that the Markov chain starting from the state $(0, 0, s)$ reaches one of the states $(W, 0, s), (W, 1, s), \dots, (W, 13, s), (W, 14, s)$. As in the previous subsection the computation leads to the following compact formula:

$$P[s \text{ wins a set serving first}] = \\ = \sum_{l=0}^{13} p(W, l, s) + p(14, 14, s) pp(Adv, s) + p(14, 14, 1-s)(1 - pp(Adv, 1-s))$$

where

$$p(W, 0, s) = pp_s^{15};$$

$$p(W, l, s) = \sum_{k=1}^l A(k, 15, l) pp_s^{15-k} pp_{l-s}^{l-k} (1 - pp_s)^k (1 - pp_{l-s})^k$$

for $l \geq 1$.

In order to compute the remaining terms, we get

$$p(14, 14, s) = \sum_{k=1}^{14} A(k, 14, 14) pp_s^{14-k} pp_{l-s}^{14-k} (1 - pp_s)^k (1 - pp_{l-s})^k;$$

$$p(14, 14, 1-s) = \sum_{k=1}^{14} B(k, 15, 14) pp_s^{15-k} pp_{l-s}^{14-k} (1 - pp_s)^k (1 - pp_{l-s})^{k-1}.$$

To conclude, in analogy to the previous case, we get

$$p(Adv, s) = \frac{pp_s^2}{pp_s^2 + pp_{l-s}^2 + pp_s pp_{l-s} - pp_s^2 pp_{l-s} - pp_s pp_{l-s}^2}.$$

In Table 2. we present the values of the probability of winning the set for different couples of probabilities p_0 and p_1 . It is now interesting to compare the winning probabilities in the two scoring systems for the same parameters p_0 and p_1 . This allows us to say if and in which case the change of the scoring system has been favorable to the stronger or the weaker team, clearly under our simplified theoretical model. Comparison of Table 1. and 2. shows that the introduction of the rally point system increased the difficulty of winning a set for the first serving team, for every choice of probabilities such that $p_0 \geq p_1$. On the other hand, if $p_1 > p_0$ and the difference $p_1 - p_0$ is substantial, then team 0 (that serves first in the set) has more chances to win the set. Hence, the change in the scoring system facilitated the weaker teams and introduced a source of randomness in the outcomes of the sets (and therefore of the matches). Clearly, in the interpretation of these probabilities, the assumption on the invariance of the probabilities p_0 and p_1 during the match is fundamental. Top teams are closer to fulfill it, while lower category teams are more subject to show dependence in the outcomes of a sequence of points.

Table 2: Side-out scoring system. Probability of winning a set by team 0 when it serves first.

p ₀	p ₁						
	0.1	0.2	0.3	0.4	0.5	0.6	0.7
0.1	0.50394	0.02017	0.00056	0.00002	0.00000	0.00000	0.00000
0.2	0.98125	0.50837	0.10241	0.01321	0.00139	0.00013	0.00001
0.3	0.99951	0.90690	0.51344	0.17214	0.03974	0.00693	0.00092
0.4	0.99999	0.98890	0.84788	0.51938	0.21975	0.06659	0.01459
0.5	1.00000	0.99894	0.96793	0.81260	0.52658	0.25127	0.08614
0.6	1.00000	0.99991	0.99501	0.94912	0.79556	0.53574	0.27064
0.7	1.00000	0.99999	0.99943	0.99038	0.93915	0.79399	0.54832

3. Winning probabilities: Match

Let us now compute the winning probabilities in the two scoring systems. In both the present- and former scoring systems, who first serves in the first set, then serves first in the third set, while the other team starts serving in the second and in the (possible) fourth set. If the teams play the deciding fifth set, a toss is carried out to determine who starts serving. The fifth, deciding set, in the rally point scoring system as in the side-out scoring system, corresponds to a rally point set ending with 15 points, whose winning probability can be derived with the same computation carried out in the previous section.

By the Markovian assumption, we get that the probability to win a match is equal to the product of the probabilities for the two teams to win the single sets. Let us denote by

$$p(W,0) = P[0 \text{ wins a set serving first}] \quad p(W,1) = P[1 \text{ wins a set serving first}],$$

while

$$p(T,0) = P[0 \text{ wins the deciding set serving first}] \quad p(T,1) = P[1 \text{ wins the deciding set serving first}]$$

Since a toss is carried out to determine who first serves the deciding set, the probability that team 0 wins this set will be equal to

$$p_T = 0.5p(T,0) + 0.5(1 - p(T,1)).$$

A simple computation gives in the rally point scoring system:

$$P[0 \text{ wins}(3,0)] = p(W,0)^2(1 - p(W,1))$$

$$P[0 \text{ wins}(3,1)] = 2(1 - p(W,0))p(W,0)(1 - p(W,1))^2 + p(W,0)^2p(W,1)(1 - p(W,1))$$

$$\begin{aligned}
& P[0\text{wins}(3,2)] = \\
& = [p(W,0)^2 p(W,1)^2 + (1-p(W,0))^2 (1-p(W,1))^2 + 4p(W,0)p(W,1)(1-p(W,0))(1-p(W,1))]p_T.
\end{aligned}$$

Therefore, the probability that team 0 wins a match when starts serving in the first set is equal to

$$P[0\text{wins}(3,0)] + P[0\text{wins}(3,1)] + P[0\text{wins}(3,2)].$$

4. Expected duration of a set

In this section we shall consider the expected duration of a set, measured in number of rallies. We shall assume again that the probabilities to win a rally could be different for the two teams, but constant along the set and independent of the previous rallies. Moreover, we shall assume that team 0 starts serving. From the Markov chain theory, it is possible to solve this problem since this is equivalent to determine the expected number of steps that the chain takes to arrive for the first time to a given state or subset of states. The problem is that this solution is finite, and therefore useful, just when the arriving subset of states includes all the closed classes of the Markov chain. In the present case, it is possible to determine the expected number of rallies needed to finish a given set, but not the expected number of rallies needed to play a set won by team 0, for example. This problem can be overcome in the rally point scoring system, but not in the side-out scoring system.

4.1 Rally point scoring system

Let us start by considering the rally point scoring system. In this case the computation is simple, since a point is scored at the end of each rally. Therefore, if the probability that team 0 or team 1 win a set with a final score $(25, l)$, $l \in \{0, \dots, 23\}$, we get that the contribution of this outcome to the expected duration of the set is equal to

$$(25+l) \times (p(W,l,0) + p(W,l,1)).$$

Slightly more complicated is the case when the score reaches $(24, 24)$. In this case we have to compute the expected number of rallies that one team needs to end the set, conditional to the fact that we arrive to the tiebreak after exactly 48 rallies. This can be easily computed thanks to the Markov chain theory if we define a suitable sub Markov chain. As before, we have to consider separately the cases that we arrive to the score $(24, 24, 0)$ or $(24, 24, 1)$, since the expected length of the tiebreak is generally different.

In order to compute the expected duration of a tiebreak, let us consider a Markov chain defined on the state space $S := \{(24, 24, 0), (24, 24, 1), (25, 24, 0), (25, 24, 1), (26, 24, 0), (26, 24, 1)\}$ with transition matrix

$$P = \begin{bmatrix} 0 & 0 & p_0 & 1-p_0 & 0 & 0 \\ 0 & 0 & 1-p_1 & p_1 & 0 & 0 \\ 0 & 1-p_0 & 0 & 0 & p_0 & 0 \\ 1-p_1 & 0 & 0 & 0 & 0 & p_1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Denoting by $E=\{(26, 24, 0), (26, 24, 1)\}$ the set of the absorbing states, an easy computation (see e.g. Norris (1998)) allows us to obtain the mean absorbing times to the set E starting from the states $(24, 24, 0), (24, 24, 1), (25, 24, 0), (25, 24, 1)$ as the (minimal) nonnegative solution k of the linear system

$$k_i = \sum_{j=1}^4 P_{i,j} k_j , \quad \text{for } i=1,\dots,4 \tag{2}$$

$$k_5 = k_6 = 0$$

where we renames the states, in the same order as before, as $\{1, 2, 3, 4, 5, 6\}$. Solving this system, we obtain that the mean duration of the tiebreak starting by $(24, 24, 0)$ is equal to k_1 , where

$$k_1 = \frac{2(p_0 + p_1 - p_0 p_1) + 2p_0(1-p_0)}{(p_0 + p_1 - p_0 p_1)^2 - p_0 p_1(1-p_0)(1-p_1)}.$$

while the mean duration of the tie break starting by $(24, 24, 1)$ is equal to k_2 , where

$$k_2 = \frac{2 + p_1(1-p_1) \times k_1}{p_0 + p_1 - p_0 p_1}.$$

Therefore, conditioning on the fact that the set reaches the $(24, 24, 0)$ or $(24, 24, 1)$ scores, respectively, the expected duration of such a set is equal to

$$k_{TB} = p(24,24,0) \times (48 + k_1) + p(24,24,1) \times (48 + k_2).$$

Collecting all these terms, we obtain that the expected duration of a set under the rally point scoring system is equal to

$$\sum_{l=0}^{23} (25+l)(p(W,l,0) + p(W,l,1)) + p(24,24,0) \times (48 + k_1) + p(24,24,1) \times (48 + k_2).$$

In Table 3. we present the values of the expected duration of the set for different couples of probabilities p_0 and p_1 . We also present, parenthetical, the estimated standard deviation of the duration of the set, obtained simulating 1,000,000 of sets for each pair (p_0, p_1) .

Table 3: Rally point scoring system. Expected duration of a set (and standard deviation, estimated by 1,000,000 replicates of played sets).

p ₀	p ₁						
	0.1	0.2	0.3	0.4	0.5	0.6	0.7
0.1	54.05265 (14.91289)	47.80129 (7.18016)	44.01562 (3.93678)	41.13975 (3.14317)	38.44668 (3.01603)	35.77786 (2.90388)	33.11061 (2.68593)
0.2	48.87572 (7.89125)	48.85170 (6.74883)	46.35838 (5.02883)	43.30936 (4.12290)	40.26262 (3.74588)	37.25137 (3.51311)	34.24906 (3.23776)
0.3	44.78652 (4.32246)	46.83601 (5.15053)	47.02174 (4.70905)	45.26924 (4.41072)	42.40624 (4.36950)	39.11653 (4.23204)	35.71018 (3.90302)
0.4	41.71137 (3.25019)	43.79625 (4.18764)	45.52038 (4.37762)	45.80878 (4.16734)	44.26045 (4.39654)	41.29771 (4.74957)	37.60288 (4.67248)
0.5	38.89249 (3.06057)	40.63788 (3.77841)	42.65790 (4.36223)	44.34614 (4.37096)	44.72994 (4.23089)	43.18093 (4.68663)	39.87165 (5.22015)
0.6	36.11127 (2.94329)	37.50153 (3.54051)	39.25614 (4.24623)	41.29776 (4.74215)	43.08409 (4.72003)	43.58975 (4.59817)	41.88578 (5.19371)
0.7	33.33334 (2.72073)	34.37508 (3.24971)	35.71170 (3.90737)	37.44397 (4.65775)	39.54475 (5.25218)	41.52885 (5.31985)	42.21003 (5.21359)

4.2 Side-out scoring system

This case is more complicated, since the side-out scoring system needs a “small tie break” to decide if a team scores a single point. Thanks to the Markov chain theory, described above, we are able to compute the expected duration of any such “small tie break”. However, this duration depends on who is serving first and so it will not be sufficient to know the expected duration of the “small tie break”, but we should know the duration of a “small tie break” won by team s . This is not possible to compute using the classical Markov chain approach and so we have two possibilities. On one side we can consider the whole set as a Markov chain and evaluate directly the expected duration solving the linear system in (2). Even if this is theoretically feasible, in practice, for the rally point scoring system this is equivalent to solve a linear system with 1254 equations or, which is equivalent, define and invert a 1254 square matrix, while for the side-out scoring system these numbers fall to 510. On the other side, it is possible, and nowadays fast, to simulate a large number of sets and estimate the expected duration of the set along with its standard deviation in a very simple way. In this subsection we will follow this approach and the results are summarized in the following Table 4., where the simulated durations of 1,000,000 sets have been obtained for each pair of parameters (p_0, p_1) .

The mean durations are lower in the rally point system as long as $p_0 \leq 0.5$ and $p_1 \leq 0.5$. Outside this range the mean durations are, generally higher in the rally point system (except for $p_1 = 0.6$ and $p_0 \leq 0.4$). Probably, this is due to the fact that, as outlined at the end of Section 2., in the rally point system it is more probable for the weaker team to reach higher scores (and possibly win the set).

Table 4: Side-out scoring system. Expected duration of a set (and standard deviation), estimated by 1,000,000 replicates of played sets.

p ₀	p ₁						
	0.1	0.2	0.3	0.4	0.5	0.6	0.7
0.1	258.89002 (59.80855)	142.48141 (36.34303)	90.00662 (22.90847)	63.61648 (15.94363)	47.78484 (11.64220)	37.20972 (8.69363)	29.67830 (6.46740)
0.2	141.33413 (36.27911)	128.11294 (28.99328)	93.00187 (22.69687)	66.74642 (16.88541)	50.00907 (12.52573)	38.74286 (9.37190)	30.71503 (6.99569)

0.3	88.87538 (22.90984)	91.99717 (22.77603)	84.42871 (18.73650)	67.93844 (15.90992)	52.37300 (13.03704)	40.64645 (10.14068)	32.03526 (7.64992)
0.4	62.53499 (15.92535)	65.47384 (16.93181)	67.00376 (16.11341)	62.46464 (13.65832)	52.90798 (12.06242)	42.52611 (10.40962)	33.67196 (8.32781)
0.5	46.68319 (11.64442)	48.76056 (12.49328)	51.06549 (13.13281)	51.95234 (12.30704)	49.13692 (10.68072)	42.86594 (9.62438)	35.23850 (8.52962)
0.6	36.11113 (8.69699)	37.49909 (9.35017)	39.23885 (10.14523)	41.08895 (10.55737)	41.84133 (9.92346)	40.05917 (8.76973)	35.53665 (7.98580)
0.7	28.56746 (6.45749)	29.46810 (6.97537)	30.60417 (7.61331)	32.05322 (8.35032)	33.59113 (8.73049)	34.41184 (8.34485)	33.33850 (7.50407)

5. The case of the Italian Volleyball League

In this section we use data from the Italian Volley League in order to check how close (or far) is our approach to the real world. In the sequel we consider the matches of the Italian Volley League in the period 2001-2012 (data available at www.legavolley.it). Our aim is to compare the probabilities of winning a set calculated via the Markov chain model presented in section 2. with the relative frequencies of the played sets.

First, we estimate the probability of winning a point after serving. We should estimate p_0 and p_1 using records of just matches between team 0 and 1, but such detail of data is not available. Nonetheless, aggregated data about points obtained by serving teams are usable. Therefore we consider the top 4 teams for scored points in the period. For such teams we assume that the estimated probabilities should not depend (too much) on the opponent. The estimated probabilities are then used to calculate the probabilities of winning a set and these are then compared with the results on the sets of the Italian Volley League. Since data about the first serving team in a set are not collected, we compute the overall probabilities of winning a set, regardless the first serving team. Since the first serving team is chosen at random, by the total probability law, we have

$$P[\text{team } s \text{ wins a set}] = 0.5 + \frac{P(s/s) - P(1-s/1-s)}{2}.$$

In Table 5. we calculate the estimated probabilities \hat{p} to score a point by the serving team, while, in Table 6., we have the estimated probabilities from data of winning a set and,

parenthetical, the exact probabilities calculated using the estimated \hat{p} . All the discrepancies between these two quantities are lower, in absolute value, than 0.05, showing that, at least for the top teams, the theoretical assumptions are reasonable and the proposed model well describes the data.

Table 5: Top 4 teams of the Italian Volley League for scored points in the period 2001/2002-2011/2012

Team	Total points	Break-points	\hat{p}
Trento	18465	6676	0.361548876
Macerata	18585	6594	0.35480226
Cuneo	18634	6568	0.352473972
Modena	18623	6322	0.339472695

Table 6: Relative frequencies (and calculated probabilities) of winning a set by team A vs. team B in the period 2001/2002 - 2011/2012.

Team A	Team B			
	Trento	Macerata	Cuneo	Modena
Trento		0.5393258 (0.5204434)	0.5454545 (0.5275133)	0.6022727 (0.5670153)
Macerata	0.4606742 (0.4795566)		0.5161290 (0.5070824)	0.5747126 (0.5467732)
Cuneo	0.4545455 (0.4724867)	0.4838710 (0.4929176)		0.5853659 (0.5397321)
Modena	0.3977273 (0.4329847)	0.4252874 (0.4532268)	0.4146341 (0.4602679)	

References

- Lee K.T., Chin S.T. (2004) Strategies to serve or receive in volleyball. *Math. Meth. Oper. Res.* **59**, 53-67.
 Newton, P.K. and Keller, J.B. (2005) Probability of winning at tennis. I. Theory and data.
Stud.Appl.Math., **114(3)**, 241-269.
 Norris J.R. (1998) *Markov chains*. Cambridge University Press.
 Simmons J. (1989) A probabilistic model of squash: strategies and applications. *Applied Statistics* **38**, 95-110.

An Olympic ranking based on sequential use of ordinal multicriteria methods

Silvio Figueiredo Gomes Júnior*, João Carlos Correia Baptista Soares de Mello** and Lidia Angulo Meza***

*Centro Universitário Estadual da Zona Oeste. Avenida Manuel Caldeira de Alvarenga, 1203, Campo Grande, 23070-200, Rio de Janeiro, RJ. silviogomes@uezo.rj.gov.br

**Departamento de Engenharia de Produção – Universidade Federal Fluminense. Rua Passo da Pátria 156, São Domingos, 24210-240, Niterói, RJ. jcsmello@pesquisador.cnpq.br

***Dep. de Engenharia de Produção – Universidade Federal Fluminense. Av. dos Trabalhadores, 420, 27255-125, Volta Redonda, RJ. lidia@metal.eeimvr.uff.br

Abstract. There is no official method to establish a final ranking for the Olympic Games. It is usual to rank the participant countries in these games in accordance with the number of medals they have won using a lexicographic multicriteria method. Furthermore, it does not take into account that the various sports may be of different importance. This work proposes a ranking model to eliminate those drawbacks. We use firstly a Lexicographic multicriteria method in each sport. After obtaining a rank for each and all sports, we build a general ranking by aggregation all the sports using a Borda multicriteria method. Our model uses the results of the London Olympic Games in 2012.

1. Introduction

The modern Olympic Games, initiated in 1896 by Baron Coubertin, tried to keep the initial spirit of individual competition. That purpose clearly failed. Ever since the very first modern Games, it became usual to play the national anthem of the winner's country (Lins *et al.*, 2003).

Despite their national character, the Olympic Committee has never issued an official ranking to pick an overall Olympic winner country. In the literature, many studies have been carried out to evaluate the results of the Olympic Games. Some of them are based on how these events can bring benefits to the host cities (Glynn, 2008; Cheng, 2009; Xiaoduo and Jianxin, 2008) and others are interested in social studies (Bernstein, 2000; Farrell, 1989; Levine, 1974; Ball, 1972).

Besides these studies, we can find researches in the environmental and health areas (Hadjichristodoulou *et al.*, 2006; Allen *et al.*, 2006; Streets *et al.*, 2007; Weiler *et al.*, 1998) some about tourism industry and others that evaluate mathematics and economic aspects of the Games (Bernard and Busse, 2004; Lins *et al.*, 2003; Li *et al.*, 2008; Heazlewood, 2006).

The mass media, however, issued a ranking. Their ranking has become a quasi-official ranking. It is based on the Lexicographic Multicriteria Method (Barba-Romero and Pomerol, 1997). This ranking orders countries regarding to the total number of gold medals, silver medals and bronze medals won by each country and the gold medal is the most important. Because of it, this ranking does not deal properly with the possible existence of countries that have won a large number of silver and bronze medals but no gold medal (Soares de Mello *et al.*, 2008), as this method over-evaluates the gold medal.

The Lexicographic Method is not the sole method used by media to rank countries in the Olympic Games. Some newspapers produce a ranking determining the total number of medals earned by each country. They simply add up bronze, silver and gold medals. This method has been widely used by American newspapers during the 2008 Beijing Olympic Games. This was done because using the Lexicographic Method China was in the first position and using the total number of medal USA was in the first place. The obvious disadvantage of this method is to under-evaluate gold medals.

An alternative approach is to make an arbitrary evaluation of each medal, for instance, 1 point for bronze, 2 for silver and 3 for gold. This is a much unsophisticated approach, as it assumes all medals to be equally desired, albeit in proportion to their value. Sitarz (2013) used the incenter of a convex cone to obtain a system of points for medals in Olympic ranking and to obtain an alternative ranking for Formula 1 motor race.

Many alternative Olympic Rankings have been proposed taking into account only medals. For instance, Soares de Mello et al (2004), Hai (2007) and Soares de Mello et al (2008) using a Data Envelopment

Analysis (DEA) approach. All those works considered that a gold medal is more important than the silver medal, the silver medal is more important than the bronze medal, and the difference between the gold and silver medal is greater than the difference between the silver and bronze medals. However, these studies are rather complex to be understood by the general public.

The aforementioned papers aim to override the distortions caused by the use of the Lexicographic Method. Another distortion pointed by Soares de Mello et al (2007) is that medals won in different competitions do not have the same value. As a matter of fact, the existing rankings do not take into account that in some sports there are more events than in others, and so there are more possibilities of winning a medal. For instance, in gymnastics there are a lot of gold medals to be earned and in football there are only two possibilities for a country to win a gold medal (one for men, the other for women). Soares de Mello et al (2009) proposed a methodology to deal with this problem. To take into account the difference in winning values for different sports, they aggregated competitions into clusters. The first clustering was directly obtained from the IOC (www.olympic.org), where each sport is a cluster. They also tried to use a second clustering based on the aggregation of the Olympic Sports done by Wallechinsky (2004), however with disappointing results.

In this work, we propose a different way to rank the countries taking into account the number of medals available in each sport. The methodology consists in a sequential use of ordinal multicriteria methods in two steps. The first one is to rank each sport independently using a Lexicographic Method. In this method, we consider that all countries participated in all sports. Thus we eliminate distortions generated among the countries participating in sports with lots of possibilities of winning medals and those who participate in sports with only two possibilities of winning a medal. In the second step, we aggregated the different rankings obtained in step one. This is achieved using a sum of the partial performances, using the Borda Method.

2. Lexicographic and Borda Methods

According to Arrow (1951), cited by Barba-Romero and Pomerol (1997), just choices do not exist, in other words, there is no "perfect" ordinal multi-criteria or multi-decision maker method. A multiple decision-maker selection method could only be considered just if it obeyed the axioms of universality, unanimity, and independence in relation to irrelevant alternatives, transitivity and totality. The Arrow theorem states that, with the exception of dictatorial methods, no choice or decision-aiding method serves all of these axioms simultaneously.

The axioms of independence in relation to irrelevant alternatives of transitivity and of universality are of special interest to this study. The first states that the order of preference between two alternatives must not depend on their preferences in relation to a third alternative. The transitivity axiom states that if one alternative is preferable to a second, and this one to a third, then the first must be preferable to the third (the fact that in the results of football matches this property is not confirmed is the reason for the popular saying that "football has no logic"). The universality axiom, meanwhile, requires the method to function, respecting all the other axioms, for any group of preferences of the decision-makers. Thus, a method that respects the axioms in some particular cases, does not respect universality.

In the Lexicographic Method criteria are ranked in the order of their importance. The alternative with the best performance score on the most important criterion is chosen. If there are ties with respect to this criterion, the performance of the tied alternatives on the next most important criterion will be compared, and so on, till a unique alternative is found (Linkov *et al.*, 2004).

The Lexicographic Methods are among those ordinal methods having specific properties. They satisfy the five axioms in Arrow's theorem, so they are dictatorial: a criterion acts as a dictator (Barba-Romero and Pomerol, 1997).

For the use of the Borda Method each decision-maker must order the alternatives according to his/her preferences (Soares de Mello *et al.*, 2005). The alternative of highest preference is assigned one point, the second two points and so on successively. In the end, the points assigned by the decision-makers to each alternative are added up and the alternative that has obtained the lowest score will be chosen (Dias *et al.*, 1996). All of the alternatives are ordered in a decreasing order of points, (which guarantees adherence to the

totality axiom). The draws are treated in the standard way as explained by Barba-Romero and Pomerol (1997). In sports, variations of the Borda method are widely used, with each competition considered a decision-maker, indicating its preferences in the final classification of the competition. It is common to make an inversion of the method, assigning a greater number of points to the preferred alternative (the victor in the competition). It is interesting to note that one of the few examples of the use of the original Borda method is in the Olympic Games yachting competitions (Soares de Mello *et al.*, 2005).

In spite of its simplicity and the widespread use of its variations, the Borda Method does not respect one of Arrow's axioms, namely that the final classification of two alternatives is not independent in relation to irrelevant alternatives. This fact can create undesirable situations, such as a vote where the last voter knows the preferences of the previous voters and alters his/her preferences so as to give greater chances to his/her preferred alternative. It can also, which is of specific interest in this case, encourage the unsporting inversion of positions in a competition to benefit a given competitor, as often occurs in Formula 1 (Soares de Mello *et al.*, 2005).

3. Modeling and Results

As we want to propose a new ranking to the 2012 Summer Olympic Games, in this section we are going to fix up the elements of the problem. Some authors as Gomes et al (2009) argue hat to organize a multicriteria problem, we should define the alternatives, the criteria and choose an appropriate method to solve the problem.

Based on the explanation in the second section, we are going to use the Lexicographic and Borda Methods, applied in two phases.

In both methods, the alternatives are all the 85 countries that won at least one medal in the 2012 London Olympic Games. For the first step, the Lexicographic Method the decision criteria are the numbers of gold, silver and bronze medals for each sport. In the second step, the Borda Method, the decision criteria are the sports of the Summer Olympic Games. As the Borda Method needs an ordinal scale we use for that the rankings obtain in each with the Lexicographic Method.

Tables 1 and 2 present the rank for Archery and Athletics using the first step, the Lexicographic Method. Also, they present the Borda points that will be used in the second step.

Table 1 - Rank for Archery using Lexicographic Method

Archery						
Rank	Borda points	Country	Gold	Silver	Bronze	Total
1	1	South Korea	3	0	1	4
2	2	Italy	1	0	0	1
3	4	Mexico	0	1	1	2
3	4	China	0	1	1	2
3	4	Japan	0	1	1	2
6	6	United States	0	1	0	1
	46	All other countries	0	0	0	0

Table 2 - Rank for Athletics using Lexicographic Method

Athletics						
Rank	Borda points	Country	Gold	Silver	Bronze	Total
1	1	United States	9	13	7	29
2	2	Russian Federation	8	5	5	18
3	3	Jamaica	4	4	4	12
4	4	Great Britain & N. Ireland	4	1	1	6
5	5	Ethiopia	3	1	3	7
6	6	Kenya	2	4	5	11
7	7	Germany	1	4	3	8

8	8	Australia	1	2	0	3
9	10,5	Dominican Republic	1	1	0	2
9	10,5	France	1	1	0	2
9	10,5	Poland	1	1	0	2
9	10,5	Turkey	1	1	0	2
13	13	China	1	0	5	6
14	14	Trinidad and Tobago	1	0	3	4
15	15	Czech Republic	1	0	1	2
16	19,5	Grenada	1	0	0	1
16	19,5	Croatia	1	0	0	1
16	19,5	Bahamas	1	0	0	1
16	19,5	Algeria	1	0	0	1
16	19,5	New Zealand	1	0	0	1
16	19,5	Kazakhstan	1	0	0	1
16	19,5	Hungary	1	0	0	1
16	19,5	Uganda	1	0	0	1
24	24	Ukraine	0	1	2	3
25	25	Cuba	0	1	1	2
26	29	South Africa	0	1	0	1
26	29	Iran	0	1	0	1
26	29	Tunisia	0	1	0	1
26	29	Slovenia	0	1	0	1
26	29	Botswana	0	1	0	1
26	29	Guatemala	0	1	0	1
26	29	Colombia	0	1	0	1
33	37	Bahrain	0	0	1	1
33	37	Canada	0	0	1	1
33	37	Puerto Rico	0	0	1	1
33	37	Qatar	0	0	1	1
33	37	Italy	0	0	1	1
33	37	Estonia	0	0	1	1
33	37	Finland	0	0	1	1
33	37	Japan	0	0	1	1
33	37	Morocco	0	0	1	1
	64	All other countries	0	0	0	0

As mentioned before, in the second step, we use the Borda Method to aggregate all the ranks obtained. The final rank according to our proposed methodology is presented in Table 3, as well as the rank obtained using only the Lexicographic Method and the discrepancy between the positions in the two ranks.

Table 3 - Rank of Borda and Lexicographic Methods and the discrepancy

Country	Borda points	Borda rank	Lexicographic rank	Discrepancy
United States	793	1	1	0
China	844	2	2	0
Russian Federation	854	3	4	1
Great Britain & N. Ireland	888	4	3	-1
Germany	1057	5	6	1
France	1100	6	7	1
Italy	1137	7	8	1
Australia	1140	8	10	2
Japan	1195	9	11	2
South Korea	1255	10	5	-5
Spain	1290	11	21	10
Canada	1328	12	36	24
Netherlands	1331	13	13	0
Ukraine	1362	14	14	0
Brazil	1377	15	22	7

Hungary	1399	16	9	-7
Cuba	1445	17	16	-1
New Zealand	1446	18	15	-3
Czech Republic	1458	19	19	0
Belarus	1459	20	26	6
Colombia	1489	21	38	17
Croatia	1500	22	25	3
Poland	1506	23	30	7
Kazakhstan	1529	24	12	-12
Denmark	1534	25	29	4
Romania	1534	25	27	2
Sweden	1542	27	37	10
Argentina	1548	28	42	14
Lithuania	1559	29	34	5
Iran	1572	30	17	-13
Switzerland	1582	31	33	2
Norway	1586	32	35	3
Azerbaijan	1586	32	30	-2
Mexico	1587	34	39	5
India	1588	35	55	20
South Africa	1590	36	23	-13
Belgium	1603	37	60	23
Slovenia	1603	37	42	5
Turkey	1617	39	32	-7
Serbia	1623	40	42	2
North Korea	1625	41	20	-21
Armenia	1632	42	60	18
Uzbekistan	1632	42	47	5
Thailand	1636	44	57	13
Mongolia	1639	45	56	11
Georgia	1658	46	39	-7
Ireland	1665	47	41	-6
Latvia	1668	48	49	1
Algeria	1669	49	50	1
Bahamas	1669	49	50	1
Slovakia	1669	49	59	10
Malaysia	1670	52	63	11
Egypt	1672	53	58	5
Tunisia	1673	54	45	-9
Cyprus	1674	55	69	14
Chinese Taipei	1675	56	63	7
Bulgaria	1676	57	63	6
Afghanistan	1677	58	79	21
Greece	1683	59	75	16
Finland	1684	60	60	0
Estonia	1685	61	63	2
Puerto Rico	1685	61	63	2
Bahrain	1686	63	79	16
Jamaica	1692	64	18	-46
Qatar	1692	64	75	11
Ethiopia	1694	66	24	-42
Kenya	1695	67	28	-39
Dominican Republic	1699	68	46	-22
Trinidad and Tobago	1703	69	47	-22
Grenada	1708	70	50	-20
Indonesia	1708	70	63	-7
Uganda	1708	70	50	-20
Venezuela	1708	70	50	-20
Gabon	1709	74	69	-5
Montenegro	1710	75	69	-6

Singapore	1711	76	75	-1
Saudi Arabia	1712	77	79	2
Hong Kong	1714	78	79	1
Portugal	1714	78	69	-9
Moldova	1714	78	75	-3
Tajikistan	1717	81	79	-2
Botswana	1718	82	69	-13
Guatemala	1718	82	69	-13
Kuwait	1719	84	79	-5
Morocco	1726	85	79	-6

Countries that in our methodology are better ranked than they are in the standard Lexicographic Method have more medals won in collective sports. Among them we can cite Brazil, this result is in line with the conclusions obtained in Soares de Mello et al (2012) when a different methodology (Data Envelopment Analysis) was used, and the data were from Beijing Olympic Games. This vindicates a tendency of Brazil to invest mainly in collective sports.

Also, among the countries that are worse ranked we can cite Jamaica. This country has a concentration of medals in individual sports mainly athletics. We shall remember that athletics is one of the sports with a larger number of distributed medals.

We may point out that in the two first positions there are no differences between the two rankings.

4. Conclusions and Future Works

The use of Lexicographic Methods for each sport altogether with the Borda Method to aggregate the various sport ranks has an important consequence: the gold medal is not so much overvalued as it is in the pure Lexicographic Method. This is an advantage of our methodology. Another advantage is that it is possible to take into account the difference among medal win in different sports in a much easier way than the method used by Soares de Mello et al (2009). The main disadvantage of our methodology is the Borda Method. As explained before this method is highly dependent of the irrelevant alternatives. For future works we may propose the use of Copeland Method or some improvement of the Borda Method.

Acknowledgements

To FAPERJ and CNPq for the financial support.

References

- Allen, T. L., Jolley, S. J., Cooley, V. J., Winn, R. T., Harrison, J. D., Price, R. R. and Rich, J. C. (2006). The epidemiology of illness and injury at the alpine venues during the salt lake city 2002 winter olympic games. *Journal of Emergency Medicine*, 30, 197-202.
- Arrow, K. J. (1951). Social choice and individual values. New York: Wiley.
- Ball, D. W. (1972). Olympic games competition: Structural correlates of national success. *International Journal of Comparative Sociology*, 12, 186-200.
- Barba-Romero, S. and Pomerol, J. C. (1997). Decisiones multicriterio: Fundamentos teóricos e utilización práctica: Universidad de Alcalá.
- Bernard, A. B. and Busse, M. R. (2004). Who wins the olympic games: Economic resources and medal totals. *Review of Economics and Statistics*, 86 (1), 413-417.
- Bernstein, E. (2000). Things you can see from there you can't see from here: Globalization, media, and the olympics. *Journal of Sport and Social*, 24, 351-369.
- Cheng, X. (2009). The urban system impact on post-games development of the olympics' venues in china. Paper read at 2009 International Association of Computer Science and Information Technology - Spring Conference, IACSIT-SC 2009.
- Dias, L. M. C., Almeida, L. M. A. T. and Climaco, J. C. N. (1996). Apoio multicritério à decisão. Coimbra: Universidade de Coimbra.

- Farrell, T. (1989). Media rhetoric as social drama: The winter olympics of 1984. *Critical Studies in Mass Communication*, 6, 158-182.
- Glynn, M. (2008). Configuring the field of play: How hosting the olympic games impacts civic community. *Journal of Management Studies*, 45 (6), 1117-1146.
- Gomes, E. G., Soares De Mello, J. C. C. B., E Souza, G. D. S., Angulo Meza, L. and Mangabeira, J. A. D. C. (2009). Efficiency and sustainability assessment for a group of farmers in the brazilian amazon. *Annals of Operations Research*, 169 (1), 167-181.
- Hadjichristodoulou, C., Mouchtouri, V., V., V., Kapoula, C., Vousourelis, A., Kalivitis, I., Chervoni, J., Papastergiou, P., Vasilogiannakopoulos, A., Daniilidis, V. D. and Kremastinou, J. (2006). Management of environmental health issues for the 2004 athens olympic games: Is enhanced integrated environmental health surveillance needed in every day routine operational. *BMC Public Health*, 6, 306.
- Hai, H. L. (2007). Using vote-ranking and cross-evaluation methods to assess the performance of nations at the olympics *WSEAS Transactions on Systems*, 6 (6), 1196-1205.
- Heazlewood, T. (2006). 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.
- Levine, N. (1974). Why do countries win olympic medals? Some structural correlates of olympic games success: 1972. *Sociology and Social Research*, 58, 353-360.
- Li, Y., Liang, L., Chen, Y. and Morita, H. (2008). Models for measuring and benchmarking olympics achievements. *Omega*, 36 (6), 933-940.
- Linkov, I., Varghese, A., Jamil, S., Seager, T. P., Kiker, G. and Bridges, T. (2004). Multi-criteria decision analysis: A framework for structuring remedial decisions at the contaminated sites. In *Comparative risk assessment and environmental decision making*, edited by Linkov, I. and Ramadan, A. B. New York: Springer, 15-54.
- Lins, M. P. E., Gomes, E. G., Soares de Mello, J. C. C. B. and Soares de Mello, A. J. R. (2003). Olympic ranking based on a zero sum gains dea model. *European Journal of Operational Research*, 148, 312-322.
- Sitarz, S. (2013). The medal points' incenter for rankings in sport. *Applied Mathematics Letters*, 26 (4), 408-412.
- Soares de Mello, J. C. C. B., Angulo-Meza, L. and Branco da Silva, B. P. (2007). A ranking for the olympic games with unitary input dea models. Paper read at First International Conference on Mathematics in Sport (IMA Sport 2007), at Manchester.
- Soares de Mello, J. C. C. B., Angulo-Meza, L. and Branco da Silva, B. P. (2009). A ranking for the olympic games with unitary input dea models. *IMA Journal Management Mathematics*, 20 (2), 201-211.
- Soares de Mello, J. C. C. B., Angulo-Meza, L. and Lacerda, F. G. (2012). A dea model with a non discretionary variable for olympic evaluation. *Pesquisa Operacional*, 32 (1), 21-29.
- Soares de Mello, J. C. C. B., Gomes, E. G., Angulo-Meza, L. and Biondi Neto, L. (2008). Cross evaluation using weight restrictions in unitary input dea models: Theoretical aspects and application to olympic games ranking. *WSEAS Transactions on Systems*, Forthcoming.
- Soares de Mello, J. C. C. B., Gomes, E. G., Angulo-Meza, L., Biondi Neto, L. and Coelho, P. H. G. (2004). A modified dea model for olympic evaluation. In XII Congreso Latino-Iberoamericano de Investigación de Operaciones y Sistemas - CLAIO 2004, at Havana.
- Soares de Mello, J. C. C. B., Gomes, L. F. A. M., Gomes, E. G. and Soares de Mello, M. H. C. (2005). Use of ordinal multi-criteria methods in the analysis of the formula 1 world championship. *Cadernos Ebape.BR*, 3 (2), 1-8.
- Streets, D. G., Fu, J. C., Jang, C. J., Hao, J., He, K., Tang, X., Zhang, Y., Wang, Z., Li, Z., Zhang, Q., Wang, L., Wang, B. and Yu, C. (2007). Air quality during the 2008 beijing olympic games. *Atmospheric Environment*, 41 (3), 480-492.
- Wallechinsky, D. (2004). The complete book of the summer olympics: Aurum Press.
- Weiler, J. M., Layton, T. and Hunt, M. (1998). Asthma in united states olympic athletes who participated in the 1996 summer games. *Journal of Allergy and Clinical Immunology*, 102 (5), 722-772.
- Xiaoduo, C. and Jianxin, Y. (2008). The factors of the urban system influenced post-development of the olympics' venues. Paper read at 2008 International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM 2008.

The effect of fatigue from the previous match in Grand Slam Tennis

D. Goossens*, J. Kempeneers, and F. Spieksma**

*KU Leuven, ORSTAT, Naamsestraat 69, 3000 Leuven, dries.goossens@kuleuven.be

** KU Leuven, ORSTAT, Naamsestraat 69, 3000 Leuven, frits.spieksma@kuleuven.be

Abstract. In this work we study whether fatigue resulting from the previous match affects a player's chances of winning his/her next match in tennis. We collect data from the 4 Grand Slam tournaments (men and women), from 1992 till 2011. For each match, we determine a priori chances of winning for both players, based on the ranking of the players and the tournament. Next, we isolate those matches where one player played more sets than his/her opponent, and verify whether there is a systematic advantage for the least tired player.

1. Introduction

To what extent does the physical effort invested in winning a match in a Grand Slam event have an effect on the probability of winning the next match? Is it indeed beneficial to close out a match in straight sets in order to be "fully fresh" for the next match? Does an extra day of rest before the final make a difference in the chance of winning the final? These questions are subject of debate, and in this contribution we intend to shed some light on this issue.

As an extreme example, recall the famous Wimbledon match between John Isner and Nicolas Mahut in 2010. This match took 11 hours and 5 minutes, spread over 3 days, making it the longest match in tennis history. Eventually, Isner won this epic match with a score of 70-68 in the fifth and final set, for a total of 183 games. In his next match, Isner was defeated by Thiemo de Bakker in 3 short sets, after a mere 74 minutes of play. Although probably no-one doubts that Isner's defeat was due to fatigue resulting from his confrontation with Mahut, we want to find out how frequently and to what extent this effect occurs in Grand Slam tennis.

Let us motivate some of the choices that we make in this empirical study. First, we concentrate on Grand Slam tournaments only. For the male players, in order to win a match in such a tournament, one must win three sets, whereas in most tournaments winning two sets suffices. Thus, in Grand Slam tournaments, matches may take up to five sets allowing for a potentially pronounced effect of fatigue on the next match. Also, Grand Slam tournaments are the most important tennis events in the year, and there is a strong drive to perform well on these events.

Second, a choice needs to be made on how to measure the effort a player invests in a match. Different options exist that capture fatigue; one can think of time spent on court, number of points played, number of games played, and number of sets played. We choose the latter. Apart from the fact that these data are available, it also allows for a straightforward way to categorize the relative effort invested, namely the difference in number of sets played in the previous match, which equals either 0, 1, or 2 in the case of male players, and equals either 0 or 1 in case of female players. Alternatively, when considering the number of games played we would need to classify matches according to boundaries that are rather arbitrary.

Third, one could argue that it is not only the effort invested in the last match that matters, but rather the effort that is invested in all previous matches of the tournament. Although this is certainly a defendable option, we choose here to focus on the last match only, since it seems reasonable that the fatigue of matches played earlier than the previous one must have been digested.

2. Related work

The effect of fatigue in sports has been the topic of numerous research articles. Due to methodological difficulties, studies that focus on tennis are far less frequent. Reid and Schneiker (2008) give an overview of training methods that best prepare a professional player for tournament play. Davey, Thorpe and Williams (2010) examine the effect of fatigue from maximal tennis hitting on skilled tennis performance. The subjects undertook two performance tests: a pre- and post-skill test of groundstrokes and service, and the Loughborough Intermittent Tennis Test (4 minutes work plus 40 seconds recovery) to volitional fatigue.

Davey et al. find that groundstroke hitting accuracy decreased by 69% from start to volitional fatigue in the intermittent test, and service accuracy declined by 30% after the intermittent tennis test. Although the results of this study suggest that fatigue provokes a decline in some tennis skills, the experimental setup does not include match or competition conditions, nor does it consider the relative effort invested.

Smekal et al. (2001) examine physiological demands of single match play in tennis. In their setup, 20 players performed 10 matches of 50 minutes. For each player, respiratory gas exchange and heart rate were measured during each game; lactate concentration was determined after each game. Furthermore, the authors monitored the duration of rallies, the effective playing time, and the stroke frequency. The results suggest that energy demands of tennis matches are significantly influenced by the duration of the rallies, and that proper conditioning is advisable especially for players who prefer to play from the baseline. Although these authors use a more realistic setting, they do not investigate the impact of fatigue on the outcome of the match.

Our work is also related to the so called *carry-over effect*, which was originally proposed by Russell (1980). If some player (or team) X plays against player A in one round, and against player B in the next round, we say that player B receives a carry-over effect from player A. This carry-over effect is particularly relevant in physical, body-contact sports. For instance, if player A is very strong, and tough-playing, one can imagine that his opponent, player X, is weakened by injuries, fatigue or lowered morale, which could be an advantage for its next opponent, player B. Goossens and Spieksma (2012) empirically studied the influence of the carry-over effect on football (soccer) matches, but found no evidence for any meaningful impact on the match outcome. In tennis, if we use the effort it takes to defeat an opponent (i.e. the number of sets played) as a proxy for the strength of that opponent, we can translate this research in terms of the carry-over effect, and consider it as an empirical study to reveal the impact of carry-over in tennis.

3. Measuring the influence of fatigue

In order to measure the influence of fatigue resulting from a player's previous match, both for men and for women, we use data from the four Grand Slam tournaments between 1992 and 2011, resulting in 20,320 matches in total. We collected this data using the TennisNavigator database, the website tennis-data.co.uk, and Wikipedia. Each Grand Slam tournament is organized with direct knock-out matches, starting with 128 players and halving the number of competitors after each round, all the way up to the final. Most of these players qualify by ranking, others through a qualification tournament or thanks to a wildcard.

The idea of our approach is to classify each game based on the difference in sets played by the players in their previous match. For each group of matches, we compare the result of each match with the result that could be expected when they would have played an equal number of sets in their previous match. From the difference between these results, we obtain insight in the influence of fatigue from the previous match (and its significance). In the first subsection, we explain how we arrive at reasonable expected match results, which we can use as a basis of comparison. Finally, we discuss the details of the comparison and set up a statistical significance test. This approach is based on the methodology proposed by Goossens and Spieksma (2012).

3.1 Deriving reasonable match results

In this section, we determine for each tennis match which result we could expect when both players had played the same number of sets in their previous game. We assume that the result of a match is determined only by the tennis court type, and the strength of both players. The tennis court type depends on the surface: grass (Wimbledon), clay (French Open), acrylic hard court (US Open), and synthetic hard court (Australian Open). The strength of a player is measured by his/her ATP/WTP ranking at the start of the tournament.

For each Grand Slam tournament, we determine a matrix, which gives the proportion of wins for the stronger player, for matches between opponents belonging to given player strength groups. We define the following 7 strength groups: players ranked 1 to 4, 5 to 8, 9 to 16, 17 to 32, 33 to 64, 65 to 128, and finally players with a ranking below 128. These groups are conveniently in line with the tournament design and seeding used in Grand Slam tournaments. Furthermore, using more strength groups would not allow us to have sufficiently many observations for each pair of different strength groups, and using less groups would

not allow us to accurately express the strength of a player. To make sure that our matrix is not influenced by fatigue, we only took into account matches between players that played the same number of sets in their previous match, as well as all first round matches (ignoring the fact that a minority of the players had already played one or more qualification matches).

Table 1. Matrix with win proportions for Australian Open (WTA)

Rank	1-4	5-8	9-16	17-32	33-64	65-128	>128
1-4	0.50	0.68	0.79	0.92	0.96	0.96	0.96
5-8	0.32	0.50	0.85	0.69	0.88	0.94	0.89
9-16	0.21	0.15	0.50	0.56	0.87	0.90	0.90
17-32	0.08	0.31	0.44	0.50	0.59	0.75	0.73
33-64	0.04	0.12	0.13	0.41	0.50	0.61	0.64
65-128	0.04	0.06	0.10	0.25	0.39	0.50	0.53
>128	0.04	0.11	0.10	0.27	0.36	0.47	0.50

Table 1 gives the matrix we obtained for the Australian Open for women. Each cell gives the probability that the player from the stronger strength group wins, for matches between players from the corresponding strength groups. For instance, if a player's WTA rank is between 17 and 32, there is an 8% chance that this player will beat a top 4 player on the Australian Open. Cells on the diagonal were set to 50%; each cell (i,j) above the diagonal and its counterpart (j,i) below the diagonal add up to 1. Ideally, one would expect win proportions to go up when playing against weaker opponents (non-increasing in rows), and to go down when confronted with stronger opponents (non-decreasing in columns). We refer to these conditions as *regularity properties*. Although it turns out that the vast majority of these regularity properties are satisfied for our dataset, a close inspection of the table reveals some anomalies. For instance, for a player ranked between 5 and 8, a win against a player ranked between 9 and 16 is less likely than a win against a weaker player, ranked between 17 and 32. Since these irregularities may well distort our search for the influence of fatigue, we made minimal adjustments to the matrices, in order to find a more reasonable estimate of the win proportions, satisfying the regularity properties. This can be done by formulating and solving a linear optimization model; we refer to Goossens and Spieksma (2012) for more details on this procedure.

3.2 Comparison and significance test

The matrices in section 3.1 provide the *expected result* when both players have played the same number of sets in their previous game. In this section, we will compare the actual result of a collection of matches with the expected result. In Table 2, as an example, we selected 5 matches from the Australian Open for women where player 1 played one set more than player 2 in her previous match. The actual result is 1 if player 1 won, and 0 if player 2 won; the expected results are based on Table 1.

Table 2: Example matches where player 1 played 1 set more than player 2 in her previous match.

Year	Round	Player 1	Player 2	Ranking player 1	Ranking player 2	Actual result	Normal result
1992	3	Fendick P.	Dechaume A.	55	72	1	0.61
1995	2	Kamio Y.	Kandarr J.	28	81	0	0.75
1999	2	Frazier A.	Fernandez MJ.	42	76	0	0.61
2004	4	Farina Elia S.	Clijsters K.	23	2	0	0.08
2009	6	Safina D.	Zvonoreva V.	3	7	1	0.68

To find out whether the actual results of these matches differ significantly from the expected results, a chi-square test is performed, where the null hypothesis is that there is no difference between actual and expected results. The outcomes are summarized in Table 3. For this particular example (which has too few observations), the test does not allow us to conclude that the actual and expected observations originate from a different probability distribution. In other words, the differences we found are purely due to coincidence, and not to the fact that player 1 played an extra set in her previous match.

Table 3. Chi-square test for the example in Table 2

	Number won	Number lost
Actual number of observations	2	3
Estimated number of observations	2.73	2.27
$\chi^2 = 0.43$		
$\chi^2_{0.95} = 3.84$ (degrees of freedom = 1)		
p-value = 0.51		

4. Results

We now give results for the four Grand Slam events for women (Table 4) and for men (Table 5). Each row in a table corresponds to the difference in sets played in the previous matches. There are three columns for each tournament, the first column gives the value of the χ^2 statistic, the second column gives the corresponding p-value, and the third column gives the number of matches in this category. Observe that in each tournament the number of matches equals 1260 (which makes sense because there are 63 non first-round matches in each tournament, and we consider a 20-year period).

Table 4. Effects of difference in sets played in the previous game for ATP (men)

Diff. in sets	Australian Open			French open			US Open			Wimbledon		
	χ^2	p	#games	χ^2	p	#games	χ^2	P	#games	χ^2	p	#games
0	0.14	0.71	438	0.52	0.47	448	0.27	0.60	478	0.62	0.43	447
1	0.03	0.86	533	1.40	0.24	527	0.68	0.41	511	0.29	0.59	507
≥ 2	2.72	0.10	289	0.95	0.33	285	1.87	0.17	271	5.72	0.02	306

Table 5. Effects of difference in sets played in the previous game for WTA (women)

Diff. in sets	Australian Open			French open			US Open			Wimbledon		
	χ^2	p	#games	χ^2	p	#games	χ^2	P	#games	χ^2	P	#games
0	0.19	0.66	723	0.03	0.86	703	1.83	0.18	730	1.65	0.20	689
≥ 1	3.66	0.06	537	4.63	0.03	557	4.51	0.03	530	5.76	0.02	571

We make the following observations:

- As expected, in case the difference in sets equals 0, there is hardly any deviation from the expected match result, both for men and women. This is not surprising since, to a large extent, we use these matches to construct the win matrices. It does confirm that the minor modifications that we carry out on these matrices in order to satisfy the regularity assumptions are indeed minor. We conclude that our win matrices are useful and valid.
- In case of the men (Table 4), it turns out that playing one additional set compared to your opponent in the previous round does not decrease your chances: the deviations are small and not significant. And except for Wimbledon, a similar statement can be made for the case of two additional sets played, although the deviations are somewhat larger, and the p-values are somewhat smaller. More

precisely, both at the US Open and the French open we do not see any significant effect, and for the Australian Open, the effect is borderline significant. However, in the case of Wimbledon it apparently matters whether one played two more sets than one's opponent. Observe also that this frequently happens: almost 1 out of 4 matches at Wimbledon, features a player who won a five setter whereas his opponent won in straight sets.

- For women (Table 5), the situation is different. For each of the four Grand Slams, having played one additional set compared to your opponent does decrease your chances. The effect is again largest at Wimbledon, but statistically significant in each of the tournaments.

One might wonder which cause is responsible for the observed phenomenon that, especially at Wimbledon, having played one (in case of women) or two (in case of men) affects your chances. One possible explanation is that due to the dominance of serve at Wimbledon compared to the other Grand Slams, a Wimbledon tennis match stands a larger chance to feature tie-breaks, and hence has a larger chance of going to five sets. In fact, this is observed in Summers (2011) who report such findings when comparing Wimbledon to the French Open. Moreover, also because of the absence of a tie-breaker in the fifth set of Wimbledon, the effort of playing a five setter at Wimbledon does decrease one's chances of winning the next match.

5. Conclusion

We consider the question whether the probability of winning a match in a Grand Slam tournament is affected by the relative effort invested in the previous match. This relative effort invested in the previous match is found by looking at the difference of number of sets played in the previous match. By constructing win matrices containing an unbiased probability of a particular match outcome, we find that:

- in case of women, there is a statistically significant event in each of the four Grand Slam events. The effect is strongest at Wimbledon.
- in case of men, having played one additional set compared to your opponent does not decrease one's chances of winning the next match. The same conclusion can be drawn when having played two additional sets, except for Wimbledon where we find a (strong) statistically significant effect.

Acknowledgements

The authors wish to thank Anne Baeten, Hilde Van Grieken, Maarten Schmook and Anneleen Steurs for their excellent contributions to the master thesis which lead to this paper.

References

- Davey, P.R., Thorpe, R.D. and Williams, C. (2002). Fatigue Decreases Skilled Tennis Performance. *Journal of Sport Sciences* **20**, 311-318.
- Goossens, D. R. and Spieksma, F. C. R. (2012). The carryover effect does not influence football results. *Journal of Sports Economics* **13**, 261-278.
- Reid, M. and K. Schneiker (2008). *Strength and conditioning in tennis: Current research and practice*. Journal of Science and Medicine in Sport **11**, 248—256.
- Russell, K. (1980). Balancing carry-over effects in round robin tournaments. *Biometrika* **67**, 127-131.
- Smekal, G., Von Duvillard, S. P., Rihacek, C., Pokan, R., Hofmann, P., Baron, R., Tschan, H. and Bachl, N. (2001). A Physiological Profile of Tennis Match Play. *Medicine and Science in Sports and Exercise* **33**, 999-1005.
- Summers, M.R. (2011). *Clay Vs. Grass: A Statistical Comparison of the French Open and Wimbledon*. American Journal of Economics and Business Administration **3**, 405-409.

Are Soccer Schedules Robust?

Dries Goossens* and Fabrice Talla Nobibon*

* PostDoc researcher for Research Foundation–Flanders, ORSTAT, Faculty of Business and Economics, Naamsestraat 69, KU Leuven, Belgium, {dries.goossens; fabrice.tallanobibon}@kuleuven.be

Abstract. The purpose of this paper is to introduce the concept of robust optimization in sport scheduling, and to study empirically the robustness of soccer schedules. We compare initial schedules, constructed before the beginning of the season, with realized schedules, i.e. the schedules as they were actually played. The notion of disruption is proposed to characterize meaningful differences between initial and realized schedules. We consider two quality measures to evaluate the robustness of soccer schedules: the number of breaks, and the balancedness of the home advantage. Finally, for ten main European soccer leagues, we identify disruptions and investigate their effects on the quality of the schedule.

1. Introduction

Each sport competition needs a schedule of play. In professional soccer, most leagues play a round robin schedule, i.e. each team plays against each other team a fixed number of times (usually twice). Constructing a suitable schedule for a professional soccer league is not an easy challenge because numerous wishes from various stakeholders (league, clubs, fans, TV, police, etc.) must be taken into account. Soccer scheduling problems in several European leagues have been studied by academics, and have been reported for the following countries: Austria and Germany [7], Belgium [18], Denmark [34], England [23], Italy [12], The Netherlands [38], and Norway [20]. In some cases, the authors managed to close a contract with the soccer association, and are providing the official schedule.

Despite hard efforts to create a good schedule at the beginning of each season, this schedule is hardly ever fully played as planned. Indeed, some matches need to be postponed due to bad weather conditions; others are rescheduled because of conflicts with the outcome of other competitions (cup, champions league, etc.). Furthermore, in some competitions teams are entitled to have their match rescheduled if they miss a number of international players (e.g. because of African Cup of Nations). In rare cases, entire matchdays are put off, as happened in Northern Ireland (2008) and Spain (2011) and due to a strike of the referees and the players, respectively. These disruptions may seriously affect the quality of the schedule.

The quality of a schedule is determined by how well it manages to satisfy a number of constraints. For instance, two teams sharing the same home venue must not have simultaneous home games, or a team A may not be able to play a home game against some team B on a particular date, because insufficient police force is available at that time to guarantee public safety. Nurmi et al. [33] provide an overview of typical constraints in sport scheduling. Generally, each constraint receives a weight relative to its importance, and the quality of the schedule is determined by the weighted sum of satisfied constraints. However, the type of constraints and their importance tends to depend heavily on traditions and geographical conditions, and hence, differ from one league to another. Moreover, the exact constraints and their weights are not publicly available. Therefore, in this paper we measure the quality of soccer schedules using two indices that do not depend on the considered league. These are: *breaks* and *balancedness*. We say that a team has a break if it has two consecutive home games, or two consecutive away games. The less breaks, the more desirable the schedule. With balancedness, we mean that the home advantage should be balanced over the

teams and over the season. The validity of our choice for these indices is illustrated by a survey of European football schedules [19], which shows that in almost all competitions, schedules are used that minimize the number of breaks and balance the home advantage.

In this paper, we focus on robust soccer scheduling; in other words, we investigate the generation of soccer schedules such that their quality is minimally impacted by the disruptions. As far as we are aware, a robust optimization approach has not been studied before in the context of sport scheduling. There are, however, numerous successful application of robust optimization in other domains, such as resource allocation, production planning, location, inventory, network design, project management, etc. [2, 8, 27, 39]

The next section presents some notation and definitions in the domain of sport scheduling and robust scheduling. We provide a formal definition of our quality indices, and a number of properties in Section 3. Section 4 provides an empirical study of the robustness of the main European soccer schedules for the seasons 2002–2012. Finally, we conclude in Section 5.

2. Notation and definitions

In this paper, we consider only leagues that have an even number of teams; we denote by $S = \{1, 2, \dots, 2n\}$ the set of teams. Each game consists of an ordered pair of teams, noted as (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. A *round* is a set of games, usually played in the same weekend, in which every team plays at most one game; the set of rounds is denoted by $T = \{1, 2, \dots, R\}$. We use the symbol $|\cdot|$ both for the absolute value of a number and for the cardinality of a set.

We now introduce the notions of *disruption* and *fictive round*. For a given initial schedule, if one or more games of round $r \in T$ were played after at least one game of round $r + 1 \in T$, then we say that there was a *disruption*. Similarly, if some games of round $r \in T$ were played before at least one game of round $r - 1 \in T$, then we also say that there was a disruption. Although a game that was not (fully) played at the scheduled date is usually postponed, it is also possible to reschedule it to an earlier date, provided of course that it is known well beforehand that it will not be possible to play the game as initially scheduled. Notice that a game which is not played as initially scheduled, but is rescheduled before any game of the next round is played (and after games of all previous rounds) is not considered as a disruption. Indeed, in this case, the order of the games remains the same. Furthermore, games that are awarded to one of both teams without rescheduling are also not considered as disruptions. Disrupted games are usually scheduled on a date on which no matches were planned in the initial schedule. All the games scheduled on that date constitute a new round, that we call *fictive round* because this round was not present in the initial schedule.

Below, we give basic notions of round robin scheduling (Section 2.1), and robust optimization (Section 2.2).

2.1 Basic notions of round robin scheduling

A round robin schedule consists of games, assigned to rounds, such that each team plays against each other team a fixed number of times. Most soccer leagues play a double round robin tournament (2RR), where the teams meet twice (once at home, once away), but triple and quadruple round robin tournaments are also used. We refer to a *phased schedule* if no team plays against any other team for the k -th time, before it played at least $k - 1$ matches against all other teams. Phased schedules can be split into equal parts, such that each part forms a single round robin tournament. Usually,

the order of the rounds in these parts is related. Most competitions use *mirroring*, i.e. the second part of the competition is identical to the first, except that the home advantage is inverted (in case of a third part, the second is mirrored, and so on). Another possibility is the so-called *French scheme*, where matches in the first and the last round are identical, as well as matches in round $2n - 1 + t$ and round $t + 1$ with $t = 1, 2, \dots, 2n - 2$ (again with the home advantage inverted).

A schedule is *compact* if it uses the minimum number of rounds required to schedule all the games; otherwise it is *relaxed*. In a compact schedule with an even number of teams, each team plays exactly one game in each round. In a relaxed round robin schedule, teams do not play in each round. We say that a team has a *bye* in some round if it does not play in that round. The sequence of home matches ('H') and away matches ('A') played by a single team is called its home-away pattern (HAP). Traditionally, rounds in which the team has a bye are ignored in the home-away pattern. Given such a HAP, the occurrence of two consecutive home matches, or two consecutive away matches is called a *break*. Teams can have consecutive breaks, causing them to play three or more home (away) games in a row. A series of consecutive away games is called an *away tour*. For team $i \in S$ and a round $r \in T$, we denote by $h_{i,r}$ (respectively $a_{i,r}$) the number of home (away) games played by that team up to round r (including round r).

The example in Table 1 shows a schedule for a double round robin tournament with 6 teams, named A to F. Notice that this schedule is phased and compact, and applies the French scheme. The HAP for team A is HAHAHHAHAA, which contains two breaks. Team F starts the competition with an away tour of 3 consecutive away games.

Table 1: A double round robin schedule for 6 teams

R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
A-B	B-E	A-F	C-B	B-F	E-B	F-A	B-C	F-B	B-A
C-D	D-A	B-D	E-A	D-E	A-D	D-B	A-E	E-D	D-C
E-F	C-F	E-C	F-D	A-C	F-C	C-E	D-F	C-A	F-E

Many of the theoretical results and algorithms in sport scheduling are based on graph theory. As far as we are aware, de Werra [11] was the first to use the complete graph K_{2n} on $2n$ nodes for constructing single round robin tournaments, where the nodes correspond with the teams, and the edges with games between the teams. A compact schedule can then be seen as an edge coloring with $2n - 1$ colors, i.e. a partitioning of the edge set into $2n - 1$ perfect matchings. For an overview of graph-based models in sports scheduling, we refer to the work by Drexl and Knust [13]. Kendall et al. [24] present an annotated bibliography of sports scheduling literature. Finally, we also mention a survey on round robin scheduling by Rasmussen and Trick [35].

2.2 Basic notions of decision making under uncertainty

Most solutions to real-world problems are computed based on estimated parameters. Furthermore, for the majority of off-line problems there is a time gap between the instant when the decision (solution) is taken and the moment when it is implemented. The soccer scheduling problem belongs to the latter class of problems; indeed, there are several months difference between the time the initial schedule is published and the moment some games are played. During this time, some information that were not known when computing the initial schedule become available and may affect the implementation of that schedule. In other words, the initial schedule is established based on incomplete and (often) unreliable data, which is mainly due to *uncertainty*. This uncertainty

can originate from several potential sources including: bad weather conditions, conflicts with the outcome of other competitions (cup, champions league, etc.), strike of the players and the referees, fans' behavior (e.g. hooliganism, violence), technical problems (e.g. lights) in the stadium, sanitary reasons (e.g. epidemic diseases).

Uncertainty is usually modeled using probability theory [22], leading to stochastic models of the considered problem (see for instance [1, 14, 28, 31]). In decision theory, there is a distinction between *risk*, *uncertainty* and *ignorance*. In a risk situation, the distribution of the outcomes under study is known with certainty whereas under uncertainty it is not possible to attribute probabilities to the possible outcomes of a decision [16, 25, 36], and ignorance pertains to the case where even the possible outcomes are not known [30]. Rosenhead et al. [36] observe that “it may be possible to convert an uncertainty problem into a risk problem, for example by the subjective estimation of probabilities, and used appropriately this can be a valuable simplification. However, some aspects of the future are genuinely unknowable, even in the probability sense. To insert notional probabilities may make the decision maker more comfortable, but that is not necessarily the objective in tackling a decision problem.”

A number of criteria are used for decision making under uncertainty; some of the most important ones for a minimization problem are (1) *minimax*: minimize the worst value of the quality measure [40], (2) *minimin*: minimize the best outcome that can occur, which is an optimistic approach, as opposed to the pessimistic minimax [21], (3) *minimax regret*: minimize the largest possible difference in quality measure between the value of the initial schedule and that of the realized schedule [37], and (4) minimize the objective *in expectation*. Within the context of this paper, the objectives (1) and (2) can be addressed by solving a specific soccer scheduling problem. For *stochastic models*, where a probability distribution is known for the possible outcomes, the common objective is to select a schedule that minimizes the expected value (4) of the schedule. In the context of this article, however, probability distributions are not available and so expected values cannot be computed.

In [3], Assavapokee et al. observe that because of incomplete information about the joint probability distribution of the uncertain parameters in the problem, decision makers are often unable to search for decisions with the best long-run average performance. Instead, they search for *robust* decisions, which perform well across all possible input scenarios without attempting to assign a fixed probability distribution to any ambiguous parameter. Daniels and Kouvelis [9] motivate the choice of regret-based objectives as follows: “a decision maker may be rightfully concerned not only with how a schedule’s performance varies with the actual realizations of the task parameters, but also with how actual performance compares with the optimal performance that could have been achieved if perfect information had been available prior to scheduling. Such comparisons against optimal performance focus the decision maker on opportunities to free short-term capacity by reducing uncertainty and efficiently utilizing resources through scheduling, . . .”. Comparable regret-based objectives have recently been examined for various combinatorial optimization problems [2–6, 29, 32, 39]. The most commonly used regret-based objectives are: the *absolute-deviation robust* and the *relative-deviation robust*; we refer to [27] for a practical motivation of these objectives.

3. Quality measures for soccer schedules

We present the quality measures that we use to evaluate soccer schedules. We first discuss the number of breaks in a schedule, and subsequently we introduce two balancedness measures.

3.1 Breaks

In order to reduce travel costs, teams may prefer to have two or more consecutive away games if its stadium is located far from the opponent's venues, and the venues of these opponents are close to each other. Scheduling away tours, however, is quite uncommon in European soccer. In fact, in most competitions breaks (and certainly successive breaks) are avoided as much as possible. The reason is pointed out by Forrest and Simmons [15], who show that scheduling home games consecutively has a negative impact on attendance. Therefore, it is desirable for each team to have a perfect alternation of home and away games; this observation has allowed to use the total number of breaks in a schedule as the measure the quality of that schedule.

It is easy to see that only two different patterns without breaks exist (HAHA...H and AHAH...A). Moreover, all teams must have different patterns (indeed, two teams with the same pattern can never play against each other), and hence, at most two teams will not have any break. Consequently, any (compact) round robin schedule for $2n$ teams will have at least $2n - 2$ breaks; de Werra [11] shows that such a schedule can be constructed for any n .

For a double round robin tournament, mirroring a compact single round robin schedule with a minimal number of breaks results in $6n - 6$ breaks. If $2n \neq 4$, this can be achieved without a team having successive breaks [11]. Sometimes, leagues prefer to equally distribute the breaks over the teams; the minimum number of breaks is then $2n$ for a single, and $4n$ for a double round robin tournament [10]. This type of schedule is called an *equitable* schedule. Starting from an equitable single round robin schedule, the French scheme is a way to create an equitable double round robin schedule. If there is no need for a phased schedule, we can limit the number of breaks to $2n - 2$ [19]. If we consider relaxed schedules, Froncek and Meszka [17] show that there exists a unique schedule in which every team has one bye and no break.

3.2 Balancedness indices

Given the advantage that the home team turns out to have in soccer, it is fair that each teams plays half of its games at home, and the other half away. Knust and von Thaden [26] call a schedule balanced if for each team the number of home and away games played at the end of the season differ by at most one. In case of an odd number of teams, this difference of one is inevitable. Nurmi et al. [33] generalize this notion of balanced schedule over all rounds (instead of just the last round), and introduce k -balancedness.

Definition 1. *A schedule is k -balanced if and only if*

$$k := \max_{r \in T} k_r := \max_{r \in T} \max_{i \in S} |h_{i,r} - a_{i,r}|. \quad (1)$$

In words, a schedule is k -balanced if and only if for each team $i \in S$ and for each round $r \in T$, the difference between the number of home and away games played by team i up to round r is at most k , and there is a team that achieves this value of k for at least one round in the schedule. Note that for each schedule we have $k \geq 1$.

Goossens and Spieksma [19] point out the importance of having a league table that offers a fair ranking after each round. In this interpretation, ideally, each team will have played the same number of home games after each round. They define g -ranking-balancedness as follows:

Definition 2. *A schedule is g -ranking-balanced if and only if*

$$g := \max_{r \in T} g_r := \max_{r \in T} \left\{ \max_{i \in S} (h_{i,r}) - \min_{j \in S} (h_{j,r}) \right\} = \max_{r \in T} \left\{ \max_{i \in S} (a_{i,r}) - \min_{j \in S} (a_{j,r}) \right\}. \quad (2)$$

Literally, a schedule is g -ranking-balanced if and only if for each round $r \in T$ and each pair of teams $i, j \in S$, the difference between the number of home games played by team i and the number of home games played by team j is at most g , and there are teams that achieve this value of g for at least one round in the schedule. We also have $g \geq 1$ for each schedule.

We assume that a high-quality schedule minimizes both balancedness indices. We next present a theoretical comparison between the indices k and g .

Proposition 1. *For any schedule we have $g \leq k$, i.e. $\max_{r \in T} g_r \leq \max_{r \in T} k_r$.*

Proof: We prove that for any round $r \in T$ we have $g_r \leq k_r$. Consider a given round $r \in T$ and $k_r := \max_{i \in S} |h_{i,r} - a_{i,r}|$. There exists a team $i_r \in S$ such that $k_r := |h_{i_r,r} - a_{i_r,r}|$. Let us assume, without loss of generality, that $h_{i_r,r} \geq a_{i_r,r}$ such that $k_r := h_{i_r,r} - a_{i_r,r}$. This implies that $h_{i_r,r} = \max_{j \in S} (h_{j,r})$. Because $h_{j,r} + a_{j,r} = r$ we infer that $h_{j,r} \geq a_{i_r,r}$ and hence $-h_{j,r} \leq -a_{i_r,r}$, for any team $j \in S$. Therefore, we have:

$$\begin{aligned} \max_{j \in S} \{-h_{j,r}\} &\leq -a_{i_r,r}, \\ -\min_{j \in S} \{h_{j,r}\} &\leq -a_{i_r,r}, \\ h_{i_r,r} - \min_{j \in S} \{h_{j,r}\} &\leq h_{i_r,r} - a_{i_r,r}, \\ \max_{i \in S} h_{i,r} - \min_{j \in S} \{h_{j,r}\} &\leq h_{i_r,r} - a_{i_r,r}, \\ g_r &\leq k_r. \end{aligned}$$

Finally, by taking the maximum over $r \in T$ we obtain the result of Proposition 1. \square

Proposition 2. *For any schedule, we have that $g = 1$ if and only if $k = 1$.*

Proof: Proposition 1 implies that any schedule with $k = 1$ has $g = 1$. Suppose that we have a schedule with $g = 1$; we prove by contradiction that $k = 1$. If $k = 2$ then there exists $r \in T$ such that $k_r := \max_{i \in S} |h_{i,r} - a_{i,r}| := 2$. Without loss of generality, we assume that there exists a team i_r such that $h_{i_r,r} \geq a_{i_r,r}$ and $k_r := h_{i_r,r} - a_{i_r,r} = 2$. Because $g = 1$ we have $h_{j,r} \geq h_{i_r,r} - 1$, for all $j \in S$. Let $S_r := \{j : h_{j,r} = h_{i_r,r}\}$ be the set of teams that have played the highest number of home games up to round r . We then have:

$$\begin{aligned} \sum_{j \in S} h_{j,r_k} &= \sum_{j \in S_r} h_{j,r} + \sum_{j \notin S_r} h_{j,r} \\ &= |S_r| h_{i_r,r} + (2n - |S_r|) (h_{i_r,r} - 1) \\ &= |S_r| + 2n (h_{i_r,r} - 1). \end{aligned} \tag{3}$$

On the other hand, because $g = 1$ and $h_{j,r} + a_{j,r} = r$ we also have that $a_{j,r} \geq a_{i_r,r} + 1$ for all $j \in S$. Therefore,

$$\begin{aligned} \sum_{j \in S} a_{j,r} &= \sum_{j \in S_r} a_{j,r} + \sum_{j \notin S_r} a_{j,r} \\ &= |S_r| a_{i_r,r} + (2n - |S_r|) (a_{i_r,r} + 1) \\ &= 2n (a_{i_r,r} + 1) - |S_r|. \end{aligned} \tag{4}$$

By subtracting (3) from (4), we obtain:

$$\begin{aligned}
\sum_{j \in S} a_{j,r} - \sum_{j \in S} h_{j,r} &= 2n(a_{i_r,r} + 1) - |S_r| - |S_r| - 2n(h_{i_r,r} - 1) \\
&= 2n(a_{i_r,r} - h_{i_r,r}) + 4n - 2|S_r| \\
&= -4n + 4n - 2|S_r| \\
&= -2|S_r| \neq 0,
\end{aligned}$$

which implies that the sum of all home games is greater than the sum of all away games. A contradiction! A similar analysis can be developed for $k \geq 3$. This completes the proof of Proposition 2. \square

Proposition 3. *For any schedule, we have that $k \leq 2g - 1$.*

Proof: For a given round $r \in T$, either (1) $h_{i,r} = a_{i,r}$ for all team $i \in S$ or (2) there exists a team $j \in S$ such that $h_{j,r} \neq a_{j,r}$. Let $T_+ = \{r : h_{i,r} = a_{i,r}, \forall i \in S\}$ and $T_- = \{r : r \notin T_+\}$. Because $k \geq 1$ we infer that $k := \max_{r \in T_-} k_r$. Also, because $g \geq 1$ we have $g := \max_{r \in T_-} g_r$. We now argue that for all $r \in T_-$ we have $k_r \leq 2g_r - 1$. Let $r \in T_-$ and $i_r \in S$ such that $k_r := h_{i_r,r} - a_{i_r,r}$; because $r \in T_-$ we have $\max_{j \in S} a_{j,r} \geq \min_{i \in S} h_{i,r} + 1$. Therefore,

$$\begin{aligned}
k_r &= h_{i_r,r} - a_{i_r,r} \\
&\leq h_{i_r,r} - \min_{j \in S} h_{j,r} - 1 + \max_{j \in S} a_{j,r} - a_{i_r,r} \\
&\leq 2g_r - 1.
\end{aligned}$$

By taking the maximum over all $r \in T_-$, we obtain the result of Proposition 3. \square

The schedule depicted in Table 2 consisting of three rounds and involving four teams has $k = 3$ and $g = 2$; this shows that the bound provided by Proposition 3 is tight.

H	H	H
A	H	A
H	A	A
A	A	H

Table 2: Illustration of a tight schedule with $k = 3$ and $g = 2$.

We complete this section by establishing a relation between the number of breaks and the above balancedness indices.

Proposition 4. *A schedule has $g = 1$ if and only if there are no breaks on even rounds.*

Proof: On the one hand, suppose that we have a schedule with $g = 1$ and a break on an even round. It is not difficult to see that this will imply that $k > 1$ and we have a contradiction with Proposition 2. On the other hand, if a schedule does not have a break on even rounds then clearly $g = 1$. \square

De Werra [11] showed that for any $2n$ teams, it is possible to construct a single round robin schedule with minimal number of breaks and without breaks on even rounds. Hence, in this case, a schedule exists for which all the above quality measures are minimized.

4. Empirical Analysis

We consider the main ten European soccer leagues and we compare the quality of initial schedules, computed before the beginning of the season, to that of realized schedules, which are known only at the end of the season, for the seasons 2002–2012. Below, we give details about the leagues and the data in Section 4.1, and we comment the obtained results in Section 4.2.

4.1 Data

For the seasons 2002–2012, we consider the following European soccer leagues: Belgium, England, France, Germany, Italy, Netherlands, Portugal, Russia, Spain, and Ukraine. We gather the information about initial schedules and realized schedules from the following websites:

<http://www.the-sports.org/soccer-foot-s1-c0-b0.html>

<http://www.rsssf.com>

In some competitions (e.g. Belgium, Netherlands), the regular round robin stage of the competition is followed by a play-off stage. Typically, play-offs involve only a subset of the teams in the league, and it is not clear from the beginning of the season which teams will take part in the play-off stage (this depends on their performance in the regular stage). Consequently, no play-off schedule can be made at the beginning of the season. Furthermore, play-off competitions are often implemented in a direct knock-out format according to a fixed schedule, which leaves little or no room for (robust) optimization. For these reasons, we chose to ignore play-offs in this paper.

In all countries except the Netherlands, the initial schedules are compact. Indeed, in the Netherlands, 2 extra rounds were used for seasons 2002-2003 and 2003-2004; the 2004-2005 season had 1 extra round scheduled. The Belgian league had two seasons in which a team was not able to finish the season due to bankruptcy, and hence, these teams' matches were awarded 5-0 to the opponent. As explained before, we do not consider this a disruption of the initial schedule.

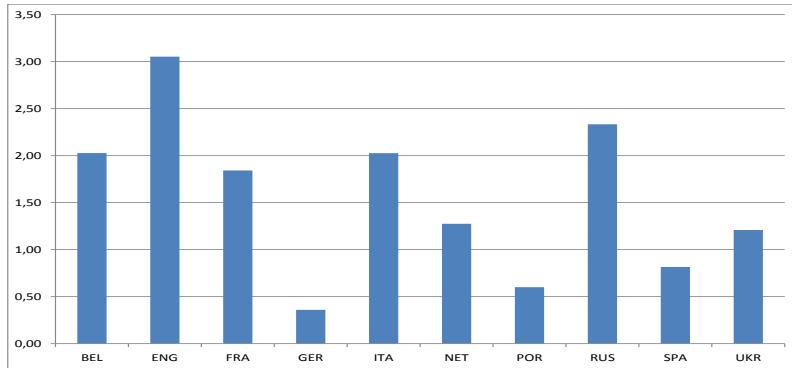
4.2 Results

We analyze the variation of the quality measures between initial and realized schedules for the ten leagues considered. Figure 1 gives an overview of the percentage of disrupted matches per league (Figure 1(a)), and per month (Figure 1(b)). We recall that a disruption is not the same as a postponed match; for instance, matches that are rescheduled before the next round starts are not considered as disruptions. It turns out that in 89% of the seasons we studied, at least one disruption occurred. Overall, on average 1.59 % of the scheduled matches are disrupted, although the frequency of disruptions differs considerably between the various leagues. For instance, England has over eight times more disruptions than Germany. Furthermore, a mild climate is no guarantee for a low number of disruptions. Nevertheless, winter clearly plays a role, as the months December, January, and February clearly cause more disruptions than the average month. July and August are also problematic, which is probably due to the large number of teams that are participating in European competitions in that period. The month of June has no disruptions, although we must mention that most leagues do not schedule matches in this month.

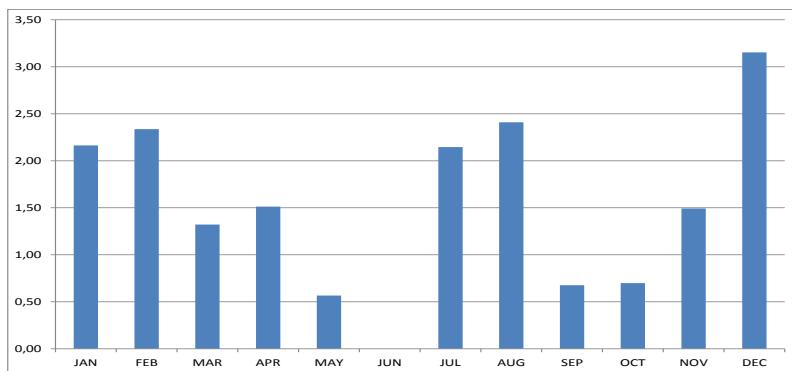
In Table 3, we display the value of each quality measure for each country during the last ten seasons. Each cell contains a triple of numbers; the first (respectively the second and the third) value represents the number of breaks (respectively the k -index and the g -index) of the schedule. The triple (25|5|4), for instance, means that the considered schedule has 25 breaks, a k -index of 5

League		2002–2003	2003–2004	2004–2005	2005–2006	2006–2007	2007–2008	2008–2009	2009–2010	2011–2011	2011–2012
Belgium	initial	(56 4 4)	(56 3 2)	(50 2 2)	(48 2 2)	(48 2 2)	(48 2 2)	(48 2 2)	(42 2 2)	(42 2 2)	(42 2 2)
	realized	(90 4 4)	(68 3 2)	(54 2 2)	(78 2 2)	(48 2 2)	(48 2 2)	(53 2 2)	(76 4 3)	(72 3 3)	(42 2 2)
England	initial	(126 2 2)	(94 2 2)	(130 1 1)	(108 2 2)	(117 0 1 1)	(134 2 2)	(130 1 1)	(172 1 1)	(124 3 3)	(126 1 1)
	realized	(140 3 3)	(118 3 3)	(160 2 2)	(152 3 3)	(186 2 2)	(138 3 3)	(140 3 3)	(204 3 3)	(160 6 5)	(134 2 2)
France	initial	(40 2 2)	(40 2 2)	(40 2 2)	(40 2 2)	(40 2 2)	(40 2 2)	(40 2 2)	(40 2 2)	(40 2 2)	(40 2 2)
	realized	(70 4 3)	(60 3 3)	(58 3 3)	(68 3 3)	(52 2 2)	(58 2 2)	(44 2 2)	(102 3 3)	(54 3 3)	(60 3 3)
Germany	initial	(48 2 2)	(48 2 2)	(48 2 2)	(48 2 2)	(48 2 2)	(48 2 2)	(48 2 2)	(48 2 2)	(48 2 2)	(48 2 2)
	realized	(52 2 2)	(56 2 2)	(52 2 2)	(54 2 2)	(48 2 2)	(56 2 2)	(50 2 2)	(48 2 2)	(52 3 3)	(50 3 2)
Italy	initial	(58 2 2)	(60 2 2)	(68 2 2)	(66 2 2)	(68 2 2)	(66 2 2)	(66 2 2)	(66 2 2)	(64 2 2)	(66 2 2)
	realized	(76 2 2)	(64 2 2)	(72 2 2)	(70 3 3)	(104 3 3)	(82 2 2)	(70 2 2)	(86 2 2)	(80 3 2)	(142 3 3)
Netherlands	initial	(122 3 3)	(140 3 2)	(110 2 2)	(116 3 2)	(120 3 3)	(100 3 2)	(116 3 2)	(114 3 3)	(112 3 3)	(124 5 4)
	realized	(144 3 3)	(148 3 3)	(118 3 3)	(122 3 3)	(124 3 3)	(118 4 3)	(116 3 2)	(116 3 3)	(120 3 3)	(130 5 4)
Portugal	initial	(48 2 2)	(48 2 2)	(48 2 2)	(48 2 2)	(42 2 2)	(42 2 2)	(42 2 2)	(42 2 2)	(42 2 2)	(42 2 2)
	realized	(54 2 2)	(58 2 2)	(50 2 2)	(48 2 2)	(48 3 3)	(42 2 2)	(46 2 2)	(48 3 2)	(46 3 3)	(44 2 2)
Russia	initial	(52 3 3)	(54 4 3)	(66 2 2)	(78 4 4)	(60 6 5)	(44 6 5)	(32 2 2)	(40 4 4)	(32 2 2)	(50 4 4)
	realized	(68 4 4)	(56 4 3)	(82 3 3)	(96 4 5)	(92 7 5)	(52 6 5)	(56 3 3)	(48 5 5)	(50 3 2)	(56 4 4)
Spain	initial	(54 2 2)	(54 2 2)	(54 2 2)	(54 2 2)	(54 2 2)	(54 2 2)	(54 2 2)	(54 2 2)	(54 2 2)	(54 2 2)
	realized	(62 2 2)	(58 2 2)	(62 3 2)	(58 2 2)	(56 2 2)	(56 2 2)	(54 2 2)	(58 2 2)	(54 2 2)	(84 3 3)
Ukraine	initial	(28 2 2)	(28 2 2)	(60 2 2)	(42 2 2)	(42 2 2)	(42 2 2)	(42 2 2)	(50 4 3)	(44 2 2)	(44 2 2)
	realized	(30 2 2)	(28 2 2)	(40 3 3)	(46 2 2)	(44 2 2)	(46 2 2)	(48 2 2)	(78 5 4)	(44 2 2)	(48 2 2)

Table 3: Value of the quality measures for the initial and the realized schedules of each league between 2002 and 2012.



(a) Average % of disrupted matches per league



(b) Average % of disrupted matches per month

Figure 1: Frequency of disruptions

and a g -index of 4. Table 3 shows that some countries consistently produce schedules that have the same quality (e.g. France, Germany, Spain). This suggest that these leagues have specific rules with respect to breaks and balancedness, which they explicitly take into account in the scheduling process. In some countries (e.g. Belgium, Portugal), the number of teams changed during the period of 10 years we studied, which explains some of the differences in number of breaks. In general, most leagues perform quite good on both quality measures for the initial schedules. Only in England, some initial schedules are computed which are perfectly balanced ($k = g = 1$), however at the expense of the number of breaks. The quality of the schedules in the Netherlands appears to be quite poor, with a high number of breaks and an unevenly balanced home advantage; we are not aware of any reasons for this phenomenon. Furthermore, out of the 100 cells in the table, only 11 have the same values for the initial and the realized schedules, corresponding to those seasons without disruptions. We analyze the impact of disruptions in more details in the next figures.

Figure 2 shows the effect of disruptions on the average number of breaks per team for each season and each league. We observe that the number of breaks is quite sensitive to disruptions;

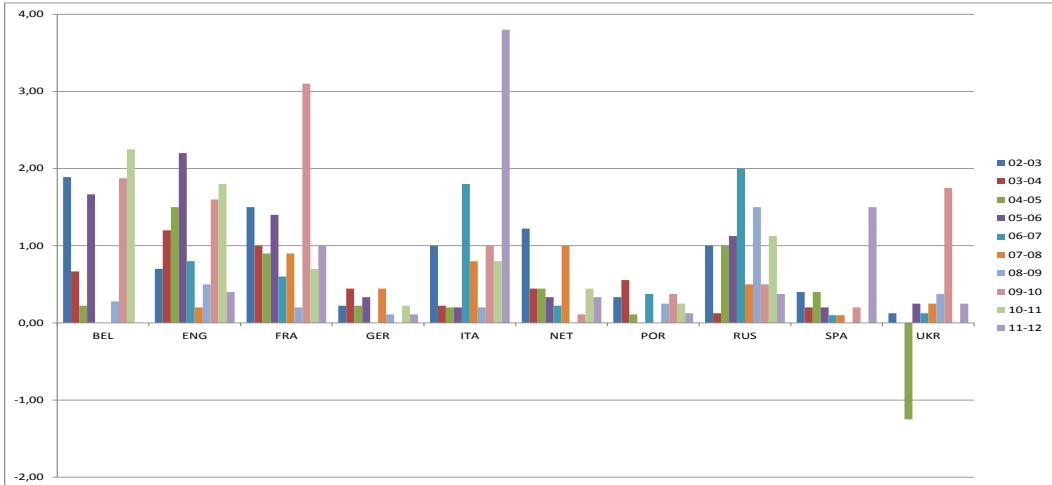


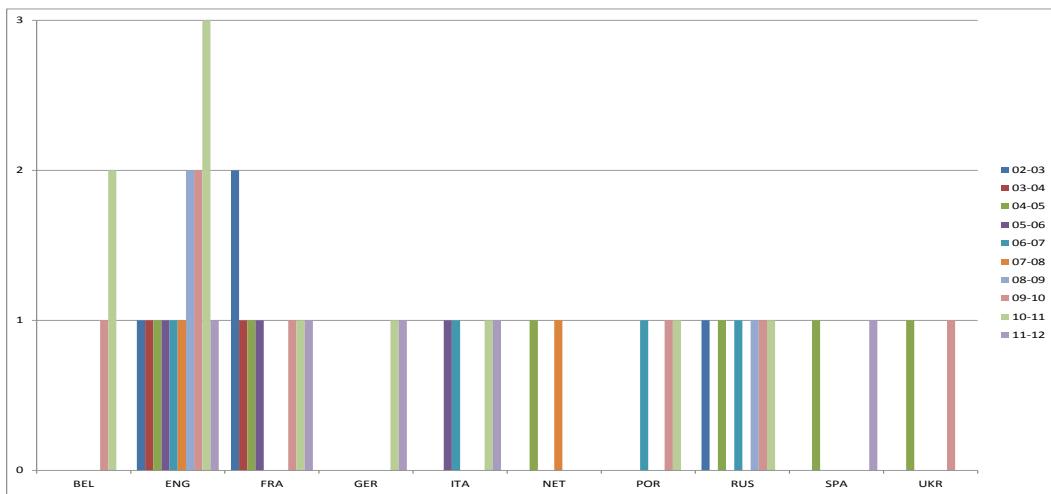
Figure 2: Effect of disruptions on the average number of breaks per team

indeed, in most cases there is a big difference between the value obtained for the initial schedule and that of the realized schedule. On a few occasions (France 2009–2010, Italy 2011–2012), the total number of breaks is even doubled. As an exception, we mention that for the season 2004–2005 in Ukraine, the number of breaks actually decreased because of disruptions. Furthermore, the fact that initial schedules in countries like England and the Netherlands already have a lot of breaks, does not imply that disruptions are handled in a better way with respect to breaks.

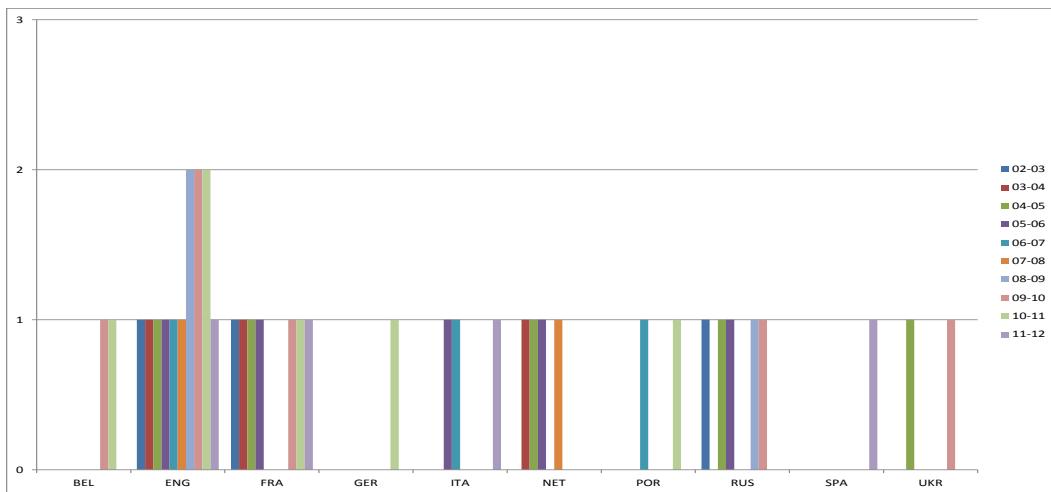
The k -index and the g -index are less sensitive to disruptions, and in some cases they still have the same values despite the occurrence of disruptions. Figures 3(a) and 3(b) show the increase of the k -index and the g -index respectively in the realized schedule compared to the initial schedule. In general, the impact on the g -index is less profound than the impact on the k -index; for both measures, the increase rarely exceeds one. England experiences a deterioration of the balancedness in each season as the competition develops, and as such, never benefits from the high-quality ($k = g = 1$) initial schedules of some seasons. Spain, Germany, and Ukraine on the other hand, have the most robust schedules.

5. Conclusions

This paper studies the robustness of soccer schedules of main European leagues. We measure the quality of a schedule by looking at the total number of breaks and the balancedness of the home advantage (through the g -index and the k -index). We establish theoretical relations among these measures, and we analyze the difference between the values of each of these measures for the initial and the realized schedules. An experimental study shows that disruptions are in general quite rare (less than 1.6 % of the scheduled matches are disrupted), however, almost 90% of the seasons are disrupted at least once. Disruptions have a profound impact on the quality of the realized schedule: in almost all cases disruptions (drastically) increased the number of breaks, and in half of the cases, the balancedness of the home advantage worsened as well. Hence, we conclude by



(a) Difference in k-index between realized and initial schedule



(b) Difference in g-index between realized and initial schedule

Figure 3: Effect of disruptions on balancedness of the home-advantage

stating that soccer schedules in Europe are not robust.

References

- [1] V.G. Adlakha and V.G. Kulkarni. A classified bibliography of research on stochastic PERT networks: 1966-1987. *INFOR*, 27:272–296, 1989.

- [2] C. Artigues, R. Leus, and F. Talla Nobibon. Robust optimization for resource-constrained project scheduling with uncertain activity durations. *Flexible Services and Manufacturing*, 25:175–205, 2013.
- [3] T. Assavapokee, M.J. Realff, and J.C. Ammons. A new min-max regret robust optimization approach for interval data uncertainty. *Journal of Optimization Theory and Applications*, 137:297–316, 2008.
- [4] T. Assavapokee, M.J. Realff, J.C. Ammons, and I.H. Hong. Scenario relaxation algorithm for finite scenario-based min-max regret and min-max relative regret robust optimization. *Computers & Operations Research*, 35:2093–2102, 2008.
- [5] I. Averbakh. Minmax regret solutions for minimax optimization problems with uncertainty. *Operations Research Letters*, 27:57–65, 2000.
- [6] I. Averbakh and V. Lebedev. Interval data minmax regret network optimization problems. *Discrete Applied Mathematics*, 138:289–301, 2004.
- [7] T. Bartsch, A. Drexl, and S. Kroger. Scheduling the professional soccer leagues of Austria and Germany. *Computers and Operations Research*, 33(7):1907–1937, 2006.
- [8] D. Bertsimas and M. Sim. Robust discrete optimization and network flows. *Mathematical Programming, Series B*, 98:49–71, 2003.
- [9] R.L. Daniels and P. Kouvelis. Robust scheduling to hedge against processing time uncertainty in single-stage production. *Management Science*, 41:363–376, 1995.
- [10] D. De Werra. Geography, games and graphs. *Discrete Applied Mathematics*, 2(4):327–337, 1980.
- [11] D. De Werra. Scheduling in sports. In P. Hansen, editor, *Studies on Graphs and Discrete Programming*, volume 11 of *Annals of Discrete Mathematics*, pages 381–395. North-Holland, Amsterdam, 1981.
- [12] F. Della Croce and D. Oliveri. Scheduling the Italian Football League: an ILP-based approach. *Computers and Operations Research*, 33(7):1963–1974, 2006.
- [13] A. Drexl and S. Knust. Sports league scheduling: Graph- and resource-based models. *Omega*, 35:465–471, 2007.
- [14] S.E. Elmaghraby. *Activity Networks: Project Planning and Control by Network Models*. Wiley, 1977.
- [15] D. Forrest and R. Simmons. New issues in attendance demand: The case of the English football league. *Journal of Sports Economics*, 7(3):247–266, 2006.
- [16] S. French. *Decision Theory. An Introduction to the Mathematics of Rationality*. Ellis Horwood Limited, 1988.
- [17] D. Froncek and M. Meszka. Round robin tournaments with one bye and no breaks in home-away patterns are unique. In G. Kendall, E. Burke, S. Petrovic, and Gendreau M., editors, *Selected papers from the 1st Multidisciplinary International Conference on Scheduling: Theory and Applications (MISTA)*, pages 331–340. Springer, 2005.

- [18] D. Goossens and F.C.R. Spieksma. Scheduling the Belgian soccer league. *Interfaces*, 39(2):109–118, 2009.
- [19] D. Goossens and F.C.R. Spieksma. Soccer schedules in Europe: an overview. *Journal of Scheduling*, 15(5):641–651, 2012.
- [20] M.D. Hausken, H. Andersson, K. Fagerholt, and T. Flatberg. Scheduling the Norwegian football league. *International Transactions in Operational Research*, 20(1):59–77, 2013.
- [21] L. Hurwicz. Optimality criteria for decision making under ignorance, 1951. Cowles Commission Discussion Paper no. 370.
- [22] T. Jørgensen. *Project scheduling as a stochastic dynamic decision problem*. PhD thesis, Norwegian University of Science and Technology, Trondheim, Norway, 1999.
- [23] G. Kendall. Scheduling english football fixtures over holiday periods. *Journal of Operational Research Society*, 59(6):743–755, 2008.
- [24] G. Kendall, S. Knust, C.C. Ribeiro, and S. Urrutia. Scheduling in sports: An annotated bibliography. *Computers and Operations Research*, 37:1–19, 2010.
- [25] F.H. Knight. *Risk, Uncertainty, and Profit*. Houghton Mifflin, Boston, 1921.
- [26] S. Knust and M. von Thaden. Balanced home-away assignments. *Discrete Optimization*, 3:354–365, 2006.
- [27] P. Kouvelis and G. Yu. *Robust Discrete Optimization and its Applications*. Kluwer Academic Publishers, Norwell, MA, 1997.
- [28] V.G. Kulkarni and V.G. Adlakha. Markov and Markov-regenerative PERT networks. *Operations Research*, 34:769–781, 1986.
- [29] V. Lebedev and I. Averbakh. Complexity of minimizing the total flow time with interval data and minmax regret criterion. *Discrete Applied Mathematics*, 154:2167–2177, 2006.
- [30] C.H. Loch, A. De Meyer, and M.T. Pich. *A New Approach to Managing High Uncertainty and Risk in Projects*. Wiley, 2006.
- [31] A. Ludwig, R.H. Möhring, and F. Stork. A computational study on bounding the makespan distribution in stochastic project networks. *Annals of Operations Research*, 102:49–64, 2001.
- [32] R. Montemanni. A mixed integer programming formulation for a single machine robust scheduling with interval data. *Journal of Mathematical Modelling and Algorithms*, 6:287–296, 2007.
- [33] K. Nurmi, D. Goossens, T. Bartsch, F. Bonomo, D. Briskorn, G. Duran, J. Kyngäs, C.C. Ribeiro, F. Spieksma, S. Urrutia, and R. Wolf-Yadlin. A framework for a highly constrained sports scheduling problem. *IAENG Transactions on engineering technologies*, 5:14–28, 2010.
- [34] R.V. Rasmussen. Scheduling a triple round robin tournament for the best Danish soccer league. *European Journal of Operational Research*, 185:795–810, 2008.
- [35] R.V. Rasmussen and M.A. Trick. Round robin scheduling – a survey. *European Journal of Operational Research*, 188:617–636, 2008.

- [36] J. Rosenhead, M. Elton, and S.K. Gupta. Robustness and optimality as criteria for strategic decisions. *Operations Research Quarterly*, 23:413–431, 1972.
- [37] L.J. Savage. The theory of statistical decision. *Journal of the American Statistical Association*, 46:55–67, 1951.
- [38] J.A.M. Schreuder. Combinatorial aspects of construction of competition Dutch Professional Football Leagues. *Discrete Applied Mathematics*, 35(3):301–312, 1992.
- [39] F. Talla Nobibon and R. Leus. Robust maximum weighted independent set problems. *Optimization Letters*, 2013. doi 10.1007/s11590-012-0563-8 (to appear).
- [40] A. Wald. *Statistical Decision Functions*. John Wiley, 1950.

Probability calculation for tournament format of the 2013 World Baseball Classic

N. Hirotsu*

*Graduate School of Health and Sports Science, Juntendo University, Inzai, Chiba, Japan :
n_hirotsu@sakura.juntendo.ac.jp

Abstract. The third World Baseball Classic (WBC) was held in March 2013. In this tournament, 16 teams play in Round 1 under a round-robin (RR) format and 8 teams which advanced to Round 2 play under a modified double-elimination (MDE) format. This 2013 WBC format is compared with the formats such as the past two WBCs held in 2006 and 2009, from the aspect of the probability of winning the tournament and the probability distribution of the number of games played by the same teams. We make the comparison by changing the relative strength of teams, and demonstrate the difference between the tournament formats.

1. Introduction

The 2013 World Baseball Classic (WBC) is an international baseball competition. This main tournament was held in March 2013, and the Dominican Republic won the tournament. The basic structure of the main tournament consists of three rounds. In Round 1, 16 teams were divided into 4 pools, each of which consists of 4 teams, and competed in each pool. The 8 teams of the top two teams in each pool advanced to Round 2. The 4 teams of the top two teams in each pool of Round 2 advanced to single-elimination in Finals.

This basic structure has not been changed since the first WBC in held 2006, but the tournament format of each pool is different between the WBCs. In the 2006 WBC, a round-robin (RR) format was employed in Rounds 1 and 2, but ties occurred to decide the top two teams in some pools. That is, 3 teams resulted in the same won-loss record in 2 out of 6 pools of Rounds 1 and 2. If ties occur, the top two teams allowing the fewest runs per nine innings in head-to-head games between the tied teams were qualified to the next round, but this tie breaking rule was controversial.

In the 2009 WBC, the RR format was replaced by a modified double-elimination (MDE) format. However, the same two teams faced each other quite often in the tournament under the MDE format, such that Japan faced South Korea in 5 out of 9 games which Japan played throughout the tournament. Further, Game 6 in Pool 1 and 2 seems to a throwaway game because Game 6 is played between the top two teams which have already been decided to advance to the next round. Actually, the coach of South Korea commented that good pitchers were not introduced to Game 6 intentionally.

In the 2013 WBC, the MDE format was replaced by a RR format only for Round 1, together with a little modification of the tie breaking rule. In the process of discussion to decide the format of the 2013 tournament, the format deleting Game 6 from the MDE format seemed to be suggested. We here denote it as modified modified-double-elimination format (MMDE), and also look at it for the comparison of formats.

In this paper, we present the formulation for probability calculation on the above formats, and compare between the 2013 tournament format and other 3 formats (the past two WBCs and the case of employing the MMDE format), from the aspect of the probability of winning the tournament and the probability distribution of the number of games played by the same teams. We make the comparison by changing the relative strength of teams, using the Bradley-Terry model for setting the probability of winning a game between two teams, and demonstrate the difference between these tournament formats. We focus on the main tournament in this paper, although a qualifying round played by 12 teams was first introduced before the main tournament in the 2013 WBC.

By the way, a variety of sport tournament formats is well studied by several researchers. In terms of the comparison between a single-elimination and a RR format (David, 1959; Appleton, 1995; McGarry and Schutz, 1997), there is a following tendency: Under a RR format, the stronger team is more likely to win the tournament, but the number of total games in the tournament is larger than single-elimination, in which the stronger team is less likely to win the tournament than a RR format. In terms of a double-elimination (DE) format, McGarry and Schutz (1997) indicate that the DE format has a good balanced feature in the selection of stronger teams with moderate total number of games.

With regards to probability calculation for the DE format, Glenn (1960) and Ladwig and Schwertman (1992) present a calculation method, although they are not considered the multi-round structure such as the 2009 WBC tournament. McGarry and Schutz (1997) analyzed a variety of tournament formats including the DE format, but they do not show the formulation of probability calculation because they use Monte Carlo simulation. We here apply the method of Ladwig and Schwertman (1992) to calculate the tournament format of WBC which includes the MDE format.

In this paper, we describe the formats employed in the WBCs in Section 2, and then we present the formulation for probability calculation on these formats in Section 3. After showing how to set a probability of winning a game in Section 4, we demonstrate the difference between these tournament formats in Section 5 and conclude in Section 6.

2. Tournament format of the WBCs

In this section, we represent the format of the main tournament of the WBCs. The format consists of three rounds. In Round 1, 16 teams are divided into 4 pools (A, B, C and D) each of which consists of 4 teams, and the top two teams in each pool advances to Round 2. In Round 2, the 4 teams from Pools A and B and the 4 teams from Pools C and D compete in Pools 1 and 2, respectively. The top two teams in each pool of Round 2 advances a single-elimination in Finals. The 4 teams cross over for the semifinals, with the winner of each pool playing against the runner-up from the other pool. Figure 1 shows the main structure of the tournament, which is consistent between the WBCs, but the format in each pool is different between the WBCs.

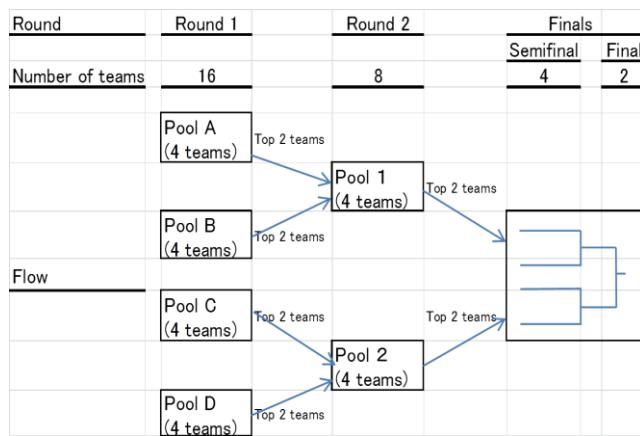
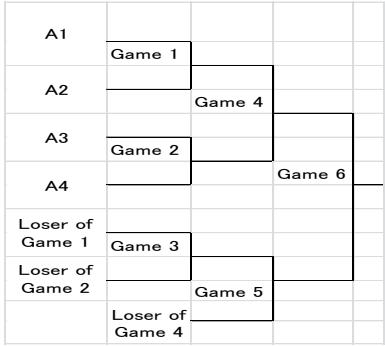


Figure 1. Main structure of the tournament format of the WBCs.

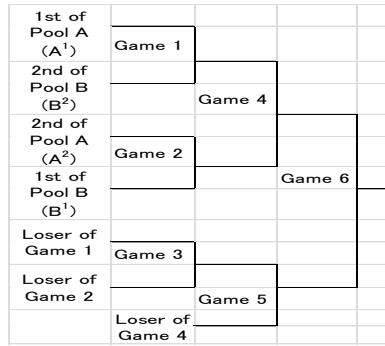
In the 2006 tournament, the RR format was employed in Rounds 1 and 2. That is, each team plays other three teams in a pool once. Teams are ranked by the winning percentage in each round, and the top two teams in each pool advance to the next round.

In the 2009 tournament, the RR format was replaced by the MDE format. This MDE format is illustrated in Figure 2. Figure 2 (a) and (b) shows the draw of Pool A of Round 1 and the draw of Pool 1 of Round 2, respectively, as an example. Figure 2 (c) shows the draw of the Finals.

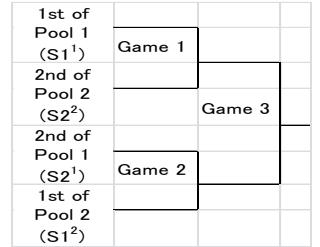
In the 2013 tournament, the MDE format was replaced by the RR format only for Round 1. Together with this replacement to the RR format, the tie breaking rule was a little modified. Roughly speaking, not only allowing runs per nine innings but also scoring runs per nine innings in head-to-head games between the tied teams are evaluated for advancing to Round 2.



(a) Pool A of Round 1



(b) Pool 1 of Round 2



(c) Finals

Figure 2. Modified double-elimination format (Rounds 1 and 2) and single-elimination (Finals) in WBC 2009

3. Formulation for probability calculation

In this section, we present a formulation for the probability calculation in terms of the RR format and the MDE format. In this paper, we assume that the result of the game is mutually independent throughout the tournament.

3.1 Round-robin format within a round

We first look at a formulation for the probability calculation in terms of the RR format of a pool within a round. Let $P(i,j)$ be the probability of team i winning a game against team j . Using this notation, for example, an event that "Team A1 wins 3 games, Team A2 wins 2 and loses 1 game, Team A3 wins 1 and loses 2 games, Team A4 loses 3 games" occurs in the probability of

$$P(A1,A2)P(A1,A3)P(A1,A4)P(A2,A3)P(A2,A4)P(A3,A4). \quad (1)$$

Here, the total number of possible win-loss events in a pool is $2^6=64$. These events are categorized into 4 patterns shown in Table 1. We can calculate the probability of each of the 4 patterns occurring by just counting the number of events in each category.

Table 1. Win-loss pattern of the round-robin format

		Win-loss pattern			
		1	2	3	4
1st	3-0	3-0	2-1	2-1	
	2-1	1-2	2-1	2-1	
	1-2	1-2	2-1	1-2	
	0-3	1-2	0-3	1-2	

3.2 Modified double-elimination format within a round

We move on to the formulation for the probability calculation in terms of the MDE format of a pool within a round. Let $P_k(i,j)$ be the probability of team i winning a game against team j in game k . To obtain this probability, we look at A1 in Figure 2(a). As there are 6 games in the draw of Figure 2(a), the total number of possible win-loss events is $2^6=64$. As there are 4 teams in a pool, a quarter of the 64 events corresponds to the events that A1 is ranked 1st in this pool. That is, there are 16 ($=64/4$) win-loss events for A1 to be ranked as the 1st in Pool A of Round 1. We arrange these 16 events according to the 2nd ranked team in this pool, as shown in Table 2. W_k and L_k in Table 2 represent a win and a loss of the regarding team in game k , respectively. For example, the win-loss event that A1 is the 1st and A2 is the 2nd in this pool appears in No.1-4 rows in Table 2, and if A1 achieves 3 wins ($W_1W_4W_6$), then A2 has to result in a 2 wins and 2 losses ($L_1W_3W_5L_6$) as shown in No.1-2 rows. In this event, there are 2 cases which correspond to the different result of Game 2, that is, either A3 or A4 wins in Game 2. Looking at the event $W_1W_4W_6$ and A3 wins in Game 2, this event occurs in probability of

$$P_{IT}(A1,A2) = P_1(A1,A2)P_2(A3,A4)P_3(A2,A3)P_4(A1,A3)P_5(A2,A3)P_6(A1,A2) \quad (2)$$

$P_{IT}(A1,A2)_2$ is also calculated by looking at the event $W_1W_4W_6$ and A4 wins in Game 2. In the case for A1 to result in the event $L_1W_3W_5W_6$, we obtain $P_{IT}(A1,A2)_3$ and $P_{IT}(A1,A2)_4$ corresponding to the case that A3 and A4 wins in Game 2, respectively. By summing them up, the probability that A1 and A2 become the 1st and the 2nd in Round 1, respectively, as

$$P_{IT}(A1,A2) = \sum_m P_{IT}(A1,A2)_m \quad (3)$$

(The subscription m corresponds to the subscription of the notation appeared in the column "Probability" in Table 2.) In the same manner, not only $P_{IT}(A1,A3), P_{IT}(A1,A4)$ but also other teams such as A2 to be the 1st in this pool can be calculated.

Here, note that the 4 terms should be randomly assigned in the draw in Pool A. So, the above calculations are conducted for the 3 different assignments, and then averaged out. That is, we consider the 3 different draws for assigning 4 teams in the MDE format in a pool such that A1 faces A2, A3 or A4 as the first game of A1, and the 3 different draws occur in 1/3 each.

Table 2. Probability of a win-loss event occurring and the number of games between two teams when A1 becomes the 1st in the draw shown in Figure 2 (a)

No.	Win-loss event				Probability	Number of games						Reamrk
	1st	2nd	A1	A2		A1-A2	A1-A3	A1-A4	A2-A3	A2-A4	A3-A4	
1	A1	$W_1W_4W_6$		$L_1W_3W_5L_6$	$P_{1T}(A1,A2)_1$	2	1	0	1	1	1	A3 wins in Game 2
					$P_{1T}(A1,A2)_2$	2	0	1	1	1	1	
3		$L_1W_3W_5W_6$	A2	$W_1W_4L_6$	$P_{1T}(A1,A2)_3$	2	1	1	1	0	1	A3 wins in Game 2
					$P_{1T}(A1,A2)_4$	2	1	1	0	1	1	
5	A1	$W_1W_4W_6$	A3	$W_2L_4W_5L_6$	$P_{1T}(A1,A3)_1$	1	2	0	1	1	1	A2 wins in Game 3
					$P_{1T}(A1,A3)_2$	1	2	0	0	1	2	
7			A3	$L_2W_3W_5L_6$	$P_{1T}(A1,A3)_3$	1	1	1	1	0	2	
					$P_{1T}(A1,A3)_4$	2	2	0	0	1	1	
8		$W_1L_4W_5W_6$	A3	$W_2W_4L_6$	$P_{1T}(A1,A3)_5$	1	2	1	0	1	1	A4 wins in Game 3
					$P_{1T}(A1,A3)_6$	2	1	1	1	0	1	
10		$L_1W_3W_5W_6$	A3	$L_2W_3W_5L_6$	$P_{1T}(A1,A3)_7$	1	0	2	1	1	1	
					$P_{1T}(A1,A3)_8$	2	1	1	0	1	2	
11	A1	$W_1W_4W_6$	A4	$W_2L_4W_5L_6$	$P_{1T}(A1,A4)_1$	1	0	2	1	1	1	A2 wins in Game 3
					$P_{1T}(A1,A4)_2$	1	0	2	1	0	2	
13			A4	$L_2W_3W_5L_6$	$P_{1T}(A1,A4)_3$	1	1	1	0	1	2	
					$P_{1T}(A1,A4)_4$	2	0	2	1	0	1	
14		$W_1L_4W_5W_6$	A4	$W_2W_4L_6$	$P_{1T}(A1,A4)_5$	1	1	2	1	0	1	A2 wins in Game 3
					$P_{1T}(A1,A4)_6$	2	1	1	0	1	1	
16		$L_1W_3W_5W_6$	A4	$L_2W_3W_5L_6$	$P_{1T}(A1,A4)_7$	1	0	2	1	1	1	
					$P_{1T}(A1,A4)_8$	2	1	1	0	1	1	

As shown in Table 2, the number of games between two teams in the pool is fixed, according to its win-loss event. For example, if A1 is the 1st and A2 is the 2nd with the event of $W_1W_4W_6$ and $L_1W_3W_5L_6$, A1 has to play A2 twice in the pool.

Using the number of games played by two teams, we can calculate the probability distribution of the number of games between two teams. Here, let $Q_I(i,j,n_I|A1,A2)$ be the probability that team i plays against team j ($i, j \in \{A1, A2, A3, A4\}$) n_I times in Round 1, under the condition that A1 and A2 results in the 1st and the 2nd in the pool, respectively. We calculate this using the probabilities shown in Table 2 such that

$$Q_I(A1,A3,1|A1,A2) = P_{IT}(A1,A2)_1 + P_{IT}(A1,A2)_3 + P_{IT}(A1,A2)_4 \quad (4)$$

Here, (4) corresponds to the probability that A1 and A3 play once in the pool, and this probability can be obtained by summing up the probabilities for resulting in the win-loss events corresponding to the number of games "1" appearing in No.1-4 rows in the column of "A1-A3" in Table 2.

3.3 Calculation between the rounds

So far, we looked at the format in a round, actually in Round 1. In order to formulate the multi-round structure, we should identify the 8 teams advanced from Round 1. Here, we denote the 1st and the 2nd team in Pool A and Pool B as $A^1, A^2 \in \{A1, A2, A3, A4\}$ and $B^1, B^2 \in \{B1, B2, B3, B4\}$, respectively.

Using this notation, in terms of the RR format in Round 2, we can calculate the probability of each of the 4 patterns occurring shown in Table 1 just by looking at A^1 , A^2 , B^1 and B^2 . However, in terms of the MDE format in Round 2, the calculation becomes a little complicated because the 4 teams are assigned into the draw of Pool 1 according to the rank in pool A and B as shown in Figure 2(b). As A^1 and A^2 are chosen from the 4 teams in Pool A in $4 \times 3 = 12$ different ways, and so as to B^1 and B^2 , there are totally $12 \times 12 = 144$ different way for the 4 teams to be assigned in the draw of Pool 1. By calculating each of these 144 different ways, we can obtain such probability that A1 advances to Finals from Pool 1.

In Finals, we should calculate in the single-elimination format shown in Figure 2(c) using 4 teams advanced from Round 2, both in the RR and MDE formats. Here, we denote the 1st and the 2nd team in Pool 1 and Pool 2 as $S1^1, S1^2 \in \{A1, A2, A3, A4, B1, B2, B3, B4\}$ and $S2^1, S2^2 \in \{C1, C2, C3, C4, D1, D2, D3, D4\}$, respectively. Using this notation, the draw of semifinal is fixed according to the rank in Pools 1 and 2. As $S1^1, S1^2$ are chosen from the 8 teams in $8 \times 7 = 56$ different ways, and so as to $S2^1, S2^2$, there are totally $56 \times 56 = 3136$ different ways for the 8 teams to be assigned in the draw of semifinal. By calculating for these 3136 different ways, we can obtain such probability that A1 wins the tournament.

In order to calculate the probability of the total number of games between team i and j occurring throughout the tournament in the MDE format, we need to sum up the number of games from Round 1 to Finals. In practice, we refer to Table 3, which shows the patterns of the number of games between two teams from Round 1 to Finals, and calculate the probability of the total number of games occurring.

Table 3. Patterns of the number of games between two teams in the MDE format

	Number of games between two teams																								
Round 1	n_1	2	2	2	2	1	2	2	1	1	1	-	2	2	1	1	-	-	-	-	0	-	-	-	
Round 2	n_2	2	2	2	1	2	1	1	2	2	1	2	0	-	1	1	2	1	2	0	-	1	1	-	-
Finals	n_F	1	0	-	1	1	0	-	0	-	1	1	-	-	0	-	0	1	-	-	0	-	0	-	
Total	n	5	4	4	4	4	3	3	3	3	3	3	2	2	2	2	2	2	1	1	1	1	0	0	0

As shown in Table 3, in order for two teams to face each other 5 times they should face twice in Round 1, and twice in Round 2, and once in Finals. We here denote this as "2 2 1". In order to face 4 times, there are 4 patterns such that "2 2 0", "2 2 -", "2 1 1" and "1 2 1", as shown in Table 3. The notation " - " denotes "not-faced" because of being assigned in a different pool, or one of two teams failing to advance to the round. The probability of each pattern occurring is calculated by taking account of the conditional probability to advance to Round 2 or Finals. For example, in order to calculate the probability that "2 2 1" occurs, we look at the following event: A1 and A2 face each other twice ($n_1=2$) in Pool A under the condition that A1 and A2 become the 1st and 2nd of the pool in Round 1. A1 and A2 then face each other twice ($n_2=2$) in Pool 1 under the condition that A1 and A2 become the 1st and 2nd of Pool 1 in Round 2. A1 and A2 finally face each other once ($n_F=1$) in Finals. This event occurs in probability which is given by the product of the above probabilities, that is,

$$Q_1(A1, A2, 2 / A1, A2) \cdot Q_2(A1, A2, 2 / A1, A2) \cdot Q_F(A1, A2, 1) \quad (5)$$

As there are 4 different rankings for A1 and A2 in advancing to the next round, we can obtain the probability that A1 and A2 faces 5 times as the sum of them, that is,

$$\begin{aligned} Q(A1, A2, 5) = & Q_1(A1, A2, 2 / A1, A2) \cdot Q_2(A1, A2, 2 / A1, A2) \cdot Q_F(A1, A2, 1) \\ & + Q_1(A1, A2, 2 / A2, A1) \cdot Q_2(A1, A2, 2 / A1, A2) \cdot Q_F(A1, A2, 1) \\ & + Q_1(A1, A2, 2 / A1, A2) \cdot Q_2(A1, A2, 2 / A2, A1) \cdot Q_F(A1, A2, 1) \\ & + Q_1(A1, A2, 2 / A2, A1) \cdot Q_2(A1, A2, 2 / A2, A1) \cdot Q_F(A1, A2, 1) \end{aligned} \quad (6)$$

By conducting the similar calculation for all patterns shown in Table 3, we can obtain the probability distribution of the number of games between two teams throughout the tournament. Further, we can obtain the probability distribution of the number of games such as between A1 and other team $i \in$

$\{A2, A3, A4, B1, B2, B3, B4, C1, C2, C3, C4, D1, D2, D3, D4\}$ throughout the tournament, under the condition that A1 wins the tournament, by looking at the probability $\Pr \{A1 \text{ faces } i \text{ times} | A1 \text{ wins the tournament}\}$ in detail.

4. Setting of strength of teams

Until now, we present the formulation of probability calculation, but we have not referred to concrete values of the probability such as $P_k(i,j)$. In actual calculation we need to set these probabilities as concrete values. In this paper, the Bradley-Terry model is used for setting the probability of team i winning the game against team j . That is, we calculate

$$P_k(i,j) = \pi_i / (\pi_i + \pi_j) = 1 / (1 + \pi_j / \pi_i), \quad (7)$$

where π_i and π_j represent the strength of team i and j , respectively. Here, we introduce the ratio r ($= \pi_{A1} / \pi_{A2} = \pi_{A2} / \pi_{A3} = \pi_{A3} / \pi_{A4}$) by arranging the relative strengths in descending order such that A1 is stronger than A2, and A2 is stronger than A3, etc. We also introduce the relative overall strength among Pools A, B, C, D as ratio s ($= \pi_A / \pi_C = \pi_C / \pi_B = \pi_B / \pi_D$) by arranging these strengths in descending order such that Pool A is relatively stronger than Pool C, and Pool C is stronger than Pool B, etc.

5. Calculation results

By changing the above ratios r and s , we have calculated such the probability that A1 wins the tournament, or the probability distribution of the number of games between A1 and others. In this section, we demonstrate the calculation results in terms of difference between the tournament formats. We also present the detail about the occurrence of ties in the RR format and the number of games between two teams in the MDE format.

5.1 Comparison between the tournament formats

Table 4 shows the comparison between the 2013 tournament format (indicated by RR&MDE) and other 3 formats (the past two classics and the MMDE format, indicated by RR², MDE² and MMDE², respectively) from the aspect of the probability of winning the tournament. In Table 4, we also show the total number of games in the tournament, the maximum number of games played by same two teams, and whether ties are possible or not.

For easy understanding, A1 is set to be the strongest team in this calculation. As shown in Table 4, if the all teams have same strength ($r=s=1$) the probability of winning the tournament is same for the all teams as 0.0625 (=1/16). However, we can see that a small difference appears between the formats according to the change of relative strength of the teams.

From Table 4, we can infer that

- MDE² is more efficient than RR².
- MDE² is more efficient than RR&MDE, but RR&MDE is better than MDE² in terms of the maximum number of games played by same two teams.
- MMDE² format seems to be more efficient than MDE². (MMDE² achieves almost the same probability of A1 winning as MDE², but the total number of games of MMDE² is 6 games less than that of MDE².)
- MMDE² is also better than MDE² in terms of the maximum number of games played by same two teams.

According to this calculation result, MMDE² seems to be the best format among them if the total number of games (33) is acceptable for conducting the WBC. In any cases, however, the probability of A1 winning the tournament is not much different between these formats. So, other factors such as the total number of the games or the maximum number of games played by same teams may be more influential than the probability of winning in terms of the decision of formats.

Table 4. Comparison between the tournament formats

Abbreviation		RR^2	MDE^2	$RR\&MDE$	$MMDE^2$
WBC No.		1	2	3	—
Format	Round 1	RR	MDE	RR	MMDE
	Round 2	RR	MDE	MDE	MMDE
Total number of games	39	39	39	39	33
Max. number of played by same teams	3	5	4	3	3
Ties	Yes	No	Yes	No	
Probability of A1 winning the tournament					
r	s				
1	1	0.0625	0.0625	0.0625	0.0625
2	1	0.2032	0.2049	0.2031	0.2041
1	1.2	0.0924	0.0973	0.0973	0.0972
2	1.2	0.2786	0.2912	0.2872	0.2897

5.2 Calculation result on occurrence of ties in round-robin format

In terms of the detail about the occurrence of ties in the RR format, we here illustrate the calculation result of how often the tie will occur in the RR format. Figure 3 shows a relationship between ratio r and the probability of ties occurring with the same win-loss record in a pool, under the condition that ratio s is set to 1. As shown in Figure 3, the probability of ties occurring in probability of 0.25, when the case that the all teams have the same strength ($r=1$). This equals to $16/64 (=0.25)$.

This probability decreases according to the increase of r , but even though r is around 3, the probability is still around 0.15. So, the fact that ties occurred 2 out of 6 pools in WBC 2006 is not curious. So, the tie breaking rule is important in the selection of top two teams in a pool.

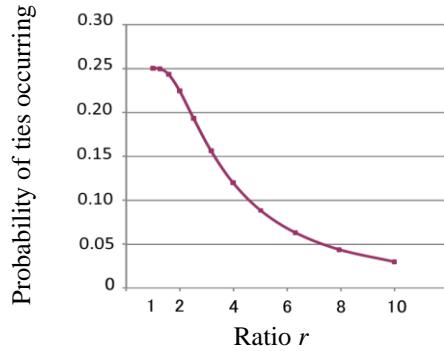


Figure 3. Relationship between ratio r and the probability of ties occurring in a pool in a round-robin format
(Ratio s is set to 1)

5.3 Calculation result on the number of games between two teams in MDE^2

In terms of the detail of the calculation result on the number of games between two teams, we compare the probability distribution of the number of games between A1 and other team throughout the tournament, under the condition that A1 wins the tournament. Here, all teams are assumed to have the same strength ($r=s=1$). As shown in Table 5, in the case of MDE^2 , A1 faces another team of Pool A (e.g. A1 faces A2) 5 times in probability of 0.031. As the event of A1 facing A2 5 times and the event of A1 facing A3 5 times are mutually exclusive, for example, the probability that A1 faces A2, A3 or A4 5 times is $0.031 \times 3 = 0.093$. In other words, looking at the winner of the tournament, the event that the winner faces another team in the same pool in Round 1 5 times in MDE^2 will occur in the probability of 0.093 (around one tenth). So, the event of two teams playing 5 times such that Japan faced South Korea 5 times in the 2009 WBC does not seem to be a rare event. If the two teams are stronger than others, it will occur more likely.

In the case of RR&MDE employed in the 2013 WBC, A1 faces another team in Pool A up to 4 times in probability of 0.042, as shown in Table 4. In the case of $MMDE^2$, A1 faces another team in Pool A up to 3 times, although the probability of not-facing is not much different from the case of MDE^2 .

Table 5. Probability distribution of the number of games between A1 and other team under the condition that A1 wins the tournament. (All teams are assumed to have the same strength ($r=s=1$))

Abbreviation	MDE ²			RR & MDE			MMDE ²		
WBC No.	2			3			—		
Number of games	A2,A3,A4	B1,B2,B3,B4	C1,C2,C3,C4, D1,D2,D3,D4	A2,A3,A4	B1,B2,B3,B4	C1,C2,C3,C4, D1,D2,D3,D4	A2,A3,A4	B1,B2,B3,B4	C1,C2,C3,C4, D1,D2,D3,D4
5	0.031			—			—		
4	0.057			0.042			—		
3	0.125	0.063		0.062	0.063		0.031	—	
2	0.182	0.141		0.146	0.141		0.234	0.125	
1	0.438	0.234	0.188	0.750	0.234	0.188	0.542	0.297	0.188
0	0.167	0.562	0.812	—	0.562	0.812	0.193	0.578	0.812
Total	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

By the way, the probability of 0.031 that A1 faces A2 5 times can be calculated easily as follows. First, in the case that all teams have the same strength, a win-loss event in 6 games expressed by (2) occurs in probability of $(1/2)^6$. And the probability that A1 faces A2 twice in Round 1 is $Q_1(A1, A2, 2/A1, A2) = (1/2)^6 \times 4 = (1/2)^4$, because there are 4 events corresponding to No.1 – 4 in Table 2. We also obtain the probability that A1 faces A2 twice in Round 2 such that $Q_2(A1, A2, 2/A1, A2) = (1/2)^4$. In Finals, the event that A1 faces A2 once should occur after both A1 and A2 advance to Finals, the probability that A1 faces A2 once in Finals is given by $Q_F(A1, A2, 1) = (1/2)^2$. Considering the rank of A1 and A2 in advancing to the next round, we obtain $Q(A1, A2, 5) = (1/2)^4 \cdot (1/2)^4 \cdot (1/2)^2 \times 4 = (1/2)^8 = 0.0039$, following the calculation of (6). We finally obtain $(0.0039/2)/(1/16) = 0.031$, because the probability of A1 winning the tournament is 1/16 and A1 wins the tournament in the probability of 1/2 under the condition that A1 and A2 faces 5 times.

The probability of 0.188 that A1 faces a team in Pool C or D once shown in Table 5 can also be calculated easily as follows. Looking at C1, for example, C1 advances to Finals in the probability of 2/8. As A1 faces C1 in semifinal in the probability of 1/2 and in the final in 1/4 (= the probability that A1 does face C1 not in the semifinal but in the final), we obtain $2/8 \times (1/2+1/4) = 3/16 = 0.188$.

6. Conclusions

In this paper, we have presented the formulation for probability calculation on the main tournament format of the 2013 WBC, and compare the 2013 tournament format to the past two tournament formats and the case of employing the MMDE format, from the aspect of the probability of winning the tournament and the probability distribution of the number of games played by the same teams, by changing the relative strength of teams. We have demonstrated the difference between these tournament formats.

As the results of calculation, a small difference appears between the formats according to the change of relative strength of the team, such that MDE² seems to be more efficient than RR², and MDE² seems to be more efficient than RR&MDE, but RR&MDE looks better than MDE² in terms of the maximum number of game faced by each other. In general, MMDE² seems to be the best format among them if the modification that the total number of games becomes 33 is acceptable for conducting the WBC, although other factors will be influenced to the decision of the formats in reality.

We also have discussed about the occurrence of ties in the RR format, and the number of games between two teams in MDE². In terms of occurrence of ties in the RR format, the fact that ties occurred 2 out of 6 pools in WBC 2006, is not curious from the aspect of probability calculation. In terms of the number of games between two teams in MDE², the event of the same two teams in the same pool in Round 1 faces 5 times in MDE² employed in WBC 2009 will occur in the probability of around one tenth, and does not seem to be a rare event.

Acknowledgement

This research is partly supported by Grant-in-aid for Scientific Research (C), #21510159.

References

- Appleton, D.R.(1995) May the best man win? *The Statistician* **44**, 529-538.
- David, H.A.(1959) Tournament and paired comparisons. *Biometrika*, **46**, 139-149.
- Glenn, W.A.(1960) A comparison of the effectiveness of tournaments. *Biometrika* **47**, 253-262.
- Knuth, D.E.(1998) *The Art of Computer Programming*, Vol.3,2nd ed. Addison-Wesley.
- Ladwig, J.A. and.Schwertman, N.C. (1992) Using probability and statistics to analyze tournament competitions. *Chance* **5**, 49-35.
- McGarry, T. and Schutz, R.W. (1997) Efficacy of traditional sports tournament structures. *Journal of the Operational Research Society* **48**, 65-74.

Developing an improved tennis ranking system

D. Irons*, S. Buckley* and T. Paulden*

* Atass Sports, Exeter, U.K. Correspondence email: David.Irons@atass-sports.co.uk

Abstract. Sports ranking systems are often viewed as inadequate for judging the quality of the teams or players involved. Meanwhile, statistical models have been shown to produce more accurate ratings for those competitors, based on their ability to forecast future results. However, whilst predictive power is a desirable property of any official ranking system, these systems must also be fair, transparent and insensitive to bias. Additional requirements may also be required, such as promoting major tournaments and deciding seedings. By considering rankings for ATP tennis players, we propose that statistical models can be used to improve the existing ranking system, in such a way that the resulting rankings are fair and usable by the governing body. In many cases, there is a trade-off between predictive power and other desirable properties, and so compromise is required to produce a system that can be implemented successfully.

1. Introduction

Sports ranking systems are designed with several competing objectives in mind. Some of these objectives are commercial - for example the rankings may be designed in a way to incentivise players or teams to appear in certain tournaments (Meyer and Pollard (2012)). At the same time, the rankings can form an integral part of how tournament draws are settled. In this instance, the sporting integrity of the competition demands that the players and fans must believe in the rankings: hence they must accurately reflect player or team ability.

In this paper we look specifically at ranking male tennis players. As well as accurately reflecting ability, a good ranking system in this context should penalize players who refuse to support important tournaments. It should be insensitive to the stage of the season – a clay court specialist whose ability remains constant throughout the year should not jump up the rankings during the clay court season only to fall back after. It should deal fairly with player injuries.

We do not set out to define a ‘best’ ranking system. Rather, we explore what aspects of established statistical models can be used to improve the current ATP system, in such a way as to not diminish important existing features; for example, encouraging players to compete and ensuring insensitivity to the stage of the season. By using statistical models as our start point we obtain alternative treatments of other factors directly affecting match outcome, such as absence through injury, playing surface and match importance. In, for example, the treatment of long term injury, one sees the difficulty of defining a best ranking system: the models maintain a player’s position through an injury period whereas the player will lose ATP ranking points, and hence fall down the official rankings, while not able to compete. Which is better?

A key point of this work is that we do not optimize models on predictive power alone, but rather show how a general modelling approach can be useful for improving a ranking system. Indeed, it is often necessary that some predictive ability is sacrificed to maintain the integrity and comprehensibility of the system. Towards this goal, we choose to use previously published modelling concepts as our starting point so as to not to cloud our main objectives.

There is a wide and varied literature on predicting sporting outcomes. For tennis in particular, there are papers that develop models from the ATP rankings. These typically take win probabilities to be a function of the difference in the respective players’ rankings, or difference in points accumulated (e.g. Boulier and Stekler (1999), Clarke and Dyte (2000)). In some cases, this approach has also been adapted to take account of basic player characteristics, say height or age (e.g. del Corrala and Prieto-Rodriguez (2010)). There are also papers that develop standalone predictive models for tennis (e.g. Glickman (1999), McHale and Morton (2011), Knottenbelt, Spanias and Madurska (2012), Dingle, Knottenbelt and Spanias (2013)), often motivated by ideas prevalent in the modelling of other sports. It is papers of this nature that have the most potential for generating improved rankings, since they provide alternative methodologies for turning match results into player ratings or ranks. They have also been shown to provide a more accurate reflection of player abilities, based on predictive power (see specifically McHale and Morton (2011)).

For this paper, our start point will be the model of McHale and Morton (2011). Their paper is based on Bradley and Terry (1952), which provides methods of analysis when presented with a set of paired comparisons (i.e. past match results in tennis). The nature of sport, specifically the change in the ability of players over time, is taken account of by weighting past results according to how long ago they occurred. This differs from the official ATP rankings that use a rolling 12 month window, with matches given varying weight (i.e. points) based only on the stage and cache of the tournament.

In Section 2 we introduce the data, the ATP ranking system and the McHale and Morton model in more detail. We also consider a series of model variants, which maintain certain aspects of the existing ATP system; this allows us to better address which features of the existing system can be replaced or improved. Section 3 compares the performance of the official and model-based ranking systems. One of the metrics is a predictive metric, but we also consider how sensitive the different systems are at a match, tournament and seasonal level. In Section 4 we discuss practical implications of the suggested models. We contrast the treatment of retirements and injuries in the models against the ATP rankings and discuss how surface-specific seeding could be implemented in practice. The paper concludes with a discussion.

2. Data and methodology

ATP data can be obtained from www.tennis-data.co.uk, which gives score, official player ranks, tournament and round for over 34000 matches between 2000 and 2013. We also get a description of the surface (hard, carpet, clay, grass) and tournament importance (equivalent to Grand Slam, World Tour Finals, Masters 1000, ATP 500 and ATP 250). We note here that this data does not have matches for challenger and futures tournaments, or qualifying games in general, which means it under-represents players further down the rankings. However, the coverage of the data for all tournaments equal to or above ATP 250 level is extremely good.

2.1 Using data to mimic ATP rankings

In the raw data, we only have information on the ranks of players during the tournaments they're competing in. So if a player takes a break, through choice or injury, there is a gap in our knowledge of the system.

Therefore we use the ATP points system to mimic the official rankings, based solely on the given data. In brief, players are awarded points for each round of a tournament they reach, with more points awarded for more prestigious tournaments (see Table 1 for details). Points then accumulate over a 12 month rolling window, with only those points from the best 18 tournaments counting. We simplify the 18 tournament rule slightly by dropping the restriction that the top 30 players from the previous year must count all Grand Slams and Masters 1000 tournaments.

Using this approach, we then have a full list of player ranks for every day, regardless of whether or not a player appears in a tournament at the same time. For comparison with other models, we update points daily (within a tournament) rather than at the end of each week. From now on we refer to these ranks as the pseudo-official ranks.

Table 1. Points used in the pseudo-official ranking system. Reaching the relevant stage of a tournament earns the player the number of points shown in the table. W represents points for tournament winner. Brackets correspond to the case where there is an extra round in the tournament. *World Tour Finals use a round robin system to start with and so have a slightly different system: 200 for each round robin match win, plus 400 for a semi-final win, plus 500 for the final win.

	W	F	SF	QF	R16	R32	R64	R128
Grand Slam	2000	1200	720	360	180	90	45	10
World Tour Finals	*1500							
Masters 1000	1000	600	360	180	90	45	10(25)	(10)
ATP 500	500	300	180	90	45	(20)		
ATP 250	250	150	90	45	20	(5)		

2.2 Basic modelling approach

As discussed earlier, the primary focus of this paper is the generation of an improved ranking system, rather than the generation of new models. Therefore, we use existing modelling approaches as our starting point, based on established work by McHale and Morton (2011) and others.

The general modelling approach is maximum likelihood estimation (MLE) of a model (f) to infer ratings θ_i for each player i . For any given day t , we maximise the pseudo-likelihood

$$L(\theta, i = 1, \dots, n) = \prod_{k \in A_t} f(\theta_{i(k)}, \theta_{j(k)}, x_k)^{w(t_k, t)}$$

with respect to player ratings (θ). Here t_k is the day on which match k was played, between players $i(k)$ and $j(k)$, with result x_k . $A_t = \{k : t_k < t\}$ is the set of matches before day t , and w is a function used to weight results dependent on how long ago they occurred. To avoid too much uncertainty for new players, we insist that a player appears 20 times in the previous 2 years before including their matches. These ratings are then simply ordered to assign rankings to all of the players on day t .

The first function, f , considered is one proposed by McHale and Morton (2011), based on a Bradley-Terry model (Bradley and Terry (1952)).

$$f_1 = \frac{\theta_i^{g_i} \theta_j^{g_j}}{(\theta_i + \theta_j)^{g_i + g_j}}.$$

This corresponds to the likelihood of the scoreline (x_k) in games, where g_i and g_j are the number of games won by players i and j in match k . For instance, if player i wins 6-2, 4-6, 6-3, $g_i = 16$ and $g_j = 11$. The likelihood also includes a weight function w , so that recent matches are more relevant than old ones. Most commonly used is an exponential decay of the form

$$w_1 = m_k * e^{\gamma(t - t_k)},$$

where γ is a constant and m_k is a match weight based on tournament importance, round or surface (1 by default).

Since a key part of our analysis is a comparison with the current ranking system, we also adapt these models to maintain some of its existing features. In particular, the current ranking system only uses match outcome (win / loss), and so we also consider a match-win likelihood

$$f_2 = \frac{e^{(\theta_i - \theta_j)}}{1 + e^{(\theta_i - \theta_j)}}.$$

This is a logistic function based on the difference between the ratings of the winning player i and losing player j . The concept of using the logistic function for predicting binary sports outcomes is again something that has been widely considered in the literature, due to the fact its inverse is the link function to the binomial distribution. Another feature is that all points gained in the past year are treated equally (no time decay) and so we also consider

$$w_2 = m_k * e^{\gamma * 365 * |(t - t_k)/365|},$$

where all matches in the past year carry an equal time weight. We note that this approach differs slightly from the official ranking system, since we include matches beyond 12 months ago. However, the concept of using a 12 month rolling window to assess the importance of matches remains.

2.3 Incorporating surface information

Again, we use the approach of McHale and Morton (2011) to consider the additional effect of playing surface. In this study, there is a separate model for each surface $s \in \{\text{hard, clay, grass}\}$, where matches on different surfaces are down-weighted. For a match taking place on surface s , model ‘ s ’ is then used to determine the abilities of the competing players.

For each surface model ‘ s ’, a weight matrix

$$M_{st} = \begin{bmatrix} 1 & 0.25 & 0.5 \\ 0.25 & 1 & 0.01 \\ 0.5 & 0.01 & 1 \end{bmatrix}$$

is used to determine the weight m_k , of match k , on surface $t \in \{\text{hard, clay, grass}\}$; for example, hard court matches get weight 0.25 in the clay model. Carpet was combined with hard (as the surface was last used in 2009).

Table 2. Comparison of predictive power with the pseudo-official ATP ranking system. N=16759 matches were used in all cases. *McHale and Morton model is equivalent to f_1 and w_1 including surface.

Likelihood function	Weight Function	Predictive power (Y_1)	Predictive power including surface
Games (f_1)	$w_1 : \gamma = 0.0035$	206.1	262.4*
Games (f_1)	$w_2 : \gamma = 0.0035$	173.2	246.8
Match (f_2)	$w_1 : \gamma = 0.003$	135.0	175.8
Match (f_2)	$w_2 : \gamma = 0.003$	118.2	166.4

3. Comparison of ranking methods

3.1 Forecasts from models and ATP rankings

When comparing rankings between the official system and the model-based approaches, we consider their predictive power as a proxy for assessing how accurately they reflect the relative abilities of players. Following on from work by previous authors (for example, Clarke and Dyte (2000) and McHale and Morton (2011)), we estimate the probability p_i of player i winning the match using the formula

$$\text{logit}(p_i) = \alpha * (\ln(r_j) - \ln(r_i)) + \beta * (r_j - r_i).$$

Here, r_i and r_j are the ranks of the players involved, while α and β are parameters fitted from a logistic regression. Note that this formula differs from previous work in that we blend a log and linear term, rather than relying on a log term alone. We found that using a log term on its own roughly captures the curvature in the rankings, but is not entirely satisfactory as it means that the probability of the 5th ranked player beating the 10th ranked player is the same as 500th ranked player beating the 1000th ranked player (and similarly for any situation where one player's rank is twice the other). We found that including a linear term as well allows us to generate much more sensible probabilities for players further down the rankings, by amplifying their difference in ability. Moreover, it is also sensible to allow α to depend on the number of sets played (3 or 5), to account for the fact that a longer contest tends to favour the stronger player. For the model f_1 and weight function w_1 , the optimal parameter values were found to be $\alpha_3 = 0.62$, $\alpha_5 = 0.9$ and $\beta = 0.0020$, where the subscripts indicate the restriction to 3-set and 5-set matches.

As can be seen in Table 2, we then use a log score to assess how accurate those probabilities are for the different models under consideration

$$Y_1 = 10000 * \left(\frac{1}{N} \sum_{k=1}^N \ln(p_{W(k)}) - \frac{1}{N} \sum_{k=1}^N \ln(o_{W(k)}) \right).$$

Here, $p_{W(k)}$ is the forecast for the player who won match k , whilst $o_{W(k)}$ is the equivalent forecast from the pseudo-official rankings (described in Section 2.1). N is the number of matches tested (those in which both players have a rank).

As was already known, the McHale and Morton model is significantly better than the official system by this metric. However, what is of more interest to us here is how different model simplifications compare, so that we can make a more informed decision on what model features to include in a ranking system. For example, even the simplest model (f_2 and w_2) is a significant improvement over the current system, despite having three of the same constraints; no surface information is used, only match results count, and a 12 month rolling window is used to weight games.

Using match outcome (f_2) rather than the more informative 'games won' approach (f_1) leads to a big drop in predictive power, indicating that using the match score would be a useful addition to the rankings. There is also a drop when using the 12 month window (w_2), rather than the exponential downweight (w_1) that puts more emphasis on recent results. However, there may be more scope for considering this simplification, since the drop in predictive power is less than half that observed for the match result (f_2 vs f_1) case. As was to be expected, removing surface information also led to a drop in predictive power.

We note that our analysis agrees with McHale and Morton (2011), when considering the disaggregation of models. We also considered the Brier score to assess differences but the results were similar (not shown).

Table 3. Comparison of how sensitive each model's ranks are to (a) an individual match, (b) an individual tournament, and (c) the time of year.

Model	Weight Function	Sensitivity to matches (Y ₂)	Sensitivity to tournaments (Y ₃)	Sensitivity to season (Y ₄)
Games (f ₁) and Surface	w ₁ : γ = 0.0035	2.25	3.91	-
Games (f ₁)	w ₁ : γ = 0.0035	1.05	1.94	2.14
Games (f ₁)	w ₂ : γ = 0.0035	0.52	1.01	1.47
Match (f ₂)	w ₁ : γ = 0.003	1.27	2.17	2.43
Match (f ₂)	w ₂ : γ = 0.003	0.59	1.14	1.64
Pseudo-official	-	0.25	1.08	1.67

3.2 Sensitivity across matches, tournaments and seasons

Another important property we want to consider is how sensitive rankings are to particular results, tournaments and situations. We start on a general level by considering the expected shift in log ranks after individual matches and tournaments. Namely,

$$Y_2 = 100 * E [(\ln(r_{ik}^*) - \ln(r_{ik}))^2],$$

$$Y_3 = 100 * E [(\ln(r_{iT}^*) - \ln(r_{iT}))^2].$$

Here, $\ln(r_{i.})$ and $\ln(r_i^*)$ are the log ranks for player i before and after match k, or tournament T. With slight abuse of notation, $E[.]$ is the mean of the value inside the brackets.

As can be seen from Table 3, the pseudo-official ranks are more stable than the model rankings following individual matches, but not always so following complete tournaments. Investigating further, we found that tournament round was the most important factor, with the early stages of the tournament having very little impact on the pseudo-official ranks (see Tables 4 and 5). However, the pseudo-official ranks were very sensitive to later matches (Quarter Finals onwards), thus impacting the variability between tournaments.

To look at variance across a whole year, we considered

$$Y_4 = 100 * E [(\ln(r_{ik}) - E_{YEAR}[\ln(r_i)])^2],$$

where we compare the log rank $\ln(r_{ik})$ for each player i in match k against their average log rank from 6 months before to 6 months after. As can be seen in Table 3, the effect discussed above is amplified as the pseudo-official ranks become more sensitive, relative to the most stable model ranks.

In summary, model ranks are liable to shift up and down throughout the tournament and year, with players generally rewarded until the point they lose. Meanwhile pseudo-official ranks are only significantly altered towards the end of the tournament when a large ranking point reward is given to a small number of competitors. The cumulative impact of a few large gains and drops in ranking points causes increased variability, as the 12 month rolling window moves along. This is evident when we consider that a player ranked between 50 and 100 in the world could get over a third of their yearly ranking points from winning an ATP250 tournament (the least important in the data).

Comparing the different models, the most sensitive models are those with the exponential time decay by day (w_1). These are most reactive to short-term effects and hence a string of good results for one player - for example, a player who has a run of good results on their favourite surface. The 12 month rolling weight strategy (w_2) is less sensitive to any short term and surface bias, since a larger range of matches are given high weight. Therefore, maintaining some form of 12 month rolling window is beneficial for preventing seasonality and surface biases from entering the rankings. However, we note that such an approach slows down how quickly a genuine change in ability is recognised by the rankings.

The above comparisons were for models that ignored surface information. It is difficult to compare rankings throughout the season when the surface method is taken into account, because each player has a different ranking on each surface (rather than an overall ranking). However, at the match and tournament level, these surface specific rankings are even more sensitive to results (see Table 3), since a match on surface 's' carries more weight in that surface model, relative to an average match in the season. In general, any seasonal bias due to surface preference is going to get bigger when we take surface into account, using the described method.

Table 4. Shifts in the log(rank) for a number of different match scenarios. We use model f_2 and weight function w_2 as this is the most comparable in terms of sensitivity to matches and tournaments.

Case	n	Expected gain in model ranks (log)	Expected gain in pseudo ranks (log)
Win match	16796	0.037	0.034
Win against better ranked player	5606	0.063	0.048
Win against worse ranked player	11190	0.024	0.027
Win opening match	6797	0.040	0.028
Win Quarter Final or above	3287	0.034	0.057
Win Grand slam match	3067	0.042	0.038
Lose match	16796	-0.036	approx 0
Lose against better ranked player	11190	-0.031	approx 0
Lose against worse ranked player	5606	-0.046	approx 0
Lose opening match	6797	-0.041	approx 0
Lose Quarter Final or above	3287	-0.034	approx 0
Lose Grand slam match	3067	-0.040	approx 0

Table 5. Shifts in the log(rank) before and after tournament, based on success in that tournament. We use model f_2 and weight function w_2 as this is the most comparable in terms of sensitivity

Round reached	Expected gain in model ranks (log)	Expected gain in pseudo ranks (log)
W	0.107	0.181
F	0.075	0.148
SF	0.058	0.099
QF	0.030	0.054
1 st Round	-0.036	0.005

Table 6. Sensitivity of ranks to Grand Slams. Grand slam effect is the measure for Grand Slam matches, relative to all other matches (1 = no difference)

Model	Weight function	Grand slam match weight	Grand Slam effect (on Y_2)	Grand Slam effect (on Y_3)
Games (f_1)	$w_1 : \gamma = 0.0035$	1	1.50	1.55
Games (f_1)	$w_2 : \gamma = 0.0035$	1	1.72	1.81
Match (f_2)	$w_1 : \gamma = 0.003$	1.4	1.54	1.50
Match (f_2)	$w_2 : \gamma = 0.003$	1.4	1.58	1.56
Pseudo-official	-	-	1.55	1.37

4. Applicability of models to other situations

4.1 Tournament importance

One important feature of the current system is that players in more prestigious tournaments are rewarded by receiving more points for competing. As can be seen from Table 6, the model based on games won (f_1) and the pseudo-official ranks react equivalently to Grand Slam tournaments. This is due to the fact that ATP Grand Slam matches are over 5 sets, and so those matches have more of an impact on the likelihood (due to more games being played). This is not true for the match outcome model (f_2) that requires matches to be up-weighted to have any additional impact. These effects are also seen when we look at the average rank move following a win in a Grand Slam match (Table 4).

For tournaments below Grand Slam level, the model does not necessarily recognise their different importance, as the number of sets would be same. Instead the model rankings are most affected by the quality of the players in each tournament, which will be roughly correlated with the quality of tournament.

One option is to weight matches (m_k) according to tournament importance; however, with the models proposed here, this will amplify the ranking drops for losing players as well as the ranking gains for winners.

It may be that a combination of prize money and penalties for non-attendance (see discussion below on withdrawals) is sufficient to ensure the best tournaments attract the best players.

4.2 Retirements and withdrawals

Another nice feature of the models based on games won (f_1) is that they provide an agreeable solution for how to treat matches ending in retirement; all the games played before a retirement can be included in the likelihood function, implying that ranking movements scale with how much of the match was completed. This compares favourably with alternatives such as giving the winner full credit (as with the current ATP system), which may overly reward a player who only won because their opponent retired. Meanwhile, down-weighting these games leaves a system that is potentially open to abuse, as players realise that it is to their advantage to retire rather than complete the match.

One new issue introduced by the model relates to walkovers, whereby a player chooses not to play. A walkover could be advantageous if a player isn't 100% fit and doesn't want to risk losing against a lower ranked opponent and suffering a big ranking drop (Table 4). A player could also be risking a seeding place in an upcoming tournament if they play a match. We note that this problem is one created by the modelling approach, rather than an issue with the existing system. One option is to put an artificial penalty on the player's model rating; whilst this may knock a player down from their 'true' ranking, it wouldn't reward their opponent and would hopefully prevent avoidable withdrawals.

4.3 Long absences through choice or injury

One major difference between the modelling approach and the current system is how long term absences are dealt with. An extreme example is shown in Figure 1, where we can see the differences in the rankings of Juan Martin Del Potro between 2010 and 2012 following a serious injury. At the lowest point, Del Potro slipped down the official rankings to number 485.

In the official ranking system, players accumulate points and so any gap in playing activity significantly damages their ranking as points from the year before disappear. In the model based approaches, injured players are not penalised for not playing. However, new results then have a much bigger impact as there are fewer recent matches to compare them to. In reality, the model is probably not perfect in the case shown in Figure 1. If this modelling approach were to be implemented, it may be that injured players' model ratings are artificially brought down a small amount for each month they are absent. However, we don't want to do anything as extreme as the current system.

Another issue that may need to be dealt with is players retiring from tennis altogether. The best solution is just to remove these players when turning ratings into rankings.

4.4 Use for tournament seeding and qualification

Whereas the ranking system should be a fair representation of a players overall ability, the same argument does not necessarily hold for decisions on tournament seedings and qualification. For example, some players perform better on a particular surface and so the seeding could better reflect this. For the general competitiveness of the tournament, one wants the best players taking part, and so the surface-specific models introduced in Section 2.3 provide a more accurate way of selecting players and seeds.

One potential issue with the surface-specific models described (based on McHale and Morton (2011)) is that grass and clay matches have virtually no impact on one another. This may be valid, but is not necessarily fair given the European clay court season directly precedes the grass court season, and a genuine improvement in the lead up to Wimbledon would not necessarily be reflected in the seedings. Therefore, it is likely that the weight matrix needs to be more even-handed in the way it decides the weightings of past tournaments, on different surfaces. For example, for some constant c ,

$$M_{st} = \begin{bmatrix} 1 & c & c \\ c & 1 & c \\ c & c & 1 \end{bmatrix}$$

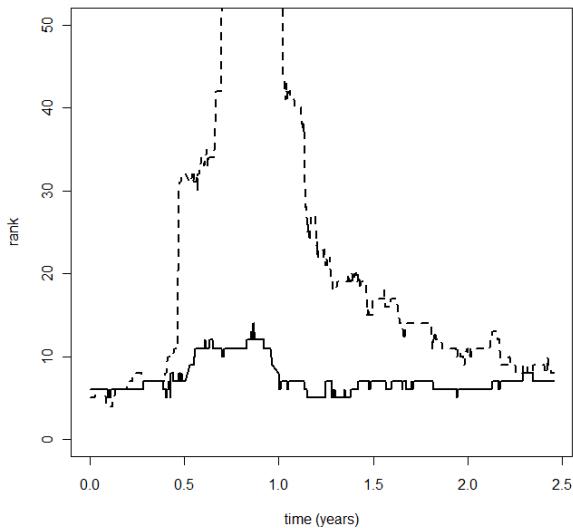


Figure 1. Juan Martin del Potro's ranking in the model (line) and pseudo-official system (dashed) between March 2010 and August 2012. Del Potro suffered an injury missing 9 months from May 2010.

5. Discussion

In this paper, we considered how a class of statistical models can be used to improve the ATP tennis ranking system. Overall, a modelling approach based on optimising pseudo-likelihoods has the ability to improve the current system across a range of measures, including predictive power, insensitivity to the stage of season and ability to cope with absences through injury. It also provides a way of assessing player ability on different playing surfaces.

In order to get a breakdown of how improvements can be obtained, we produced a series of models, separated by features of the existing ATP system - for example, dropping the constraints that (i) only match outcome (win / lose) affects rankings, (ii) games are time independent within a 12 month rolling window and (iii) all surfaces are treated equally.

Considering match score in terms of games won (model f_1), rather than just match outcome (model f_2), was better against all metrics tested. Moreover, the method of using games won provides a sensible approach to 5 set Grand Slam matches, as more data means those matches have a greater impact on the likelihood function, and hence future ranks. This is opposed to artificially upweighting matches in these tournaments, as is necessary in alternative systems. Similarly, the model based on games won (f_1) deals fairly with retirements, since matches that end early due to a genuine injury have less of an impact on the future ranks of the players involved. Also, since the final scoreline is taken into account, a player who retires whilst behind does not gain any unfair advantage. Having a model based on games won and lost, rather than match outcomes, does have one minor issue when it comes to fairness, since a player can win a match but lose more games than their opponent. For example, winning a match 0-6, 7-6, 7-6, is counted as losing 14-18 in games. However, we believe the inclusion of games can make a significant improvement to rankings and encourages players to take every game seriously. It also makes it fair to lesser ranked players who do better than expected, but still go on to lose. We therefore believe that the use of the scoreline provides a key improvement to the current system.

With respect to the second restriction, we found a trade-off between predictive power and sensitivity, with a greater emphasis on short-term form (weight function w_1) leading to more accurate player rankings but greater variability in player ranks. In some cases this is absolutely fine, but the clustering of tournaments on the same surface, at different times of the season, implies that we move too much for some players. For example, a clay court specialist's rank may go up during the clay court season and then down later in the year, rather than representing their 'true' overall ability. This is not a problem when all games are treated equally within a 12 month rolling window (weight function w_2). The incorporation of surface information

amplifies the above mentioned trade off; also, each player ends up with a separate ranking for each surface, rather than an overall rank, which potentially makes the system more difficult for the average tennis fan to follow and appreciate.

Overall we feel that constraint (i), where only match outcome is used, can be dropped without damaging the standing and success of any new ranking system. However, using a time downweight such as an exponential or incorporating surface information could be problematic, as the models are not currently sophisticated enough to prevent seasonal biases creeping into the overall rankings. Since a key aspect of any successful ranking system is that it is seen as fair by the general public, we do not believe that dropping constraints (ii) or (iii) are a good idea with the currently available models. In the future, improvements to the models may well address the above issues and lead to better rankings than those proposed here. Despite this, the models that only drop the first constraint are still a major improvement over the current system, and provide a solid platform for further improvements in the future. Moreover, the surface-specific models could still be extremely useful for working out qualification and seeding for tournaments, by ensuring that the players with the greatest chance of success get to take part.

We also looked at where the major differences lie between the models and official ranking system. It is interesting to note that the model with the 12 month rolling time window (f_1 and w_2) was less sensitive to the time of season than the official rankings. Part of the reason for this may be that the official rankings are still biased towards surface preference or a short burst of good form. A player can pick up a lot of points by reaching the later stages of a few tournaments on their favoured surface, and then not be adversely affected during the remainder of the season (as bad results don't get explicitly penalised). Meanwhile, the models penalise any bad results and don't reward tournament success to the same extent. However, the models still give sensible rewards to players as they progress through the tournament, and still reward Grand Slam matches over those from other tournaments. In particular the model based on games won and lost naturally increases the importance of 5 set Grand Slam matches, because more data is available. One minor issue that remains is whether tournaments below Grand Slam level need to be weighted.

One thing that has not been mentioned so far, is how understandable such a model-based approach might be to the general public. Conceptually though, what we are proposing is fairly simple. Players have ratings, which best reflect their past match results. If a player then does better / worse than expected in an upcoming game, this helps / hinders their rating. Also, there is only one formula to put forward, which could be accompanied by simple examples to show how ratings change in response to a handful of games. To this end, keeping the models simple is an important part of the process and shouldn't be sacrificed for relatively small gains on other metrics. One complicated thing to communicate is that all of the ratings are inferred at the same time, so that there is not a simple relationship between one player's result and their change in rating. However, perhaps we can take some hope from the example of the Duckworth-Lewis method in cricket, which gained acceptance despite being both a complex model and proprietary.

In terms of implementing such a model-based system, naturally the models would require adaptation in response to any concerns from the governing body. For example, weight functions and tournament weights would need to be decided upon. Also, at the moment players remain unrated until they have played 20 games; this rule may need to be adapted to ensure new talent is not stifled. There may be some cases where the model ratings have to be altered to penalise players as well. For example, it may be necessary to penalise the ratings of players that withdraw from a tournament, are out for a long time through injury, or fail to compete in enough tournaments throughout the year. We could use the mean and variance of the model rating distribution to set sensible levels for these penalties and the severity could be judged by the governing body on a case-by-case basis. Also, players that are known to have retired from tennis need to be discarded before rankings are calculated. It should also be emphasised that the data used here does not have matches for challenger and futures tournaments, or qualifying games in general, which means it under-represents players further down the rankings. Therefore, before implementation, the models would need to be optimised and tested on data from all official games, which the governing body would be able to provide.

Finally, we point out that not only can these approaches be applied to women's WTA rankings, they can also be applied to other ranking systems in other sports. However, each system would have to be developed to deal with its own specific needs. For example, Grand Slams in the WTA may need to be upweighted since

they are only 3 sets. Moreover, exponential time decays could easily be used in sports in which there are no obvious seasonal biases.

In conclusion, we believe that statistical models should be more widely used in sports ranking systems, and that it is possible to adapt them to address any concerns from governing bodies. A compromise on predictive power is likely to be necessary for this, but huge improvements can still be made.

References

- Boulier, B. L., and Stekler, H.O. (1999). Are sports seedings good predictors? An evaluation. *International Journal of Forecasting*, **15**, 83-91.
- Bradley, R. A., and Terry, M. E. (1952). Rank analysis of incomplete block designs I: the method of paired comparisons. *Biometrika*, **39**, 324–345.
- Clarke, S. R., and Dyte, D. (2000). Using official ratings to simulate major tennis tournaments. *International Transactions in Operational Research*, **7**, 585–594.
- Del Corrala, J. and Prieto-Rodriguez, J. (2010). Are differences in ranks good predictors for Grand Slam tennis matches? *International Journal of Forecasting*. **26**, 551-563.
- Dingle, N., Knottenbelt, W., and Spanias, D. (2013). On the (Page) Ranking of Professional Tennis Players. *Computer Performance Engineering, Lecture Notes in Computer Science* **7587**, 2013, 237-247
- Glickman, M.E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics*, **48**(3), 377-394.
- Knottenbelt, W.J., Spanias, D., and Madurska, A.M. (2012). A Common-Opponent Stochastic Model for Predicting the Outcome of Professional Tennis Matches. *Computers and Mathematics with Applications*, **64-12**, 3820-3827.
- McHale, I. and Morton, A. (2011). A Bradley-Terry type model for forecasting tennis match results. *International Journal of Forecasting*, **27**, 619-630
- Meyer, D., and Pollard, G. (2012). Fame and fortune in elite tennis.
In *Proceedings of the 11th Australasian Conference on Mathematics and Computers in Sport*.

Relative importance of the offensive and defensive efficiencies in the FIVB Men's Volleyball World League

Elsuida Kondo*, Raymond T. Stefani** and Anthony Bedford***

*School of Mathematical and Geospatial Sciences, RMIT University, elsuida.kondo@rmit.edu.au

**California State University, Long Beach, USA

***School of Mathematical and Geospatial Sciences, RMIT University

Abstract. In volleyball as in tennis, the successful team is the one winning the most sets. Points in a set may be scored while serving or when receiving, in which case the point-scoring team takes over serving duties. The box score evaluates three types of offensive phases where points may be scored or lost (serves, blocks and spikes) as well as three types of defensive phases where points may be lost or activity (sets, digs and receptions) supports a subsequent offensive phase. Data were extracted from the 2011 season of the FIVB Men's Volleyball World League, including more than 100 matches with more than 200 sets of team data. The fraction of points scored by each team was the dependent variable while six standard and six improved efficiencies (one for each of the six phases) formed two sets of independent variables. A regression analysis showed that the improved efficiencies explained 38% more variance than the standard efficiencies. The most important phases of the games were identified as to winning. The results provide volleyball coaches and players with information as to which phases of the game to concentrate upon and which match statistics are the most important when evaluating an opponent or that team's own strengths and weaknesses.

1. Introduction

The data that are collected in a sport tend to depend on the pace of the sport. A continuous action sport like rugby and Association football (soccer) leave little time for data recording as the game progresses. On the other hand, a more structured sport like American baseball and cricket, with many starts and stops, lends itself to data collection. With that data, arises a question about which data or combinations thereof are most useful for demarcating how the game actually was played and also which are best for the purposes of coaching and predicting future contests. An entire science has arisen regarding baseball called Sabremetrics, for the Society for American Baseball Research, founded by American Bill James; see for example James (1982). A recent movie, *Moneyball*, was devoted to that subject, Stefani (2011). We intend to apply a Sabremetric-like approach to volleyball, another structured sport with a wealth of collected data per match. The International Volleyball Federation FIVB routinely summarizes data into certain statistics intended to summarize phases of play. We intend to examine the fit of those statistics to past match results and seek variations of those statistics that better fit past match results.

The paper begins with a general discussion of the game of volleyball. Next, we define six phases of play and the data collected for each phase. We list the current standard efficiency statistics that are published by FIVB after each match and then we suggest better efficiency statistics that correct for faults. Our Results section summarizes a regression analysis of the standard and improved efficiencies for the 2011 FIVB Men's Volleyball World League. We summarise our findings in the Conclusion section.

2. The Basics of Volleyball

Volleyball is an exciting and challenging sport that has witnessed significant growth over the past few years and has developed into a professional spectator event. The sport originated in the United States, and is now just achieving the type of popularity in the U.S. that it has received on a global basis, where it ranks behind only soccer among participation sports. With the success of world competitions such as the FIVB World Championships, Olympic Games, the US20\$ million FIVB World League, FIVB World Grand Prix, FIVB World Cup, and FIVB Grand Champions Cup the level of participation at all levels internationally continues to grow.

It is now one of the big five international sports, and the FIVB, with its 220 affiliated national federations, is the largest international sporting federation in the world. Understanding the rules, technical skills, and strategies of competitive volleyball is essential for its full appreciation. The FIVB Men's and Women's

World Championships are a truly global competition, contested every four years. A noteworthy 215 men's and women's national teams registered to take part in the 2010 event. The first FIVB Men's World Championship, held on a repurposed outdoor tennis court in Prague, Czechoslovakia, in 1949 was for all practical purposes a European Championship, with all 10 teams from Europe. In theory, the objective of volleyball is to "ground" the ball on the opponents' side of the net. Accomplishing this objective in a consistent manner requires the highest levels of speed, agility, power, concentration and teamwork.

At its World Congress in October 1998, the FIVB ratified the "rally point system." Every rally would earn a point. Matches are played on a best of best of five set format. The first four sets are played until one team scores 25 points, but the winning team must be ahead by at least two points. The fifth set is played until one team scores 15 points, and again the winner must have a two-point margin. The system was designed to make the scoring system easier to follow and games faster and more exciting. There are six players on court for a Volleyball team, who each must rotate one position clockwise every time their team wins back service from the opposition. Only the three players at the net positions can jump and spike or block near the net. The backcourt players can only hit the ball over the net if they jump from behind the attack line, also known as the three-metre line, which separates the front and back part of the court. Most teams will include in their starting line-up a setter, two centre blockers, two receiver-hitters and a universal spiker. Only certain players will be involved with service reception. Players will also have specialist positions for attack and defence. Substitutions are allowed during the game.

Volleyball may seem to be a simple sport, but the skills required are challenging. The ball is spiked from up to 60 cm above the height of a basketball hoop and the ball then takes fractions of a second to travel from the spiker to the receiver. That means the receiver must assess incoming angle, decide where to pass the ball and then control the pass in the blink of an eye. A purely rebound sport (you can't hold the ball), Volleyball is a game of constant motion. Power and height have become vital components of international teams, but the ability of teams and coaches to devise new strategies, tactics and skills has been crucial for continued success.

This study provides volleyball coaches and players with information as to which phases of the game to concentrate upon and which match statistics are the most important when evaluating an opponent or that team's own strengths and weaknesses.

3. Phases of Play

Table 1 summarizes the six phases of play. Play begins with the serve. As the ball crosses the net on a serve, the receiving team has at most three touches. The first touch is called a reception. If that is successful, the next touch becomes a set (an overhead pass made with the hands), and if that is successful the third touch is called a spike (the overhead attacking shot) which should cross the net. If any of the first two touches crosses the net, a spike is assigned along with a reception or set. If play continues, on a subsequent rally, a team can try to block the opponent shot as it crosses the net. A block into the same court does count as one of the maximum of three touches in Beach Volleyball, but does not count as a touch for indoor Volleyball which is being studied herein. Excluding a block, the first touch on a rally is designated a dig (appearing the same as a reception on a serve), followed, if successful, by a set and spike. If a ball crosses the net on a dig or set, a spike is also assigned. Thus, the six phases in Table 1 are the serve, block, reception, dig, set and spike. A play ends with a fault, that is, the ball touches the ground on the receiving side, goes out of a boundary on the other team's side, or one of many errors are made, which are beyond the scope of this paper to describe. For details, the FIVB website provides the rules: http://www.fivb.org/EN/Refereeing-Rules/RulesOfTheGame_VB.asp

Table 1: Six Phases of Play in Volleyball, play ends with a fault, scored for the opponent.

<i>Order of Action</i>	<i>Phase of Play</i>
Play begins	Serve
Ball intercepted at the net (not counted for three touch limit)	Block
First touch	Reception (on a serve), Dig (on a rally)

Second touch	Set
Third touch (or any shot over the net)	Spike
Rally continues	Block, Dig, Set, Spike

A qualitative description is assigned for each of the six phases of play. Table 2 summarizes the six phases and three options for each phase, in the order of good, bad and indifferent (OK) from left to right, followed by total attempts, the sum of the three options. For the three offensive phases, the spike, block and serve, the good option results in a point for that team, the bad option, a fault, gives a point to the opponent and play continues for the indifferent option. For the three defensive phases, the dig, set and reception, the bad option, the fault, gives a point to the opponent, while the good and indifferent (OK) options are intended to rate the quality of that successful (non fault) defensive action. The sum of three “good” offensive options, spikes, kill blocks and aces plus opponent errors when receiving a serve adds to a team’s score. The sum of faults adds to the opponent’s score. The term “spike” is used both for a phase and for a scoring shot during the spike phase.

Table 2: Volleyball Offensive and Defensive Data for the Six Phases of Play

Phase	Offense			Total
	Score a Point	Give Up a Point	Continue	
Spike	Spikes	Faults	Shots	Total Attempts
Block	Kill Blocks	Faults	Rebounds	Total Attempts
Serve	Aces	Faults	Serve Hits	Total Attempts
	Opponent Errors	Team Faults		
Phase	Defense			Total
	Good Defense	Give Up a Point	OK Defense	
Dig	Digs	Faults	Receptions	Total Attempts
Set	Running Sets	Faults	Still Sets	Total Attempts
Reception	Excellent	Faults	Serve Reception	Total Attempts

4. Standard and Improved Efficiencies

FIVB calculates and posts on a match “box score” what we will call the “standard efficiencies” for each of the six phases. Each such efficiency shown below represents the “good” outcome for each phase divided by total attempts.

- $ESp1 = \text{spikes}/\text{total attempts}$
- $EB1 = \text{kill blocks}/\text{total attempts}$
- $ESr1 = (\text{aces} + \text{opponent errors}) / \text{total attempts}$
- $ED1 = \text{digs}/\text{total attempts}$
- $ESt1 = \text{running sets}/\text{total attempts}$
- $ER1 = \text{excellent}/\text{total attempts}$

Six (hopefully) improved efficiencies are shown below, each of which involves subtracting faults from the good result and then dividing the adjusted value by total attempts. We hypothesize that these improved (adjusted) efficiencies will better explain match results and thus serve as better guidance to coaches and to those wanting to rank teams and predict outcomes of future matches.

- $Sp2 = (\text{spikes} - \text{faults}) / \text{total attempts}$
- $EB2 = (\text{kill blocks} - \text{faults}) / \text{total attempts}$
- $ESr2 = (\text{aces} + \text{opponent errors} - \text{faults} - \text{team faults}) / \text{total attempts}$
- $ED2 = (\text{digs} - \text{faults}) / \text{total attempts}$
- $ESt2 = (\text{running sets} - \text{faults}) / \text{total attempts}$
- $ER2 = (\text{excellent} - \text{faults}) / \text{total attempts}$

The standard and improved efficiencies are now analysed by extracting data from the 2011 FIVB Men's Volleyball World League box scores

5. Results

5.1 The data set

Data were extracted from the 2011 season of the FIVB Men's Volleyball World League, including 111 of the 112 matches with 222 sets of team data. Data were taken from

http://en.wikipedia.org/wiki/FIVB_Volleyball_World_League

Data were missing for one match. The competition format and teams involved may be found at the website. For each match, 34 quantities were extracted. Data were cross checked by ensuring the total attempts for each phase was equal to the sum of the various phase options. Further, score for each team was checked to add to the sum of good offensive options while score against was checked to add to the sum of team faults. After correcting typing errors, a persistent error was found for all matches played in the USA and for all matches played in Italy. Matches played in the USA had insufficient assigned dig errors and matches played in Italy had insufficient team faults on serve. Sufficient values were added to compensate. These errors were found to continue in 2012, so the FIVB will be notified.

5.2 Standard and Improved Efficiency Analysis

For the set of 111 team data, the fraction of points scored was assumed to be a dependent variable to be regressed with each of the two sets of efficiencies (standard and improved) as the independent variables. Table 3, Table 4 and Figure 1 are for the standard efficiencies, while Table 5, Table 6 and Figure 2 are for the improved efficiencies.

Model		Unstandardized Coefficients		Standardized Coefficients	t	P value
		B	Std. Error			
1	(Constant)	.181	.021		8.446	.000
	ESp1	.403	.031	.580	12.853	.000
	EB1	.264	.026	.446	10.087	.000
	ESr1	.229	.039	.258	5.949	.000
	ED1	.008	.012	.032	.675	.500
	ESt1	.010	.017	.026	.552	.582
	ER1	.001	.015	.002	.044	.965

Table 3: Standard Efficiency Summary for ESp1, ED1, ESr1, ESt1, ED1, ER1

Table 4:
standard
Coefficients

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.784 ^a	.615	.605	.030155

Summary of
Efficiency

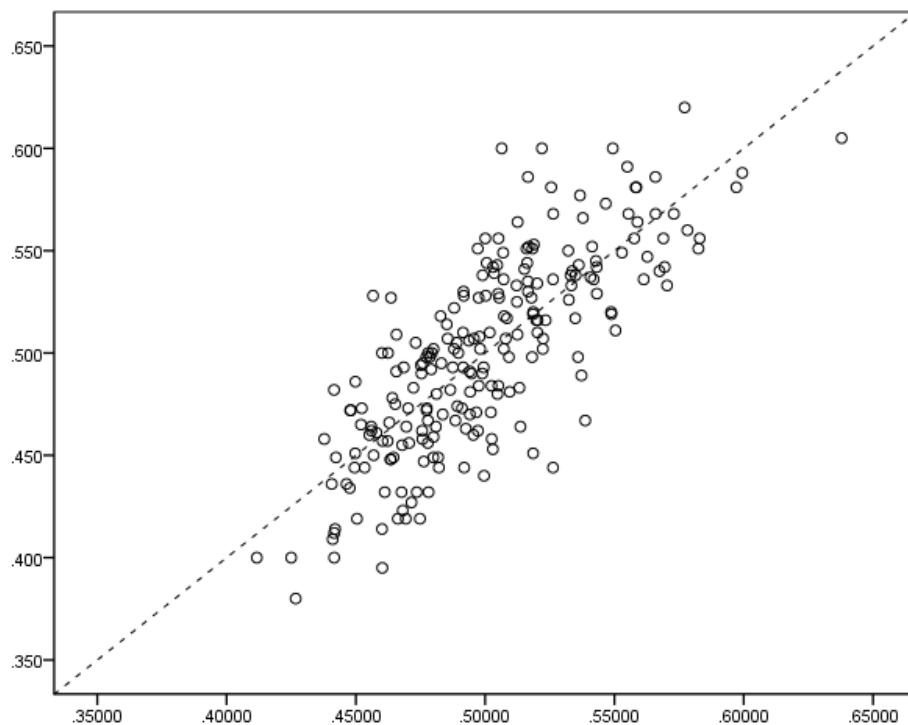


Figure 1: Illustration of a scatter plot for goodness of fit for the standard efficiencies

For the standard efficiencies, 60.3% of the variance is explained based on the adjusted R squared (Table 3). The three offensive standard efficiencies (for spikes, blocks and serves) have coefficients that are significant at more than the 99% level (p value less than .01) as regressed with the fraction of points scored (Table 4). The three defensive standard efficiencies are not significant. Figure 2 shows a scatter plot for goodness of fit.

For the improved efficiencies, 83.2% of the variance is explained (Table 5). That is 38% more than for the standard efficiencies

Table 5: Improved Efficiencies Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.914 ^a	.836	.832	.019684

a. Predictors: (Constant), ESp2, EB2, ESr2, ESt2, ED2, ER2

In Table 6, the three offensive improved efficiencies (for spikes, blocks and serves) are joined by the dig defensive improved efficiency with coefficients that are significant at more than the 99% level (p value less than .01). The reception defensive improved efficiency is now significant at the 90% level. The set improved efficiency is not significant, but p value is less (.186 versus .582) than for the standard efficiency. Simply, distinguishing a running set from a standing set is not a strong measure of quality. In the USA, the American Volleyball Coaches Association uses a good set measure as being a set leading to a point scored (an assist) while the other option is called a “zero-assist”. If those designations would be used by FIVB, it is likely that such an improved efficiency would become significant. Figure 2 shows a better goodness of fit for the improved efficiencies (tighter cluster than in Figure 1).

Table 6: Improved Efficiencies Model Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients Beta	t	P value
		B	Std. Error			
1	(Constant)	.433	.006		67.691	.000
	ESp2	.319	.014	.674	23.243	.000
	EB2	.147	.009	.489	15.990	.000
	ESr2	.122	.011	.313	11.230	.000
	ED2	.045	.006	.240	7.893	.000
	ESt2	-.015	.012	-.042	-1.328	.186
	ER2	-.016	.009	-.053	-1.719	.087
a. Dependent Variable: s%						

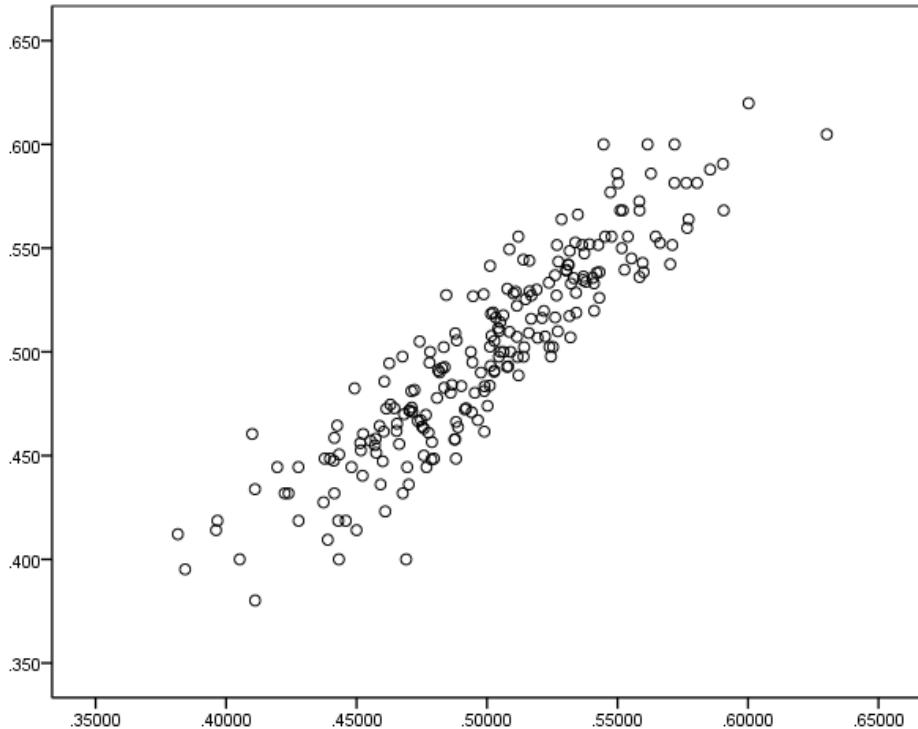


Figure 2: Illustration of a scatter plot for goodness of fit for the improved efficiencies.

5.3 Efficiencies versus winning

Table 7 illustrates an example of a volleyball match against Russia and Brazil by taking for-against points in each phase for the winning team, and for the losing team. We take $\frac{1}{2}$ of the winner difference – the loser difference (because every point is counted twice). Those differences add to the match difference. Table 8 shows the fraction of games where the winning team had a positive score difference in each phase. From the results it clear to see that the winning team was better at the spike in 86.5% of the games and better at the block in 72.1%. The winner was superior in the serve, dig and reception at about the same fraction of games. All of those five phase results are significant at the 99% level (p less than 1%). That leaves the set as not significant because so few faults occur in the set.

Table 7: 2011 Match Championship Match: Scoring Summary

	Russia			Brazil			$\frac{1}{2}(Russia\ Diff - Brazil\ Diff)$
Activity	For	Against	Diff	For	Against	Diff	
Spike	63	20	+43	54	19	+35	+9
Block	12	41	-29	12	54	-42	+6.5
Serve	37	27	+10	43	18	+25	-12.5
Dig		14	-14		12	-12	-1
Set		1	-1		1	-1	0
Reception		6	-6		8	-8	+1
Total	112	109	+3	109	112	-3	+3

Table 8: Fraction of Games Where the Winning Team has a Higher Phase Score Difference

* = Significant at the 99% level

Phase	Average Score Difference: Winning Team	Fraction of Games Where Winning Team has Higher Score Difference (N = 111)
Spike	4.9	0.865*
Block	2.4	0.721*
Serve	2.6	0.631*
Dig	1.3	0.676*
Set	0.0	0.315
Reception	0.7	0.640*
Total	11.9	

6. Conclusions

FIVB should publish the improved efficiencies, which were 38% better at explaining the fraction of points won in a match, compared to the standard efficiencies commonly shown in the match box scores of the World League. The scoring of a dig should be changed to the USA system of assists and zero-assists rather than running sets and standing sets respectively. All matches scored in the USA had insufficient dig faults relative to points given and all matches scored in Italy similarly had insufficient team serving faults. FIVB should make adjustments to provide uniform scoring.

References

- James, B. (1982) *The Bill James Baseball Abstract*, New York: Ballantine Books
http://www.fivb.org/EN/Refereeing-Rules/RulesOfTheGame_VB.asp
http://en.wikipedia.org/wiki/FIVB_Volleyball_World_League
Stefani, R. (2011) Moneyball: Brad Pitt, the statistician and the movie, *Significance*, December, 185-186.

Career lengths and age of retirement of ATP singles players during the Open Era

Stephanie Kovalchik*

*National Cancer Institute, kovalchiksa@nih.gov

Abstract. What were the odds that Andre Agassi would have a career lasting 20 years? How likely is it that Roger Federer will compete at Rio in 2016? Questions about retirement are often raised by tennis sports writers, players, and fans, yet few studies have examined the career lengths of professional tennis players. With a database of over 2,000 top 500 ATP players, I used survival analysis techniques to characterize career lengths and quantify retirement probabilities by player's age for retirements occurring between 1990 and 2012. During the Open Era, nearly 20% of ATP players have had career lengths longer than 15 years and the same proportion have retired after age 32 years. However, players who earn a higher spot in the rankings during their career, who enter the Tour at an earlier age, and who have periods of inactivity of no more than four years can expect to have the longest professional careers. The chance of retirement after age 30 is primarily dictated by a player's age, which supports the existence of an “aging cliff” in professional tennis. Contrary to concerns players have expressed about shortening career lengths, career lengths of ATP players have generally increased since the mid-1990s.

1. Introduction

In 1969, one year after the beginning of the Open Era, Rod Laver became the only male singles player in tennis history to win a second Grand Slam. Sports writers covering Laver's consecutive sweep of the majors were in almost unanimous agreement about his dominance of the game and few could envision the end of his reign. In the eyes of the press, Laver was the ‘southpaw court killer’¹, an ‘odds on favorite’² of winning a major, ‘tough’, ‘unbeatable’, and ‘unapproachable’³. From the perspective of today's age-obsessed tennis culture, where the approach of a player's 30th birthday immediately raises talk of performance decline and imminent retirement, it is remarkable that writers in Laver's era made little todo about the fact that the Rocket blew out 31 candles during his 1969 season.

Flash forward forty years and few players in their thirties are able to escape the shadow cast by the aging factor. Even the greatest of the game is no exception. In 2011, when Roger Federer turned 30 years-old, numerous headlines seemed to foretell an end to the Swiss's supremacy—‘Federer at 30 must heed aging’⁴, they said, and 2011 would be his ‘last opportunity to win a Grand Slam’⁵. Despite the press' conviction of a looming decline, Federer would go on to win the 2011 ATP World Tour Finals and regain the World No. 1 position the following summer.

At some point over the course of the Open Era the notion of competitive decline at age 30 became an idée fixe in tennis circles. Was this a response to shortening career lengths? Or a downward shift in retirement ages of ATP players?

Although age's influence on career longevity has become an increasing concern of tennis players, sports writers, and fans, few scientific studies have examined the retirement characteristics of professional tennis players. When studies have, their focus has been on economic considerations. Coate and Robbins (2001) investigated the influence of prize earnings on the probability of retirement for 236 male and 216

¹ “Ashe, Roche, Laver Win In U.S. Open Tourney.” *Post Herald and Register*. September 1969.

² “Laver odds on favorite at Wimbledon.” *The Southeast Missourian*. June 1969.

³ “Another redheaded league.” *Sports Illustrated*. July 1969.

⁴ “Roger Federer not to focus on turning 30.” *Live Tennis Guide*. August 2011.

⁵ “Roger Federer at 30 must heed aging.” *Los Angeles Times*. August 2011.

female top 50 professional tennis players competing during 1979 and 1994. Geyer (2010) assessed the effect of prize money on the career lengths of professional tennis players in a cohort of 614 professional male tennis players competing between 1985 and 2007. Both studies concluded that lower prize earnings are associated with earlier retirement.

The prize money a tennis player earns is not within his or her direct control but is largely dictated by mandatory entry requirements, tournament organizers, and tournament draw. In contrast, the age and year a player enters the tour are professional decisions. Consequently, how these and other potentially modifiable variables affect career longevity may be of more immediate relevance to players and coaches than prize earnings. Using survival analysis techniques and a database of over 2,000 ATP players, the present study therefore undertook to quantify the effects of age turned professional, the year turned professional, the highest career rank, and the longest period of inactivity on the career lengths and retirement probabilities of elite male tennis players.

2. Materials and Methods

2.1 Data

A database was constructed containing demographic and retirement characteristics for ATP players who had a year-end singles ranking in the top 500 for any year between 1990 and 2010. Year-end ATP World Tour rankings were the standings as of December 24th for each year and are available for download from the Tour website (www.atpworldtour.com). R software was written to download demographic information and retirement status from each player's profile page maintained by the Tour (R Development Core Team 2012). The extracted data were birth date, birthplace, height, weight, playing hand, date turned professional, career highest rank, date of retirement, and longest period of inactivity before retirement. All characteristics pertained to singles play.

Players' date turned professional, date of retirement, and longest period of inactivity are not directly reported by the ATP World Tour. Instead, this information was inferred from a competitor's history of playing activity. The date of the first match in which a player competed that could earn points toward his Tour ranking was taken to be the date a player turned professional. Exceptions to this rule were made if a player had a period of inactivity of four years or more following the year of his first Tour match. In these instances, the year turned professional was taken to be the date of the first Tour match following the period of inactivity. A similar procedure was used to determine a player's date of retirement, setting the day of a player's last match that could earn points toward the Tour ranking as the date of retirement. However, if a player had a period of inactivity of four years or more after he entered the Tour, the date of retirement was set equal to the day of the last Tour match that occurred before the gap in activity. Players whose 'date of retirement' occurred in 2012 or later were considered still active, and their time-to-retirement was censored at the day of their most recent Tour match. The longest period of inactivity between entering and retiring from the Tour was defined as the maximum number of days between consecutive Tour matches.

The decision to omit ATP playing activity that preceded or followed a four-year or longer gap in Tour play was made in order to separate career play from periods of 'dabbling' in the sport. Although somewhat arbitrary, the four-year length is likely to be a conservative lower bound for entrance and exit from professional tennis as a three-year period of inactivity is the length after which an ATP player loses his entry protection for World Tour events.

Based on Tour standings as of December 2012, 2,175 top 500 ATP players were identified for inclusion in the study database. Forty players were excluded from this cohort because their date of birth was not reported on their ATP profile page nor could be obtained from other web sources. Of the

remaining players, there were 44 whose date of birth occurred after reported match dates within the player's career history. Because these inconsistencies could not be resolved, these players were omitted from the study sample. Finally, 30 additional players were removed who were younger than 15 years when they became professional singles players.

2.2. Statistical Analysis

Kaplan-Meier curves were used to describe the time-to-retirement in the study sample. Separate curves were constructed for the time scales of career year and retirement age. For career years, time was measured from time zero that began at the date a player turned professional. For retirement age, retirement was left-truncated at the age a player joined the ATP Tour. Time-to-retirement for active players was right-censored at the date (age) of their last Tour match at the end of the 2012 season. The complement of the Kaplan-Meier survival estimate at a given career year (age) is the overall probability that an ATP player will retire from singles play by that career year (age).

I investigated how several career player characteristics influenced career length using a proportional hazards model of years-to-retirement. A parametric baseline was specified, which utilized a piecewise exponential baseline hazard with distinct hazard rates for every one-year interval. The characteristics whose associations with retirement were examined were age turned professional, highest career rank, and longest gap of inactivity in years. Each characteristic was a time-independent variable. The representation of each variable (categorical, linear, quadratic, etc.) was guided by the pattern of effects in a fitted model in which each variable was evenly split into ten categories.

For all the proportional hazards modeling, calendar year turned professional was included as an adjustment variable. Throughout, it was assumed that censoring times for retirement were non-informative conditional on the year a player joined the Tour.

Estimated retirement probabilities for each year on Tour, for lengths between 10 and 20 years, were obtained from a model containing age turned professional, year turned professional, rank and longest period of inactivity. Retirement curves were calculated for selected ages turned professional and career highest rank, in order to investigate how the relative effects of these factors influenced the cumulative probabilities of retirement. Sensitivity analyses were also performed, which assessed the robustness of the findings from the main retirement analysis to the proportional hazards assumption. This was done by recomputing the retirement curves using a model with time-varying effects for rank and age turned professional.

3. Results

Of the 2,083 ATP singles players included in the study sample, 75% were born between 1965 and 1985, 30% were a top 100 player at one time, and over 50% became professional singles players before age 20 years (Table 1) For the 71% of the cohort who retired from professional singles play, all retired by 2010 and approximately 31% retired in the last half of the 2000s. Ages between 25 and 30 years were the most common retirement ages.

Table 1. Description of the study sample of ATP singles players (N=2,083).

Characteristics	Number (Percentage)
Birth Year	
1950—1964	137 (6.6)
1965—1969	310 (14.9)

1970—1974	427 (20.5)
1975—1979	424 (20.4)
1980—1984	437 (21.0)
1985—1989	285 (13.7)
1990—1992	63 (3.0)
Height	
5ft 0in—5ft 9in	165 (7.9)
5ft 10in—5ft 11in	318 (15.3)
6ft 0in—6ft 1in	411 (19.7)
6ft 2in—6ft 3in	405 (19.4)
6ft 4in—6ft 9in	72 (3.5)
Missing	712 (34.2)
Highest Career Rank	
1—99	635 (30.5)
100—199	472 (22.7)
200—299	473 (22.7)
300—399	360 (17.3)
400—500	143 (6.9)
Age Turned Pro	
15—17	415 (19.9)
18—19	902 (43.3)
20—21	416 (20.0)
22—23	252 (12.1)
24—29	98 (4.7)
Retired	1,488 (71.4)
Retirement Year	
1989—1993	226 (15.2)
1994—1997	245 (16.5)
1998—2001	277 (18.6)
2002—2005	286 (19.2)
2006—2009	454 (30.5)
Retirement Age	
16—20	29 (1.9)
21—22	73 (4.9)
23—24	192 (12.9)
25—26	280 (18.8)
27—28	302 (20.3)
29—30	249 (16.7)
31—32	204 (13.7)
33—34	94 (6.3)
35—44	65 (4.4)

Since the 1990s, retirement in the ATP cohort as a whole indicates that 45% of ATP singles players retired within 10 years of competing on the Tour and 81% within 15 years (Figure 1). The cumulative probability of retirement increased more rapidly with years on tour than a constant exponential rate. For example, an exponential distribution fit to the retirement data would predict 70% retired by 15 years,

which was approximately 10% less than the observed proportion retired. More than half of the ATP cohort retired by age 30 years and all but 19% of singles players retired by age 33 years (Figure 2).

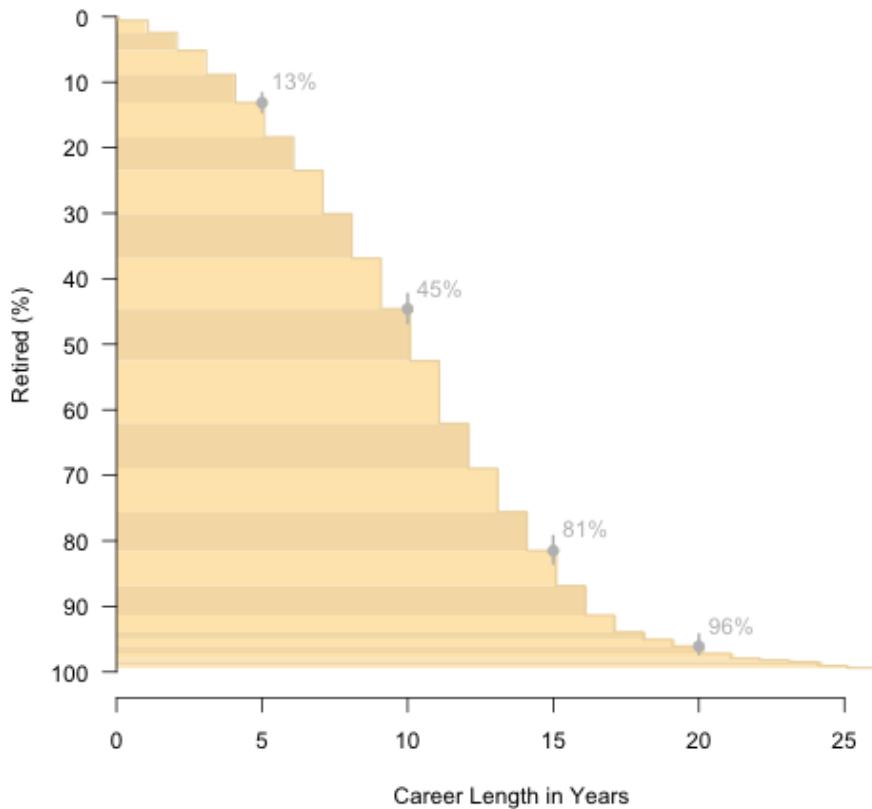


Figure 1. Kaplan-Meier plot of cumulative retirement probabilities for ATP singles cohort by a specified career length. Lines denote the 95% confidence interval for the labeled retirement probability.

Several career characteristics had a significant influence on time to retirement. The probability of retirement after any specified number of years was higher for players who joined the Tour at an older age (Table 2). Players with a lower career highest rank were also more likely to retire than more highly-ranked players. Every additional year older a player was when he joined the Tour was associated with a statistically significant 21% increase in his relative likelihood of retirement, after controlling for rank, year turned professional, and longest period of inactivity. Similarly, according to the multivariable model, a drop of 10 places in the career highest rank was associated with a 5% increase in the rate of retirement. However, more years of inactivity (of no more than 4 years) was associated with a lower rate of retirement. Specifically, each additional year of inactivity was associated with a 35% reduction in the rate of retirement.

Table 2. Proportional hazards regression model^b of ATP singles career lengths.

Variable	Rate Ratio	P-value
Age Turned Pro	1.214	<0.001
Year Turned Pro		
1970—1980	1.000	Reference
1981—1985	1.655	0.0082
1986—1990	1.450	0.0384
1991—1995	1.289	0.1589
1996—2000	0.630	0.0118
2001—2005	0.339	<0.001
2006—2010	0.147	<0.001
Highest Career Rank	1.005	<0.001
Maximum Years Inactive ^a	0.647	<0.001

a Less than four years.

b The baseline hazard was a piecewise exponential with one-year intervals.

In comparison to the generation of players who joined the Tour between 1970 and 1980, players who became professionals between 1981 and 1995 were 30 to 65% more likely to retire in any given career year (Table 2). This association reversed for all later ATP cohorts, suggesting that career lengths have generally increased for the most recent generations of players. As a verification that this pattern was not due to informative censoring resulting from the shorter period of follow-up for players who began playing professionally after the mid-1990s, I examined the association between the year turned pro and the probability of retirement within 10 years for players who entered the Tour prior to 2001 only. Thus, all players in this analysis had a minimum of 10 years of follow-up. The results from the logistic analysis confirmed the previous analysis, showing a decreased chance of career lengths shorter than 10 years beginning in the mid 1990s and becoming increasingly less likely for more recent cohorts (Table 3).

Table 4 shows the implications of the effects of rank and age turned professional (as described in Table 2) on the actual probability of retirement. Among players who attained a highest career ranking within the top 50, age turned professional had a bigger impact on the chance of retirement than differences in the career highest rank. For example, if one considers the variation in the probabilities of retirement after 10 years by age for a fixed rank level, the lowest and highest retirement probabilities for players who entered the Tour between 18 and 21 years differed by approximately 20%. However, the same comparison across career highest rank within the top 50 for players who entered the Tour at the same age showed a difference in the lowest and highest probabilities of approximately 7%.

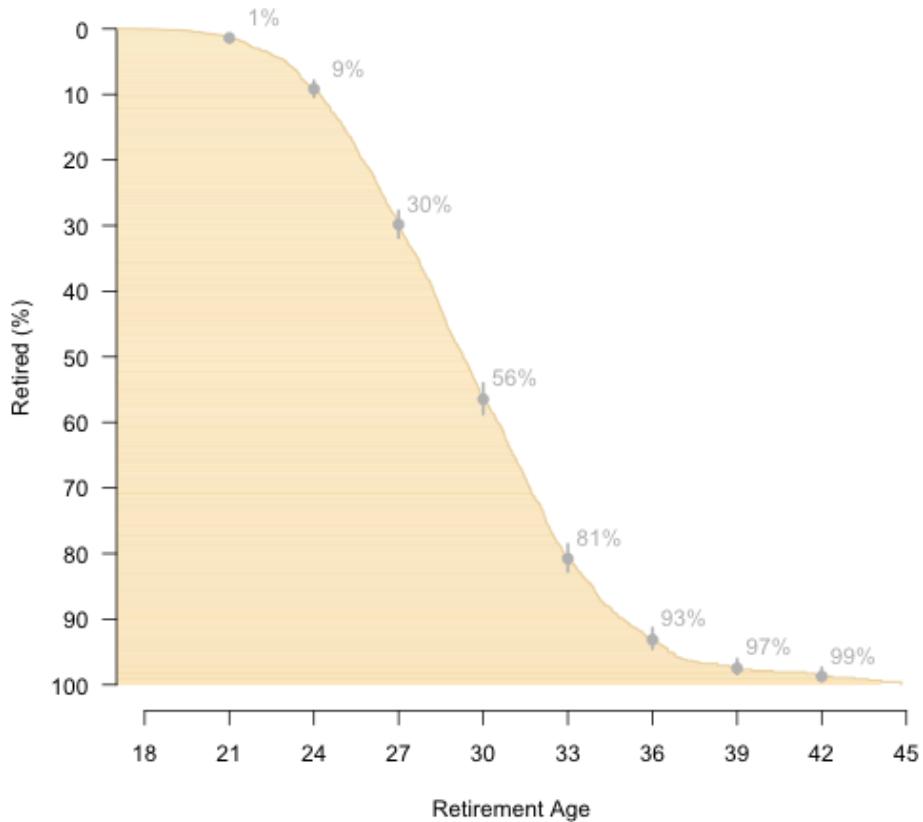


Figure 2. Kaplan-Meier plot of cumulative retirement probabilities for ATP singles cohort by a specified age. Lines denote the 95% confidence interval for the labeled retirement probability.

Table 3. Logistic regression model for the probability of retirement within 10 years for ATP singles players who entered the Tour before 2001.

Year Turned Pro	Number (Percent)	Percent Retired	Odds Ratio	P-value
1970—1987	306 (20.1)	54.2	1.00	Reference
1988—1990	259 (17.1)	65.3	1.58	0.008
1991—1993	228 (15.0)	68.0	1.79	0.001
1994—1996	239 (15.7)	54.0	0.99	0.949
1997—1998	292 (19.2)	57.5	1.14	0.419
1999—2000	196 (12.9)	57.5	0.67	0.031

The association analyses were made under the proportional hazards model. As a sensitivity analysis for the proportionality assumption, the retirement model was refit with time-varying effects for age and highest career rank. The time-varying model resulted in a general increase in the probabilities of retirement for any given career length or retirement age (data not shown). Among top 10 players, for example, retirement after 14 years on tour was between 83 and 96% and retirement by age 32 ranged from 75 to 83%. Despite this upward shift in retirement probabilities, the comparative effects of age and rank were not substantively changed.

The dependence of career length on a player's age of entry on the Tour suggest that probabilities of retirement may be more uniform on the age time scale than the career-years time scale. Indeed, not only were probabilities of retirement for ages 28 through 38 less influenced by the age a player became professional but, unlike career length, retirement at any given age was more likely for players who started their professional career at a younger age (Table 5). Moreover, by the thirties the effects of age turned professional became increasingly negligible. By age 32, there was less than a 1% difference in the probability of retirement for players of the same rank who were between 18 and 21 years when they became a professional singles player.

Table 4. Estimated percent of ATP singles players retired by a specified career length.

Highest Career Rank	Years On Tour	Age Turned Professional			
		18	19	20	21
1-10	10	43.5	50.0	56.9	64.0
	12	60.4	67.5	74.4	80.9
	14	75.1	81.5	87.1	91.6
	16	87.4	91.9	95.3	97.5
	18	92.9	96.0	98.0	99.1
	20	96.0	98.0	99.1	99.7
21-30	10	47.0	53.7	60.7	67.8
	12	64.3	71.3	78.0	84.1
	14	78.7	84.7	89.7	93.7
	16	90.0	93.9	96.6	98.4
	18	94.7	97.2	98.7	99.5
	20	97.2	98.7	99.5	99.8
41-50	10	50.6	57.5	64.6	71.7
	12	75.6	75.1	81.5	87.1
	14	82.0	87.6	92.0	95.4
	16	92.3	95.5	97.7	99.0
	18	96.2	98.1	99.2	99.7
	20	98.1	99.2	99.7	99.9

4. Discussion

During the Open Era, nearly 20% of ATP players have had career lengths longer than 15 years and the same proportion have retired after age 32 years. However, these characteristics do not apply equally to all players. The age a player turned professional, the highest career rank, and the longest period of inactivity were each found to have a significant independent influence on a player's career length. The effects of these factors indicate that players who earn a higher spot in the rankings during their career, who enter the Tour at an earlier age, and who have periods of inactivity of no more than four years can expect to have the longest professional careers.

Table 5. Estimated percent of ATP singles players retired by a specified age.

Highest Career Rank	Retirement Age	Age Turned Professional			
		18	19	20	21
1-10	28	43.5	41.3	38.6	35.9
	30	60.4	60.0	56.9	54.4
	32	75.1	74.9	74.4	74.1
	34	87.4	87.4	87.1	87.0
	36	92.9	94.7	95.3	95.2
	38	96.0	97.1	98.0	98.7
21-30	28	47.0	44.7	41.9	39.0
	30	64.3	63.9	60.7	58.2
	32	78.7	78.5	78.0	77.7
	34	90.0	90.0	89.7	89.6
	36	94.7	96.2	96.6	96.6
	38	97.2	98.0	98.7	99.2
41-50	28	50.6	48.3	45.3	42.3
	30	75.6	67.8	64.6	62.1
	32	82.0	81.9	81.5	81.2
	34	92.3	92.3	92.0	91.9
	36	96.2	97.4	97.7	97.7
	38	98.1	98.7	99.2	99.5

The study findings have important insights for strategies players can take to lengthen their careers. Better performance, as measured by the highest attained rank, is unsurprisingly more predictive of greater career longevity. Doing what it takes to increase one's rank may therefore have the added benefit of delaying retirement. The specific reason for the association between rank and career longevity is unclear. As a player's rank is highly correlated with prize earnings and may also be strongly correlated with a player's commitment to and passion for the sport, a number of factors could be driving rank's observed effect.

Whatever an ATP player's skill level, the results of this study suggest that he may improve his chances of a long career by entering the Tour at an earlier age and also being willing to take time off to

recuperate from minor injuries, supposing that injury is the main reason for gaps in Tour play. However, these conclusions presume a causal link between the observed associated factors and retirement outcomes. Another possibility is that age turned pro and gaps in activity are correlated to true causal factors not considered in the present study because these factors are more difficult to measure. For example, players who enter the Tour at an earlier age may have a stronger support system (a dedicated “team”) or superior fitness as compared to players who turn professional at an older age.

Another explanation for the positive effect a player's age turned pro has on career is aging. As suggested by the mythical proportions age 30 has taken in the tennis media, an ‘aging cliff’ may exist for professional singles players whereby a physiological decline that sets in after a certain age results in a decline in competitive performance and marks the beginning of the exit from the Tour. The evidence from the present study supports the existence of an ‘aging cliff’ in professional tennis. By the thirties, I found that career longevity became an increasingly less relevant factor on the probability of retirement than a player's age. Although the exact age depended on a player's skill level (as measured by career highest rank), the uniformity in retirement probabilities after age 30, for any given skill level, were striking. Among players who achieved a top 10 ranking during their career, for instance, the probability of retiring by age 34 years was 87% whether a player had been playing on Tour for 13 years or as long as 16 years. Because the present study focused on career trajectories rather than performance trajectories, it is not possible to say at which age a downturn in performance typically occurs for players. Further analyses are needed to clarify the relationship between age and performance over the course of an ATP player's career.

In recent years, a number of players have been critical of the ATP for not making changes to the Tour schedule. Since the introduction of the Tennis Master Series in 2000, players have argued that the increasing demands of the schedule⁶ and the increasing physical demands of the sport⁷ are shortening players' careers⁸. At the same time in the past decade commentators have noted that the Tour has gotten older⁹ and it has become more and more difficult for young players to break into the highest echelon of the game¹⁰, which would suggest that careers are getting longer rather than shorter. So, which view is the correct one?

⁶ In an interview for *Sport Illustrated* on February 12, 2013 entitled “Nadal: ATP not concerned about players' health”, Rafael Nadal said “The ATP has to start thinking about ways to lengthen the players' careers...I can't imagine any other sport involving aggressive movements such as tennis being played on such aggressive surfaces such as ours. We are the only sport in the world making this mistake and it won't change.”

⁷ In the article “Andy Roddick claims current tennis schedule is ‘ridiculous’”, reported by the *Telegraph* on October 13, 2009, Andy Roddick gave the following summary of the current Tour schedule: “We finish around 30 November and have to be pretty much Grand Slam-ready by 4 January, year after year after year...No sport can do that and it means your career is shorter.”

⁸ *Setanta* posted an article on November 10, 2010 entitled “Murray queries ATP Tour schedule” in which Andy Murray made the following comments about injury and the grind of the Tour season: “There's no time for you to take a break to get rid of an injury...Instead players end up playing through it and that actually shortens careers.”

⁹ On October 9, 2012, in an article on www.tennis.com entitled “Golden Era, or Groundhog Era?”, Steve Tignor wrote that “Twenty-seven seems to be when players are entering their primes.”

¹⁰ In “Tommy Haas and his time machine”, which was posted on the web on August 18, 2012, Nick Bollettieri said the following in response some recent wins of 34 year-old Haas: “There's been lots of discussion lately about the tour getting older—teenagers breaking through at Wimbledon, by winning Wimbledon a lá Boris Becker seems like ancient history. But if the tour is getting older, it certainly isn't getting this old, is it? Maybe the answer is yes, as other tour players will want to emulate Haas' longevity.”

The present study of trends in career lengths found no evidence of a decrease in career lengths over time. To the contrary, since the mid-1990s career lengths of tennis players have gotten longer in general. How can this apparent between the study's findings and players' expectations be explained?

The shift in career longevity appears to have begun in 1995, five years after the establishment of the ATP World Tour. This was a dramatic turning point in tennis history. With the new Tour, players gained an equal role in the oversight of the professional circuit and there was an immediate 50% increase in prize money at Tour events (ATP World Tour 2012a) Tennis successfully crossed a crossroads (ATP World Tour 2008) when it created the ATP Tour, and the benefits to players in this new era have been numerous. The ATP Tour has given tennis a worldwide audience and attracted new and wealthier sponsors, all of which has helped to make the sport more lucrative and increase the depth of the competitive field. Thus, the sum of these positive effects may well have overshadowed any detrimental effects of an increasingly tougher schedule.

Still, the present study does not fully address the concerns current players have expressed to the press regarding the Tour schedule. At the root of the players' criticism is the hypothesis that competitors who play more often and play much tougher matches will have shorter careers. To test the validity of this hypothesis, further research is needed to examine how factors directly related to play activity, such as total matches played and total match duration, affect career lengths.

References

- ATP World Tour (2013). <http://www.atpworldtour.com/Corporate/History.aspx>, Accessed: February 24, 2013.
- ATP World Tour (March 21, 2008): "Former ATP Tour CEO Hamilton Jordan passes away", <http://www.atpworldtour.com/News/Tennis/2008/05/jordan.aspx>, Accessed: February 24, 2013.
- Coate D, Robbins D (2001): "The tournament careers of top-ranked men and women tennis professionals: Are the gentlemen more committed than the ladies? *Journal of Labor Research*, 22, 185—193.
- Geyer H. (2010): "Quit behavior of professional tennis players," *Journal of Sports Economics*, 11, 89—99.
- R Development Core Team (2012): *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, <http://www.R-project.org>.

The PEAST Algorithm in Sports Scheduling - Evolution of the Major Finnish Hockey League Schedules

N. Kyngäs*, D. Goossens**, K. Nurmi* and J. Kyngäs*

* Satakunta University of Applied Sciences, Pori, Finland; {nico.kyngas, cimmo.nurmi, jari.kyngas}@samk.fi

** Center for Operations Research and Business Statistics, University of Leuven, Belgium;

dries.goossens@kuleuven.be

Abstract. PEAST (Population, Ejection, Annealing, Shuffling, Tabu) is a population-based local search method that has been used for solving various real-world scheduling problems such as school timetabling, shift generation, staff rostering, days-off scheduling and sports scheduling. We have used it to schedule the Finnish Major Ice Hockey League since the 2008-2009 season. The League wants to continuously improve its schedule format, and as a result the scheduling problems are ever-increasing in complexity. We introduce some of the specific real-world cases we have encountered along the years and make them available for the scheduling community as benchmarks. We compare the performance of PEAST and a basic CPLEX implementation on one of these cases.

1. Introduction

In the past decades professional sports leagues have become big businesses; at the same time the quality of the schedules has become increasingly important. This is not surprising, since the schedule directly impacts the revenue of all involved parties. For instance, the number of spectators in the stadiums, and the traveling costs for the teams are influenced by the schedule. TV networks that pay for broadcasting rights want the most attractive games to be scheduled at commercially interesting times in return. Furthermore, a good schedule can make a tournament more interesting for the media and the fans, and fairer for the teams. Nurmi et al. (2010) report a growing number of cases where academic researchers have been able to close a scheduling contract with a professional sports league owner. Excellent overviews of sports scheduling can be found in Easton, Nemhauser and Trick (2004), Dinitz et al. (2006), Drexl and Knust (2007) and Rasmussen and Trick (2008). An extensive bibliography can be found in Knust (2012) and an annotated bibliography in Kendall et al. (2010).

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 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 opponent's 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.

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 (1981) 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 an *M2RR*, it is always possible to construct a schedule with exactly $3n - 6$ 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 (2001). In *TTP* the teams do not return home after each away game but instead travel from one away game to the next. However, excessively

long away trips as well as home stands should be avoided. Thielen and Westphal (2011) recently showed 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 which *satisfies given constraints* the *Constrained Sports Scheduling Problem (CSSP)* as defined by Nurmi et al. (2010). The goal is to find a feasible solution that is the most acceptable for the sports league owner - that is, a solution that has no hard constraint violations and that minimizes the weighted sum of the soft constraint violations.

Scheduling the Finnish major ice hockey league is an example of a *CSSP*. It is very important to minimize the number of breaks. The fans do not like long periods without home games, consecutive home games reduce gate receipts and long sequences of home or away games might influence the team's current position in the tournament. It is also very important to minimize the travel distances. Some of the teams do not return home after each away game but instead travel from one away game to the next. There are also around a dozen more other criteria that must be optimized.

We have used the PEAST algorithm (Kygäs, Nurmi and Kyngäs, 2013) to schedule the Finnish major ice hockey league since the season 2008-2009. Over time, scheduling the league has grown more complicated. We will give an overview on how the constraints of the league schedule have changed over time. In Section 2 we describe the problem setting. Section 3 introduces the general constraints used for the league schedules for different seasons. Section 4 introduces the PEAST algorithm and a small comparison between PEAST and CPLEX on a real-world instance. In Section 5 we consider some of the practical aspects of the scheduling process.

2. The Finnish major ice hockey league

Ice hockey is the biggest sport in Finland, both in terms of revenue and number of spectators. The Finnish major ice hockey league is a private company with fifteen shareholders, one for each team in the league and one for the Finnish Ice Hockey Association. Each team is also a private company. The CEO of the team is responsible for getting the best possible schedule for his team.

The CEO of the league is responsible for producing the schedule. Prior to the 2008-2009 season, the schedule was produced manually. After making the schedule for the 2007- 2008 season with an increasing number of requirements and requests, the CEO realized that they were no longer able to handle the schedule manually.

Seven of the teams in the league are located in big cities (over 100,000 citizens) and the rest in smaller cities. One team is quite a long way up north, two are located in the east and the rest in the south (see Figure 1).

The schedule format for the league is quite stable. The base of the schedule is a quadruple round robin tournament resulting in 52 games for each team. In addition, the teams are divided into two groups of seven teams to get a few more games to play. These teams play a single round robin tournament resulting in 6 games. Therefore, there are 58 games for each team and a total of 406 games to be scheduled.

Since the 2010-2011 season every team has played two additional games due to the "January leveling", a concept we proposed to the league administration. 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 by maximizing the winning possibilities or by maximizing the ticket sales. Thus the seasons since 2010-2011 have included 60 games per team and 420 games in total.

The competition starts with a regular season in September and ends with the playoffs from mid-March to mid-April. The league fixes the dates on which the games can be played. Since the 2008-2009 season the last team of the regular season has played best out of seven elimination games against the best team of the Finnish 1st division ice hockey league. The six best teams of the regular season proceed directly to quarter-finals. Teams placed between 7th and 10th play preliminary playoffs best out of three. The two winners take

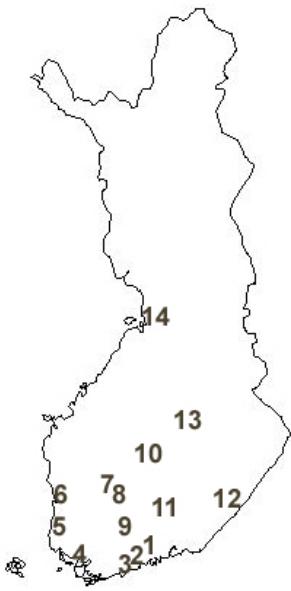


Figure 1. The map of Finland and the fourteen teams in the Finnish major ice hockey league.

the last two quarter-final slots. Teams are paired up for each playoff round according to the regular season standings, so that the highest-ranking team plays against the lowest-ranking, and so on. The playoffs are played best out of seven. The winner of the playoffs receives the Canada Bowl, the championship trophy of the League.

Often there are parties other than the league and the teams involved in the scheduling process. Examples of such parties include TV networks and other leagues. In the case of the Finnish major ice hockey league, the TV network chooses the games to show from the final schedule, thus not affecting the scheduling process.

In the 2008-2009 season two of the teams also played in the Champions Hockey League that was dismantled after only one year. This translated into four additional Wednesday games for both involved teams, Kärpät and Blues. The CHL play-offs resulted in two more games for Blues, but those were not considered at the time of scheduling.

3. Scheduling the Finnish major ice hockey league

Minimizing breaks is very important to the league. There are three main reasons for this: the fans do not like long periods without home games, consecutive home games reduce gate receipts, and long sequences of home or away games might influence the team's current position in the tournament.

The standard game days used to be Tuesday, Thursday and Saturday. From the 2011-2012 season the league decided to change Thursdays to Fridays to get more spectators. Friday games have had about 10% more spectators. However, playing at home both on Fridays and on Saturdays is not allowed. The games that cannot be scheduled on Fridays are played on Thursdays. This on the other hand means that some teams play two consecutive games and some teams have a rest day before the Saturday game. In the last ten seasons the probability for a home team to defeat an away team that has had a rest day is 10% smaller. Likewise, the probability for an away team to defeat a home team that has had a rest day is even 85% smaller.

Some teams desire away tours because of the traveling distances between their venue and some of the opponents' venues. In the last ten seasons the probability for the team to win its second away game has been 30% smaller than to win any away game, which indicates that away tours may be an exchange of decreased sports-related success for financial benefits.

The full real-world instances for the listed seasons along with our solutions are available at Nurmi et al. (2013).

3.1. Pervasive constraints

The following types of constraints and/or goals were present in all the examined seasons.

- Certain teams have a lot of forbidden home game days due to venue unavailability every season.
- The number of breaks must be minimized, while two home games on consecutive days is prohibited.
- The schedule should be as compact as possible, i.e. use as few rounds as possible.
- On some days, no games can be played (holidays etc.).
- Two teams share a venue, so they cannot play at home at the same time.
- The two biggest capital area teams shouldn't play at home at the same time. The third capital area team's home games should coincide equally with the home games of the other two teams.
- The first two rounds should be full, i.e. all teams should play on the same days.
- The northern team should have away tours when visiting the southern teams.
- 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 equal at any point in the tournament.
- Certain games must be played at given dates (celebratory matches etc.).

3.2 Season-specific constraints: 2008-2009

- Two teams participate in the Champions Hockey League.
- The preferred minimum for the number of rounds between two games featuring the same teams is eight.
- One team cannot play at home during the first 13 rounds due to venue renovation.

3.3 Season-specific constraints: 2012-2013

- The preferred minimum for the number of rounds between two games featuring the same teams is six.
- Each team has a preferred number of home games for each weekday.
- January leveling is introduced.
- Back-to-back games are introduced: a full round played on Friday followed by an inverted round one day later, i.e. the home teams of Friday become the away teams on Saturday.

3.4 Season-specific constraints: 2013-2014

- The preferred minimum for the number of rounds between two games featuring the same teams is five.
- The practice of January leveling is continued.
- Back-to-back games are played again.
- Maximize local rival games in the first two rounds.
- Maximize certain local rival games on Fridays and Saturdays.
- The number of away games played on Fridays and Saturdays by the 2 capital area teams that yield the most revenue should be evenly distributed for the home teams.
- Away tours for four additional teams are introduced.

4. The PEAST algorithm and CPLEX

4.1 The PEAST algorithm

The PEAST algorithm is a population-based local search method that has been used to solve several kinds of real-world scheduling problems (Kyngäs, Nurmi and Kyngäs, 2013). It is also in industrial use. The heart of the algorithm is the local search operator called GHCM (greedy hill-climbing mutation) that was first introduced by Nurmi (1998). The GHCM operator is used to explore promising areas in the search space to find local optimum solutions. Another important feature of the algorithm is the use of shuffling operators. They assist in escaping from local optima in a systematic way. Furthermore, simulated annealing (van

Laarhoven and Aarts, 1987) and tabu search (Glover, McMillan and Novick, 1987) are used to avoid staying stuck in the promising search areas too long. No crossover operators are applied to the population of schedules. Every c iterations the least fit individual is replaced with a clone of the fittest individual. The PEAST algorithm uses ADAGEN, the adaptive genetic penalty method introduced by Nurmi (1998). For the detailed discussion we refer to Kyngäs, Nurmi and Kyngäs (2013). The pseudo-code of the algorithm is given in Figure 2. We have used the PEAST algorithm and its predecessors to schedule the league since the 2008-2009 season.

```

Set the iteration limit  $t$ , cloning interval  $c$ , shuffling interval  $s$ , ADAGEN update interval  $a$  and the population size  $n$ 
Generate a random initial population of schedules  $S_i$  for  $1 \leq i \leq n$ 
Set  $best\_sol = null$ ,  $round = 1$ 
WHILE  $round \leq t$ 
     $index = 1$ 
    WHILE  $index++ \leq n$ 
        Apply GHCM to schedule  $S_{index}$  to get a new schedule
        IF  $Cost(S_{index}) < Cost(best\_sol)$  THEN Set  $best\_sol = S_{index}$ 
    END REPEAT
    Update simulated annealing framework
    IF  $round \equiv 0 \pmod{a}$  THEN Update the ADAGEN framework
    IF  $round \equiv 0 \pmod{s}$  THEN Apply shuffling operators
    IF  $round \equiv 0 \pmod{c}$  THEN Replace the worst schedule with the best one
    Set  $round = round + 1$ 
END WHILE
Output  $best\_sol$ 

```

Figure 2. The pseudo-code of the PEAST algorithm.

4.2 Comparison with CPLEX on half of season 2012-2013

IBM ILOG CPLEX is a “high-performance mathematical programming solver for linear programming, mixed integer programming, and quadratic programming” (IBM, 2013) that is often used in academic papers to solve various combinatorial optimization and scheduling problems. We used it as a simple benchmark for PEAST in one real-world instance.

The problem instance originates from the latter half of the season 2012-2013 (30 games per team). There are only 32 rounds available, hence the schedule is quite close to compact. The basis of the schedule is a double round robin with some additional group-based games (see Section 2). The problem is to assign each game to a round, taking into account the following (type of) hard constraints:

- Each team plays at most once on a given round.
- A given team cannot play a home game on a given round.
- A given team cannot play an away game on a given round.
- Two given teams cannot play at home on the same round.
- A given game must be played on a given round.
- No team should have 3 or more home games in a row.
- No team should have 3 or more away games in a row.
- No team should have consecutive byes.

Furthermore, there are 5 soft constraints:

- Breaks are to be avoided. A break incurs one soft constraint violation.
- There should be at least 6 rounds before two teams meet again. If the next encounter is scheduled within $k < 6$ rounds, a total of $6 - k$ soft constraint violations are incurred.
- Each team has a preferred number of home matches for each weekday. For each weekday, the team should play at least this number of home matches. For each team and each weekday for which this threshold is not reached, one soft constraint violation is incurred.
- Two given teams request not to play home games on the same round. For each round where they do play a home game simultaneously, one soft constraint violation is incurred.

- A given game cannot be played before round r . If this game is scheduled on some round $k < r$, a total of $r - k$ soft constraint violations are incurred.

The PEAST algorithm was implemented in Java. The machine used for our PEAST experiments was an Intel Core i7 X980 3.33GHz with 6GB of RAM running Windows 7 Professional Edition and Java 7 Update

Table 1. Results of the PEAST algorithm versus CPLEX. Best bounds are given in parentheses.

Configuration	PEAST			CPLEX		
	Min	Avg	Time	10min	60min	120min
C_1^1	30	37	60min	(19)	30 (20)	27 (20)
C_2^2	50	69	80min	(19)	(19)	(19)
C_3^3	128	148	120min	(36)	(38)	(38)

¹ Hard + breaks.

² Hard + breaks + at least 6 rounds between games between same teams.

³ All constraints.

9. For CPLEX experiments we used IBM ILOG CPLEX 12.3 on a system with 4 Intel Core 2.70 GHz processors and 4GB RAM memory, using a Windows 7 environment. We set the MIP emphasis in CPLEX to feasibility. Even so CPLEX had great trouble producing feasible solutions to the whole problem, so we also experimented on easier versions of the problem. The results are in Table 1. CPLEX was not able to produce a feasible solution for the real-world sport scheduling problem after 2 hours of computation time. The break-related constraints are the most important ones. When considering only these soft constraints (in addition to the hard constraints), CPLEX comes up with a decent solution in a relatively short time, and it manages to improve on this solution continuously with more computation time. CPLEX can also handle adding another soft constraint, except for the one used in configuration C_2 . In practice all the soft constraints given in the problem description are necessary, i.e. it is not possible to consistently produce solutions good enough for practical use without considering all of them. PEAST does this without a problem within reasonable computation time. The results are briefly introduced in Table 1.

5. Some practical considerations

Constructing a single, double or quadruple round robin tournament is quite an easy task nowadays, but when we introduce requirements and requests, the problem becomes intractable. Furthermore, being able to produce an acceptable schedule is not only about first defining the requirements and requests and then developing a suitable solution method. An essential part of the problem is the process of consulting with the various parties.

The process of scheduling the league takes about two months. First, we discuss the possible improvements to the format with the league's competition manager. Then, the format is accepted by the team CEOs. Next, all the restrictions, requirements and requests by the teams are gathered. After the importance (penalty value) of the constraints has been decided, we run the PEAST algorithm for a week with different configurations and choose the best solution, which we send to the league. This results in a good schedule that every involved party should be happy with – in theory. Unfortunately in practice this is not necessarily the case.

In 2007 we interviewed four team CEOs (Jokerit, Kärpät, Tappara and Ässät) and the CEO of the league. The teams were chosen so that we would get a clear picture of the requirements and requests: Jokerit is a big team with a home venue that is used for a lot of other events, Kärpät has to travel the most, Tappara plays at the same venue as another team and Ässät is a good representative of a small team located in a small city.

The interviews gave us quite a good picture of the different requests the teams have and might have. The rest of the teams gave their requests directly to the CEO of the league. He, in turn, sent them to us. At this stage we thought we knew all the requests the teams had.

Every team CEO agreed that it is very important that the requirements and the requests are considered in the final schedule. However, after we generated the first schedule for the 2008-2009 season the team CEOs “discovered” that some of their requests had not been considered. The simple reason was that we were not aware of them. For some reason, not all of the requests had reached us. One reason could be that the team CEO had not actually given the requests to the CEO of the league. Another reason could be that the CEO of the league did not inform us.

Therefore, we had to generate a second schedule for the 2008-2009 season. We were somewhat surprised that the same thing occurred again. We got some new requests, but not all of them. We believed then that the team CEOs are used to getting an unsatisfactory schedule, which they then try to modify to better fit their requests. We based this thought on two things: first, we did not get the requests from the team CEOs, and second, the schedules prior to that season were not so good. We vowed to concentrate on this problem in the future.

However, after six years we still suffer from the same problems to some extent. The basic requests are made known to us before we make the schedule. After we have made the schedule the CEOs of various teams introduce a host of new venue unavailabilities to the final schedule, or even celebratory matches that require a certain opponent on a certain day – i.e. constraints that should have been perfectly clear at the time of the gathering of the constraints. This happens to such an extent that it cannot be an accident – they simply pay no heed to what their changes do to the final schedule. This raises some questions, the foremost ones being whether they even care what the final schedule looks like for their team and if the things they claim to be crucial are actually important at all. In any case the games continually draw in more spectators on average, except for the seasons when the total number of games has been altered (see Table 2).

Table 2. Spectators totals per season with play-offs omitted.

Season	Spectators	Spectators per game
2005-2006	1 958 843	4997
2006-2007	1 943 312	4957
2007-2008	1 964 626	5012
2008-2009	1 997 114	4919
2009-2010	2 015 080	4976
2010-2011	2 036 915	4850
2011-2012	2 145 462	5108
2012-2013	2 189 350	5213

References

- de Werra, D. (1981) Scheduling in Sports. In *Studies on graphs and discrete programming* (Amsterdam and Hansen, Eds.), Netherlands, pp. 381-395.
- Dinitz, J.H., Froncek, D., Lamken, E.R. and Wallis, W.D. (2006) Scheduling a tournament. In *Handbook of Combinatorial Designs* (Colbourn and Dinitz, Eds.), CRC Press, Florida, USA, pp. 591-606.
- Drexl, A. and Knust, S. (2007) Sports league scheduling: graph- and resourcebased models. *Omega* **35**, 465-471.
- Easton, K., Nemhauser, G. and Trick, M. (2001) The traveling tournament problem: description and benchmarks. In *Proc. 7th. International Conference on Principles and Practice of Constraint Programming*, Paphos, pp. 580-584.
- Easton, K., Nemhauser, G. and Trick, M. (2004) Sports scheduling. In *Handbook of Scheduling* (Leung, Ed.) CRC Press, Florida, USA, pp. 52.1-52.19.
- Glover, F., McMillan, C. and Novick, B. (1985) Interactive Decision Software and Computer Graphics for Architectural and Space Planning. In *Annals of Operations Research* **5**, 557-573.
- IBM (2013). IBM ILOG CPLEX Optimizer. [online] Available at: <http://www-01.ibm.com/software/integration/optimization/cplex-optimizer>. [Accessed 28.3.2013].

- Kendall, G., Knust, S., Ribeiro, C.C. and Urrutia, S. (2010) Scheduling in Sports: An annotated bibliography. In *Computers and Operations Research* **37**, 1-19.
- Knust, S. (2012) Sports Scheduling Bibliography. [online] Available at: http://www.inf.uos.de/knust/sportssched/sportlit_class. [Accessed 26.11.2012].
- Kyngäs, N., Nurmi, K. and Kyngäs, J. (2013) Crucial Components of the PEAST Algorithm in Solving Real World Scheduling Problems. In *Proc. 2nd International Conference on Software and Computer Applications*, Paris, France.
- Nurmi, K. (1998) *Genetic Algorithms for Timetabling and Traveling Salesman Problems*. Ph.D. dissertation, Dept. of Applied Math., University of Turku, Finland. Available at: <http://www.bit.spt.fi/cimmo.nurmi>.
- Nurmi, K., Goossens, D., Bartsch, T., Bonomo, F., Briskorn, D., Duran, G., Kyngäs, J., Marenco, J., Ribeiro, C.C., Spieksma, FCR., Urrutia, S. and Wolf-Yadlin, R. (2010) A Framework for Scheduling Professional Sports Leagues. In *IAENG Transactions on Engineering Technologies Volume 5* (Ao, Katagiri, Xu and Chan, Eds.), Springer, USA.
- Nurmi, K. et al. (2013) Sports Scheduling Problem. [online] Available at: <http://www.samk.fi/ssp>.
- Rasmussen, P. and Trick, M. (2008) Round robin scheduling - A survey. In *European Journal of Operational Research* **188**, 617-636.
- Thielen, C. and Westphal, S. (2011) Complexity of the traveling tournament problem. In *Theoretical Computer Science* **412(4-5)**, 345-351.
- van Laarhoven, P.J.M. and Aarts, E.H.L. (1987). *Simulated annealing: Theory and applications*, Kluwer Academic Publishers.

Scheduling a Sports League with Divisional and Round-robin Play

Jeffrey Larson* and Mikael Johansson**

*School of Electrical Engineering, KTH, Stockholm, Sweden, email address: jeffreyl@kth.se

**School of Electrical Engineering, KTH, Stockholm, Sweden, email address: mikaelj@kth.se

Abstract. Sports leagues may wish to increase their season length by adding games to their double round-robin tournament. One possibility, chosen by the top Swedish handball league Elitserien, is to split the teams into two divisions and have each hold an additional single round-robin tournament to start the season. This introduces new concerns since some teams meet three times during the season while others only meet twice. In this paper, we analyze and generate schedules which satisfy Elitserien's concerns. Since these concerns are common to many leagues, our analysis should also be useful in other situations. We enumerate the number of minimum break home/away pattern sets that satisfy the requirements, not all of which are schedulable. To reduce the search space further, we propose a sequence of necessary conditions to remove unschedulable home/away pattern sets. Lastly, we discuss the final steps of assigning teams to the numbers of a tournament template; such an assignment is being used to construct the 2013-14 Elitserien schedule.

1. Introduction

One of the most common sports league schedules is a double round-robin tournament (DRRT) where each pair of teams plays twice, once at each team's home venue. Such leagues may consider expanding their schedule for a variety of reasons, including increasing league exposure and revenue. In such a situation, many options exist. Adding additional teams to the league is a possibility, but finding team owners and talented players can make league expansion difficult. Extending the season to a triple round-robin tournament (TRRT), as has been done in the top Danish soccer league Rasmussen [2008], may introduce too many games. In this paper, we address a third option: splitting the league into two divisions which will hold concurrent, single round-robin tournaments (SRRTs) in addition to the existing DRRT. (Significant work has been devoted to designing sports schedules; an overview of which is beyond the scope of this work. For a good review, see Briskorn [2008b], Kendall et al. [2010] and the amazingly well-maintained website Knust [2013].)

Elitserien, the top Swedish handball league, is one such league. The 14 team owners considered its traditional 26-game DRRT to have too few games, while a possible 39-team TRRT was deemed too long. The schedule was therefore augmented by dividing teams into two seven-team divisions (based on geographic proximity) and having each play an intra-division SRRT before the traditional DRRT. The league's schedule is separated into three parts: Part I (periods 1-7) consists of two concurrent, intra-division SRRTs; Part II (periods 8-20) consists of an RRT between all teams, and Part III (periods 21-33) is the mirror of Part II. Therefore, teams in the same division meet three times, once in each part; teams in different divisions would meet in Part II and Part III. This, of course, yields new concerns about the schedule. In addition to previously studied constraints (e.g., minimizing consecutive home or away matches, called *breaks*; ensuring equal number of home and away games; etc.) the league wants to insure that any consecutive meetings between pairs of teams alternated venues. We call this constraint the *Alternating Venue Requirement* (AVR). For two teams from separate divisions, this was readily ensured since Part III mirrors Part II. For teams in the same division, no obvious solution exists.

Many sports leagues have formed divisions of teams who meet each other more often than other teams. To the best of our knowledge, no other league has a format where a divisional RRT precedes a DRRT to form a complete season. Nevertheless, there exists some pertinent research relating to groups within an RRT. Though the groups considered in the literature are often *strength groups* (teams of similar ability), such results could obviously be applied to groups constructed based on geography. For example, Briskorn [2009], Briskorn and Knust [2010] analyze schedules where no team plays against a team from the same group in consecutive matches (*group-changing*) and schedules where no team plays a team from the same group within a stretch of p games (*group-balanced*). The combinatorial properties of these strength groups are summarized in Briskorn, [2008b], Chapter 4. Multiple strength groups are considered in van't Hof

et al. [2010] where they construct a minimum break SRRT, where some teams never play at home during the same period (complementary schedules), and teams from different strength groups are sufficiently spread throughout the tournament. Ensuring desirable home-away patterns when playing teams from different groups has yet to be addressed.

In this paper we present a methodology for scheduling general tournaments where each division holds an RRT before a season long DRRT through a specific example, the Swedish Elitserien. We first show that it is always possible to construct a tournament satisfying the Alternating Venue Requirement (if we relax other requirements). Then, we construct home-away patterns (HAPs) for each team such that the league can always be scheduled (if we relax the Alternating Venue Requirement). A particular effort is devoted to constructing and characterizing HAP sets where a maximal number of pairs of teams never play at home during the same period. We further present some necessary conditions for a HAP set to be scheduled in a manner satisfying the Alternating Venue Requirement. This allows us to reduce the number of possible HAP sets to a reasonable number, which, after a quick search, locates a schedule satisfying all of the league's requirements. The results are presented as a tournament template, a full-season schedule with generic team numbers rather than specific teams (e.g., Team 1 plays Team 2 in period 4). To form a schedule each year, we show how teams can be assigned numbers to further increase the schedule's attractiveness. While we only address one league, we consider the league requirements to be reasonable; we hope this makes the problem and its solution relevant to any other league looking to add games to a DRRT schedule without including additional teams in the league.

The outline of the paper follows: In Section 2, we define the constraints of the Elitserien schedule and justify the use of HAP sets to construct a solution. We construct tournament templates satisfying the AVR in Section 3.1 (possibly not satisfying other requirements) and construct schedulable HAP sets in Section 3.2 (possibly not satisfying the Alternating Venue Requirement). In Section 3.3, we outline necessary conditions for when a HAP set from Section 3.2 will satisfy the AVR and then search over the remaining HAP sets to construct a tournament template satisfying every league requirement. Lastly, in Section 4 we outline the league's concerns when assigning teams to the template. Section 5 concludes the paper.

2. Problem Statement

The requirements on the Elitserien schedule can be broadly classified into two categories: the first addresses schedule structure and fairness in terms of breaks, periods without games (called *byes*), and the sequence of home and away games; the second concerns stadium and referee availabilities, the desire to support various match-ups (such as rivalries), and wishes from the media. Historically, Elitserien has determined their schedule by first proposing a tournament template which addresses the fairness constraints. This tournament template has numbers in place of actual teams in the schedule. Once the clubs have agreed to the template, it can be used for many years. Every year, the league collects information about unavailabilities and particular wishes from the clubs and assigns teams to numbers in the tournament template to form the season schedule.

We find the intermediate process of constructing a tournament template to be a useful step. In a relatively straightforward fashion, one can convey the strengths and weaknesses of a schedule in a single table without involving any actual team names. This generality allows individuals to articulate what they desire in a schedule more easily than when team names (and memories/biases that such names invoke) are involved. The template also provides clarity when analyzing properties that can be difficult to optimize over. For example, we find that complementary schedules can be more easily recognized when analyzing a template with numbers than when inspecting a schedule with team names.

The assignment of actual teams to the numbers of the tournament template is rather straightforward and will be dealt with in Section 4. The majority of this paper addresses the more interesting and challenging problem of constructing and characterizing tournament templates which satisfy the Elitserien requirements. We will now state this problem in detail by declaring the Elitserien schedule requirements in Table 1.

Table 1: Elitserien schedule requirements.

-
1. Each 7-team division must hold a SRRT to start the season.
 2. This must be followed by two SRRTs between the entire league, the second SRRT being a mirror of the first.
 3. There must be a minimum number of breaks in the schedule.
 4. Each team has one bye during the season (to occur during the divisional RRT).
 5. At no point during the season can the number of home and away games played by a team differ by more than 1.
 6. Any pair of teams must have consecutive meetings occur at different venues.
 7. Each division must have 3 pairs of complementary schedules.
-

Before constructing tournament templates, we first provide a lower bound on the minimum number of breaks required to schedule a league such as Elitserien.

Proposition 2.1 *In an n-team league (n even) with a schedule consisting of two concurrent divisional RRTs followed by two consecutive full-league RRTs, if only one bye is allowed and it must occur during the divisional RRT, any schedule must have at least $2n - 4$ breaks.*

Proof: Since we are looking to minimize breaks in the schedule two results apply. First, de Werra [1981] proves an n-team RRT (n even) must have at least $n - 2$ breaks and also constructs schedules meeting this lower bound. Second, Froncek and Meszka [2005] construct unique RRT with one bye and no breaks (and also shows such RRTs are unique). Therefore the divisional RRTs can be scheduled without any breaks, since we allow a bye in Part I, and the two full-league RRTs must each have $n - 2$ breaks, resulting in a total of $2n - 4$ breaks.

We will show that it is possible to construct schedules achieving this lower bound. That is, one can transition from one tournament to another without introducing breaks. Before doing so, we wish to justify the use of HAP sets to solve this problem. Since Froncek and Meszka [2005] proves the HAP set for the divisional tournament is unique, we only must show that the 14-team HAP set is unique if 1) the difference between total home and away games is at most 1 during any period of the season (so breaks must occur during odd weeks), and 2) the schedule has $n - 2$ breaks, with each team having at most one break. See Figure 1.

BAAHAAH
HBAHAAA
AHBAHAH
HAHBAHA
AHAHBAH
HAHAHBA
AHAHAHB
Or
BHAHAAA
ABHAHAAH
HABHAAA
AHABHAAH
HAHABHAA
AHAHABH
HAHAHAB

(a) Two unique HAP sets for a 7-team, zero-break RRT.

AHAHAHAHAHAHA
AHAHAHAHAHAHH
AHAHAHAHAHHAH
AHAHAHAHHAHAAH
AHAHAHHAHAAHAAH
AHAHHAHAAHAAH
AHHAAHAAHAAHAAH
HAAHAHAAHAAHAA
HAHAAHAAHAAHAA
HAAHAAHAAHAAH
HAAHAAHAAHAAHA
HAAHAAHAAHAAHAA
HAAHAAHAAHAAHAA

(b) Unique HAP set satisfying the Elitserien requirements for a 14-team, 12-break R

Figure 1: The basic home-away patterns used to construct the schedule.

3. Constructing Tournament Templates

We first show that we can construct a HAP set which can be scheduled as a tournament satisfying every constraint above, except for possibly the Alternating Venue Requirement. We then show that we can construct a tournament satisfying the AVR, but possibly not the other constraints. We lastly demonstrate some necessary conditions which remove many HAP sets as unschedulable.

3.1 Alternating Venue Schedulability

Proposition 3.1 *It is always possible to construct a schedule with satisfies the AVR (which might not satisfy other requirements).*

Proof: If we are only looking to satisfy the Alternating Venue Requirement, we can construct a tournament in the following fashion. Take the unique 1-bye tournament on $n/2$ teams for one division and its mirror for the other division. This will be Part I of the full schedule. Let the first $n/2$ periods of Part II be the mirror of Part I, except if a team is scheduled for a bye. In that case, pair it with the team from the other division also scheduled for a bye (assigning home arbitrarily). Complete the remaining $n/2 - 1$ matches of Part II by cycling through the remaining teams, assigning home arbitrarily. Lastly, let all periods of Part III be a mirrors of the same period in Part II.

An example of one such tournament for $n = 6$ is in Table 2. In Table 2, a positive (negative) j entry (i,p) denotes team i plays at home (away) against team j in period p .

Table 2: A tournament template satisfying the Alternating Venue Requirement

0	3	-2	4	-3	2	-5	6	-4	3	-2	5	-6
-3	0	1	3	5	-1	6	-4	-3	-5	1	-6	4
2	-1	0	-2	1	-6	4	-5	2	-1	6	-4	5
0	-6	5	-1	6	-5	-3	2	1	-6	5	3	-2
6	0	-4	-6	-2	4	1	3	6	2	-4	-1	-3
-5	4	0	5	-4	3	-2	-1	-5	4	-3	2	1

3.2 Home-Away Pattern Set Construction

For each team, we can take one row of Figure 1(a), append one row of Figure 1(b), and then append the reflected, mirrored copy of the same row from Figure 1(b) to yield a HAP for a given team. An example of a possible HAP set for the Elitserien is given in Table 3.

Table 3: One possible HAP set.

B	H	A	H	A	H	A	H	A	H	A	H	A
H	A	B	H	A	H	A	H	A	H	A	H	A
H	A	H	A	B	H	A	H	A	H	A	H	A
H	A	H	A	H	A	B	H	A	H	A	H	A
A	B	H	A	H	A	H	A	H	A	H	A	H
A	H	A	B	H	A	H	H	A	H	A	H	A
A	H	A	H	A	B	H	A	H	A	H	A	H
B	A	H	A	H	A	H	A	H	A	H	A	H
A	H	B	A	H	A	H	H	A	H	A	H	A
A	H	A	H	B	A	H	A	H	A	H	A	H
H	B	A	H	A	H	A	H	A	H	A	H	A
H	A	H	B	A	H	A	H	A	H	A	H	A
H	A	H	A	H	B	A	H	A	H	A	H	A

This schedule has many attractive properties. Taking the unique no-break, 7-team tournament HAP set and its mirror ensures that 7 teams play at home and 7 teams play away in period 8 without introducing a break. If we did not take the mirror, we would have 8 teams needing to play at home in period 8, an impossibility. Since we are reflecting and mirroring Part II to schedule Part III, and breaks only occur during odd periods (to ensure the cumulative home and away games never differ by more than 1 at any point in the season), there are no breaks in period 9. This implies no breaks to end the season; in other words, each team plays at home one of the last two periods of the season.

Even though the HAP sets generated in this fashion have some appealing features, one cannot guarantee they can be scheduled so that the Alternating Venue Requirement is satisfied. In fact, there is no way to schedule the HAP set in Table 3 so the AVR is satisfied. To see this, notice that it is impossible for team 13 to host team 14 in Part I or Part II of the schedule.

At first glance, the reflecting and mirroring Part II to form Part III forces teams to play the same team in periods 20 as they do in period 21 (at the opposite venue). This could be undesirable, depending on the league, but it is a non-issue for the Elitserien. Period II ends before Christmas, allowing for a month-long break for Champions League competitions before Period III starts at the beginning of February.

3.3 Maximally Complementary Home-Away Pattern Sets

Any HAP set created with one row of Figure 1(a), one row from Figure 1(b), and that same row of Figure 1(b) reflected mirror can be scheduled satisfying the constraints from the Elitserien, except for (possibly) the Alternating Venue Requirement. Since there are $14!$ (over 87 billion) possible combinations of rows from Figure 1, a search over all of them to find a HAP set satisfying the AVR would be prohibitively expensive. That is, we can form an integer program to assign games to the HAP set and see if a feasible solution can be found which satisfies the AVR. Of course, calling an integer program for 87 billion HAP sets is impractical computationally. Luckily, we can rule out many of the $n!$ combinations *a priori*, for example by ensuring there are many complementary HAPs within each division.

When $n/2$ is even (resp. odd), this means we desire $n/4$ (resp. $(n-2)/4$) pairs of complementary teams in each division. The following two propositions enumerate the number of HAP sets which satisfy this requirement.

Proposition 3.2 *For an n-team tournament, n/2 even, with a divisional RRT before full-league DRRT, there are $n/2!$ unique HAP sets satisfying the list of requirements in Table 1, except for possibly the AVR, with $n/4$ pairs of complementary schedules within each division.*

Proof: When $n/2$ is even, and each division must have $n/4$ pairs of complementary schedules in Parts I, II, and III, the analysis is straightforward. Each Part I HAP is complementary with only one other HAP within its division. Therefore, once any of the Part II HAPs is appended to any Part I HAP, the complementary pair is uniquely determined. There are $n/2!$ ways to assign each complementary pair from Part II to any pair from Part I. Therefore, there are $n/2!$ possible HAP sets.

Proposition 3.3 *For an n-team tournament, n/2 odd, with a divisional RRT before full-league DRRT, there are*

$$\frac{n}{2} P_{\frac{n-2}{4}} \times \left(\frac{n+2}{4}\right)^3 \times \frac{n-2}{4}! \quad (1)$$

unique HAP sets satisfying the list of requirements in Table 1, except for possibly the AVR, with $(n-2)/4$ pairs of complementary schedules within each division.

Proof: It is always possible to order the divisional HAP sets so each HAP is complementary with the team above and below it. (See Figure 1(a) as an example for $n = 14$.) The first and last HAPs are complementary, but both end with the same type of game (home in the top HAP set, away in the bottom HAP set of

Figure 1(a)). They therefore cannot be made complementary throughout the season without introducing additional breaks into the schedule.

If a division will have $(n - 2)/4$ pairs of complementary teams, the team without a complementary counterpart inside the division must be numbered odd since the first and last divisional HAPs cannot be completed to complementary HAPs for the entire season. In a 7-team division for example, if 3 is not complementary with anyone, then 1 must be complementary with 2, 4 with 5, and 6 with 7.

So there are $(n + 2)/4$ unique ways in which the Part I HAP set can start with $(n - 2)/4$ pairs of complementary teams. This start must be continued in Part II for the schedule to be completely complementary. Say team 1 is the team “left unpaired.” Each HAP in Part II has a complementary counterpart. (See Figure 1(b) as an example for $n = 14$.) So there are $n/2$ pairs of complementary teams, so

$$\binom{\frac{n}{2}}{\frac{n-2}{4}}$$

ways to assign these to the pairs $(2,3), \dots, (n/2 - 1, n/2)$, each with $(n - 2)/4!$ ways to permute them. A more succinct way is to just say there are $\frac{n}{2}P_{(n-2)/4}$ ways to permute the $(n - 2)/4$ pairs of the $n/2$ total complementary pairs. There are $n - (n - 2)/4$ Part II patterns remaining to assign to team 1, $(n + 2)/4$ of which begin with away. There are therefore

$$\frac{(n + 2)}{4} \times \frac{n}{2}P_{(n-2)/4} \times \frac{(n + 2)}{4}$$

ways to generate $(n - 2)/4$ pairs of complementary HAP sets for the first division.

For the second division, again the unpaired team must be odd (for the same reason as described above). There are $(n + 2)/4$ ways do this, and then $[(n - 2)/4]!$ ways to permute the remaining pairs of complementary Part II HAPs. Therefore, the number of HAP sets with $(n - 2)/4$ pairs of complementary teams in each division is:

$$\frac{n}{2} \frac{P_{n-2}}{4} \times \left(\frac{n + 2}{4}\right)^3 \times \frac{n - 2}{4}!$$

(By design, the odd teams in the second division end in away games, whereas the odd teams in the first end in home games. This ensures that no break is introduced when assigning the “unpaired” second division team the counterpart to the Part II HAP from the “unpaired” first division team.)

For the Elitserien with $n = 14$, Proposition 3.3 leaves 80640 HAP sets which satisfy the league’s schedule requirements, except for possibly the AVR. Next, we will demonstrate that also accounting for the AVR allows us to rule out even more HAP sets.

3.4 Necessary Conditions for AVR Schedulability

Several researchers have attempted to address when a schedule can be assigned to a HAP set. A necessary condition for general round-robin tournaments, proposed in Miyashiro et al. [2003], is that there must be “enough opportunities” for any subset of teams to play each other. Explicitly,

$$\sum_{p \in P} \min(c_A(T', p), c_H(T', p)) \geq \binom{|T'|}{2} \quad (2)$$

where c_A (resp. c_H) counts the number of away (resp. home) games in the HAPs for teams in the subset T' in period p . Therefore, the left side of (2) counts the number of possible matches that could possibly be played between teams in T' , and the right side of (2) is the number of necessary matches between teams in T' . Using condition (2), we see that the 6-team HAP set Table 4 cannot be scheduled. (Let $T' = \{1, 5, 6\}$). It should be noted that the HAP sets in Figure 1 are schedulable and therefore satisfy (2).

Table 4: A possible HAP set that is detected to be unschedulable using (2).

Team 1	AHAHA
Team 2	AAAH
Team 3	AHHAH
Team 4	HAHAA
Team 5	HHAHA
Team 6	HAAHA

In the following, we propose some possible generalizations of the necessary condition (2) to account for the Alternating Venue Requirement. For example, since all pairs of teams in the same division must play home-away or away-home in Parts I and II, an obvious modification of (2) is:

$$\sum_{p \in \{\text{Part I}\} \cup \{\text{Part II}\}} \min(c_A(T', p), c_H(T', p)) \geq 2 \binom{|T'|}{2}, \quad T' \subseteq \{\text{Division 1}\} \text{ or } T' \subseteq \{\text{Part II}\}.$$

That is, any subset of teams in the same division must be able to meet the required number of times. Our HAP sets satisfy this condition by design, so we do not need to check it. Another necessary condition would be that there are “enough” home and away games to satisfy the AVR requirement. For an arbitrary HAP set S , define

$$S(t, p) = \begin{cases} H: & \text{if team } t \text{ plays home in period } p, \\ A: & \text{if team } t \text{ plays away in period } p, \\ B: & \text{if team } t \text{ has a bye in period } p. \end{cases}$$

Then for S to be schedulable, for any two teams t_1 and t_2 in the same division, there must be two periods p_1 in Part I and p_2 in Part II such that

$$\begin{aligned} S(t_1, p_1) &= H \quad \text{and} \quad S(t_2, p_1) = A \\ S(t_1, p_2) &= A \quad \text{and} \quad S(t_2, p_2) = H \end{aligned} \tag{3}$$

It should be noted that HAP set in Table 3 does not satisfy this requirement. We can therefore determine it is unschedulable.

Since the divisional tournament (who plays whom in each period) is unique, we can improve the pairwise bound by specifying which type of games are required in the Part I RRT. This simple condition allows to rule out 30720 of the 80640 HAP sets (roughly 39%) generated by accounting for the other tournament template requirements. Table 5 demonstrates the efficiency of the condition for other league sizes.

Table 5: Measure of efficiency of a simple necessary condition.

n	HAP Sets from (1)	HAP sets removed by simple condition (3)	% removed
6	24	8 (of 20 unschedulable)	40%
10	1080	396 (of 998 unschedulable)	$\approx 40\%$
14	80640	30720 (of 79024 unschedulable)	$\approx 39\%$

In addition to this simple condition, we can check if i or j is already “committed” to play another team in every period when they could possibly meet. This is only slightly more expensive computationally to check than condition (3), but it catches many “deeper” contradictions. For example, if for a given HAP set, i

can only play j in periods p_1 or p_2 , but i must play k_1 in p_1 and j must play k_2 in p_2 , this test will determine such a HAP set is unschedulable. This condition allows us to remove 46944 of the 80640 HAP sets (59%).

Rather than dealing with conditions on the HAP set, we can instead build an $n \times n$ array, where each entry (i, j) is a vector of periods when it is possible for teams i and j to meet. We can then use various logical arguments to update this array. This reduces the problem of determining if a HAP set can be scheduled so it satisfies the AVR into the problem of completing a Latin square. We can then check the following logical conditions iteratively to determine if a HAP set is unschedulable.

- If (i, j) has only one entry, remove that value (if possible) from any vector (i, k) , $k \neq j$ and any vector (k, j) , $k \neq i$.
- If (i, j) has more than one entry, see if any value is unique in a row or column. Replace (i, j) by that value.
- Stop if any (i, j) is empty, or no change is observed after checking the above two conditions for all (i, j) .

Checking all of these conditions can result in three possibilities. If an entry (i, j) of the Latin square becomes an empty vector, then there is no feasible period for team i to meet team j and the HAP set can be ruled unschedulable. If every entry of the Latin square becomes a 1-dimensional vector, then the corresponding HAP set is schedulable and we have discovered unique tournament template. Lastly, if the Latin square remains unchanged after checking every entry (i, j) , the HAP set cannot be ruled unschedulable.

This problem formulation has many advantages, namely that we can often determine when HAP set is unschedulable. Even if the logic does not decide a HAP set to be infeasible, further constraints on the schedule are nearly always discovered (e.g., when attempting to see if a HAP set can be scheduled to satisfy the AVR, it is determined that i must play j in some period p_1). This information can then be used to greatly reduce the search space for an integer program attempting to schedule a HAP set. Sadly, the above logic is insufficient for determining if a Latin square has a completion; in fact, the problem in general is shown to be NP-complete Colbourn [1984].

It is not surprising that this Latin square approach is both powerful and time consuming. In Table 6, we see that such a condition is able to remove almost all unschedulable HAP sets from the search space for the 14-team case, but the time required to check all 80640 HAP sets is approximately 1 day in MATLAB. This method still has the ability to quickly remove many HAP sets if we do not attempt to completely determine a given Latin square is unschedulable. If we instead only check the logical conditions for each entry (i, j) of the Latin square one time (rather than revisiting all previous entries whenever an entry of the Latin square changes), we can still determine many HAP sets to be unschedulable. This relaxation requires approximately an hour in MATLAB for the 14-team case, but still detects a large portion of unschedulable HAP sets, as seen in Table 7.

Table 6: Measure of efficiency of the complete Latin square condition.

n	HAP Sets from (1)	HAP sets removed by complete Latin square condition.	% removed
6	24	20 (of 20 unschedulable)	100%
10	1080	998 (of 998 unschedulable)	100%
14	80640	75995 (of 79024 unschedulable)	$\approx 96\%$

Table 7: Measure of efficiency of simplified Latin square condition.

n	HAP Sets from (1)	HAP sets removed by simplified Latin square condition.	% removed
6	24	10 (of 20 unschedulable)	50%
10	1080	504 (of 998 unschedulable)	$\approx 51\%$
14	80640	51946 (of 79024 unschedulable)	$\approx 66\%$

Table 8: One possible tournament template satisfying the requirements of Elitserien.

0	-2	3	-4	5	-6	7	-8	9	-5	6	-7	10	-11	12	-13	4	-14	2	-3	3	-2	14	-4	13	-12	11	-10	7	-6	5	-9	8
-7	1	0	-3	4	-5	6	-9	7	-12	10	-6	5	-4	11	-14	13	3	-1	8	-8	1	-3	-13	14	-11	4	-5	6	-10	12	-7	9
-6	7	-1	2	0	-4	5	-10	8	-7	9	-5	11	-13	6	-12	14	-2	4	1	-1	-4	2	-14	12	-6	13	-11	5	-9	7	-8	10
-5	6	-7	1	-2	3	0	-11	10	-6	5	-13	12	2	-14	8	-1	9	-3	7	-7	3	-9	1	-8	14	-2	-12	13	-5	6	-10	11
4	0	-6	7	-1	2	-3	13	-14	1	-4	3	-2	8	-7	9	-10	-11	6	-12	12	-6	11	10	-9	7	-8	2	-3	4	-1	14	-13
3	-4	5	0	-7	1	-2	12	-13	4	-1	2	-8	14	-3	7	-9	10	-5	-11	11	5	-10	9	-7	3	-14	8	-2	1	-4	13	-12
2	-3	4	-5	6	0	-1	14	-2	3	-8	1	-9	-10	5	-6	11	-12	13	-4	4	-13	12	-11	6	-5	10	9	-1	8	-3	2	-14
0	9	-10	11	-12	13	-14	1	-3	-9	7	-11	6	-5	10	-4	12	-13	14	-2	2	-14	13	-12	4	-10	5	-6	11	-7	9	3	-1
14	-8	0	10	-11	12	-13	2	-1	8	-3	-10	7	-12	13	-5	6	-4	11	-14	14	-11	4	-6	5	-13	12	-7	10	3	-8	1	-2
13	-14	8	-9	0	11	-12	3	-4	14	-2	9	-1	7	-8	-11	5	-6	12	-13	13	-12	6	-5	11	8	-7	1	-9	2	-14	4	-3
12	-13	14	-8	9	-10	0	4	-12	13	-14	8	-3	1	-2	10	-7	5	-9	6	-6	9	-5	7	-10	2	-1	3	-8	14	-13	12	-4
-11	0	13	-14	8	-9	10	-6	11	2	-13	14	-4	9	-1	3	-8	7	-10	5	-5	10	-7	8	-3	1	-9	4	-14	13	-2	-11	6
-10	11	-12	0	14	-8	9	-5	6	-11	12	4	-14	3	-9	1	-2	8	-7	10	-10	7	-8	2	-1	9	-3	14	-4	-12	11	-6	5
-9	10	-11	12	-13	0	8	-7	5	-10	11	-12	13	-6	4	2	-3	1	-8	9	-9	8	-1	3	-2	-4	6	-13	12	-11	10	-5	7

3.5 From HAP Set to Tournament Template

Once a large number of HAP sets have been ruled out, it remains to populate the sets with games to form the tournament template. This can be done either by integer programming as in Briskorn [2008b], attempting to find the best schedule with respect to a certain metric (*e.g.* the schedule with the best group-balance) or by a simple search. A representative tournament template is shown in Table 8.

4. Assigning Teams to Numbers

The final step of the scheduling process is to assign actual teams to the numbers used in the template(s). This step accounts for a number of factors in a priority order agreed by the Swedish Handball Association and representatives from the Elitserien clubs. Such an assignment of teams to templates is currently being used to construct the 2013-14 Elitserien season schedule.

Currently, the most important requirement is stadium and referee availability. Luckily for Elitserien, such constraints are rarely “hard.” This is due to the fact that each period of play where teams are scheduled to meet spans multiple days. For example, if two teams are scheduled to meet in period 7, this means that the teams must play on one of several possible game dates (*e.g.* Friday, Saturday, or Sunday). In the rare occasion that a venue is occupied every day of a period p_k or no team or referees will be available during p_k , it is possible to ensure that the relevant teams are not assigned to certain numbers in the template. More often than not, teams have “soft” preferences (collected by the league before the season) where they would prefer not to play at home.

Equally important is to assign teams that share the same stadium to numbers with complementary HAP-sets. In the current Elitserien, there are two pairs of teams that must be assigned complementary HAP-sets.

Next in priority, the league wishes to support wishes and initiatives from the clubs which raise the visibility of handball. This could include the preference of a club to play home in a certain period to inaugurate a new stadium or in connection with an important local event (*e.g.*, a large youth tournament in the area). This could also be the desire to organize derby games. In addition to standard derbies, the Elitserien has earlier accommodated events such as the “Battle of Scania”, where the four Elitserien teams from the region of Scania play each other on the same day.

The last priority concerns placing certain “desired games” throughout the schedule. On the one hand, there is a wish to spread interesting games over the season, to make sure that there are good games to air on TV each week. On the other hand, there is also a wish to schedule many games among the teams that are likely to fight for playoff positions to occur late in the season, so that the league remains undecided for as long as possible. By a similar argument, games between teams that are likely to fight against relegation should also be scheduled late in the season.

Explicitly addressing all of these concerns can be somewhat cumbersome. For the interested reader, we formally declare the model which we use to assign teams to numbers in the Appendix.

5. Conclusion

In this paper, we analyzed the situation where a league augments a traditional DRRT schedule by forming two divisions of teams, each of which hold an SRRT to start the season. This asymmetry (pairs of teams play three times if they are in the same division, twice otherwise) makes constructing feasible schedules an interesting problem. We highlighted the concerns of Elitserien, which we consider to be general enough to apply to many other leagues, and enumerated the number of HAP sets which satisfy these concerns. We next constructed necessary conditions for a HAP set to satisfy the Alternating Venue Requirement; this allowed us to remove many unschedulable HAP sets from the search space. After constructing a schedule template, we finally highlighted the principles for how we assign teams to the template to form the 2013-14 Elitserien schedule.

Acknowledgments

We want to thank Lars Westman with the Swedish Handball Association for sharing his extensive experience of scheduling the Elitserien. His hard work in carefully outlining the league's concerns ensured that this research was useful in practice.

Appendix - IP model

To solve this problem, we can form an IP model of small enough size to be easily solved for any size league (at most, a few hundred binary variables). Let T be the set of teams (indexed by t), let P be the set of periods (indexed by p), and let N be the set of numbers in the template (indexed by n). Let T_1 (T_2) be the first (second) division of teams and let N_1 (N_2) be the first (second) group of numbers, each containing 3 pairs of complementary teams. Let $M_{p,1}, \dots, M_{p,n/2}$ each be the ordered pair of numbers which play during period p . (E.g., $M_{20,1} = (3,8)$ if numbers 3 and 8 play in period 20). Lastly, let $D(t_1, t_2, p)$ be 1 if team t_1 playing team t_2 in period p is desired.

- **Variables:**

1. A variable to determine which number each team is assigned to.

$$x_{tn} = \begin{cases} 1: & \text{if team } t \text{ is assigned to number } n, \\ 0: & \text{otherwise,} \end{cases} \quad \forall t \in T, n \in N.$$

2. A variable to determine which group of numbers each division is assigned to.

$$\Delta = \begin{cases} 1: & \text{if team } T_1 \text{ is assigned to the group of numbers number } N_1, \\ 0: & \text{otherwise.} \end{cases}$$

3. A variable to count the number of soft violations.

$$y_{tn} = \text{The number of soft violations incurred from assigning } t \text{ to } n, \quad \forall t \in T, n \in N.$$

4. A variable indicating if a desired game occurs during a given period.

$$\delta(M_{p,k}, p) = \begin{cases} 1: & \text{if matchup } M_{p,k} \text{ in period } p \text{ is desired,} \\ 0: & \text{otherwise,} \end{cases} \quad \forall M_{p,k}, p \in P.$$

- **Parameters:**

1. A parameter denoting hard venue unavailabilities.

$$A_{tp} = \begin{cases} 0: & \text{if venue } t \text{ is unavailable in period } p, \\ 1: & \text{otherwise,} \end{cases} \quad \forall t \in T, p \in P.$$

2. A parameter denoting soft venue unavailabilities.

$$S_{tp} = \begin{cases} 0: & \text{if venue } t \text{ would prefer not to host in period } p, \\ 1: & \text{otherwise,} \end{cases} \quad \forall t \in T, p \in P.$$

3. A numerical value for HAP set entries.

$$H_{np} = \begin{cases} 1: & \text{if number } n \text{ plays at home during period } p, \\ 0: & \text{if number } n \text{ has a bye during period } p, \\ -1: & \text{if number } n \text{ plays at away during period } p, \end{cases} \quad \forall n \in N, p \in P.$$

- 4. Define $\alpha \in (0,1)$ as the trade-off between minimizing soft conflicts and maximizing desired games.
- **Constraints:**
 1. Ensure each number is assigned a team.

$$\sum_t x_{tn} = 1 \quad \forall n \in N.$$

2. Ensure each team is assigned a number.

$$\sum_n x_{tn} = 1 \quad \forall t \in T.$$

3. Ensure teams in T_1 are in the same subgroup of numbers.

$$\sum_{t \in T_1} \sum_n x_{tn} \leq \Delta|N|.$$

4. Ensure hard venue unavailabilities are not violated.

$$\sum_n x_{tn} H_{np} \leq A_{tp} \quad \forall t \in T, p \in P.$$

5. Ensure y_{tn} counts the soft violations of assigning t to n .

$$y_{tn} = \sum_p (H_{np} S_{tp})^+ \quad \forall t \in T, n \in N.$$

6. Ensure $\delta(M_{p,k}, p)$ indicates when matchup $M_{p,k}$ is desired.

$$\begin{aligned} x_{t_1n_1} + x_{t_2n_2} + x_{t_1n_2} + x_{t_2n_1} &\leq 2\delta(M_{p,k}, p) \quad \forall M_{p,k}, p \in P, \text{ such that } D(t_1, t_2, p) = 1, \\ x_{t_1n_1} + x_{t_2n_2} + x_{t_1n_2} + x_{t_2n_1} &\geq 2\delta(M_{p,k}, p) \quad \forall M_{p,k}, p \in P, \text{ such that } D(t_1, t_2, p) = 1, \end{aligned}$$

- **Objective Function**

$$\text{minimize } (1 - \alpha) \sum_t \sum_n x_{tn} y_{tn} - (1 - \alpha) \sum_p \sum_k \delta(M_{p,k}, p).$$

References

- Briskorn, D. (2008a) Feasibility of home-away-pattern sets for round robin tournaments. *Operations Research Letters*, **36**(3):283–284.
- Briskorn, D. (2008b) *Sports Leagues Scheduling*, volume 603 of *Lecture Notes in Economics and Mathematical Systems*. Springer Berlin, Heidelberg.
- Briskorn, D. (2009) Combinatorial properties of strength groups in round robin tournaments. *European Journal of Operational Research*, **192**(3):744–754.

- Briskorn, D. and Knust, S. (2010) Constructing fair sports league schedules with regard to strength groups. *Discrete Applied Mathematics*, **158**(2):123–135.
- Colbourn, C.J. (1984) The complexity of completing partial Latin squares. *Discrete Applied Mathematics*, **8**(1):25–30.
- de Werra, D. (1981) Scheduling in Sports. In P. Hansen, editor, *Annals of Discrete Mathematics (11) Studies on Graphs and Discrete Programming*, pp. 381–395. North-Holland.
- Froncek, D. and Meszka M. (2005) Round robin tournaments with one bye and no breaks in home-away patterns are unique. *Multidisciplinary Scheduling: Theory and Applications*, pp. 331–340.
- Kendall, G., Knust, S., Ribeiro, C., and Urrutia, S. (2010) Scheduling in sports: An annotated bibliography. *Computers & Operations Research*, **37**(1):1–19.
- Knust, S. (2013) Classification of Literature on Sports Scheduling. http://www.inf.uos.de/knust/sportssched/sportlit_class/.
- Miyashiro, R., Iwasaki, H., and Matsui, T. (2003) Characterizing feasible pattern sets with a minimum number of breaks. *Practice and Theory of Automated Timetabling IV*, pp. 78–99.
- Rasmussen, R.V. (2008). Scheduling a triple round robin tournament for the best Danish soccer league. *European Journal of Operational Research*, **185**(2): 795–810.
- van't Hof, P., Post, G., and Briskorn, D. (2010). Constructing fair round robin tournaments with a minimum number of breaks. *Operations Research Letters*, **38**(6):592–596.

Assignment of swimmers to events in a multi-team meeting for team global performance optimization

S. Mancini*

*Control and Computer Engineering Department, Politecnico di Torino, Italy, simona.mancini@polito.it

Abstract. Assigning swimmers to events in order to maximize the global performance of a team in a multi-team meeting is not a trivial matter for coaches. Months of hard work and training are often wasted if a mistake is made in the line-up decision process. Expert coaches use their long experience to take the correct decisions, but they often fail to reach an optimal assignment. Preferences of certain athletes can also affect the decision process and can make the coaches jobs even harder. Furthermore, the actual goal that has to be achieved may vary from situation to situation. Two different integer programming models which are based on an estimation of the opponents performances, are proposed in this paper. These models are constructed on the basis of two different philosophies and addresses two different situations. The first model just maximizes the total score obtained by the team, while the second model has the aim of optimizing the position achieved by the team in the final ranking of the meeting and the advantage over the next team in the ranking. A detailed analysis of good and bad points of the two approaches and of situations in which one approach could be preferable to the other is reported. A real case example, taken from an Italian Regional Master Meeting, has been analyzed in depth and a discussion is given on a comparison of the results obtained with the assignment provided by the two models and the actual lineup proposed by the coach.

1. Introduction

Assigning swimmers to certain events in order to maximize the global performance of the team is not a trivial issue for coaches who must decide which athletes should compete in which event. In a sport in which every point counts, these decisions are extremely important. Months of hard work and training are often wasted if even only one mistake is made in the lineup decision process.

Expert coaches use their long experience in order to make correct decisions, but often without obtaining an optimal assignment. At the Master level (the category in which all athletes who do not play swim at the professional level and who are over 25 years of age can compete), apart from the capability of the athletes to perform in a given event, their willingness to compete in that event also strongly influences the decision process.

In fact, swimmers prefer to compete in events in which they know they can obtain a better personal result, and at the same time to maximize their own performance without considering the benefits for the team. Furthermore, the desire to compete in an event can also depend on the physical efforts required, (long distance events are generally less populated than short ones), and on the quantity and competitiveness of the athletes from the opponent teams (if the probability, considering the prevision of the opponents performances, to win a medal in an event is high, swimmers are willing to participate in it).

Other factors, such as the low frequency of an event in the seasonal agenda, may also effect swimmers preferences. In fact, particular events, for instance 200 butterfly, are not available at each meeting, but are only scheduled a few times per season. For this reason, most swimmers prefer to participate in these events, when available, and therefore neglect other events.

This can result in a lineup in which not all the events are covered by the team components, with a consequent negative effect on the team's performance. Opponent team lineups are not known in advance, but a very precise prevision of the assignment can be obtained from an analysis of the previous meetings. However, even when a coach is fully aware of the assignment of the opponent teams athletes and their performance capabilities, the challenge of creating a competitive lineup, without the help of a formal model, is prohibitive. In this paper, two different integer programming models able to obtain the optimal lineup of swimmers assignment are proposed.

A similar problem has been addressed in Nowak et al. (2006), where only dual meet events were considered and a model that is capable of maximizing team's score was provided. 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, Freeze (1974), cricket, Normann and Clarke (2007) and Clarke(1998), and ice hockey Washburn (1991) and for single player sports like tennis, Normann(1985), long jump, Sphicas and Ladany (1977), and pole vaulting, Hersh and Ladany (1989), 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 Kendall et al. (2010). Both integer programming (for instance, Van Voorhis (2002) for basketball, Saltzman and Bradford (1996), and Urban and Russell (2003) for American football) and heuristics (Russel and Leung (1994) for baseball, Amstrong and Willis (1993) for cricket and Schonberger et al. (2010), 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 Wright (2009).

The paper is organized as follows. A detailed description of the problem is given in Section 2, while the two integer programming models are presented in Section 3. The computational results obtained testing the two models on a real case 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 presentation

The problem treated in this paper consists in determining an optimal lineup in order to maximize the performance of a team without violating any constraint established by federation rules.

The size of the team is given by the number of athletes eligible to compete in the meeting, considering that all the athletes are supposed to be able to take part in each event. This does not imply a loss of generality because, in cases in which one or more athletes cannot compete in one discipline because of physical or technical limitations, the issue can be simply overcome by adding one or more further constraints to rule out the forbidden assignments. The swimmers roaster must be decided before the meeting and cannot be modified during the competition. The assignment of swimmers is limited by the following constraints established by the Italian federation and which are valid for most European federations at the professional, semi-professional, recreational and school levels:

- Each swimmer can compete in at most two events at the same meeting
- Each team can enter as many swimmers as it likes in the same event, but only the best placed one is awarded with points
- Positions after the 15th place do not yield to any awarding of points
- The score for positions between the 4th and the 15th place is computed according to the following formula $SCORE = 16 - POSITION$
- The first, second and third places are awarded with 20,17 and 15 points, respectively

All the decisions on the assignments are made considering that estimated times of each swimmer in the team (hereafter referred to as *the team*) and in opponents team are known. The opponents lineups are estimated using data from previous meetings, as they are generally available to a coach.

A prevision of the team times is made using information from earlier meetings, training sessions and previous seasons. Even though the actual event times are random variables and any information about past performance can only lead to estimate parameters, such as the means and variances of their probability distributions, the prevision we have are very precise and the experience in this sport at the master level has pointed out that the variance is very small and, given the heterogeneity of the athletes, times can be very different from each other (which does not happen at the professional level in which the times are very homogeneous); for this reason, the real time can be considered equal to estimated time, and the stochastic component of the problem can be neglected without introducing a relevant mistake.

3. Model formulation

In this paper two different integer programming models able to determine the optimal swimmers roasters are proposed. The first model allows to maximize the score obtained by the team, while the second one can be

used to determine the assignment which optimize the position of the team in the final meeting ranking maximizing the distance, in points, between the team, and its first follower in the ranking. The two approaches are slightly different. In the first case, attention is only focused on the team lineup and the approach neglect how this lineup has an impact on the performances of the opponents. In the second model takes into account the effect that team lineup has on the opponents final score.

3.1. Definitions and notations

The variables and parameters used in the two models are defined in the following:

- \mathbf{I} is the set of athletes in the team
- \mathbf{J} is the set of events
- \mathbf{K} is the set of opponent teams
- $p(i, j)$ is the position that athlete i would reach if he competes in event j which is computed considering the prevision of the opponents lineups as he were the only athlete in the team which takes part in the event (given as input data)
- $q(k, j)$ is the position that the best placed athlete of team k would reach in event j , which is computed considering the prevision of the opponents lineups as no other athletes from the team took part in the event (given as input data)
- $h(i, j, k)$ is a constant given as input data which is equal to 1 if $p(i,j) \leq q(k, j)$ and otherwise is zero. This represents a flag which is on if and only if athlete i is competing in event j and he/she reaches a better position than the best placed athlete in team k
- $X(i, j)$ is a binary variable which is equal to 1 if athlete i in *the team* competes in event j
- $Z(i, j)$ is a binary variable which is equal to 1 if athlete i in the team competes in event j and results to be the best placed athlete in the team
- $P(j)$ is the position of the best placed athlete in the team in event j (which is expressed as a linear function of the Z variables)
- $a(j)$ is a binary variable which is equal to 1 if the best placed athlete in *the team* reaches first place in event j
- $b(j)$ is a binary variable which is equal to 1 if the best placed athlete in *the team* reaches second place in event j
- $c(j)$ is a binary variable which is equal to 1 if the best placed athlete in *the team* reaches third place in event j
- $S(j)$ is the score obtained by the team in event j (which is expressed as a linear function of $P(j), a(j), b(j)$ and $c(j)$)
- T is the total score obtained by *the team* at the meeting
- $Q(k, j)$ is the position of the best placed athlete in team k in event j (which is expressed as a linear function of the X variables)
- $\alpha(k, j)$ is a binary variable which is equal to 1 if the best placed athlete in team k reaches first place in event j
- $\beta(k, j)$ is a binary variable which is equal to 1 if the best placed athlete in team k reaches second place in event j
- $\gamma(k, j)$ is a binary variable which is equal to 1 if the best placed athlete in team k reaches third place in event j
- $Y(k, j)$ is the score obtained by team k in event j (which is expressed as a linear function of $Q(k, j), \alpha(k, j), \beta(k, j), \gamma(k, j)$)
- $W(k)$ is the total score obtained by team k at the meeting
- F represents the score of the team which reaches the next place in the ranking with respect to *the team* at the meeting
- $O(k)$ is a binary variable which is equal to 1 if team k obtains a better or equal score to the one obtained by the team

- \mathbf{N} is the number of teams that obtain a better or equal placement in the meeting ranking with respect to *the team*
- ϵ is a very small constant
- \mathbf{M} is a very large constant

In the case in which the estimated time of athlete i in event j is equal to the estimated time of an opponent teams athlete, who is supposed to compete in the same event, i is supposed to be places behind his/her opponent in the ranking.

3.2. An Integer Programming Model for Team Score Maximization

The model used to maximize the team score, named MOD1, can be formulated as follows:

$$\max T \quad (1)$$

$$\sum_{j \in J} X(i, j) \leq 2 \quad \forall i \in I \quad (2)$$

$$\sum_{i \in I} Z(i, j) \leq 1 \quad \forall j \in J \quad (3)$$

$$Z(i, j) \leq X(i, j) \quad \forall i \in I \quad \forall j \in J \quad (4)$$

$$P(j) = \sum_{i \in I} p(i, j)Z(i, j) + 16 \left(1 - \sum_{i \in I} Z(i, j) \right) \quad \forall j \in J \quad (5)$$

$$S(j) = 16 - P(j) + 5a(j) + 3b(j) + 2c(j) \quad \forall j \in J \quad (6)$$

$$a(j) + b(j) + c(j) \leq 1 \quad \forall j \in J \quad (7)$$

$$a(j) \leq 1 + \epsilon(1 - P(j)) \quad \forall j \in J \quad (8)$$

$$b(j) \leq 1 + \epsilon(2 - P(j)) \quad \forall j \in J \quad (9)$$

$$c(j) \leq 1 + \epsilon(3 - P(j)) \quad \forall j \in J \quad (10)$$

$$T = \sum_{j \in J} S(j) \quad (11)$$

$$X(i, j) \in \{0, 1\} \quad \forall i \in I \quad \forall j \in J \quad (12)$$

$$Z(i, j) \in \{0, 1\} \quad \forall i \in I \quad \forall j \in J \quad (13)$$

$$P(j) \in \mathbf{Z}^+ \quad \forall j \in J \quad (14)$$

$$S(j) \in \mathbf{Z}^+ \quad \forall j \in J \quad (15)$$

$$T \in \mathbf{Z}^+ \quad (16)$$

$$a(j) \in \{0, 1\} \quad \forall j \in J \quad (17)$$

$$b(j) \in \{0, 1\} \quad \forall j \in J \quad (18)$$

$$c(j) \in \{0, 1\} \quad \forall j \in J \quad (19)$$

The objective function maximizes the total score obtained by the team. Constraints (2) and (3) impose, respectively, that each swimmer in the team may compete in at most two events and that in each event only the best placed athlete of the team is awarded with points. Constraint (4) ensures that an athlete in the team can only obtain points in an event if he takes part in it. A relation between the position reached by the team in an event and the lineup is given in (5), while constraint (6) indicates according how the score reached by a team in an event is correlated to its position, and ensures that, if no athletes in the team take part in an event,

the score achieved by the team for that event is null. Bonuses related to the achievement of one of the three best places are managed by constraints (7),(8),(9) and (10). Constraint (11) imposes that the total score of the team should be equal to the sum of the score obtained in each event. Finally, (12)-(19) specify the domains of the variables.

Since this model has the aim of maximizing the total score of the team, it discards information on the effect of the teams assignment on the opponent teams final score. In fact, only $|J|$ assignments are made by the model (those that allow one to obtain points) while the remaining $2|I| - |J|$ are left free.

This means that MOD1 leaves a great deal of freedom to the athletes to choose which events to participate in, and leaves most of the team member free to make their own choices. On the other hand, MOD1 only provides an assignment which allows the team to reach the best score possible with its athletes, without considering the position in the ranking. In the following another model is presented, that is able to provide an assignment which allows to obtain the best available position with the greatest advantage with respect to the next place in the ranking, is presented.

3.3. An Integer Programming Model for team position optimization

The model to optimize the team's position in the ranking, named MOD2, can be formulated as follows:

$$\min N - \varepsilon(T - F) \quad (20)$$

$$\sum_{j \in J} X(i, j) \leq 2 \quad \forall i \in I \quad (21)$$

$$\sum_{i \in I} Z(i, j) \leq 1 \quad \forall j \in J \quad (22)$$

$$Z(i, j) \leq X(i, j) \quad \forall i \in I \quad \forall j \in J \quad (23)$$

$$P(j) = \sum_{i \in I} p(i, j)Z(i, j) + 16 \left(1 - \sum_{i \in I} Z(i, j) \right) \quad \forall j \in J \quad (24)$$

$$S(j) = 16 - P(j) + 5a(j) + 3b(j) + 2c(j) \quad \forall j \in J \quad (25)$$

$$a(j) + b(j) + c(j) \leq 1 \quad \forall j \in J \quad (26)$$

$$a(j) \leq 1 + \varepsilon(1 - P(j)) \quad \forall j \in J \quad (27)$$

$$b(j) \leq 1 + \varepsilon(2 - P(j)) \quad \forall j \in J \quad (28)$$

$$c(j) \leq 1 + \varepsilon(3 - P(j)) \quad \forall j \in J \quad (29)$$

$$T = \sum_{j \in J} S(j) \quad (30)$$

$$Q(k, j) = q(k, j) + \sum_{i \in I} X(i, j)h(i, j, k) \quad \forall k \in K \quad \forall j \in J \quad (31)$$

$$Y(k, j) = 16 - Q(k, j) + 2\alpha(k, j) + \beta(k, j) + 2\gamma(k, j) \quad \forall k \in K \quad \forall j \in J \quad (32)$$

$$\alpha(k, j) \leq \varepsilon + \varepsilon(1 - Q(k, j)) \quad \forall k \in K \quad \forall j \in J \quad (33)$$

$$\beta(k, j) \leq \varepsilon + \varepsilon(2 - Q(k, j)) \quad \forall k \in K \quad \forall j \in J \quad (34)$$

$$\gamma(k, j) \leq \varepsilon + \varepsilon(3 - Q(k, j)) \quad \forall k \in K \quad \forall j \in J \quad (35)$$

$$W(k) = \sum_{j \in J} Y(k, j) \quad \forall k \in K \quad (36)$$

$$O(k) \geq \varepsilon(W(k) + 1 - T) \quad \forall k \in K \quad (37)$$

$$F \geq W(k) - M \cdot O(k) \quad \forall k \in K \quad (38)$$

$$N = \sum_{k \in K} O(k) \quad (39)$$

$$X(i, j) \in \{0,1\} \quad \forall i \in I \quad \forall j \in J \quad (40)$$

$$Z(i, j) \in \{0,1\} \quad \forall i \in I \quad \forall j \in J \quad (41)$$

$$P(j) \in \mathbf{Z}^+ \quad \forall j \in J \quad (42)$$

$$S(j) \in \mathbf{Z}^+ \quad \forall j \in J \quad (43)$$

$$T \in \mathbf{Z}^+ \quad (44)$$

$$a(j) \in \{0,1\} \quad \forall j \in J \quad (45)$$

$$b(j) \in \{0,1\} \quad \forall j \in J \quad (46)$$

$$c(j) \in \{0,1\} \quad \forall j \in J \quad (47)$$

$$Q(k, j) \in \mathbf{Z}^+ \quad (48)$$

$$\alpha(k, j) \in \{0,1\} \quad \forall k \in K \quad \forall j \in J \quad (49)$$

$$\beta(k, j) \in \{0,1\} \quad \forall k \in K \quad \forall j \in J \quad (50)$$

$$\gamma(k, j) \in \{0,1\} \quad \forall k \in K \quad \forall j \in J \quad (51)$$

$$Y(k, j) \in \mathbf{Z}^+ \quad \forall k \in K \quad \forall j \in J \quad (52)$$

$$W(k) \in \mathbf{Z}^+ \quad \forall k \in K \quad (53)$$

$$O(k) \in \{0,1\} \quad \forall k \in K \quad (54)$$

$$N \in \mathbf{Z}^+ \quad (55)$$

$$F \in \mathbf{Z}^+ \quad (56)$$

The objective function is given in (20). Constraints (21)-(30) play the same role, as (2)-(11) play in MOD1. A relation among the position reached by an opponent team in an event and the lineup of the team, is given in (31), while constraint (32) indicates how the score reached by a team in an event is correlated to its position. Bonuses related to the achievement of one of the three best places are managed by constraints (33),(34) and (35). Constraint (36) imposes that the total score obtained by an opponent team must be equal to the sum of the score obtained in each event. Constraint (37) ensures that, for each opponent team k , $O(k)$ is equal to one, if, and only if, k obtains a better or equal score to the one obtained by the team. In Constraint (38), F is forced to be larger than the maximum score obtained by an opponent team classified at a lower rank than the team while, the number of opponent teams which obtain an higher score respect to the one obtained by the team, is computed in (39). Finally, (40)-(56) specify the domains of the variables.

This model is able to determine a line-up which allows to reach the best position possible in the ranking, which is the actual goal of a team, with the greatest advantage over the next opponent in the ranking. This is an important issue, because, the higher this gap, the stronger the robustness of the solution. In fact, since stochasticity is not explicitly considered, a greater advantage allows the position in the ranking to be maintained even if the opponent teams swimmers perform better, or one of the swimmers in the team, does not perform as well as expected.

The assignment given by this model obviously has to be considered fixed because, even just one change in the lineup could yield a worse final result. The philosophy on which MOD2 is based is different from that of MOD1. In fact, while MOD1 tries to have a more flexible approach, MOD2 aims to obtain the best for the team without considering the personal wishes and preferences of the athletes, reveling a more aggressive and competitive approach.

Both models could be extremely useful in real life applications, as they could allow coaches to choose a more aggressive or flexible approach. Furthermore, MOD2 could also be used, just changing the objective function, also for specific purposes, like the maximization of advantage points respect to a given team. This situation is very common in seasonal leagues, in which, especially in the last meetings of the season, a team is more interested in improving its advantage over its nearest opponent or in reducing the gap between the team that precedes it in the seasonal ranking, than in obtaining the best reachable position in the meeting.

4. A real case analysis

The model has been tested on data taken from an Italian Regional Master Meeting which took place in 2009, with access restricted to teams from Piedmont, in which 5 events were scheduled (50 Buttery, 50 Backstroke, 50 Breaststroke, 50 Freestyle and 100 Medley). In Master meetings athletes are divided, other than by sex, into several age categories: M25, to which athletes in the 25-29 age range belong, M30 to which athletes in the 30-34 age range belong, and so on with five year steps. Since each athlete belongs to one and only one category and categories can be independently scheduled, in this example we have focused on a single category (M25 Female).

The team (TEAM) is composed of 8 athletes, each one of which is able to compete in each event. Seven opponent teams are considered (A,B,C,D,E,F,G), each one of which has a different number of athletes varying between 4 and 9. The predicted times of the team are reported in Table 1 while the opponents predicted times are shown in Table 2. The values of $p(i,j)$ and $q(k,j)$ have been computed using an estimation of the opponent teams line-ups. This estimation has been done by analyzing data from past editions of the meeting.

From this analysis, it has been noticed that athletes are usually competing in the same events in a given meeting every year, because of their training strategies and workloads, which vary during the season, and allow them to reach their performance peak, in each event, in different moments of the season. Furthermore, athletes at the professional level are very specialized and often compete in the same events at each meeting.

The rule adopted to determine the expected line-up is the following.

If an athlete took part in the same meeting in the previous year, he/she is supposed to compete in the same events he/she had competed in. If he/she did not compete at that meeting the previous year, an analysis of other meetings he/she took part in is performed, and if he/she is an athlete who likes to compete in different events throughout the year, he/she is assumed to compete in the events he/she took part in, during the previous year, in the same part of the season in which the meeting takes place. Instead, if he/she is only used to competing in some events, for instance, there is a particular style in which he/she excels or he/she prefers short (or long) distances, he/she is supposed to compete in the events in which he/she is most used to participate in. If no information is available on an athlete, because it is his/her first competitive season, the estimations are based on information about his/her performances, in each event, obtained from training trials, if available. If no data is available, even from training trials, the athlete is supposed to compete in the events which, on average, have a higher number of participants. The list of events ordered on the basis of a decreasing number of participants, and determined on the basis of an analysis of the past seasons is: 50 FREE, 100 FREE, 50 BREAST, 50 BACK, 100 MEDLEY, 50 FLY, 200 FREE, 100 BACK, 100 BREAST, 200 MEDLEY, 100 FLY, 400 FREE, 800 FREE, 1500 FREE, 200 BACK, 200 BREAST, 400 MEDLEY and 200 FLY.

5	35"45	37"10	42"28	30"98	1'18"83
6	37"71	41"82	44"11	33"63	1'27"11
7	35"00	47"99	43"21	31"56	1'18"89
SWIMMER	50 FLY	50 BACK	50 BREAST	50 FREE	100 MEDLEY
1	41"20	41"93	41"80	32"80	1'23"43
2	47"10	48"39	54"20	40"48	1'43"43
3	52"00	1'00"11	55"10	38"80	2'08"12
4	36"65	39"11	57"11	33"69	1'26"00
5	35"45	37"10	42"28	30"98	1'18"83
6	37"71	41"82	44"11	33"63	1'27"11
7	35"00	47"99	43"21	31"56	1'18"89
8	37"00	37"89	40"49	32"82	1'22"21

21

Results obtained by MOD1 and MOD2 are compared with those obtained by a Single Profit Strategy (SPS), in which each swimmer tries to maximize his/her own profit, choosing to compete in the pair of events that allows him/her to achieve the highest personal score, computed as the sum of the scores obtained in the two events in which him/her takes part, and, finally, with the actual lineup presented at the meeting (REAL). The resulting rankings and scores, that would be obtained if the actual performances are considered equal to the predicted ones are reported in Table 3, while the related team lineups are shown in Table 4. The latter table can be read as follows: if athlete i competes in event j , the corresponding cell in the table contains 1, otherwise it is blank. If a cell is underlined it means that the related assignment is not compulsory, i.e. athlete i could be assigned to any other event without any change in the objective function, which means that athlete i is free to choose in which event to compete.

Both models, run under XPRESS (Release 2008), were able to solve the problem to the optimality in a very short time (less than 0.1 seconds) on a machine with an Intel Dual Core processor at 2.5 GHz with 4 GB of RAM. The reported results show that both MOD1 and MOD2 perform better than SPS or REAL. In fact, both models obtain the first position in the ranking, while the team only reaches the third place with SPS and REAL. MOD2 is able to provide a more robust solution than to MOD1, since the lineup provided by MOD2 allows the athlete to win with a larger margin (9 points) than the one obtained with MOD1 (5 points).

This is an important issue, because, with a larger margin over the next competitor in the ranking, in the case in which opponents perform better, or the team athletes perform worse than predicted, the probability of losing the position in the ranking becomes lower. Nevertheless, even though MOD2 seems to outperform MOD1, it should be pointed out that MOD1 provides a very good lineup, and leaves more freedom to the swimmers in the event choice. In fact, as reported in Table 4, only 5 assignment out of 16 (8 athletes who can compete in 2 events each) are compulsory in the MOD1 lineup, while in the solution provided by MOD2 there are ten compulsory assignments.

This means that in some cases, especially at the master level, in which the performances of the team and the personal satisfaction of the athletes should be balanced, MOD1 may be the most suitable tool to address the swimmers assignment problem.

Table 2. Opponents predicted time

		50 FLY	50 BACK	50 BREAST	50 FREE		100 MEDLEY
1	A1	30"88	1	D2	32"28	1	C2
2	A2	32"22	2	D3	36"00	2	C3
3	D1	36"23	3	D4	37"80	3	E3
4	C1	41"00	4	B3	38"70	4	E2
5	C2	44"49	5	B4	39"95	5	F3
6	D2	49"89	6	B5	40"17	6	F1
7	B1	50"11	7	F2	40"98	7	E5
8	B2	51"22	8	E1	41"12	8	E4
9	E1	55"11	9	F3	41"23	9	A5
10	G1	59"11	10	G1	43"11	10	A6
11	G2	1'01"89	11	G2	44"11	11	G4
			12	G3	44"89	12	A3
			13	A1	45"85	13	G5
			14	A2	46"96	14	A4
			15	E4	47"11		
			17	E6	48"59		
			18	B1	49"99		
			19	B2	50"01		
			20	C5	56"11		
			21	C6	59"95		
			22	C4	1'00"00		
			23	A3	1'01"38		
			24	A4	1'01"43		
						25	C4
						26	C5
							46"00

4.1. A further analysis

In order to offer a more stronger proof of the advantages obtained when using the proposed approach, it could also be interesting to analyze what happens if one of the other teams is considered as the team. In Table 5 are reported, for each team k , the position in the final ranking, the advantage on the next competitor in the ranking and the total score obtained using MOD1, MOD2 and SPS, if k is considered as the team.

The last three columns report the actual achievement reached in the meeting. A null advantage value means that the team obtained the same position of one or more opponent teams, while, in case in which the team has been classified last, a negative number is reported, whose absolute value represents the dis-

advantage points with respect to the nearest team in the ranking. In this way, values with a smaller absolute value indicate a better performance of the team.

As can be observed from the results, the use of MOD1 and MOD2 yields a great advantage on the team global performances, allowing in the most cases to improve the team placement in the ranking, and, in the other cases, to strongly increase the advantage on the nearest competitor in the ranking. The fact that almost all the teams may reach the first place was expectable, because, at the master level, performances capability are strongly homogeneous among teams, differently than at the professional level, and the final result strongly depends on the decisions made by the coach when choosing the line-up.

Table 3. Comparison of ranking and scores obtained with different approaches

RANKING	MOD1	MOD2	SPS	REAL
1	TEAM:63	TEAM:63	D:58	D:56
2	D:58	D:54	F:55	F:55
3	F:55	F:53	TEAM:51	TEAM:54
4	E:49	E:48	E:46	E:48
5	B:42	B:44	A:42	B:45
6	A:41	A:41	B:41	A:40
7	C:39	C:41	C:38	C:40
8	G:12	G:13	G:10	G:10

Table 4. Team lineups provided by different approaches

MOD1					
SWIMMER	50 FLY	50 BACK	50 BREAST	50 FREE	100 MEDLEY
1	1				1
2		1		1	
3		1		1	
4	1				1
5		1		1	
6	1				1
7	1				1
8	1		1		

MOD2					
SWIMMER	50 FLY	50 BACK	50 BREAST	50 FREE	100 MEDLEY
1			1	1	
2	1		1		
3	1	1			
4		1		1	
5		1		1	
6	1			1	
7	1				1
8		1		1	

SPS					
SWIMMER	50 FLY	50 BACK	50 BREAST	50 FREE	100 MEDLEY
1	1		1		
2	1		1		
3	1		1		
4	1	1			
5	1	1			
6	1		1		
7	1				1
8	1				1

REAL					
SWIMMER	50 FLY	50 BACK	50 BREAST	50 FREE	100 MEDLEY
1			1	1	
2		1			1
3			1	1	
4	1			1	
5				1	1
6		1			1
7	1				1
8		1	1		

Table 5. Comparison of rankings and scores obtained, considering each team as the team, with different approaches

THE TEAM	MOD1			MOD2			SPS			REAL		
	RANK	ADVANTAGE	SCORE									
A	1	2	58	1	3	58	3	0	40	6	0	40
B	5	7	52	5	8	52	6	1	38	5	5	45
C	1	4	63	1	4	63	2	1	57	6	0	40
D	1	13	84	1	13	84	1	8	67	1	1	56
E	1	4	70	1	4	70	3	0	55	4	3	48
F	1	19	81	1	21	81	1	6	64	2	1	55
G	8	-5	35	8	-4	35	8	-7	34	8	-30	10

5. Conclusions and future perspectives

The two models presented in this paper are able to solve real instances to the optimality in a very short time. The computational results show that both approaches used together are strongly preferable to a Single Profit Strategy, according to which each athlete is assigned to the events in which he is more competitive, which is the classical rule applied by coaches to determine a lineup.

A comparison with the actual lineup that took part in the meeting, decided by the coach, has shown that finding the best swimmers assignment is not a trivial task and it is advantageous to address this task with the aid of a mathematical tool. MOD2 guarantees the most robust solution, and its use is preferable, but there are situations in which the coach also has to take into account swimmers preferences, in which case the use of MOD1 is preferable because it leaves more freedom to swimmers choices, since only a small number of the proposed assignment are compulsory. In this way, coaches can choose, between a more flexible tool or a more efficient one, and decide which is the more suitable for their case.

If predictions are not very precise, as it occurs at the Junior level in which athletes performances capability can improve remarkably in a very short time, opponents predicted time may be decreased by 1% or 2% thus allowing the model to provide a new assignment. Furthermore, MOD2 can also be used, just modifying the objective function, in order to address particular tasks, such as the maximization of advantage points with respect to a given team. This is a very common in seasonal leagues, in which the main goal of the team could be to improve its advantage over its nearest opponents (or to reduce the gap between the team that precedes it) in the seasonal ranking.

Last but not least, the same approach could be also used to assign athletes to track and field events as they suffer from similar constraints and have similar objectives.

Future developments in this field could concern the incorporation of the prohibition against swimming in consecutive events, in all cases, or only when there are long distance events for which the time required to recover from the physical effort may be longer. At the master level, especially for events like national championships which are scheduled over five days, athletes may not be available for such a long period and could ask to be assigned to two events in the same day.

This issue could be addressed by introducing further constraints. Another interesting development could be the consideration of a degradation of an athlete's performances, when competing in two consecutive events (or in two long distance ones) which would depend on the events and on the athlete's characteristics.

References

- Armstrong, J. and Willis, J. (1993) Scheduling the cricket world cup: a case study. *Journal of the Operational Research Society* **44**(11), 1067-1072.
- Clarke, S.(1998) Dynamic programming in one-day cricket optimal scoring rates. *Journal of the Operational Research Society* **39**(4), 331-337.
- Della Croce, F. and Oliveri, D. (2006) Scheduling the italian football league: An ilp-based approach. *Computers & Operations Research* **33**(7), 1963-1974.
- Freeze, R. (1974). An analysis of baseball batting order by monte carlo simulation. *Operations Research*, **22**(4), 728-735.

- Hersh, M. and Ladany, S. (1989) Optimal pole-vaulting strategy. *Operations Research* **37**(1), 172-175.
- Kendall, G., Knust, S., Ribeiro C., and Urrutia, S. (2010). Scheduling in sports: An annotated bibliography. *Computers & Operations Research*, **37**(1), 1-19.
- Normann, J. (1985) Dynamic programming in tennis: When to use a fast serve. *Journal of the Operational Research Society*, **36**(1), 75-77.
- Normann, J. and Clarke, S. (2007) Dynamic programming in cricket: Optimizing batting order for a sticky wicket. *Journal of the Operational Research Society* **58**(12), 1678-1682.
- Nowak, M., Epelman, M. and Pollock, S. (2006) Assignment of swimmers to dual meet events. *Computers & Operations Research* **33**, 1951-1962.
- Russell, R. and Leung, J. (1994) Devising a cost effective schedule for a baseball league. *Operations Research* **42**(4), 614-625.
- Saltzman, R. and Bradford, R. (1996) Optimal realignments of the teams in the national football league. *European Journal of Operational Research* **93**(3), 469-475.
- Schonberger, J., Mattfeld, D. and Kopfer, H. (2004) Memetic algorithm timetabling for non-commercial sport leagues. *European Journal of Operational Research* **153**(1), 102-116.
- Sphicas, G. and Ladany, S. (1977) *Optimal Strategies in Sport*. North-Holland, chapter Dynamic policies in the long jump. 101-112.
- Urban, T. and Russell, R. (2003) Scheduling sports competitions on multiple venues. *European Journal of Operational Research* **148**(2), 302-311.
- Van Voorhis, T. (2002) Highly constrained college basketball scheduling. *Journal of the Operational Research Society* **53**(6), 603-609.
- Washburn, A. (1991) Still more on pulling the goalie. *Interfaces* **21**, 59-64.
- Wright, M. (2009) 50 years of or in sport. *Journal of the Operational Research Society* **60**, 161-168.

Fair Referee Assignment for the Italian Soccer Serie A

S. Mancini* and A. Isabelllo**

*Control and Computer Engineering Department, Politecnico di Torino, Italy, simona.mancini@polito.it

**Independent consultant, andrea.isabelllo@siti.polito.it

Abstract. The Referee Assignment Problem (RAP) is a novel arising problem in sports management, in which a limited number of referees with different qualifications and availabilities should be assigned to a set of games already scheduled, in order to respect a list of constraints. The number and the nature of constraints may significantly vary for sports, nation and type of league. Almost each tournament has its own particular set of constraints to be satisfied, therefore it is very difficult to generalize this problem. The goal of the problem is to find a feasible assignment, i.e. a configuration which allows to respect all the constraints given. An extension of the RAP is the Fair Referee Assignment Problem (FRAP), in which the objective is to minimize the violation of a set of soft (optional) constraints, while satisfying all the hard (mandatory) ones. In this work, the Italian Major Soccer League, the so-called SERIE A, is addressed, and an integer programming model for the related FRAP is proposed. Soft and hard constraint have been formulated according to the rules suggested by the AIA (Italian Referee Association) which is in charge of referee assignment for the SerieA. The model has been tested on a real instance taken from the season 2011/2012. Results obtained show the efficacy and the effectiveness of the model.

1. Introduction

Operation research in sports is a field of increasing interest, and different optimization techniques have been applied to solve problems arising from different sports. For a complete overview of the subject we refer the reader to Wright (2009).

The main areas of study are in the analysis of tactics and strategy, scheduling and forecasting. Several papers addressing tactical and strategic issues, may be found in literature, both concerning team sports, like baseball, Freeze (1974), cricket, Normann and Clarke (2007) and Clarke(1998), and ice hockey Washburn (1991) and single player sports like tennis, Normann(1985), long jump, Sphicas and Ladany (1977), and pole vaulting, Hersh and Ladany (1989). The literature on sports league scheduling is somewhat large and covers almost all team sports; a complete survey can be found in Kendall et al. (2010). Both integer programming (for instance, Van Voorhis (2002) for basketball, Saltzman and Bradford (1996), and Urban and Russell (2003) for American football) and heuristics (Russel and Leung (1994) for baseball, Amstrong and Willis (1993) for cricket and Schonberger et al. (2010), who address sport league scheduling for different sports) have been proposed. In particular, soccer tournament scheduling is one of the most addressed issue in literature. Although, many constraints are common for almost all soccer schedule problems, each national league has its own specific requirements. Several papers deals with scheduling games in different leagues (Austrian and German, (Bartsch et al. (2006)), Brazilian (Riberio and Urrutia (2006) and Bjaioli et al. (2004)) , Chilean, (Duran et al. (2006)), Danish, (Rasmussen (2008)), Dutch, (Schreuder (1992)), and Italian, Della Croce and Oliveri (2006)) requiring considerably different objective functions and constraints. For a complete overview of soccer scheduling in Europe, please refer to Goossens and Spieksma (2012). Tournament scheduling are not the only issue to be addressed dealing with sport leagues. Also officials assignment problems arises in this context. All sports have their officials, whether known as referees as in soccer, football or rugby, umpires as in cricket and hockey, judges in equestrian events, etc. Scheduling these officials so as to ensure fairness, high quality, suitable experience and low cost is of great importance to any kind of sport competition.

A novel arising problem in sports management is the Referees Assignment Problem (RAP), in which a limited number of referees with different qualifications and availabilities should be assigned to a set of games already scheduled. in order to respect a list of constraints. The goal of the problem is to find a feasible assignment, i.e. a configuration which allows to respect all the constraints given. Number and the nature of constraints may significantly vary for sports, nation and type of league. Almost each tournament has its own particular set of constraints to be satisfied, therefore it is very difficult to generalize this problem.

In this paper we deal with an extension of this problem, the Fair Referees Assignment Problem, (FRAP), in which the goal is to minimize the violation of a set of soft (optional) constraints, while satisfying all the hard (mandatory) ones. This problem has been introduced for the first time in Yavuz et al. (2008), where the authors propose a non-linear model, but they don't explicitly solve it, and some lower bounds. They solve the problem by a local search based heuristic. Computational tests are carried out on instances obtained by the Turkish Premier League (TPL).

In this paper we present a integer programming model for the FRAP and we apply it on the Italian SerieA, season 2010/2011. The model can be solved to the optimality in very short time. Furthermore, a comparison with the real assignment decided by the Italian Referees Association (AIA), showing the effectiveness of our approach, is reported.

The paper is organized as follows. A detailed description of the problem is given in Section 2, while the integer programming models is presented in Section 3. Computational results obtained testing the model on a real case taken from the Italian Serie A are reported and a detailed analysis of these results is provided in Section 4. In Section 5, a further analysis, concerning the utility of the model in cases in which the difficulty level of a match, the availability or the skill level of a referee changes during the season, is presented. Finally, Section 6 is devoted to the conclusions and possible future developments.

2. Problem definition

The problem we address in the paper consist into finding the referees assignment, for all season matches, which aims to minimize the soft constraints violation, while satisfying all the hard constraints. The list of constraints has been created following the guidelines provided by the Italian Referees Association (AIA) which is in charge of referees assignment for the SerieA. Most of the constraints hold for the others main national league. We tried to generalize, as much as possible, the constraints exploitation, using parametric constants, in order to make the model more general and easily applicable for different leagues. The list of hard and soft constraints is reported in the following:

Hard Constraints

- A referee cannot officiate more than one match in the same matchday
- A referee can officiate the same team at most P times at home and P times away
- A referee can officiate a match involving the same couple of teams at most once a season
- A referee can officiate a match only if he hold the minimum skill level required
- A referee cannot officiate a team coming from the same province he comes from, neither at home nor away
- A referee cannot officiate in a matchday if it is not available (for private reasons, or because he has officiated in a European or Continental match during the week and it is not able to come back on time or to completely recover from the physical effort)

Soft Constraints

- v1: Once a referee has officiated a team, he cannot officiate it again for q matchdays (q is equal to 5 for the SerieA, but may take different values for other leagues, depending also on the number of team taking part of the league)
- v2: A referee cannot officiate in more than s consecutive matchdays

- v3: A referee cannot officiate a match if he doesn't hold a skill level strictly higher than the minimum required
- v4: A referee from the SerieB (lower division) cannot officiate in SerieA

Soft constraints may be violated with a unitary penalty in the objective function. List of referees is fixed and known in advance. For each referee, we know the province of belongings (which is the province in which the referees currently live and may be different from the province in which he was born), the skill level, the availability and the category to which he belongs (SerieA or SerieB). For each matchday we know all the matches scheduled in that matchday, and the difficulty level of each match.

3. An Integer Programming model for the FRAP

In this section we present the mathematical formulation of the proposed Integer Programming Model. Before to do that, we introduce some definitions and notations which are used in the following.

3.1 Definitions and notations

We define the following parameters which will be used in the model formulation

- I is the set of team in the league
- K is the set of available referees
- R is the set of matchdays in the season. Generally $|R|$ is function of $|I|$, and in particular is equal to $(|I|-1)*2$, since each team has to play twice, once home and once away, against each other team in the league, but for some leagues, for instance the Scottish Premier League (SPL), due to the small number of teams taking part of the tournament, each team plays four times against each other team. For this reason, we prefer to consider R as a separate parameter.
- M is the set of matches to be played at each matchdays $|M|$ is function of $|I|$, generally, it is equal to $|I|/2$, but in cases of leagues with an odd number of participants it could be at most equal to $\lfloor |I|/2 \rfloor$.
- Q is a parameter which indicates for how many matchdays a referee should not officiates a team after he has officiated it
- S is a parameter which indicates for how many consecutive matchdays a referee may officiate, without violating a soft constraint
- P is a parameter which indicates how many times a referee may officiate a team
- $a(r,m,i)$ is a given constant which is equal to 1 if team i plays, away, in match m of matchday r and is equal to 0 otherwise
- $h(r,m,i)$ is a given constant which is equal to 1 if team i plays, at home, in match m of matchday r and is equal to 0 otherwise
- $d(r,m)$ indicates the difficulty level of match m in matchday r (1 corresponds to the most difficult match while 5 to the easiest one)
- $l(k)$ is the skill level of referee k (1 corresponds to the higher skill level while 5 to the lower one)

- $e(k,i)$ is a given constant which is equal to 1 if referee k can officiate team i when it plays away and is equal to 0 otherwise
- $f(k,i)$ is a given constant which is equal to 1 if referee k can officiate team i when it plays at home and is equal to 0 otherwise
- $g(r,k)$ is a given constant which is equal to 1 if referee k is available at matchday r and is equal to 0 otherwise
- $u(k)$ is a given constant which is equal to 1 if referee k belongs to a lower division and is equal to 0 otherwise.

We also define the following variables which will be used in the model formulation:

- $Z(r,m,k)$ is a binary variable which is equal to 1 if referee k officiates match m on matchday r and is equal to 0 otherwise
- $T(k,i,j)$ is a binary variable which is equal to 1 if referee k officiates the match between teams i and j , where i is playing at home, and is equal to 0 otherwise
- $V1(r,k,i)$ is a binary variable which is equal to 1 if referee k officiates team i at least twice in q consecutive matchdays, starting from matchday r and is equal to 0 otherwise
- $V2(r,k)$ is a binary variable which is equal to 1 if referee k officiates on $s+1$ consecutive matchdays starting from matchday r and is equal to 0 otherwise
- $V3(r,m,k)$ is a binary variable which is equal to 1 if referee k , at matchday r , officiates a match, m , with the maximum difficulty level he is allowed to officiate r and is equal to 0 otherwise
- $V4(r,m,k)$ is a binary variable which is equal to 1 if referee k , officiating match m on matchday r , belongs to a lower division r and is equal to 0 otherwise

Please note that each variable $V1$, $V2$, $V3$ and $V4$ taking value equal to 1, represents a violation of a soft constraint of type $v1, v2, v3$ and $v4$, respectively. Variables $V1$ are defined only for $r \leq |R| - q$, while $V2$ are defined only for $r \leq |R| - s$. Furthermore, since each team takes part of at most one match at each matchday (exactly one if there is an even number of teams in the league), the following relations hold:

$$\sum_{m=1}^{|M|} a(r, m, i) \leq 1 \text{ and } \sum_{m=1}^{|M|} h(r, m, i) \leq 1 \quad \forall r \in R \quad \forall i \in I.$$

3.2 Model Formulation

The integer programming model we propose to minimize the number of soft constraints violation is the following:

$$\min \sum_{r=1}^{|R|-q} \sum_{k=1}^{|K|} \sum_{i=1}^{|I|} V1(r, k, i) + \sum_{r=1}^{|R|-s} \sum_{k=1}^{|K|} V2(r, k) + \sum_{r=1}^{|R|} \sum_{m=1}^{|M|} \sum_{k=1}^{|K|} V3(r, m, k) + \sum_{r=1}^{|R|} \sum_{m=1}^{|M|} \sum_{k=1}^{|K|} V4(r, m, k) \quad (1)$$

$$\sum_{m=1}^{|M|} Z(r, m, k) \leq g(r, k) \quad \forall r \in R \quad \forall k \in K \quad (2)$$

$$\sum_{k=1}^{|K|} Z(r, m, k) \leq 1 \quad \forall r \in R \quad \forall m \in M \quad (3)$$

$$\sum_{r=1}^{|R|} \sum_{m=1}^{|M|} Z(r, m, k) * h(r, m, i) \leq P \quad \forall k \in K \quad \forall i \in I \quad (4)$$

$$\sum_{r=1}^{|R|} \sum_{m=1}^{|M|} Z(r, m, k) * a(r, m, i) \leq P \quad \forall k \in K \quad \forall i \in I \quad (5)$$

$$T(k, i, j) = \sum_{r=1}^{|R|} \sum_{m=1}^{|M|} Z(r, m, k) * a(r, m, i) * h(r, m, j) \quad \forall k \in K \quad \forall i \in I \quad \forall j \in I \quad (6)$$

$$T(k, i, j) + T(k, j, i) \leq 1 \quad \forall k \in K \quad \forall i \in I \quad \forall j \in I \quad (7)$$

$$Z(r, m, k) \leq 1 + 0.1 * (d(r, m) - l(k) + 1) \quad \forall r \in R \quad \forall m \in M \quad \forall k \in K \quad (8)$$

$$Z(r, m, k) \leq \sum_{i=i}^{|I|} e(k, i) a(r, m, i) \quad \forall r \in R \quad \forall m \in M \quad \forall k \in K \quad (9)$$

$$Z(r, m, k) \leq \sum_{i=i}^{|I|} f(k, i) h(r, m, i) \quad \forall r \in R \quad \forall m \in M \quad \forall k \in K \quad (10)$$

$$V1(r, k, i) \geq \frac{1}{Q} \left(\sum_{q=0}^Q \sum_{m=1}^{|M|} Z(r+q, m, k) * a(r+q, m, i) * h(r+q, m, i) \right) \quad \forall r \in [1, |R| - Q] \\ \forall k \in K \quad \forall i \in I \quad (11)$$

$$V2(r, k) \geq \sum_{s=0}^S \sum_{m=1}^{|M|} Z(r+s, m, k) - S \quad \forall r \in [1, |R| - S] \quad \forall k \in K \quad (12)$$

$$V3(r, m, k) = (l(k) - d(r, m)) * Z(r, m, k) \quad \forall r \in R \quad \forall m \in M \quad \forall k \in K \quad (13)$$

$$V4(r, m, k) = Z(r, m, k) * u(r, m, k) \quad \forall r \in R \quad \forall m \in M \quad \forall k \in K \quad (14)$$

$$Z(r, m, k) \in \{0, 1\} \quad \forall r \in R \quad \forall m \in M \quad \forall k \in K \quad (15)$$

$$T(i, j, k) \in \{0, 1\} \quad \forall k \in K \quad \forall i \in I \quad \forall j \in I \quad (16)$$

$$V1(r, k, i) \in \{0, 1\} \quad \forall r \in [1, |R| - Q] \quad \forall k \in K \quad \forall i \in I \quad (17)$$

$$V2(r, k) \in \{0, 1\} \quad \forall r \in [1, |R| - S] \quad \forall k \in K \quad (18)$$

$$V3(r, k, m) \in \{0, 1\} \quad \forall r \in R \quad \forall m \in M \quad \forall k \in K \quad (19)$$

$$V4(r, k, m) \in \{0, 1\} \quad \forall r \in R \quad \forall m \in M \quad \forall k \in K \quad (20)$$

The objective function is reported in (1). Constraint (2) imposes that a referee cannot officiate on a matchday for which he is not available, while constraint (3) ensures that each match of each matchday is officiated by one and only one referee. Constraints (4) and (5) define how many times a referee may officiate a team, when it plays at home and when it plays away, respectively. In constraint (6), we compute the value of flags $T(i, j, k)$, indicating if a referee is scheduled to officiate a match between team i and team j , or not. Constraints (7) prevent that a referee officiate twice a match between the same teams. Constraint (8) implies that a referee can officiate a match only if he holds the necessary skill level. The minimum skill level required to officiate a match with difficulty “x” is “x-1”, but if the referee hold a skill level equal to “x-1”, a soft constraint is violated. (A referee of level 1 can officiate all matches without penalties, a referee of level 2 can officiate all matches, but a penalty is added if he officiates a match of level 1, a referee of level 3 may officiate matches of level 3,4 and 5 without penalties, matches of level 2 with penalties, and cannot officiate matches of level 1, and so on..).

Constraints (9) and (10) prevent the assignment of referees which are not allowed to officiate a team taking part of the match, at home and away, respectively. Constraints (11) - (14) exploit the soft constraints.

More in details, Constraint (11) is the so called *spacing constraint*, which impose that once a referee has officiated a team he cannot officiate a match, which the same team take part of, for a given number of matchdays, Q, otherwise a soft constraint is violated. Constraint (12) impose that a soft constraint violation occurs when a referee officiates for more than S consecutive matches. The same role is played by Constraints (13) when a referee officiates a match for which it holds the minimum requested skill level, and by Constraints (14) when a referee belonging to a lower division officiates a match. Finally, (15)-(20) specify the domains of the variables.

4. Referees assignment for the Italian SerieA

The above presented model has been tested on a real instances taken from the Italian Soccer Serie A, season 2010/2011. Data have been provided by the Italian Referees Association (AIA).

The Italian Soccer SerieA, to which, from now on, we refer simply as SerieA, is the major Italian soccer league and it is composed by 20 teams. Each team must play against each other team in the league, twice, once at home and once away. Thus, the number of matchdays in the season, R is equal to 38, and each matchday is formed by 10 matches, (M=10). The referees set is composed by 28 referees, 20 of which belongings to the SerieA and the remaining 8 to the lower division, SerieB. Referees should not officiates for more than two consecutive matches, (S=2), and, once they have officiated a team, they should not officiate it again for at least 4 matches (Q=4), otherwise a soft constraint violation occurs.

The number of soft constraints, belonging to each category can be computed as follows:

- V1: $(|R|-Q)*|K|^*|I| = 34*28*20 = 19040$
- V2: $(|R|-S)*|K| = 36*28 = 1008$
- V3: $|R|^*|M|^*|K| = 38*10*28 = 10640$
- V3: $|R|^*|M|^*|K| = 38*10*28 = 10640$

Thus, the total number of soft constraints amount to 41328. The model run under Xpress 7.3 on a machine with 8 GB of RAM and a processor CORE i7 working at 2 GHz, and took 10 seconds to close the problem to the optimality. In the optimal solution 72 soft constraints are violated, corresponding to the 0.17% of the total constraints, while the AIA declared that their referees assignment, performed by hand, and matchday by matchday, generally produce solutions in which the 3% of constraints is violated. These results underline the complexity of the problem, which cannot be fruitfully solved by hand, and show the efficiency and effectiveness of our approach.

5. Scenarios analysis

Planning the referee assignment for the whole season at the same time, as we do with our model, yields strong advantages, whit respect to performing the assignment operation matchday by matchday, because it allows to exploit better the available resources. In fact, planning one matchday at the time is very easy at the beginning of the season when all the resources (referees) are available, but may become very complex at the end of the season, when it could happen that most of the high skill level referees may be not assigned to a match because they have already officiate for the maximum number of times allowed one of the team taking part of the match, and we are obliged to assign that match to a less qualified referee with a consequent soft constraints violation. On the other hand, when some changes on the initial conditions occur, i.e. a referee skill level decrease because of some negative performances on the previous matches, or the difficulty level of a match increases because the two opponents are competing for the same position in the ranking or, although the end of the season is very near, they are still involved into reaching their seasonal goal (qualification for the European competition, avoiding to be relegated in the last positions of the ranking..), the assignment provided at the beginning of the season could be not anymore feasible because some hard constraints violations may occur. In the following, through a scenarios analysis, we show how we can easily overcome this issue, using a slightly modified version of the model.

5.1 Initial conditions perturbation

Analyzing historical data related to the past seasons of the SerieA, we define the following perturbations and their probability of realization:

- **Event1:** a match has a difficulty level higher than the expected one. (it happens when the two opponents are competing for the same position in the ranking or, although the end of the season is very near, they are still involved into reaching their seasonal goal). The probability of realization is almost null at the beginning of the season, and sensibly increase in the central part of the season and can be described by the following equation:

$$P(\text{Event1 at matchday } r) = \frac{(r-1)*e^{\frac{-r}{|I|}}}{50} \quad (21)$$

As shown in Figure 1, at the beginning of the season the probability of realization of this event is very low, while the peak is reached at the 22nd matchday, while the probability slightly decrease.

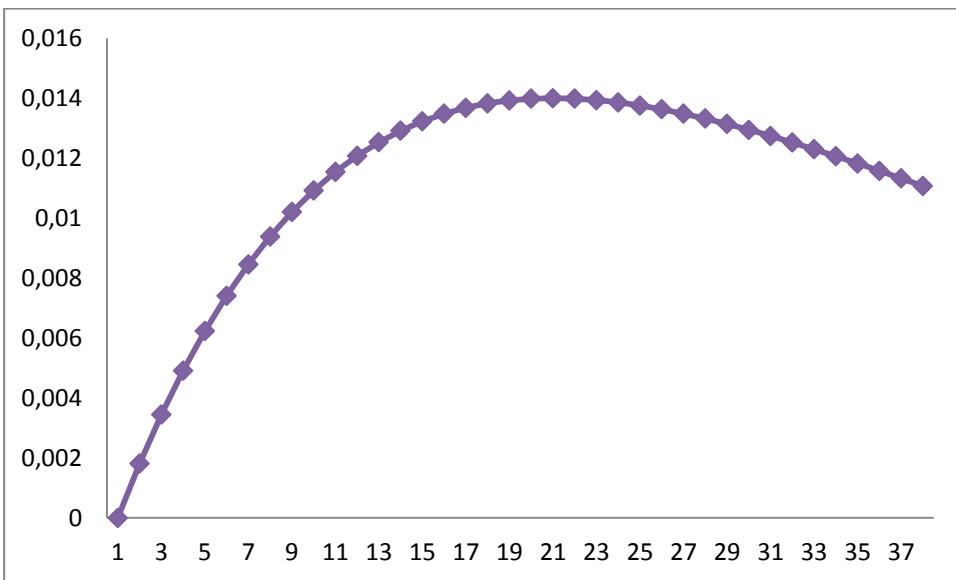


Figure 1. Graphical representation of Event 1 realization probability

- **Event 2:** a referee skill level may decrease during the season due his bad performances in some consecutive matches, and he could result to be not anymore qualified to officiate a match to which it has been assigned. The probability of realization of this event depends on the difficulty level of the match and the referee skill level, and can be expressed as follows:

$$P(\text{Event2 for referee } k \text{ in match } m \text{ of matchday } r) = \frac{5 - (l(k) - d(r, m))}{500} \quad (22)$$

- **Event 3:** a referee is not anymore available for a matchday in which he is supposed to officiate. In this case the probability, which has been computed observing how many times an injury or, more generally, a health problem occurs to the referees and which does not depend on the referee skill level, nor on the part of the season, is

$$P(\text{Event3 for a referee } k) = 0.005 \quad (23)$$

which means that, this event occurs, on average, twice a season.

- **Event 4:** a referee is not anymore allowed to officiate a given team (home, away or both) because last time he officiated the team there has been some arguments about the decisions he took during the match. The probability of realization of this event is:

$$P(\text{Event3 for a referee } k) = 0.01 \quad (24)$$

which means that, this event occurs, on average, 4 times a season.

According to the above defined probabilities, 10 scenarios were generated. The number of occurrences of events, for the different categories are reported in Table 1.

Table 1. Perturbations occurrences

SCENARIO	EVENT 1	EVENT 2	EVENT 3	EVENT 4
1	3	5	1	3
2	2	4	2	2
3	1	3	3	2
4	1	3	2	4
5	0	3	2	2
6	1	1	2	3
7	1	1	1	5
8	1	1	0	1
9	2	2	1	4
10	3	3	1	0

We address each scenario, run again the model, at each matchday in which an event occurs, taking fixed the assignment related to the past matchdays, while optimizing the remaining ones. These procedure, named MODEL allowed to recover from the infeasibility in all the analyzed scenarios, and to close to the optimality all the problems.

Another approach has been tested in which we reoptimize only the matchday in which the event occurs, instead of reoptimizing all the remaining matchdays. We named this procedure LS, (local search). Objective functions and computational times, related to the application of the model are reported in Table 2. From these results it is easy to notice that both the MODEL and the LS are able to overcome infeasibility problems and to find solution in which only 0.2% of soft constraints is violated, while the averaged percentage of violation 3% which occurs in the reality. Furthermore, applying LS, we may strongly reduce computational times without a significant loss of quality in the results.

Table 2. Scenarios analysis

SCENARIO	MODEL O.F.	LS O.F.	MODEL TIMES (s)	LS TIMES (s)
1	81	81	45	4
2	76	78	50	5
3	74	76	40	3
4	76	76	33	3
5	74	74	30	2
6	72	73	28	2
7	72	75	29	2
8	72	72	32	1
9	82	83	45	3
10	78	81	45	3
AVERAGED	75.7	76.9	37.7	2.8

6. Conclusions and future developments

In this paper we address the Referee Assignment Problem (RAP) is a novel arising problem in sports management, in which a limited number of referees with different qualifications and availabilities should be assigned to a set of games already scheduled, in order to respect a list of constraints. The goal of the problem is to find a feasible assignment, i.e. a configuration which allows to respect all the constraints given. An extension of the RAP is the Fair Referee Assignment Problem (FRAP), in which the objective is to minimize the violation of a set of soft (optional) constraints, while satisfying all the hard (mandatory) ones. We propose an integer programming model for the related FRAP. We apply it on an instance taken from the Italian Soccer SerieA, taken from the season 2011/2012. The problem can be solved to the optimality in very short time. The optimal assignment yield to a violation of only the 0.17% of the soft constraints, with respect on the average of 3% observed in the reality. These results strongly encourage the use of the proposed model in practical application. Moreover, a scenarios analysis in which perturbation on referees availability and match difficulty is performed and the model shows its capability to overcome this issue without a significant increase of soft constraint violations. Finally, a local search (LS) procedure, consisting in reoptimizing, each time a change in the referees availability or match difficulty occurs, only the involved matchday assignment. This method reach almost the same results, in terms of number of violated constraints, as the model, but within much smaller computational times. Further developments in this field could address the consideration of all four referees that are assigned to a soccer game. In this case both the modeling and solving the problem could become challenging tasks due to the increase in the number of variables and constraints. Similarly, referee assignments in other sports with different referee characteristics is another possible research direction. Furthermore, referee assignment in different contexts, as amateur and junior leagues, for which different fairness criteria may hold, could be addressed.

References

- Armstrong, J. and Willis, J. (1993) Scheduling the cricket world cup: a case study. *Journal of the Operational Research Society* **44**(11), 1067-1072.
- Bartsch T., Drexel A. and Kroger S. (2006) Scheduling the professional soccer leagues of Austria and Germany. *Computers and Operations Research* **33**(7), 1907–1937.
- Biajoli F., Chaves A., Mine O., Souza M., Pontes R. and Lucena A. (2004), Scheduling the Brazilian soccer championship: a simulated annealing approach. In: *Fifth international conference on the practice and theory of automated timetabling*, PATAT 2004.
- Clarke, S. (1998) Dynamic programming in one-day cricket optimal scoring rates. *Journal of the Operational Research Society* **39**(4), 331-337.
- Della Croce, F. and Oliveri, D. (2006) Scheduling the italian football league: An ilp-based approach. *Computers & Operations Research* **33**(7), 1963-1974.
- Duran G., Noronha T., Ribeiro C., Souyris S. and Weintraub A. (2006) Branch-and-cut for a real-life highly constrained soccer tournament scheduling problem. In: *Proceedings of the sixth international conference on the practice and theory of automated timetabling*, PATAT 2006.
- Freeze, R. (1974). An analysis of baseball batting order by monte carlo simulation. *Operations Research*, **22**(4), 728-735.
- Goossens, D.R. and Spieksma R. (2012) Soccer schedules in Europe: an overview, *Journal of Scheduling*, **15**(5), 641-651.
- Hersh, M. and Ladany, S. (1989) Optimal pole-vaulting strategy. *Operations Research* **37**(1), 172-175.
- Kendall, G., Knust, S., Ribeiro C., and Urrutia, S. (2010). Scheduling in sports: An annotated bibliography. *Computers & Operations Research*, **37**(1), 1-19.
- Normann, J. (1985) Dynamic programming in tennis: When to use a fast serve. *Journal of the Operational Research Society*, **36**(1), 75-77.
- Normann, J. and Clarke, S. (2007) Dynamic programming in cricket: Optimizing batting order for a sticky wicket. *Journal of the Operational Research Society* **58**(12), 1678-1682.

- Rasmussen, R. (2008). Scheduling a triple round robin tournament for the best Danish soccer league, *European Journal of Operational Research* **185**, 795-810.
- Ribeiro C. and Urrutia, S. (2006) Scheduling the Brazilian soccer championship. In: *Proceedings of the sixth international conference on the practice and theory of automated timetabling*, PATAT 2006.
- Russell, R. and Leung, J.(1994) Devising a cost effective schedule for a baseball league. *Operations Research* **42**(4), 614-625.
- Saltzman, R. and Bradford, R.(1996) Optimal realignments of the teams in the national football league. *European Journal of Operational Research* **93**(3), 469-475.
- Schonberger, J., Mattfeld, D. and Kopfer, H. (2004) Memetic algorithm timetabling for non-commercial sport leagues. *European Journal of Operational Research* **153**(1), 102-116.
- Schreuder J. (1992) Combinatorial aspects of construction and competition Dutch professional football leagues. *Discrete Applied Mathematics*, **35**, 301–312.
- Sphicas, G. and Ladany, S. (1977) *Optimal Strategies in Sport*. North-Holland, chapter Dynamic policies in the long jump. 101-112.
- Urban, T. and Russell, R. (2003) Scheduling sports competitions on multiple venues. *European Journal of Operational Research* **148**(2), 302-311.
- Van Voorhis, T. (2002) Highly constrained college basketball scheduling. *Journal of the Operational Research Society* **53**(6), 603-609.
- Washburn, A. (1991) Still more on pulling the goalie. *Interfaces* **21**, 59-64.
- Wright, M. (2009) 50 years of or in sport. *Journal of the Operational Research Society* **60**, 161-168.

Using Probabilistic Models to Simulate Tennis Matches, with Applications to Betting Strategies

Viktoras Mogilenko* and Gordon Hunter**

School of Mathematics, Kingston University, Penrhyn Road, Kingston upon Thames, KT1 2EE, U.K.

* viktoras.mogilenko@gmail.com

** (Author for correspondence) : G.Hunter@kingston.ac.uk

Abstract. Modelling and analysing the progress of tennis matches is a topic of considerable interest, both to players & their coaches, and observers with different interests in predicting outcomes, including bookmakers and gamblers. In this paper, we describe our computer simulations based on probabilistic models, and using point-by-point statistics for actual professional players on specific playing surfaces, to predict match outcomes. We also obtain results from “virtual bets” based on our simulations, using two different strategies – a fixed stake bet on a player to win a match and a Kelly “odds overlay” bet where the stake is modified according to the odds offered by the bookmaker and the probability of the player winning predicted by our simulations. We compare the results of these over a large number of such virtual bets, taking the actual outcomes of the matches into account, with those which would have been obtained from the odds offered pre-match by a commercial on-line betting exchange.

1. Introduction

Tennis is a very popular sport – both in terms of personal participation and as a spectator sport – across the World, and many fans follow their favorite players’ performances in the major professional tournaments. In recent years, there has been increased interest in gambling on the outcomes of professional tennis matches. In order to bet successfully on such results, one needs to have a method of predicting who will win a given match or tournament which is at least as reliable as the methods used by bookmakers and/or betting exchanges to calculate the odds of any particular player winning that event. (We will use “event” to mean point, game, set, match or tournament, as appropriate to the context.) In order to gain an advantage over the bookmaker (or exchange), gamblers need both a method of predicting match outcomes and a strategy for placing their bets in sensible ways. This paper will discuss both of these.

Tennis singles matches are a relatively simple, controlled situation to model : during each game, the two competitors alternately play strokes in a sequence called a “rally” until either one of them fails to strike the ball, or else hits it outside the area of play (the “court”) or into the net which divides the court in two. A game is won by the player who is the first to win at least 4 distinct points, and leading his/her opponent by at least 2 clear points, during that game. The same player initiates each rally of a particular game by “serving”, with the two players alternately taking turns to serve (i.e. player A will serve in the odd numbered games, with player B serving in the even numbered ones). A sequence of several games is called a set, with a player needing to win at least 6 games (and more than his/her opponent) to win the set. A match is won by the first player to win a specified number of sets – normally 2 in women’s tennis, and either 2 or 3, depending on the tournament, in men’s tennis. Thus, a player needs to win several points to win a game, several games to win a set, and several sets to win a match. However, in any given singles match, there will always be a player who wins, with the other player losing – there are no “draws” or “tied matches” in singles tennis.

The remainder of this paper is structured as follows. In section 2, we discuss several previous approaches to apply probabilistic models to tennis matches, and decide on one on which to base our computer simulation software. We describe the development of this software in section 3, including how data from real tennis tournaments was included in an attempt to make it produce realistic outcomes. In section 4, we discuss two different strategies, simple betting and Kelly “odds overlay” betting, for gambling on the outcomes of tennis matches. We implement both these strategies, using win probabilities estimated using our software and pre-match odds offered by a commercial betting exchange, to place “virtual bets” on the outcomes of all the men’s singles matches played over four professional tournaments from the Spring of 2012. The results of doing this are discussed in section 5. Finally, we present our conclusions and propose some possible directions for future work.

2. Probabilistic Models of Tennis Matches

Various previous authors have applied Markov-like probabilistic models to tennis matches, both at the level of the individual rally or point (Hunter et al 2007, 2008, Hunter & Zienowicz 2009) and at the level from the point upwards, modeling games as a collection of points, sets as a collection of games and matches as a collection of sets, according to the rules of tennis in relation to scoring (Barnett & Clarke 2005, Newton & Aslam 2009, Spanias & Knottenbelt 2012). In the latter type of model, previous authors have used statistics of how successful individual World-class players are at winning points on or against serve. This is then used to estimate the probability of a given player winning any particular point against a particular opponent. In the simpler models, unless “head to head” statistics are available for matches between that particular pair of players, these probabilities are based on averaging the players’ individual probabilities in an appropriate way. These can either be used in computer simulations of matches, or fed into recurrence relations in order to estimate the probabilities of winning any given game, set or match.

Even the most simple models, in which it is just assumed that each player has a particular fixed probability of winning or losing each point whilst serving or receiving, illustrate that quite a small advantage in win probabilities at the level of individual points on serve can result in serious “mis-matches” where the player holding the advantage wins easily over his/her opponent (Parramore et al 2005). This has also been observed in more sophisticated models (Barnett & Clarke 2002, O’Malley 2008, Knottenbelt et al 2012).

3. Computer Simulations

Our computer simulation program was written in VisualBasic and based on the models of Barnett & Clarke (2005) and Spanias & Knottenbelt (2012), where the probability of a given named player winning a point on serve when player a specific opponent is calculated using statistics of the server’s probability of winning a service point against an “average” opponent on the ATP circuit, combined with their opponent’s probability on winning a point when receiving serve, again with respect to an average opponent on the circuit. The statistics for players were obtained from the ATP website <http://atpworldtour.com>, which provides statistics on proportions of points won on and against serve by each ATP tour player in all ATP recognized tournaments, including breakdowns according to the type of surface each match was played on. The statistical data for a number of players, appropriate to the four tournaments – namely the ATP500 Barcelona Open, the ATP250 BMW Open, the ATP250 Estoril Open and the ATP2250 Serbian Open – from the Spring of 2012 investigated in this study. All these tournaments were played on clay courts. The computer program allowed that the probabilities used could be based on player statistics available for clay courts alone, for all surfaces, or “combined statistics”, where the probability of a player winning a point until particular conditions on a surface for which very limited data is available is estimated based on their statistics for each type of surface, weighted according to the number of matches played on each. For some players, insufficient statistics were available for them playing on clay courts (a minimum of 20 games each of serving and receiving were required), so in these cases only the “all surface” and “combined statistics” were used.

The results of simulating each of the matches in the 2012 Barcelona Open, using each of the types of available data on the appropriate players, were compared with both the “BetFair favorite” for each match and the player who eventually won that match. A further evaluation of the utility of the software from the perspective of betting on match outcomes was investigated by placing “virtual bets” on each match simulated, in accordance with each of the two betting strategies discussed in section 4 below. The results of the “winner predictions” and of the use of the program in the virtual betting are presented and discussed in section 5.

Further details of the design and implementation of the computer program can be found in Mogilenko (2012).

4. Betting Strategies

In this paper, we experiment with two different betting strategies, only using a fixed bet, the other adjusting the stake according to Kelly’s criterion for optimizing asymptotic return. In both cases, the odds used were obtained from the on-line betting exchange BetFair (<http://www.betfair.com/sport>). The computer

simulations made use of “decimal odds”, which for a particular bet is the reciprocal of the probability which the bookmaker is claiming the bet has of winning – i.e. if the bookmaker is claiming that a given bet has a probability q of winning, then the corresponding decimal odds would be $d = 1/q$. Decimal odds can readily be converted to traditional fractional odds (and vice versa), since a bet given traditional odds of “ α to 1” (i.e. placing a bet of 1 currency unit would, if the bet is won, result in the return of the stake and winnings of α units) corresponds to a (bookmaker’s) probability of it winning of $q = 1/(\alpha + 1)$, and so the equivalent decimal odds are $d = \alpha + 1$. Note that the probabilities effectively offered by bookmakers (i.e. those corresponding to the odds offered) do not normally sum to 1 over all possible outcomes, in order to allow the bookmaker a margin for profit.

4.1 Simple betting

In this approach, we assume the gambler stakes a fixed amount on each bet. This remains the same regardless of the odds offered by the bookmaker or the gambler’s predicted probability of winning the bet. However, the gambler will place the bet on whichever player he/she expects to win the match, based on the probability of each player winning given by our computer simulations. Given decimal odds d , if the bet is won, then the gambler gets back d currency units (including the original stake) for each currency unit gambled, otherwise the gambler loses his/her stake. For the purposes of this paper, there is no possibility other than “win” or “lose” on each bet made.

4.2 Kelly “Odds Overlay” betting

This approach, originally proposed by Kelly (1956), allows the gambler to adjust his/her stake according to the probability he/she believes the bet has to succeed and the odds offered by the bookmaker, in a way which should maximise the expected return in the asymptotic limit as the total number of bets placed tends to infinity. If the probability of a single bet resulting in a win is p , and the net odds offered by the bookmaker is “ α to 1”, then the gambler should bet a fraction f of his/her current capital, where $f = (p(\alpha + 1) - 1)/\alpha$. The gambler will back the named player if the “edge”, defined as $\alpha - ((1-p)/p)$, is positive, not bet anything if the edge is zero, and back the other player if the edge is negative. The Kelly approach can be shown to be equivalent to choosing to bet in a way which maximizes the geometric mean of the possible outcomes and has been discussed in the context of gambling on card games and other sports, and of investing in stock markets (Thorpe 1962, 1997). However, some authors have also issued caveats to strictly relying on this strategy in an attempt to maximize returns over a short period (e.g. Samuelson 1971), since the Kelly approach could lead to problems for the gambler, particularly if his/her estimates of the winning probabilities are not accurate. In our experiments, we once again make use of estimates of win probabilities from our own computer simulations of matches.

5. Results and Discussion

The initial evaluation of our computer program was in using it to predict the outcomes of matches in the 2012 Barcelona Open. Where appropriate data for each appropriate player was available, point winning probabilities calculated from clay court match statistics, all surface statistics, and “combined surface” statistics were each used to simulate each match of the tournament. The predicted winners of each match from these simulations were compared with the actual winners, and with the player predicted to win on the basis of the pre-match odds offered on BetFair. The results of predictions of the 40 first and second round matches played are given in table 1. The results show that our simulations perform comparably to use of the BetFair favorites. It has been noted (Wozniak 2011) that odds on on-line “betting exchanges” such as BetFair are effectively set by gamblers placing bets on the possible outcomes, and that there is good evidence that a high proportion of such on-line betting is done by people who are making use of simulation software. Thus, it is not surprising that the BetFair favorite provides a good prediction of match outcomes. The slightly disappointing result obtained from the clay court statistics, which might have been expected to provide the best predictions for this tournament, can possibly be explained by noting that many players do not play a very high proportion of all their matches on clay, so point winning probabilities based on statistics for this surface alone may not be all that reliable.

Table 1 : Prediction of 2012 Barcelona Open First and Second Round Match Winners using BetFair Odds, and our computer simulations using point winning probabilities calculated using statistics from different types of playing surfaces.

	BetFair Favorite	Clay Court Stats	“Combined Stats”	All Surface Stats
Total Matches Predicted	40	32	35	38
Correct Predictions	34	27	30	33
Percentage Correct	85.00%	84.38%	85.71%	86.84%

The program was also tested on match results from the ATP 250 Serbian Open, BMW Open and Estoril Open, all held in Spring 2012. The match prediction success rates, using point win probabilities calculated from the clay court and all surface statistics for each player, for first round matches across these three tournaments are shown in Table 2, where they are compared with the corresponding results from using the favorite (based on the pre-match BetFair odds) of each match to predict the winners.

Table 2 : Overall match winner prediction success rates for our program (using point probabilities based on clay court and all surface statistics) and for predictions using the BetFair favorite for the first rounds of the 2012 Serbian Open, BMW Open and Estoril Open.

	Using BetFair Odds	All Surface Statistics	Clay Court Statistics
Total Matches	27	27	18
Correct Predictions	15	17	13
Success Rate	55.56%	62.96%	72.22%

The predictions of the program were also used to investigate how effective its use could be for gambling, in conjunction with each of the betting strategies discussed in section 4 above. The probabilities of each player winning a match were estimated using our computer model, whereas the odds used were those provided pre-match on the BetFair website. The virtual bet was won, and a return calculated using the BetFair odds, if the player backed won the given match. Otherwise, the amount staked was lost. The results are shown in Tables 3 to 6 below, with ROI denoting Return On Investment, i.e. the profit made relative to the total amount staked.

Table 3 : Results of virtual betting on 2012 Barcelona Open (Men’s Singles only)

	Simple Betting			Total Amount Bet	Kelly Odds Overlay Betting		
	Clay Stats	Combined Stats	All Surface Stats		Clay Stats	Combined Stats	All Surface Stats
Number of 100 unit bets placed	39	33	35		621	501	497.26
Profit	855	904	1588		Profit	-20	46
ROI (%)	21.92	27.39	45.37		ROI (%)	-3.22	9.18
							43.11

Table 4 : Summary results of virtual betting on 2012 Serbian Open (Men’s Singles only)

	Simple Betting		Total Profit	Kelly Odds Overlay Betting		
	All Surface Stats	Clay Court Stats		All Surface Stats	Clay Court Stats	
Total Profit	61	-107	71.17	47.29		
ROI (%)	7.63	-13.38	66.11	30.08		

Table 5 : Summary results of virtual betting on 2012 BMW Open (Men's Singles only)

Simple Betting			Kelly Odds Overlay Betting		
	All Surface Stats	Clay Court Stats		All Surface Stats	Clay Court Stats
Total Profit	324	400	Total Profit	110.12	87.29
ROI (%)	27.00	80.00	ROI (%)	71.47	78.02

Table 6 : Summary results of virtual betting on 2012 Estoril Open (Men's Singles only)

Simple Betting			Kelly Odds Overlay Betting		
	All Surface Stats	Clay Court Stats		All Surface Stats	Clay Court Stats
Total Profit	-44	28	Total Profit	-40.36	30.94
ROI (%)	-6.29	4.67	ROI (%)	-22.14	19.12

It is somewhat difficult to make simple concluding statements from these observations. However, over the four tournaments studied, it would appear that use of our computer software in conjunction with point winning probabilities calculated from the ATP player statistics does provide a more reliable way of predicting the match winners than simply choosing the favorite (based on the BetFair odds). Furthermore, when applied to gambling, due to the exchange's allowed profit margins, backing the BetFair favorite when being given the BetFair odds cannot hope to yield a positive long term return. The success rate obtained using probabilities computed using the smaller amount of data available from ATP clay court matches – most appropriate for these tournaments - mostly yielded better predictions than those using the larger all surface datasets. The situation is less clear regarding the best set of ATP statistics to use when gambling. However, in most cases the Kelly odds overlay gambling strategy outperformed the simple betting approach in terms of return on investment obtained, although there were examples where this did not occur. This illustrates that, although the Kelly approach is likely to yield the best return in the very long run, it cannot be guaranteed to give a short term profit, especially if the gambler's assessment of winning probabilities is poor.

4. Conclusions and Future Work

We have shown that computer simulations, based on probabilistic models (Barnett & Clarke 2005, Spanias & Knottenbelt 2012) and point winning probabilities for individual player computed using ATP statistics can be used to predict match outcomes more reliably than the use of pre-match favorites based on betting exchange odds. Use of these simulations to place bets on the match results can, on the evidence of the tournaments studied here, usually result in a modest but positive return on investment. The Kelly odds overlay betting strategy (Kelly 1956), where the stake is modified according to the odds offered and the gambler's expected probability that the bet will win, was in most cases studied found to give superior returns than a simple fixed stake betting approach.

Future work could investigate the reliability of our findings by investigating a wider range of tournaments, and perhaps modelling women's tournament data. The models used could also be made more sophisticated. At present, the point winning probabilities for each player are estimated by comparing each player's serving and receiving performance relative to an average player on the ATP circuit. This could be refined by using statistics from previous "head-to-head" matches between the two players of interest (where available), and/or estimating the probabilities using statistics from matches where the players of interest have both played some common opponents – i.e. if the current players are A and B, the model would use data from matches between players A and C and between players B and C, where C has played each of A and B in the past. This approach has been investigated by Knottenbelt et al (2012). Furthermore, these models working "at the level of the point and above" could be combined with those working at the level of the individual rally or point

(Hunter et al 2007, 2008, Hunter & Zienowicz 2009) to produce stroke-by-stroke simulations of entire matches.

Acknowledgements

We would like to thank Dr. Nigel Atkins and Dr. Peter Soan, both of the School of Mathematics, Kingston University, for some valuable comments on this work, particularly relating to their in-depth knowledge of the sport of tennis.

References

- Barnett, T.J. & Clarke, S.R. (2002) Using Microsoft Excel to model a tennis match, *Proceedings of 6th Australian Conference on Mathematics and Computers in Sport* (ed. G. Cohen), Bond University, Queensland, Australia, pp. 63–68.
- Barnett, T.J. & Clarke, S.R. (2005) Combining player statistics to predict outcomes of tennis matches, *IMA Journal of Management Mathematics*, Vol. 16 (2), pp 113–120.
- Hunter, G. , Shihab, A. & Zienowicz, K. (2007) Modelling tennis rallies using information from both audio and video signals, *Proceedings of the I.M.A. International Conference on Mathematics in Sport*, pp. 103-108, Salford, Manchester, U.K., June 2007
- Hunter, G. , Zienowicz, K. & Shihab, A. (2008) 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 (JASA)*, Vol. 123 (5), pp 3431 and *Proceedings of International Conference on Acoustics (Acoustics '08)*, Paris, France, June 2008
- Hunter, G. & Zienowicz, K. (2009) Can Markov models accurately simulate lawn tennis rallies ? *Proceedings of the 2nd I.M.A. International Conference on Mathematics in Sport*, Groningen, Netherlands, June 2009
- Kelly, J.L. (1956) A New Interpretation of Information Rate, *Bell System Technical Journal*, Vol. 35, pp 917-926
- Knottenbelt, W.J., Spanias, D. & Madurska, A.M. (2012) A common-opponent stochastic model for predicting the outcome of professional tennis matches, *Computers & Mathematics with Applications*, Vol. 64, Issue 12, December 2012, pp 3820–3827, <http://dx.doi.org/10.1016/j.camwa.2012.03.005>
- Mogilenko, V. (2012) Tennis match prediction software using statistical modelling, BSc Computing with Statistics project report, Kingston University, U.K.
- Newton, P. & Aslam, K. (2009) Monte Carlo Tennis : A Stochastic Markov Chain Model, *Journal of Quantitative Analysis in Sport*, Vol. 5, No. 3, article 7.
- O'Malley, A.J. (2008) Probability formulas and statistical analysis in tennis, *Journal of Quantitative Analysis in Sports*, Vol. 4 (2), article 15.
- Parramore, K., Stephens, J. & Compton, C. (2005) Simulation : spreadsheets and repitions, Chapter 8, pp 216-218, in *Decision Mathematics 2 and C* (3rd Ed), MEI Structured Mathematics, Hodder Murray Education
- Samuelson, P.A. (1971) The fallacy of maximising the geometric mean in long sequences of investing or gambling, *Proceedings of the National Academy of Science*, Vol. 68, pp 2493 - 2496
- Spanias, D. & Knottenbelt, W.J. (2012) Predicting the outcomes of tennis matches using a low-level point model, *IMA Journal of Management Mathematics* (advance on-line access), doi:10.1093/imaman/dps010
- Thorpe, E.O. (1962) *Beat the dealer: a winning strategy for the game of twenty-one. A scientific analysis of the world-wide game known variously as blackjack, twenty-one, vingt-et-un, pontoon or Van John*, Blaisdell Publishing Corporation
- Thorpe, E.O. (1997) The Kelly Criterion in Blackjack, Sports Betting and the Stock Market, *10th International Conference on Gambling and Risk Taking*
- Wozniak, J. (2011) Inferring Tennis Match Progress from In-Play Betting Odds, MEng Information Systems Project Report, Imperial College London, U.K., <http://www.doc.ic.ac.uk/teaching/distinguished-projects/2011/j.wozniak.pdf>

Multivariate analysis of Heptathlon results

A. Murphy* and P. Bidgood**

* Kingston University (UK), asjmurphy@gmail.com

** Kingston University (UK), penelope.bidgood@kingston.ac.uk

Abstract. The notion that the heptathlon is the ultimate test of a female athlete was investigated and challenged. Factor analysis was used to extract common factors from the results in order to investigate the underlying athletic qualities that the events in the heptathlon measure. Athletes were clustered, methods of cluster analysis were compared and the members of clusters were analyzed to produce profiles of different types of athletes. The profiles and common factors were used to identify different classes of athletes and their success in the heptathlon. It was found that athletes did not need to excel in all events or abilities in order to rank highly in the heptathlon. The athletes who exhibited equally good ability in all of the events were arguably those who were the best athletes, despite not ranking top overall.

1. Introduction

The heptathlon is arguably the ultimate test of a female athlete, the multi disciplinary event tests athletes performance in seven athletics events covering running, jumping and throwing, in an attempt to assess overall athletic ability. The aim of this paper was to use multivariate analysis methods to further investigating the relationship between performance in different events and total score. It sought to identify common profiles of ability of athletes and identify the athletic abilities that cannot be directly observed but are tested through the seven events.

Cluster analysis had previously been used in the analysis of Olympic heptathlon results from 1992 (Dawkins, Andreea, O'Connor, 1994) and found clusters which each contained athletes who exhibited similar strengths in events. The results of athletes who competed in the 2004 Olympic heptathlon were placed into four clusters and three groups of athletes identified: the largest group contained athletes who achieved their lowest scores in throwing events; the smallest group contained athletes who achieved similar scores across all events; athletes who achieved comparatively high scores in throwing events but often ranked low overall.

The underlying athletic qualities measured by the heptathlon were examined through factor analysis. A principal component method was used initially to identify potential common factors and the maximum likelihood estimation of factor analysis was used to confirm the suspected common factors and further analyze the influence of each event in each factor. It was discovered that the Heptathlon measured three athletic qualities: agility, strength and fitness, throwing ability.

2. A first look at the Olympic Heptathlon

The outdoor Heptathlon which women compete in at the Olympics consists of the seven events completed over two days. They occur in the following order with the first four on day one and the last three on day two: 100m hurdles; high jump; shot put; 200m; long jump; javelin; 800m. The raw results from each event are converted into an overall points score using the tables in the appendix. athletes are then ranked by the sum of their scores across all events with the highest scorer being the winner. We focus on the results of the Women's heptathlon at the 2004 Summer Olympics, shown in the appendix, considering only athletes who completed all of the events.

Figure 1 shows a box plot of the scores achieved in each event. The hurdles had the highest mean score and one of the smallest ranges. In contract the throwing events had the lowest mean scores but the greatest ranges, suggesting athletes' ability in these events varies greatly but they also provide an opportunity to gain a large points advantage over other athletes. The upper body strength required to

perform well in throwing events could be considered cumbersome in the running and jumping events that make up much of the heptathlon.

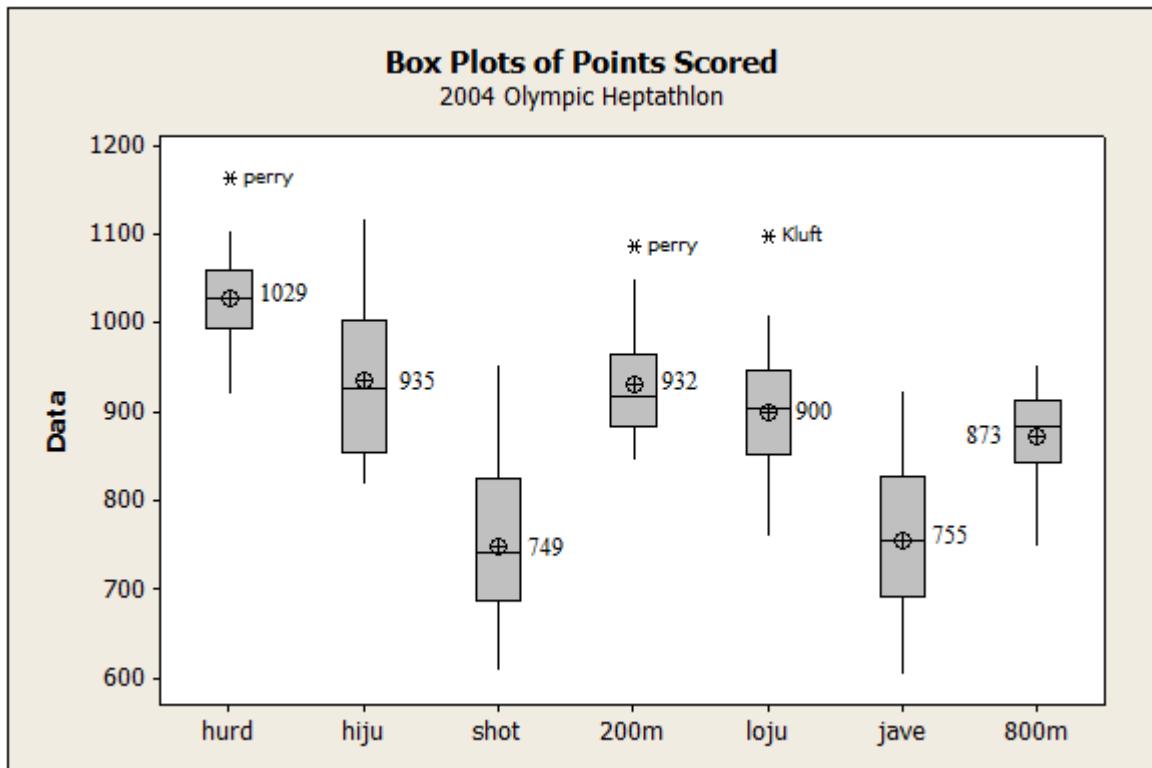


Figure 1. Box plots of the scores achieved in each event with means and outliers labeled.

The outliers in the hurdles and 200m events are the scores of Michelle Perry, who scored exceptionally high scores in these events but much lower in others which resulted in her finishing in 14th place. The outlier in the long jump is due to gold medal winner Carolina Kluft who also competed in the long jump as an individual event at the Olympics. Heptathlon world record holder Jackie Joyner-Kersee also previously held the world record for the long jump and numerous previous Olympic medal winners have competed in both events at an international level. Five of the top six ranking athletes have also competed in the Pentathlon. Many heptathletes also compete in the Pentathlon or individual jumping events but few compete in both.

From looking at figure 1 and the background of some heptathletes, we expect the points scored in the long jump and total points scored to be positively correlated. At the 5% level this is true and there is also some positive correlation between points scored in the high jump and total points. Surprisingly the shot put is also positively correlated with total score and the throwing events only have small negative correlation with some of the running and jumping events. The presence of correlation coefficients exceeding 0.32 (Tabachnick and Fidell, 2007) suggests that the model may be reduced using some method of factor analysis. We could explore the number of potential common factors through exploratory factor analysis.

3. Cluster analysis of Olympic Heptathlon data

Cluster analysis aims to form meaningful, homogenous groups of objects based on multiple attributes and it is hoped it can be used to group together athletes with similar athletic strengths and weaknesses. We expect that the athletes who also compete in jumping events and other multi-disciplinary events will fall into different clusters.

The complete linkage method, also known as further neighbor method, of cluster analysis is used as it obeys the conditions of space conservation and clump admissibility. The SAS Aceclus procedure is used on the athletes raw results and cluster analysis is carried out on the output to ensure the data being used has a spherical covariance matrix and meets the normality assumption of cluster analysis.

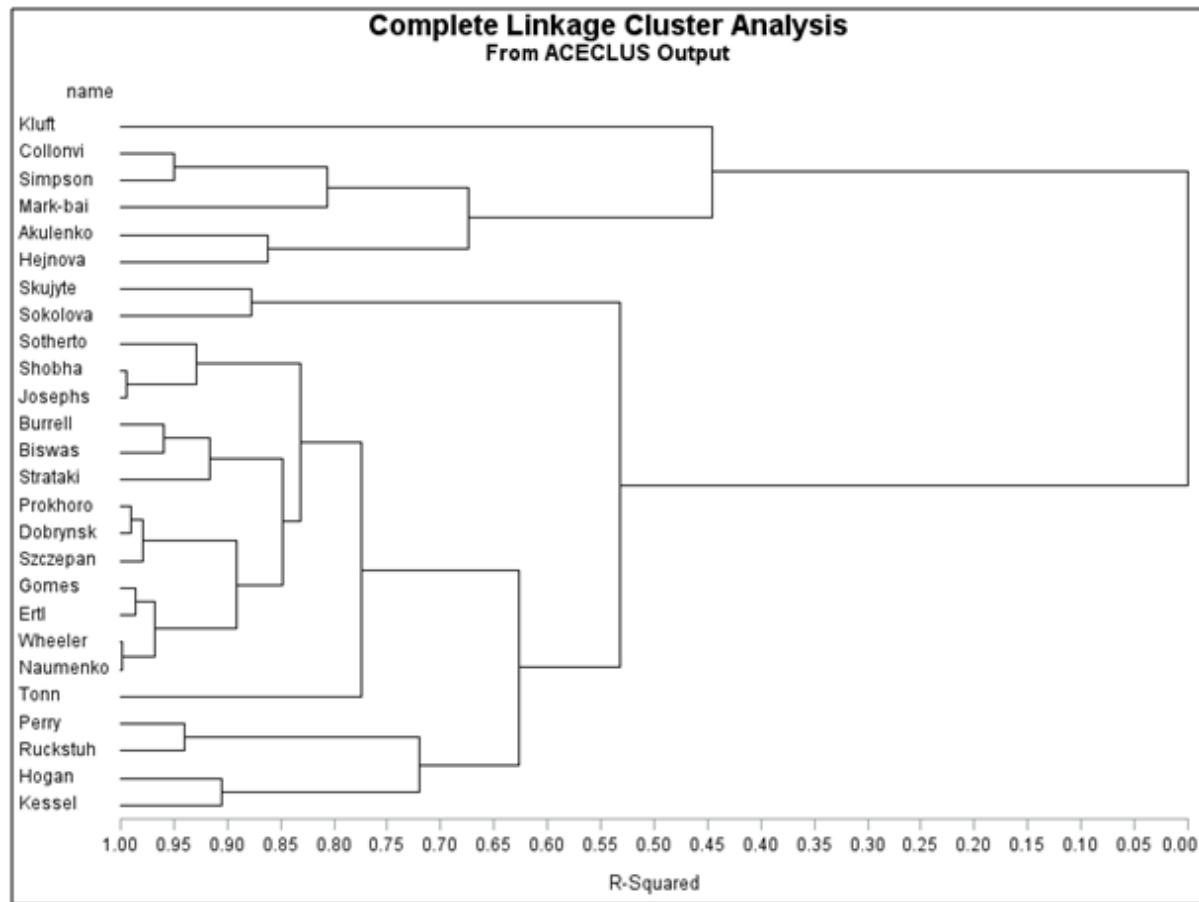


Figure 2. Dendrogram showing the proportion of total variance accounted for at each stage of clustering.

In the resulting dendrogram shown in figure 2 Kluft again stands out from the other athletes, being the last to remain in a single member cluster. Four or seven groups of athletes are suggested by the output statistics. Four clusters accounts for a smaller proportion of the variance within the results but produces better defined clusters that are more easily interpreted.

The athletes are clustered into four groups and the values of their first and second canonical variables are plotted against each other in figure 3. The four clusters are well defined but of unequal sizes, as emphasized by the shading around the clusters. The single member cluster is Kluft with her unusual, exceptionally good performance across all the events.

The two member cluster, A, contains Skujyte and Sokolova who performed similarly well in all events including throwing, having a small range of points scored across the events is very unusual but suggests a balance of all athletic abilities. The five member cluster labeled C contains athletes who did significantly better in throwing events than jumping, the opposite of typical heptathlon performance. The largest cluster, B, contains most of the athletes who competed; they achieved their worst scores in throwing events and score highly in running or jumping events and finished in a range of rankings overall.

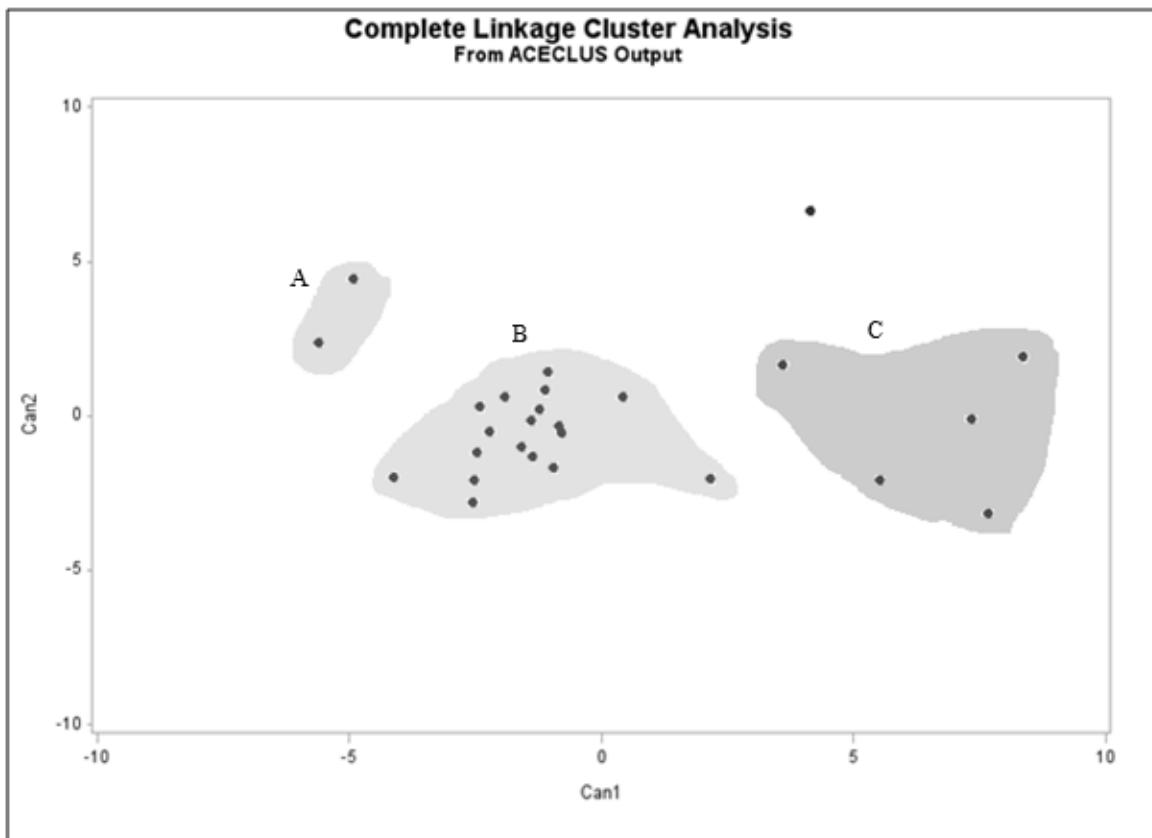


Figure 3. Scatter plot of the 1st and 2nd canonical variables with clusters outlined and labeled.

4. Identifying athletic qualities by extracting common factors

As found section 2, the correlation between the scores in many events in the heptathlon suggests that heptathlon results are factorable. It is hoped that factor analysis may identify the number of athletic qualities the heptathlon attempts to measure and help us to define these qualities that cannot be directly observed. A larger sample size than one occurrence of the Olympic heptathlon is required for factor analysis, the results of 107 athletes who completed all seven events at four international athletic events are used.

Exploratory factor analysis is carried out in SAS and the output used to determine the number of factors we wish to extract. The Kaiser-Guttmann criterion that the number of factors is equal to the number of eigenvalues greater than one suggests three factors, as shown in figure 4, accounting for 71.74% of the variance. The maximum likelihood estimation method of factor analysis is carried out in SAS and the chi-square statistics produced confirm that three factors is the smallest number of factors that sufficiently describes the variance within the model.

Three factors are extracted, the orthogonally rotated factor loadings and correlation of the factors with the events are shown in figure 5. The three factors are athletic qualities that cannot be directly observed and can be considered to be,

- Agility, related to performance in the 200m, long jump and hurdles. These are events with the highest mean scores and many athletes also compete in as individual events or achieve their highest scores in.
- The second factor can be considered to be an athlete's general all round strength and fitness. Many events have moderate loadings or are slightly positively correlated to this factor.
- Throwing ability and upper body strength. This factor is inversely related to the 200m and 800m. Most heptathletes perform much better in running events than throwing events and the

upper body strength required to achieve high scores in throwing events has a small negative effect on performance in running events.

	Eigenvalue	Proportion	Cumulative
1	2.66194584	0.3803	0.3803
2	1.32301361	0.1890	0.5693
3	1.03678188	0.1481	0.7174
4	0.67922380	0.0970	0.8144
5	0.59105409	0.0844	0.8989
6	0.39888391	0.0570	0.9558
7	0.30909686	0.0442	1.0000

Figure 4. Table of eigenvalues and the proportion of the variance within the results accounted for.

	Factor1		Factor2		Factor3	
100m hurdles	87*	0.79*	9	0.28*	18	0.27*
high jump	16	0.14	78*	0.92*	14	0.04
shot put	11	0.16	45	0.42*	67*	0.49*
200m	87*	0.95*	11	0.11	-22	-0.29*
long jump	72*	0.64*	38	0.57*	16	0.16
javelin	0	-0.03	-24	-0.09	84*	0.85*
800m	19	0.28*	74*	0.54*	-25	-0.35*

Figure 5. Rotated factor loadings followed by Pearson's correlation coefficients, asterisk denotes significance at a 5% level.

Orthogonal rotation of the factor loadings means the factors are not correlated with each other but there is correlation between factors and events as shown in figure 5. There is significant negative correlation between the 200m and 800m events and the third common factor, this confirms our previous speculation of the inverse relationship throwing ability on running ability. There are many significant positive correlations which indicate that an athlete's performance in most of the seven events is related to all aspects of athletic ability measured by the common factors. All athletes have a base level of fitness, strength and skill but it is in their specialist events and additional athletic abilities where they gain a points advantage against their competitors.

5. Conclusion

The clusters formed and factors extracted allow us to define and explain three broad profiles of heptathlete and the athletic characteristics of each,

- Athletes that are good runners and jumpers with throwing events their weakest, the majority of heptathletes fit this profile and their overall rankings vary greatly.
- Athletes who achieve their highest scores in throwing events but often do poorly in running events and rank low overall. These athletes are strong and good at throwing but are the least agile group.
- The smallest group of athletes is those who perform equally well in all events demonstrating well rounded athletic ability. Athletes that fall within this group do not always rank highly overall but a good ability in all events suggests a very good all round athlete.

If the heptathlon aims to be the ultimate test of a female heptathlete then arguably the best heptathlete is one with balanced abilities who excels in all events. This last, small group of athletes fits this description best of the three groups. Skujyte and Sokolova may not have won gold in the 2004 Olympic heptathlon but they best fit this description of an all round athlete. These well rounded athletes will possess greater all round strength and throwing ability than the large group of runner-jumper athletes whose total scores depend heavily on their agility

Appendix

Formula and coefficients for converting raw scores to points.

Event Type	Formula
Running	$P = a \cdot (b - T)^c$
Jumping	$P = a \cdot (M - b)^c$
Throwing	$P = a \cdot (D - b)^c$

Event	a	b	c
200m	4.990870	42.5	1.81
800m	0.111930	254	1.88
100m hurdles	9.230760	26.7	1.835
High jump	1.845230	75.0	1.348
Long jump	0.188807	210	1.41
Shot put	56.021100	1.50	1.05
Javelin throw	15.980300	3.80	1.04

Raw result followed by points scored and total points scored by Athletes in the 2004 Women's Olympic Heptathlon.

	100m hurdles		High jump		Shot put		200m		Long jump		Javelin		800m		Tot
kluft	13.2 1	109 3	19 1	111 9	14. 77	84 5	23. 27	105 2	67 8	109 9	48. 89	83 9	134. 15	90 5	695 2
skujyte	14.0 3	974	17 6	928	16. 4	95 5	24. 82	903	63 0	943	49. 58	85 2	135. 92	88 0	643 5
sotherton	13.4 4	105 9	18 5	104 1	13. 29	74 7	23. 57	102	65 1	101	37. 0	61 19	132. 3	93 27	642 5
burrell	13.1 7	109 9	17 0	855	13. 14	73 7	24. 06	975	62 5	927	47. 69	81 5	135. 32	88 8	629 7
prokhorova	13.8 4	100 1	17 9	966	13. 67	77 2	24. 71	914	62 1	915	45. 58	77 5	131. 31	94 6	628 9
kessel	13.3 8	106 8	17 6	928	14. 53	82 9	25. 23	866	64 2	981	42. 99	72 5	135. 21	89 0	628 7
collonville	13.6 5	102 8	18 5	104 1	12. 35	68 4	25. 26	863	61 9	908	49. 14	84 3	133. 62	91 2	628 0
dobrynska	13.8 9	994 2	18 3	100	14. 7	84 1	25. 02	885	62 3	921	44. 08	74 6	137. 01	86 5	625 5
simpson	13.5 6	104 1	17 9	966	12. 41	68 8	24. 62	922	60 2	856	53. 32	92 5	137. 72	85 5	625 3
sokolova	13.7 1	102 0	17	855	14. 61	83 5	24. 21	961	58 4	801	47. 86	81 9	133. 23	91 8	621 0
shobha	13.5 3	104 6	16 7	818	12. 52	69 6	23. 41	103	63 8	962	44. 36	75 1	137. 28	86 1	617 3

tonn	13.9	993	18 2	100 3	11. 92	65 6	24. 84	902	63 5	959	41. 42	69 5	130. 77	95 3	616 1
gomes	13.5 8	103 9	18 5	104 1	14. 71	84 1	25. 46	845	61 0	880	40. 75	68 2	140. 05	82 3	615 1
perry	12.7 4	116 4	17 0	855	11. 28	61 4	22. 91	108	60 2	856	38. 36	63 6	133. 69	91 1	612 4
strataki	13.6 5	102 8	17 9	966	13. 52	76 2	24. 57	927	59 7	840	43. 87	74 2	137. 9	85 2	611 7
ruckstuhl	13.4 4	105 9	18 5	104 1	13. 37	75 2	24. 59	925	59 0	819	36. 7	60 4	133. 95	90 8	610 9
ertl	13.5 2	104 7	17 3	891	13. 92	78 9	24. 71	914	60 3	859	44. 45	75 3	138. 68	84 2	609 5
wheeler	13.8 8	995 9	17 9	966	13. 18	73 9	24. 35	947	63 6	962	37. 77	62 5	137. 65	85 6	609 0
josephs	13.6 9	102 3	17 0	855	12. 48	69 3	23. 37	104	62 1	915	41. 8	70 2	138. 47	84 4	607 4
hogan	13.1 3	110 5	16 7	818	14. 43	82 3	24. 99	888	61 5	896	45. 84	78 0	145. 1	75 6	606 6
szczepan ska	14.4 1	921 6	17	928	13. 79	78 0	25. 29	860	59 8	843	44. 8	76 0	133. 08	92 0	601 2
naumenko o	14.1 6	956 9	17	966	12. 95	72 4	24. 88	898	61 6	899	39. 5	65 8	134. 57	89 9	600 0
akulenko	14.1 1	963 3	17	891	13. 15	73 7	24. 57	927	60 2	856	48. 62	83 3	142. 58	78 9	599 7
biswas	13.8 6	998 0	17	855	12. 01	66 2	24. 5	933	59 2	825	44. 84	76 0	132. 27	93 2	596 6
mark- baird	13.5 8	103 9	17 0	855	11. 2	60 8	25. 11	877	62 2	918	49. 9	85 8	141. 21	80 7	596 3
hejnova	13.8 2	100 4	17 0	855	12. 13	67 0	25. 36	854	57 0	759	48. 22	82 6	145. 68	74 8	571 6

References

- Dawkins, B.P., Andreae, P.M. and O'Connor, P.M. (1994) 'Analysis of Olympic Heptathlon Data', *Journal of the American Statistical Association*, vol. 89, no. 427, p.1100-1106.
- Zapnas, K. G., and Zeller, R. A. (2002) 'Minimizing Sample Size When Using Exploratory Factor Analysis for Measurement', *Journal of Nursing Measurement*, vol. 10, no.2, p135-153.
- Tabachnick, B.G. and Fidell, L.S. (2007) *Using Multivariate Statistics*. 5th Edition. Pearson Education Inc.
- Everitt, B.S, Landau, S. and Leese, M. (2009) *Cluster Analysis*. 4th Edition. Wiley.
- (2004)IAAF Scoring Tables For Combined Events [Online] Available from:
http://www.iaaf.org/mm/Document/Competitions/TechnicalArea/ScoringTables_CE_744.pdf
[Accessed 1st September 2011].

Market making with an inverse Kelly strategy

E. Noon* and W. Knottenbelt**

*Imperial College London, South Kensington Campus, London, SW7 2AZ. en108@doc.ic.ac.uk

** Imperial College London, South Kensington Campus, London, SW7 2AZ. wjk@doc.ic.ac.uk

Abstract. Kelly's celebrated staking system calculates the optimal fraction of wealth to bet on each of a series of favourable bets. We previously extended Kelly's ideas to allow for the much wider range of bets available today, in particular laying as well as backing, and including previously placed bets. Here we extend this further by suggesting that it may be inverted to provide a market making tool. For an event we set the odds to be those which Kelly would have chosen for a fixed fraction. We show how the fraction used here may be thought of as a measure of how tightly the prices will be quoted. As previously matched bets are included in the calculation the quoted prices will be able to respond to errors in the model used; the market making system is independent from the model. Prices are able to respond to trades made in related markets. In order to provide data for a future simulation we ran a number of tests of this method on a live exchange, using different fractions and different sizes for several different markets in the lower English football leagues. We present some preliminary findings from this data.

1. Introduction

Kelly's fractional staking has been much discussed in literature in the 50 years since first publication. Many are convinced: Breiman (1961); Thorp (1969); Markowitz(1976) and Bell and Cover (1980). Others are not: Samuelson (1971) and Merton and Samuelson (1974). Much of the published work over the recent decades has concentrated upon financial markets, for example, Cover (1991) and Bell and Cover (1980) . For many years after Kelly's work was published there was little innovation by bookmakers. Recently in some countries there has been a relaxation of some of the strict laws which control gambling, and this has brought about change.

The introduction of betting exchanges, in particular, has been responsible for much of this. Exchanges are peer to peer betting platforms, (individuals place bets with each other rather than a bookmaker). The exchange is responsible for settling bets and taking and making payments, but not for the risk. Now when considering our portfolio a there are new choices available. Two of the most interesting are the ability to place lay bets (i.e. betting that something will not happen) and taking bets off before the race has finished or even started.

We, Noon et al (2012), have previously extended Kelly's staking to include these new possibilities and now we turn our attention to a new possibility. We are able to place bets which won't immediately match, but if they do they will trade at a better price than that previously available. In particular, we consider the case when we are the first to place a price in a particular market and we intend to place not just one price, but all possible prices. This activity is often referred to as market making.

2. Market Making

On the popular exchanges major events such as horse racing at the leading courses and top flight football have substantial activity. A considerable time before the start of such an event there are usually market prices with little over-round (beyond the exchange commission). The markets in lower profile events such as football matches several divisions from the top are less well served. On a recent Wednesday the over-round for betting on the correct score for the following Saturday's League One matches was about 12%, and there is no realistic price in any market for matches on the Tuesday. (English League One is the third level, after the Premiership and the Championship.) Anyone planning to place a bet would need to choose a price, rather than merely deciding to trade, or not to trade with an existing price. We suggest choosing a notional sum of money, C , and then selecting the decimal odds o such that Kelly's fractional stake f gives the desired bet size b . Then if p is the probability of the backed event from Kelly we have:

$$f = p - \frac{(1-p)}{(o-1)},$$

rearranging this gives:

$$o = 1 + \frac{1-p}{p-f}. \quad (1)$$

From this we see that we need $p > f$. If offering lay bets as well as back bets there is another important consideration. Each of the entries is considered on its own, but it is possible that more than one will trade at once. For example, if you are offering back and lay prices on a tennis match it is possible that someone may match your back on one player and your lay bet on the other player. With some exchanges this is in fact likely; Betfair's web client automatically does this calculation and a client will see size which is generated by combining a direct transaction with appropriate combinations of other bets. So b should be chosen to allow for this.

To be practical this method combines existing bets into the calculation and at the same time we consider multiple markets. So if an event has n possible outcomes and the probability of outcome i is p_i , and if there are m possible bets which depend solely upon those outcomes, then bet j has odds o_j and traditionally we are choosing the fraction of our wealth x_j to wager on bet j . For each of the j we can take into account previously made bets with size s_j and odds e_j . M is a matrix with $M_{ij} = 1$ if bet j is considered winning if outcome i occurs and $M_{ij} = 0$ otherwise. Previously we would have wanted to maximise A where:

$$A = \sum_i p_i \log \left(1 + \sum_j M_{ij} (o_j x_j + e_j s_j - x_j + s_j) \right).$$

In this case we find for each j in turn the value of o_j such that maximising A gives $x_j = f$. Again we must ensure that $f < p_i$ for all i . Other than that constraint we are free to choose f as necessary to control the desired over-round. From equation 1 we can see that $1/o = (p - f) / (1 - f)$, so the lay under-round if we place m back bets is

$$\frac{1 - mf}{1 - f}.$$

The choice of C becomes a decision about the confidence of the model used. If making markets on one side only (all backs, or all lays) we might choose it such that $f = b/C$, or similarly when offering bets on both

sides $f = 2b/C$, where b is the size of the bet placed in the market. In this case the prices will move so as to unwind the first bet placed at our fair value, or the middle of our previous market. With higher confidence in a model a higher value of C would seem suitable as prices would change more slowly with trading activity.

3. Practical Test – Preliminary Results

In previous work we have been able to create a test using historical data. Suitable assumptions, such as a test size much smaller than the market so as not to impact it needed to be made. The significant point in this technique is that prices aren't available at the time the bets are placed into the exchange. Whether the bets will be matched immediately, improved upon, or ignored is not immediately obvious. In the near future we will build a simulation tool to assist with this work, but we will need some data for calibration. One of the advantages of studying betting markets rather than financial markets is that it is possible to make tests in real markets with a considerably small capital outlay; the minimum bet on Betfair is £2.

Using a model based upon Dixon and Coles (1997), fitted to the previous matches of the current season and the two prior seasons, but with a value for tau of zero we estimated the probabilities of final scores for English Football League One and Two (the third and fourth level respectively). When a match was added to

Betfair the market making tool randomly assigned it a category. One category was to ignore the market, placing no bets, offering these markets as controls. Other categories had various choices for b , C and f . A final group, category 8, reduced f as confidence in the current prices increased. For those active matches the tool quoted markets in Match Odds, Over/Under 2.5 and Over/Under 3.5. We had hoped to quote in Correct Score markets but the cost of doing so was deemed too high for the initial test. The amount of capital needed to quote back and lay bets of size b in each of the m runners of a market is $b(m + l - 2)$, where l is the highest lay odds placed. The market making tool ran, with some interruptions, from 11th February to 28th March 2013, a period of 215 football matches. Over this time we placed 6 415 bets.

In several cases the software was interrupted and in some others there were human interventions which introduced errors. Sometimes the thread detecting new markets in Betfair was too slow and by the time it placed prices in the market there were already prices there. Occasionally these matched immediately. One of the parameters our future simulation will need to fit is the inter-arrival times for the bets. These almost instant matching bets skew the data significantly. Some of these we can easily detect and have excluded reducing our total bet count to 5722 bets. Others are less obvious. We are still in the early stages of cleaning the data and need to combine several log files to sort out those occasions when we are less confident with the data. Here are some preliminary results.

Table 1: Showing the spread of data across the two leagues we considered, and the average volume of matched bets.

	Matches Ignored	Category 1-7	Category 8	Total Matches
League One	45	49	13	107
League Two	46	47	15	108
Market Size (£)	39 682	40 689	84 669	

These are not uniformly distributed across this period. In the first half of the time period there is a higher proportion of ignored matches than there is later in the period. The market size is the average of the total volume of matched bets in each of our market categories. We started quoting in a low proportion of matches and increased this as there were no significant losses. As previously stated, quoting in the Correct Score market was particularly expensive and was, therefore, assigned a low probability. This happened only once, matching 19 lay bets not including the winning score.

Table 2: The number of back and lay bets for each type of market considered.

	Match Odds	Over/Under 2.5	Over/Under 3.5	Correct Score
Back	1065	577	216	0
Lay	2864	749	232	19

It is immediately obvious for markets with more than two outcomes that customers prefer to make back bets (so we make lay bets). Perhaps for binary markets customers still favour back bets, but the Betfair web client converts our back bets into lay bets on the other outcome.

We had speculated that quoting prices in a market might increase the total activity in that market (beyond our own activity). From the above table it can be seen that in general it doesn't, at least not significantly. The exception is Category 8. This type of market maker quoted tight prices ($f=0.015$) after an initial settling down phase. Performing an F-test on the data shows that the variances are significantly different with a high degree of confidence. We are investigating further.

We had also expected that bets which matched quickly might be more likely to be losing bets. To estimate this we calculated for each market a money weighted average odds, normalised so that the reciprocal gives a market implied probability. We use this to calculate an expected profit on each bet. We anticipated a positive correlation between this and the time between placing a bet and it being matched. Surprisingly the calculated correlation is -0.07. It is possible that when the data has been further cleaned this will change, but it is surprising. It is possible we need to consider how far in advance of the match we are placing bets. If we place bets too far ahead it may be that no-one notices for many hours, or days, and this will probably include the bet with the highest negative expected value.

We look forward to investigating this data further, to try to reveal some hidden information.

References

- Bell, R. M. and Cover , T. M. (1980) Competitive Optimality of Logarithmic Investment. *Mathematics of Operations Research, INFORMS* **5**, 161-166.
- Breiman, L. (1961) Optimal Gambling Systems for Favorable Games. *Proceedings Fourth Berkeley Symposium*, **1**, 65-78.
- Cover T. M. (1991) Universal Portfolios. *Mathematical Finance*, **1**, 1-29.
- Dixon, M. J. and Coles, S. G. (1997) Modelling Association Football Scores and Inefficiencies in the Football Betting Market. *Journal of the Royal Statistical Society. Series C*, **46**, 265-280
- Markowitz, H. M. (1976) Investment for the Long Run: New Evidence for an Old Rule. *The Journal of Finance*, **31**, 1273-1286.
- Merton, R. C. and Samuelson, P. A. (1974) Fallacy of the log-normal approximation to optimal portfolio decision-making over many periods. *Journal of Financial Economics*, **1**, 67 – 94.
- Noon, E., Knottenbelt, W. J. and Kuhn, D. (2012) Kelly's fractional staking updated for betting exchanges. *IMA Journal of Management Mathematics* 2012; doi: 10.1093/imaman/dps015.
- Samuelson, P. A. (1971) The "Fallacy" of Maximizing the Geometric Mean in Long Sequences of Investing or Gambling. *Proceedings National Academy Sciences USA*, **68**, No. 10, pp 2493-2496.
- Thorp, E. O. (1969) Optimal Gambling Systems for Favorable Games. *Review of the International Statistical Institute*, **37**, 273-293.

Elite Players' Perceptions of Football Playing Surfaces: An Ordinal Logistic Regression Model of Players' Overall Opinions

A. Owen*, A. Smith**, P. Osei-Owusu**, J. Roberts**, A. Harland** and S. Larman***

*Mathematics Education Centre, Loughborough University, LE11 3TU, UK. a.j.owen@lboro.ac.uk

**Sports Technology Institute, Loughborough University

***Member Associations and Development Division, FIFA

Abstract. Since 2004, the use of modern artificial surfaces in football (soccer) has been sanctioned by FIFA (the world governing body for football) and UEFA (the European Football Confederation) for use in competitive matches. In 2012, the Sports Technology Institute at Loughborough University collaborated with FIFA, its member associations and FIFPro (the professional players' organisation), to collect data from over 1,000 elite players world-wide in order to investigate players' perceptions of different playing surfaces and the influences these surfaces may have on various aspects of the game. The aim of this paper is to communicate some of the initial results from the study, but also to illustrate the application of ordinal logistic regression to data on players' perceptions of surfaces in world football. A Principal Components Analysis (PCA) is also undertaken in order to address dimensionality and multi-collinearity issues with the data, which also considers the use of polychoric correlations within the PCA due to the ordinal nature of the data. Players' overall opinions were found to depend on their experience of different surfaces during their junior and senior careers, and also to some extent on the FIFA confederation where they are currently playing.

1. Introduction

Following a Commission of Enquiry into Playing Surfaces conducted in 1989, the English Football League banned the use of artificial surfaces for competitive league matches. As well as concerns over ball behaviours and player injuries on these surfaces at that time, the commission also considered concerns regarding an unfair home advantage outlined in Barnett and Hilditch (1993). Since that time however, there have been ongoing improvements in the quality of grass surfaces, and also developments in artificial surfaces which have led to the production of Third Generation (3G) pitches (more commonly referred to by FIFA as Football Turf) that better replicate the qualities of a natural grass surface (referred to in this study as Natural Turf, although this surface is now commonly referred to as Natural Grass by FIFA). Since 2004, FIFA has sanctioned the use of Football Turf pitches in international and professional league matches, where the surface meets a minimum 1-star standard (although some leagues and tournaments require a minimum 2-star standard) outlined in FIFA's Quality Concept for Football Turf (www.FIFA.com).

As part of the ongoing work being undertaken by FIFA to understand the influence of the playing surface on the game of football, the Sports Technology Institute at Loughborough University has collaborated with FIFA, its member associations and FIFPro, to collect data via a questionnaire from over 1,000 elite players world-wide. The questions covered their personal experiences of different surfaces and injuries, along with their perceptions of how surface characteristics affect injury risk, style of play, footwear choice and ball-surface interactions.

There is evidence, for example in Nigg and Yeadon (1987), supporting the view that lab-based material assessments of biomechanical aspects of playing surfaces cannot fully take account of the overall experience of players during the game. Therefore players own assessments and perceptions are an important factor when considering differences between playing surfaces. However, the published research relating to such player perceptions in this context is limited. Exceptions to this include Anderson et. al. (2008) which reports on elite players perceptions of Football Turf (referred to as artificial turf in their work) versus Natural Turf, but their study is restricted to players only from the top leagues in Swedish football and only to a relatively small sample size of just 93 players (72 male and 21 female). Zanetti (2009) however, considers a much larger sample (1,671) of male players' perceptions of Football Turf (again referred to there as artificial turf), but this only relates to amateur players in Italy. The present study therefore represents a significant addition to the available data in this area, since it provides information on the perceptions of different playing surfaces from over 1,000 elite players worldwide, almost all of whom are professional.

Initial descriptive summary analyses of the data have been reported to FIFA and its member associations and this paper presents some of these initial results, but in addition it reports on the results of further investigations. The aim is to both communicate some of the key results from the study to as wide an audience as possible, and also to illustrate the application of ordinal logistic regression to data on players' perceptions of surfaces in world football. However, whilst the data derived from the current study represents a large and potentially rich source of information on players' perceptions and opinions, the limit on time and space means that only one or two aspects of this study can be reported on here. This paper therefore focuses on players' overall opinions on the use of Football Turf compared to Natural Turf. An overview of the study and the data is first provided in Section 2 followed by an examination of players' overall opinions of Football Turf compared to Natural Turf in Section 3. It is anticipated that geographical location may have a bearing on their perceptions of these surfaces, and therefore Section 3 also presents comparisons between the confederations in which they are currently playing. Previous surface experience is also anticipated as having a bearing on players' perceptions of different surfaces, and therefore data on players' career surface experience is discussed in Section 4. These factors are then combined in Section 5 which considers an ordinal logistic regression model to assess the impact of confederation and surface experience on players' perceptions.

2. Study and Data Overview

Data was collected via a questionnaire from a total of 1,129 elite players (1,018 male and 111 female), almost all playing at a professional level. An informal and pragmatic non-random cluster sampling approach was used to identify a total of 45 countries across the six FIFA confederations which consist of AFC (Asia), CAF (Africa), CONCACAF (North and Central America and the Caribbean), CONMEBOL (South America), OFC (Oceania) and UEFA (Europe). Within each of the countries chosen, a convenience sample of clubs was identified, along with a small number of tournaments organized by FIFPro, from where data could be collected locally. The total number of countries and the number of players that took part in the study in each confederation are shown in Table 1. The overall sample size allows for a margin of error of approximately +/-3% for the overall estimated proportions from the study. The sample sizes within confederations were derived from a balance between achieving a reasonable level of accuracy and the practicalities of collecting large amounts of data from a wide range of geographical locations. Note that the number of players in the OFC included in the study is lowest amongst the confederations and so any inferences will need to be treated with the greatest amount of caution for this confederation.

Table 1: Number of countries and players by confederation

	AFC	CAF	CONCACAF	CONMEBOL	OFC	UEFA	Total
Countries	8	8	2	8	3	16	45
Players	161	199	99	115	50	505	1,129

The age distribution of the players ranged from 18 to 39 years and these were similar across each confederation. However, whilst data on male players was collected from almost all countries, data on female players was only collected from just six countries; England (37), France (17), Germany (25) and Iceland (7) within UEFA, and Japan (25) and Singapore (1) within AFC. Gender is therefore not considered here due to the small sample size but further work is currently ongoing to gather additional data from female players.

3. Player Opinions

3.1 Overall Player Opinions

Prior to the commencement of the study, interviews and focus groups were conducted with a small group of players, which highlighted a number of key statements representing commonly reported issues. In the subsequent study these issues were presented to the players in Part 6 of the questionnaire as a series of statements. Players were asked to indicate the extent to which they agreed with each of these statements on a five-point ordinal scale consisting of "Strongly Disagree", "Disagree", "Neutral", "Agree" and "Strongly Agree".

This paper considers three of these statements which were as follows:

Q6.1C “Teams that play on Football Turf pitches have a big advantage for home games”

Q6.1E “All top level fixtures should be played on Natural Turf”

Q6.1F “I would rather play on a modern Football Turf pitch rather than a poor quality Natural Turf pitch”

A total of 1,119 players responded to each of these questions and the distribution of their responses to these statements are summarised in Table 2. This also includes the results of Wilcoxon Signed Rank tests, based on scoring the responses to each statement using 1 = “Strongly Disagree”, 2 = “Disagree”, 3 = “Neutral”, 4 = “Agree”, and 5 = “Strongly Agree”, in order to assess the significance of the median score being different from the neutral position of 3. In all three cases there was evidence ($P<0.001$) that overall players tend to agree with the statements. The general agreement with Statement C, that Football Turf offers a big home advantage, as well as Statement E, that all fixtures should be played on Natural Turf, is consistent with the overall negative impression of Football Turf reported in Andersdon et. al. (2008). However, the general agreement with Statement F indicates that the majority of players would rather play on a modern Football Turf pitch than a poor quality Natural Turf pitch, which suggests positive support for Football Turf. This would appear to offer a contradiction with the results for Statements C and E and so the emerging picture of players’ perceptions of Natural Turf and Football Turf appears not to be so straight forward.

Table 2: Distribution of players’ responses to statements

Statement	Strongly Disagree (%)	Disagree (%)	Neutral (%)	Agree (%)	Strongly Agree (%)	Wilcoxon S.R. Test	
						Median Score	p
C	4.8	12.0	17.4	43.4	22.4	4	<0.001
E	4.7	8.7	10.7	25.0	50.9	5	<0.001
F	15.2	14.4	15.6	36.2	18.7	4	<0.001

3.2 Player Opinions by Confederation

One aspect to consider is the variation in the types and quality of surfaces currently being used by players in different parts of the world. This could be governed, for example, by local climatic conditions, finances and national league policies with respect to the use of Football Turf. One simple way of assessing the potential impact of these local factors on players’ opinions, is to make comparisons between the six member confederations. The cumulative percentage distributions of responses for the three statements by confederation are shown in Figure 1. For Statement C, that Football Turf offers a big home advantage, Figure 1(a) illustrates that apart from players currently playing in AFC there is quite a degree of consistency of opinion across almost all other confederations. The higher curve for AFC in Figure 1(a) appears due to more players from this confederation taking a neutral view or disagreeing with the statement and fewer players agreeing with it. Figure 1(b) indicates a slightly greater level of variation between confederations with respect to Statement E, which referred to the view that all fixtures should be played on Natural Turf. Again players currently playing in AFC appear to take more of a neutral view with fewer players disagreeing with this statement, whilst players in CONMEBOL appear to agree most often in this case. Figure 1(c) indicates a much broader range of opinion between confederations with regard to Statement F, which referred to a preference for playing on a modern Football Turf pitch rather than a poor quality Natural Turf pitch. Players currently playing in CAF and OFC agreed most often with this statement, whereas CONCACAF, CONMEBOL and UEFA agreed least often, with AFC falling between these two groups.

One approach to assessing the statistical significance of these differences between confederations could be to use Kruskal-Wallis tests. However, the equivalent approach of an ordinal logistic regression model is considered instead. This is delayed until Section 5 where an extended ordinal logistic regression model is considered to facilitate the comparisons between confederations, in conjunction with additional potential explanatory factors based on players’ surface experience during their junior and senior careers.

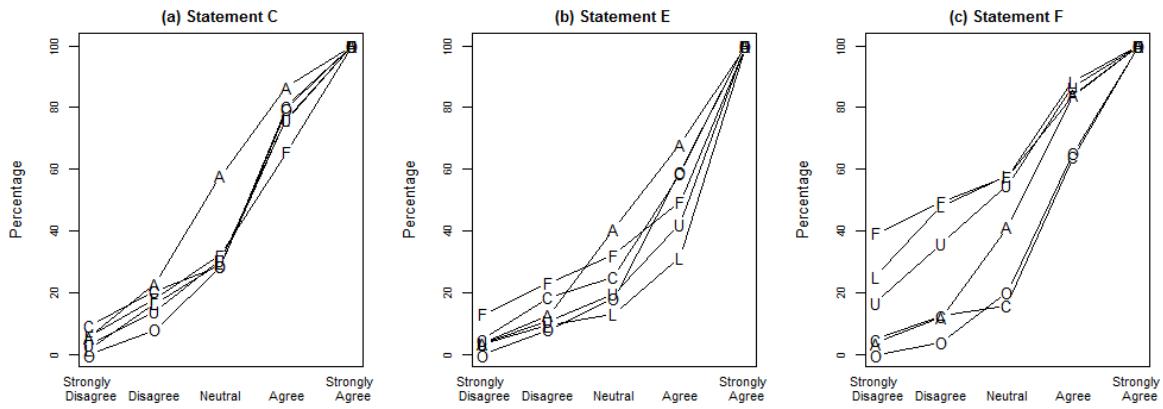


Figure 1: Cumulative percentage responses to statements
(A=AFC, C=CAF, F=CONCACAF, L=CONMEBOL, O=OFC, U=UEFA)

4. Surface Experience

4.1 Questionnaire Responses

In Part 3 of the questionnaire players answered the following four questions:

- Q3.1 “Which surfaces did you TRAIN on as a JUNIOR player (under 18 years)?”
- Q3.2 “Which surfaces did you PLAY on as a JUNIOR player (under 18 years)?”
- Q3.3 “Which surfaces did you TRAIN on as a SENIOR player?”
- Q3.4 “Which surfaces did you PLAY on as a SENIOR player?”

In each case, they were asked to indicate the extent to which they trained or played on each of the four surfaces; “Natural Turf”, “Football Turf,” “Gravel or similar hard surface” and “Indoor Sports Hall”. For each surface they were asked to indicate their response on a five point ordinal scale consisting of “Never”, “Rarely”, “Sometimes”, “Usually” and “Always”. Hence there were four parts to Q3.1 which related to surface experience **training** as a **junior** on Natural Turf (Q3.1a), Football Turf (Q3.1b), Gravel or similar hard surface (Q3.1c) and Indoor Sports Hall (Q3.1d). Similarly there were also four parts to Q3.2, Q3.3 and Q3.4, relating to players’ experience **playing** as a **junior** (Q3.2a-d), **training** as a **senior** (Q3.3a-d) and **playing** as a **senior** (Q3.4a-d). Players therefore answered a total of 16 questions. The cumulative percentage distributions of responses for players’ surface experience on each surface are summarised in Figure 2.

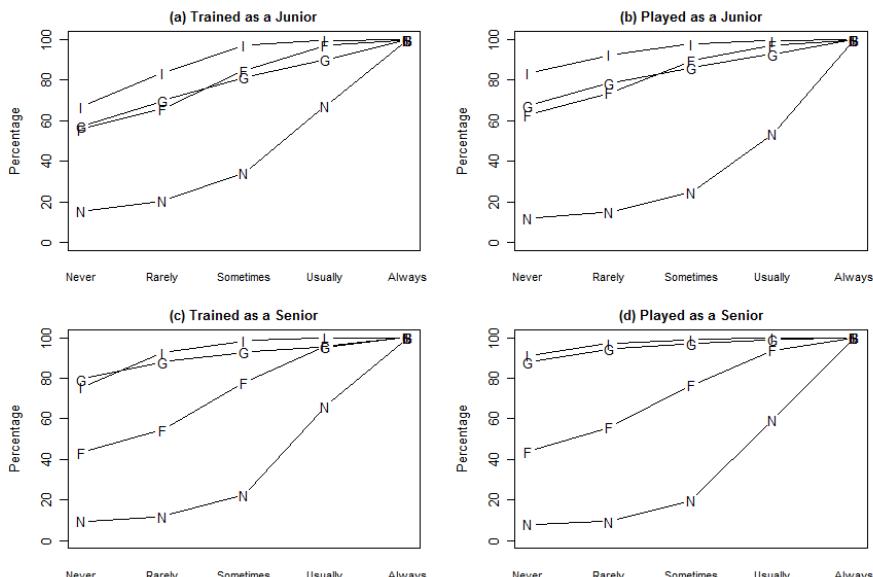


Figure 2: Cumulative percentage distributions of responses for players’ surface experience

(N=Natural Turf, F=Football Turf, G=Gravel or similar, I=Indoor)

Figure 2 shows that, perhaps as expected, Natural Turf dominates as indicated by the lower curves for Natural Turf, with higher proportions of players responding with “Usually” or “Always” on this surface. However, Figure 2 also illustrates that Football Turf features more prominently as a senior, both in training and play situations, compared to as a junior, and also highlights the limited amount of time that indoor surfaces feature.

4.2 Principal Components Analysis (PCA) of Surface Experience Scores

Section 5 will consider an ordinal logistic regression model that relates players’ responses to Statements C, E and F to their confederation and career surface experience. Players’ career surface experience could be measured using their responses to the 16 questions discussed above. However, in order to explore surface experience further there are two main problems. Firstly, there is the issue of dimensionality in that there are 16 variables to consider which makes the task a rather complex one, and secondly and perhaps more importantly, these 16 variables are not independent and so using these as potential predictor variables in any subsequent ordinal logistic regression model could potentially be problematic. The lack of independence between players’ responses to the surface experience questions can be illustrated, for example, by the fact that if a player selected “Always” to Q3.1a, then Q3.1b, Q3.1c and Q3.1d were all forced to be recorded as “Never”. Similarly, responding “Usually” to Q3.1a meant that Q3.1b, Q3.1c and Q3.1d could not be recorded as “Always”. This was the case for Q3.1 through to Q3.4 and so the responses to the four questions **within** any one of the four play-train/junior-senior situations are negatively correlated. As an illustration, the empirical (Spearman) correlation between Q3.4a and Q3.4b was -0.849. This lack of independence is further exacerbated by the fact that positive correlations exist **between** different play/train-junior/senior situations for the same surface, since players often experience the same surfaces throughout their careers. As an illustration, the empirical (Spearman) correlation between Q3.1c and Q3.2c was +0.75.

One approach that can resolve both the dimensionality and independence issues is to consider Principal Components Analysis (PCA). This technique forms new variables (principal components) from a linear combination of the original variables that are orthogonal and hence independent. In addition, it is often possible that much of the information contained in a large number of original variables can be explained by a much smaller number of the new principal components, which therefore reduces the dimensionality of the data. PCA is most effective when used with scale data, since it typically relies on the Pearson correlation matrix amongst the original variables to determine the principal components. However, since our data is ordinal, we instead use the correlation matrix amongst the original variables formed using polychoric correlations rather than the usual Pearson correlations. Polychoric correlation essentially assumes that the two-way contingency table formed from a pair of ordinal variables, is a discretization from a pair of variables measured on a continuous scale which follow a bivariate normal distribution. Hence the polychoric correlation is the correlation resulting from a bivariate normal distribution which is fitted to the contingency table. Further details on polychoric correlation can be found for example in Olsson (1979) and also Drasgow (1986).

In order to compare the impact of using polychoric rather than Pearson correlation, both methods are illustrated here. A PCA analysis was undertaken on the responses to just the 12 questions contained in Q3.1a-c, Q3.2a-c, and Q3.4a-c. The data from Q3.1d, Q3.2d, Q3.3d and Q3.4d were excluded since there was very little information contained in relation to the Indoor Sports Hall surface. Responses to the questions were scored using the scale 1=“Never”, 2=“Rarely”, 3=“Sometimes”, 4=“Usually” and 5=“Always”. The analysis was undertaken using the *princomp* function within the statistical software R. The cumulative proportion (%) of the total variation in the data from these 12 questions that is explained by increasing numbers of principal components is shown in Table 3 (note that just the first eight principle components are listed although the technique will produce a total of 12 orthogonal principal components). Table 3 highlights the advantage to be gained by the use of polychoric correlations in this case, since with polychoric correlation the first four principal components explain 87.3% of the total variation, whereas with Pearson correlation we would require six principal components to achieve this.

Table 3: Cumulative proportion (%) of variation in surface experience scores

Correlation	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Polychoric	46.2%	66.3%	80.4%	87.3%	93.4%	96.3%	98.3%	99.4%
Pearson	37.0%	55.1%	68.0%	76.3%	83.1%	87.5%	91.1%	94.2%

The choice of how many principle components to use is a subjective one but here we concentrate on the first four principle components. This is because between them they explain almost 90% of the variation in the data, and also this choice was supported by the characteristics displayed in the Scree plot (not shown) which is often used as a tool for determining the number of components to use. The first four principal components are in effect linear combinations of the original 12 variables (after scaling to have zero means and unit variances), The coefficients in these linear terms, which are usually referred to as the component loadings, are shown in Table 4 below. Components with greater weighting (above 0.2) are highlighted in bold in Table 4. For example, the first principal component (PC1) is derived as a linear combination of the 12 questions (after scaling to have zero means and unit variances) using the coefficients in the first column of Table 4. Histograms of the resulting principal component scores for these four principal components are summarised in Figure 3 overleaf. Interpreting the component loadings from Table 4 allows potentially meaningful interpretations to be attached to these four principal components as follows:

- PC1: Larger positive values are associated with players who have more experience of Natural Turf and less experience of Football Turf or Gravel, and vice-versa giving larger negative values. Hence this principal component appears to reflect a measure of players' experience on Natural Turf.
- PC2: Larger positive values are generally associated with players who have more experience of Gravel and less experience of Football Turf, and vice-versa giving larger negative values. Hence this principal component appears to reflect a contrast between players with more Gravel experience (positive values) versus those with more Football Turf experience (negative values).
- PC3: Larger positive values are mainly associated with players who as a junior had more experience of Natural Turf, but as a senior had more experience of Football Turf. Larger negative values are mainly associated with players who as a junior had more experience of Football Turf but as a senior had more experience of Natural Turf. Hence this principal component appears to mostly reflect a measure of the extent to which players' surface experience changed between Natural Turf and Football Turf, and in which direction during the transition between a junior to a senior. Gravel does also feature somewhat in this component but to a lesser extent.
- PC4: Larger positive values are mostly associated with players who as a senior trained more on Natural Turf but played more on Football Turf. Larger negative values are mostly associated with players who have in the main trained on Football Turf, but have more experience of playing on Natural Turf. A similar pattern is evident as a junior where players played on Football Turf but didn't train as much on that surface. Hence this principal component reflects a measure of the extent to which players' surface experience differs between training and playing on Football Turf.

Table 4: Principal component loadings

Question	PC1	PC2	PC3	PC4
Q3.1a Trained as a Junior on Natural Turf	+0.338	+0.840	+0.383	+0.050
Q3.1b Trained as a Junior on Football Turf	-0.093	-0.536	-0.233	-0.241
Q3.1c Trained as a Junior on Gravel or similar	-0.333	+0.218	-0.271	+0.133
Q3.2a Played as a Junior on Natural Turf	+0.369	+0.084	+0.244	-0.263
Q3.2b Played as a Junior on Football Turf	-0.184	-0.480	-0.141	+0.192
Q3.2c Played as a Junior on Gravel or similar	-0.359	+0.211	-0.194	+0.228
Q3.3a Trained as a Senior on Natural Turf	+0.332	+0.085	-0.214	+0.506
Q3.3b Trained as a Senior on Football Turf	-0.204	-0.398	+0.302	-0.252
Q3.3c Trained as a Senior on Gravel or similar	-0.312	+0.350	-0.012	-0.333
Q3.4a Played as a Senior on Natural Turf	+0.295	-0.008	-0.470	-0.239
Q3.4b Played as a Senior on Football Turf	-0.261	-0.090	+0.482	+0.425
Q3.4c Played as a Senior on Gravel or similar	-0.252	+0.281	+0.132	-0.310

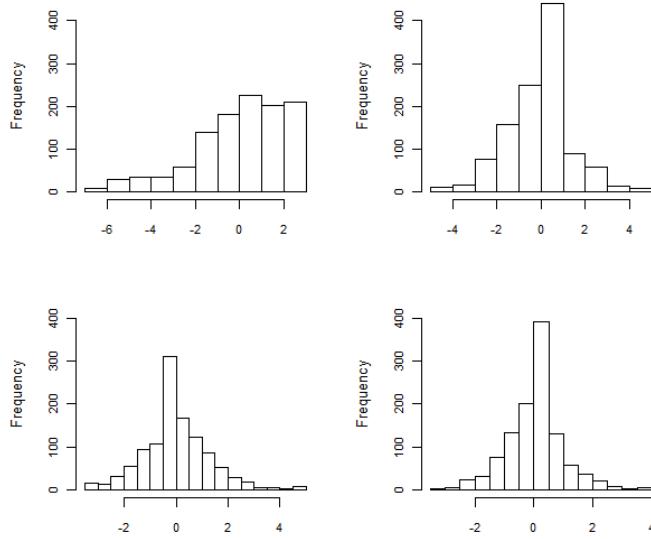


Figure 3: Histograms for PC1, PC2, PC3 and PC4

5. An Ordinal Logistic Regression Model

The model is specified by letting y_i represent the response for player i to a particular statement, scored using three ordinal category scores; 1 for “Strongly Disagree” or “Disagree”, 2 for “Neutral”, and 3 for “Agree” or “Strongly Agree”. These three categories are referred to here as “Disagreeing”, “Neutral” and “Agreeing” respectively. The original scores on the scale of 1 to 5 (for “Strongly Disagree”, “Disagree”, “Neutral”, “Agree” and “Strongly Agree”) could have been used, however the 1-3 scaling not only facilitates an easier interpretation of the model outputs, it also effectively increases the sample size when estimating the parameters in the model. Furthermore the overall conclusions are similar irrespective of which scale is used.

The probability of player i responding to a statement with a category score of j or lower ($j=1, 2$) is then defined as $\pi_j = \text{prob}(y_i \leq j)$. The usual (ordinal) odds can then be defined as $\theta_j = \pi_j / (1 - \pi_j)$, which represents the odds of player i responding to a statement with a score of j or lower ($j=1, 2$). Note that since $\pi_3 = 1$, odds are only defined for $j \leq 2$. The ordinal logistic regression model is then be specified by relating the log-odds to a linear combination of the predictor variables as follows:

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

The matrix $\mathbf{X} = [\mathbf{C}_1, \mathbf{C}_2, \mathbf{C}_3, \mathbf{C}_4, \mathbf{C}_5, \text{PC1}, \text{PC2}, \text{PC3}, \text{PC4}]$ contains the observations for the predictor variables, where the \mathbf{C}_k ($k=1, \dots, 5$) are specified such that $C_{k,i}=1$ if a player is from Confederation k and zero

otherwise, and hence reflect a series of indicator variables for confederation. Note that one of the C_k is redundant and so needs to be set to zero; in the above specification $C_6=0$. Therefore the C_k reflect comparisons against a “baseline” confederation which in the above specification is Confederation 6 (UEFA). PC1, PC2, PC3 and PC4 are the principal component scores discussed previously in Section 5.2. The vector $\beta = [\beta_1, \beta_2, \dots, \beta_9]^T$ contains the model parameters to be estimated which describe the effect of the predictors on the log-odds. Finally, the α_j are “threshold” parameters which indicate the log-odds of a category score of j or lower ($j=1, 2$) associated with the baseline confederation. The α_j are not of particular interest here as they simply serve a similar role as the intercept in a linear regression model.

Note that the manner in which the model is defined and the inclusion of a negative sign in (1), allows positive values for any of the β_i to indicate that players with increasing values for the associated predictor variable are more likely to respond with higher category scores, and hence are more likely to take a neutral view or agree with the statement and so less likely to disagree with the respective statement.

The model given by (1) reflects the common form of the ordinal logistic regression model which assumes proportional odds, such that the relationship between the log-odds and the predictor variables is the same for both $j=1$ and $j=2$. However, model assessments for our data provided evidence to suggest that this assumption was doubtful, in the manner in which the log-odds are related to Confederation (no details are shown here for brevity). Hence a slightly more complex model is required which allows the relationship between the log-odds and Confederation to differ for $j=1$ and $j=2$. This simply requires the vector β to be re-stated so that it is specific to j as $\beta_j = [\beta_{1j}, \beta_{2j}, \beta_{3j}, \beta_{4j}, \beta_{5j}, \beta_6, \beta_7, \beta_8, \beta_9]^T$, for $j=1, 2$. The first five parameters which relate to Confederation are specific to j , but this is not required for PC1, PC2, PC3 or PC4 since no problems with the proportional odds assumption were found in relation to those variables.

The model was implemented using the statistical software R, making use of the *ordinal* package which allows the mixed non-proportional and proportional ordinal odds model to be implemented. UEFA was chosen as the baseline confederation, but this choice has no impact on the resulting predicted outcome probabilities arising from the model. Table 5 below shows the resulting parameter estimates and some elements of the model fit assessments undertaken for each Statement. This includes the odds-ratios $\exp(\beta_j)$, where for example $\exp(\beta_{11})$ indicates the associated multiplicative change in the odds of a response of “Neutral/Agreeing” versus “Disagreeing” for players in AFC compared to UEFA, and $\exp(\beta_{12})$ indicates the associated multiplicative change in the odds of a response of “Agreeing” versus “Neutral/ Disagreeing” for players in AFC compared to UEFA. For Confederation these two sets of odds ratios are not assumed to be the same, whereas for PC1, PC2, PC3 or PC4 they are assumed to be the same since in this later case we assume proportional odds. The significance of individual parameters in the model are assessed using the usual Wald tests, although the overall significance of Confederation is assessed using the usual likelihood ratio test.

In terms of overall model fit, Table 5 summarises the change in the Log_e-Likelihood compared to a null model containing no predictor variables (with just the threshold parameters). For all three statements this change is significant ($p<0.001$) which suggest the variables included in the model explain a significant amount of the variation in players’ responses in each case. The values for R^2 (Nagelkerke) suggest however that there is still a large of unexplained variation in player’s responses that would merit further investigation. The results for the parameter estimates in Table 5 suggest that Confederation and PC1 are significant predictors of players’ responses to all three statements. PC2 is also a borderline significant predictor for Statements E ($p=0.047$) and F ($p=0.056$), whilst for Statement F there are additional significant predictors in PC3 ($p<0.001$) and PC4 ($p=0.015$).

As an illustrative example of how to interpret the parameter estimates, the positive parameter estimate for PC1 (β_6) associated with Statement C, suggests that players with more experience of Natural Turf agree more with the idea of Football Turf offering a home advantage. The larger negative estimates for β_{11} and β_{12} (AFC) for Statement C suggests that players currently playing in AFC are more likely to disagree with this statement and less likely to take a neutral view or agree compared to players from other confederations. The odds ratio of 0.61 (associated with β_{11}) indicates that the odds of players in AFC agreeing or taking a neutral view with this statement are almost half the odds for players in UEFA. Note that this lower agreement with Statement C amongst players in AFC suggested by the model is consistent with Figure 1(a) discussed earlier,

but this conclusion also holds even after accounting for players' career surface experience. The mixed positive and negative estimates for CAF (β_{21} and β_{22}) with Statement C could indicate that players in CAF took less of a neutral view and instead took a position of either disagreeing or agreeing more with this statement compared to UEFA. However this later contrast between CAF and UEFA was not statistically significant and so there is no evidence that this difference exists.

For Statement E, which referred to the view that all fixtures should be played on Natural Turf, the positive parameter estimates for PC1 (β_6) and PC2 (β_7) suggests that players with more experience on either Natural Turf or on Gravel or a similar hard surface and less experience of Football Turf agree more with this statement. This might suggest a possibility that players' opinions are biased towards the surface they have more familiarity with. The larger negative estimates for β_{31} and β_{32} (CONCACAF), β_{12} (AFC) and β_{21} (CAF) with Statement E suggests that players currently playing in those confederations tend to agree less with this statement.

For Statement F, which referred to a preference for playing on a modern Football Turf pitch rather than a poor quality Natural Turf pitch, the negative parameter estimates for PC1 (β_6) and PC2 (β_7), suggests that the players that are more likely to agree with this statement are those with less experience on Natural Turf or on Gravel or similar hard surface and hence those with more experience of Football Turf. The positive parameter estimates for PC3 (β_8) and PC4 (β_9) however, suggest that players who started out on Natural Turf as a junior but experienced Football Turf more so as a senior are more likely to agree with this statement, as are players who as a senior have trained on Natural Turf but have played as a senior more on Football Turf. These conclusions would appear to be consistent with the view that players have a bias towards the surface that they have more experience of. This may also explain at least in part the contradiction between Statements C, E and F in terms of the overall view of players opinions, reported earlier in Section 3.1. Finally for Statement F, the larger positive estimates for β_{11} and β_{12} (AFC), β_{21} and β_{22} (CAF) and β_{51} and β_{52} (OFC) suggest that players in those confederations appear to agree more with this statement compared to the other confederations.

A useful approach to visualizing the effect of these parameter estimates is given overleaf in Figure 4. This shows the predicted probabilities of players' responses to Statement C falling within the "Agreeing", "Neutral" or "Disagreeing" score categories, plotted against increasing values of PC1 (with PC2, PC3 and PC4 all remaining constant at zero) for each confederation. This illustrates clearly the increasing agreement with this statement with increasing values of PC1, and also the lower general level of agreement with AFC players.

Table 5: Parameter estimates

Model Term	Model Parameter	Statement C			Statement E			Statement F		
		Estimate (β_j)	p	Odds Ratio $\exp(\beta_j)$	Estimate (β_j)	p	Odds Ratio $\exp(\beta_j)$	Estimate (β_j)	p	Odds Ratio $\exp(\beta_j)$
Confederation	β_{11} (AFC)	-0.498	0.035	0.61	-0.216	0.451	0.81	1.312	<0.001	3.71
	β_{12} (AFC)	-1.065	<0.001	0.35	-1.081	<0.001	0.34	0.496	0.012	1.64
	β_{21} (CAF)	-0.202	0.430	0.82	-0.577	0.032	0.56	1.012	<0.001	2.75
	β_{22} (CAF)	0.398	0.080	1.49	-0.259	0.282	0.77	1.544	<0.001	4.69
	β_{31} (CONCACAF)	-0.449	0.146	0.64	-1.164	<0.001	0.31	-0.321	0.184	0.73
	β_{32} (CONCACAF)	-0.201	0.431	0.82	-0.919	0.001	0.40	0.175	0.475	1.19
	β_{41} (CONMBEOL)	-0.289	0.331	0.75	-0.057	0.873	0.94	-0.419	0.065	0.66
	β_{42} (CONMBEOL)	-0.009	0.970	0.99	0.290	0.348	1.34	-0.006	0.980	0.99
	β_{51} (OFC)	0.460	0.403	1.58	0.315	0.613	1.37	2.833	<0.001	16.99
	β_{52} (OFC)	0.050	0.888	1.05	-0.077	0.853	0.93	1.857	<0.001	6.41
Overall		$\chi^2=64.53$ df=9 p<0.001			$\chi^2=68.38$ df=9 p<0.001			$\chi^2=126.62$ df=9 p<0.001		
Surface Experience	β_6 (PC1)	0.147	<0.001	1.16	0.085	0.021	1.09	-0.176	<0.001	0.838
	β_7 (PC2)	0.015	0.765	1.02	0.106	0.047	1.11	-0.100	0.056	0.905
	β_8 (PC3)	0.041	0.435	1.04	0.041	0.477	1.04	0.215	<0.001	1.24
	β_9 (PC4)	0.084	0.235	1.09	0.037	0.628	0.096	0.175	0.015	1.19

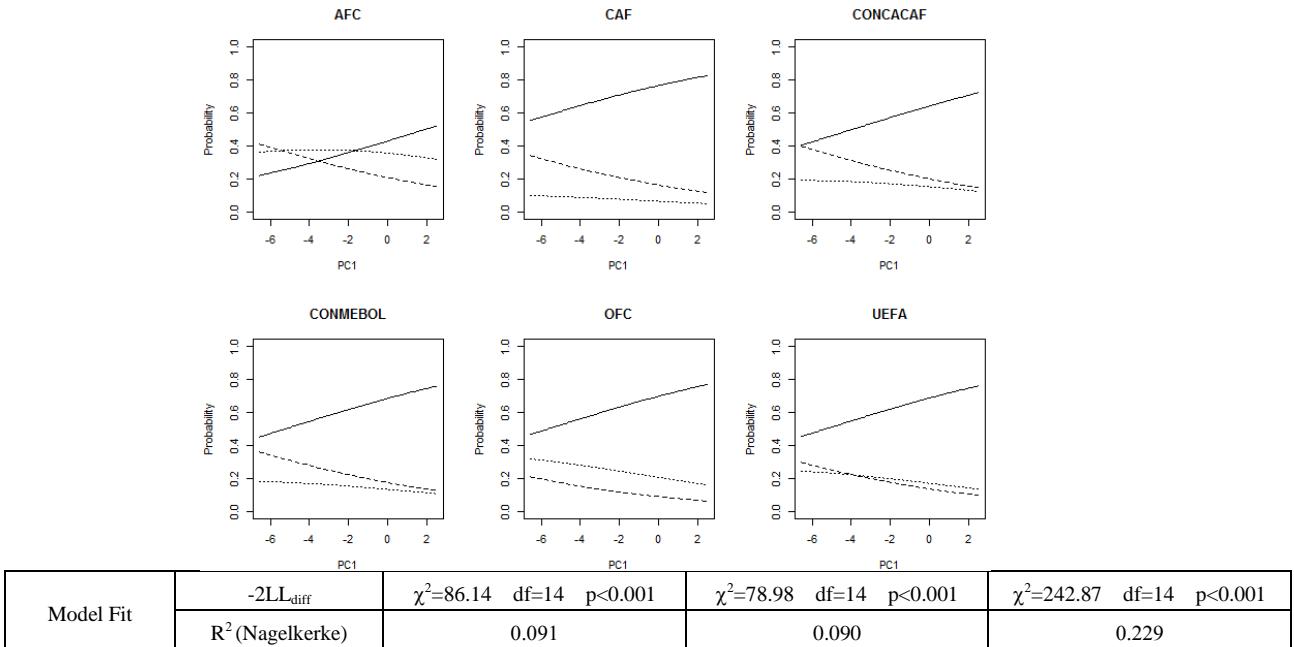


Figure 4: Model predicted probabilities for Statement C against PC1 by Confederation
 “Agreeing” (—) “Neutral” (---) “Disagreeing” (-·-·-)

6. Discussion

The present study has enabled a potentially rich and extensive database on elite players’ perceptions and opinions of the use of surfaces in football to be compiled, although space here has only permitted the consideration of a few aspects of player’s opinions. Overall players tend to agree with Statement C that Football Turf affords a big home advantage, and also Statement E that all fixtures should be played on Natural Turf. However this appears to be contradicted by the fact that the majority of players also agreed with Statement F that they would rather play on a modern Football Turf pitch than a poor quality Natural Turf pitch. However, players’ surface experience during their career both as a junior and as a senior had a significant impact on their responses. There appears to be a possibility that players’ perceptions are to some extent biased towards the surface they have more familiarity with. Players with more experience on Natural Turf appear to be more likely to agree with Statements C and E but disagree with Statement F, and hence take a negative view of Football Turf. Players with more experience of Football Turf are more likely to disagree with Statements C and E but agree with Statement F and hence take a positive view of Football Turf. This would seem to explain the contradiction noted above. There were two other surface experience groups that are more likely to agree with the view that they would rather play on a modern Football Turf pitch than a poor quality Natural Turf pitch, these are players who started out on Natural Turf as a junior but experienced Football Turf more so as a senior and players who as a senior trained on Natural Turf but have played more as a senior on Football Turf. There were also differences in opinions between confederations, even after accounting for players’ surface experience. Work continues on the project.

Acknowledgements

We would like to thank FIFA, its member associations and FIFPro for their assistance in this project.

References

- Andersson, H., Ekblom, B. and Krstrup, P. (2008) Elite football on artificial turf versus natural grass: Movement patterns, technical standards, and player impressions. *Journal of Sports Sci.*, **26(2)**, 113-122.
- Barnett, V. and Hilditch, S. (1993) The effect of an artificial pitch surface on home team performance in football (soccer). *Journal of the Royal Statistical Society A*, **156**, 39-50.
- Drasgow, F. (1986) Polychoric and polyserial correlations. In *The Encyclopedia of Statistics, Volume 7* (Kotz and Johnson, Eds.), Wiley, pp. 68–74.

- Nigg, B.M. and Yeadon, M.R. (1987) Biomechanical aspects of playing surfaces. *Journal of Sports Sci.* **5**, 117–145.
- Olsson, U. (1979) Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* **44**, 443-460.
- Zanetti, E.M. (2009) Amateur football game on artificial turf: Players' perceptions. *Applied Ergonomics*, **40**, 485-490.

A DEA Evaluation and Financial Resources Reallocation for Brazilian Olympic Sports regarding their results in the 2011 Pan-American Games

Renato Pescarini Valério* and Lidia Angulo-Meza**

*Av. dos Trabalhadores 420, 27255-125, Volta Redonda, RJ, Brazil, renato_pv@yahoo.com.br

** Av. dos Trabalhadores 420, 27255-125, Volta Redonda, RJ, Brazil, lidia_a_meza@pq.cnpq.br

Abstract. This paper proposes the use of Data Envelopment Analysis (DEA) to evaluate the efficiency of Brazilian Olympic sports and to reallocate the financial resources received by each them. The sports selected were those that received financial resources from the Agnelo/Piva Law in 2011. Previous research showed that results obtained by Brazil in the Olympic Games were scarce to determine the sports efficiency, as there are many null results for many sports. The solution found was using data from the Pan-American Games, specifically from the 2011 Guadalajara Pan-American Games. The inputs in the model are the funds from the Agnelo/Piva Law that were distributed among the Olympic sports by the Brazilian Olympic Committee in 2011. Also, the number of gold medals offered by each sport in the 2011 Guadalajara Pan American Games is considered as input, as a proxy for difficulty measure in winning a medal. The outputs are the number of gold, silver and bronze medals won by each sport during the 2011 Guadalajara Pan American Games. A DEA non-radial model with weights restrictions is formulated to perform the Olympic sports efficiency evaluation. With these results a reallocation of the financial resources is proposed using a ZSG-DEA non-radial approach.

1. Introduction

Brazil will be soon the host country of two major world sporting events: the 2014 FIFA World Cup and the 2016 Olympic Games. It represents a unique opportunity for the country to take advantage of the large investment that will be made and to leave a great impression all over the world, whether the events are well organized and achieved. Moreover, in this position of great international visibility, the country performance in both events is a growing concern. In order for the country to achieve a good performance during the sporting events, it is necessary a high investment in sports.

With the aim of contributing to the country on the improvement of its sporting performance, this paper proposes the use of Data Envelopment Analysis (DEA) to firstly evaluate some Brazilian Olympic sports efficiency, based on the country results in the 2011 Guadalajara Pan-American Games. Posteriorly a financial resources reallocation is made considering the funds transferred to each sport committee as defined by the Agnelo/Piva Law in 2011, in order to improve the DMUs efficiency.

This paper is divided into six sections. In section 1 an introduction and the motivation of the work is presented. Posteriorly, in section 2, the theoretical explanation of the methodology used in the paper and its main features are exposed. Still in this section, some studies using DEA concerning financial resources destined to sports are presented. In section 3, it is shown how Data Envelopment Analysis was used to reach the results, presented in section 4, where it is found also a discussion about these results. Finally, in section 5, final comments are made.

2. Data Envelopment Analysis

Data Envelopment Analysis (DEA) (Charnes et al., 1978) is a mathematical technique used to evaluate the efficiency of a productive units group, called Decision Making Units (DMUs). The DEA method involves the use of Linear Programming (LP) to determine the relative efficiency of each DMU. A group of DMUs represents productive units that, with the same targets and with the use of the same kind of resources (inputs), generate products (outputs). Many DEA models have been created and all of them can have two kinds of orientation: input orientation, used when the target is decreasing the inputs keeping the outputs constant, and output orientation, used when the target is increasing the outputs keeping the inputs constant.

In this paper, in order to evaluate the sports efficiency it is used a DEA non-radial model. The same one can be found in Banker and Morey (1986). It is very similar to the BCC model (Banker et al., 1984), a DEA classical one. The BCC model allows variable return of scale, avoiding possible problems caused by

imperfect competition situations. The DEA non-radial model has this same BCC feature. However, there is a big difference between them: the DEA non-radial model accepts the existence of non-controllable variables, variables that cannot be modified by the decision maker, while the BCC model does not accept these variables. In (1) it is presented the non-radial input oriented model, the same used in this paper. In this Problem of Linear Programming the term h_0 consists on the efficiency of the DMU o ; x_{ik} and y_{jk} represent, respectively, the value of inputs i and outputs j of a DMU k ; λ_k represents the contribution of each DMU k in the composition of the target of DMU o . The inputs having C as an index are the controllable ones, while those having NC as an index are the non-controllable ones. We can note that the first restriction concerns only the controllable inputs, that are multiplied by the term h_0 in the left part of the equation. However, the second restriction, very similar to the first one, concerns only the non-controllable inputs, that are not multiplied by the term h_0 in the left part of the equation.

$$\begin{aligned}
& \text{Min } h_0 \\
& \text{Subject to} \\
& h_0 x_{i0}^C \geq \sum_{k=1}^n \lambda_k x_{ik}^C, \forall i \\
& x_{i0}^{NC} \geq \sum_{k=1}^n \lambda_k x_{ik}^{NC}, \forall i \\
& y_{j0} \leq \sum_{k=1}^n \lambda_k y_{jk}, \forall j \\
& \sum_{k=1}^n \lambda_k = 1 \\
& \lambda_k \geq 0, \forall k
\end{aligned} \tag{1}$$

Santos et al. (2011) pointed out that additional information about the variables was translated into weight restrictions (Allen et al., 1997) and included in the model, as we can see in (2). In this model $A^t \gamma$ represents the coefficients matrix of the outputs weights restrictions, $Au \leq 0$, as presented in Lins et al. (2003) and also used in Fonseca et al. (Fonseca et al., 2010). This model was used iteratively by Fonseca et al. (2010) until all DMUs became efficient and the resource totally distributed among DMUs”.

$$\begin{aligned}
& \text{Min } h_0 \\
& \text{Subject to} \\
& h_0 x_{i0}^C \geq \sum_{k=1}^n \lambda_k x_{ik}^C, \forall i \\
& x_{i0}^{NC} \geq \sum_{k=1}^n \lambda_k x_{ik}^{NC}, \forall i \\
& y_{j0} \leq \sum_{k=1}^n \lambda_k y_{jk} - A^t \gamma_i, \forall j \\
& \sum_{k=1}^n \lambda_k = 1 \\
& \lambda_k \geq 0, \forall k
\end{aligned} \tag{2}$$

So far, the model presented is used for the performance evaluation. Posteriorly this evaluation, this work proposes also a financial resources reallocation using the input targets from the non-radial model. This reallocation is made based on a ZSG non-radial approach. The DEA Zero Sum Gains model (DEA-ZSG) was proposed to solve problems where the total sum of some inputs or outputs values must be constant (Lins et al., 2003, Gomes et al., 2003). As the sum of the financial resources received by each sport in this study must be constant, the reallocation is made based on this approach. The equation 3 shows how the input reallocation was calculated. The term $x_{io}^{\text{reallocated}}$ is the new value of the input i for the DMU o ; x_{io}^{target} is the

target of the input i for the DMU o obtained with the non-radial model; $x_{ik}^{original}$ is the original input i for a DMU k ; x_{ik}^{target} is the target of the input i for a DMU k obtained with the non-radial model; n is the total number of DMUs.

$$x_{io}^{reallocated} = x_{io}^{target} \times \left(\sum_{k=1}^n x_{ik}^{original} \div \sum_{k=1}^n x_{ik}^{target} \right) \quad (3)$$

1.1 DEA for financial resources in sports

The application of DEA having as a variable the financial resources destined to sports can have two different approaches. The first one is the sports efficiency evaluation, which analyses how well each sport uses the financial resources received based on the results obtained by them. A summary of works using this approach can be found in Soares de Mello et al. (2008). Other works with this same approach can be found in Zhang et al. (2009), Wu et al. (2009bb) and Wu et al. (2009aa). The second one is the financial resources reallocation, which proposes better ways of the resources redistribution among the sports, optimizing their efficiencies. Works containing this approach can be found in Santos et al. (2011), Wu et al. (2009bb) and Villa and Lozano (2004).

3. Brazilian Olympic sports – Evaluation and Financial Resources Reallocation

The Olympic sports taken as DMUs in this paper are the ones that could participate to the 2011 Guadalajara Pan American Games and that also received funds from the Agnelo/Piva Law in 2011. This Law was sanctioned in 2001 and determines that 2% of the gross revenues from the Brazilian federal lotteries must be destined to the Brazilian Olympic Committee, which receives 85% of the amount, and to the Brazilian Paralympic Committee, which receives the 15% remaining. Both these Committees must invest 10% of the amount received in school sport, 15% in university sport and the 75% in the Brazilian Olympic Confederations. Some sports, as Soccer and Bowling, despite having Brazilian competitors in the Games, they are not considered in this paper since they didn't receive any funds coming from the Agnelo/Piva Law in 2011. Even not having Brazilian competitors in the Games, the Hockey on Grass is considered as DMU because it received funds coming from the Agnelo/Piva Law in 2011. In total there are 26 Sports Confederations considered as DMUs in this study.

Moreover, the study didn't use results of Brazilian Olympic sports in the Olympic Games because previous research, as in Santos et al. (2011), showed that results obtained by Brazil in these Games were scarce to determine the sports efficiency, as there are many null results for many sports. As Brazil has always a better performance in the Pan American Games, the solution found were using data from these Games.

The model was formulated using two inputs and three outputs. The first input is represented by the funds coming from the Agnelo/Piva Law that were transferred to each Olympic sport by the Brazilian Olympic Committee in 2011. This input measures the amount of money available for each sport investment. The second one is the number of gold medals offered for each sport in the 2011 Guadalajara Pan American Games, as a proxy for difficulty measure in winning a medal. This second output represents the non-controllable variable of the problem. The outputs are the number of gold, silver and bronze medals won by each sport during the same sporting event.

We also included in the non-radial model three weight restrictions regarding the importance of each medal (Lins et al., 2003). These weight restrictions can be translated as: a gold medal is more important than a silver one; a silver medal is more important than a bronze one; and the importance difference between a gold medal and a silver medal is greater than the importance difference between a silver one and a bronze one.

4. Results and discussions

This section is divided into two parts: firstly we analyse the results concerning the efficiency of each sport regarding the financial resources received and the results obtained by each sport in the 2011 Pan American Games. Also, based on the efficiency index we perform the financial resources reallocation. These results are depicted in Table 1.

Table 1 – Efficiency, Original Resource and Reallocated Resource for each DMU

Sports	Efficiency Score	Original Resources (R\$)	Reallocated Resources (R\$)
Athletics	1.0000	3,000,000.00	4,534,431.80
Badminton	0.4487	1,300,000.00	881,695.17
Basketball	0.3297	2,100,000.00	1,046,406.29
Boxing	0.6373	1,700,000.00	1,637,433.96
Canoeing	0.3623	2,300,000.00	1,259,564.94
Cycling	0.2174	2,300,000.00	755,737.58
Water Sports	1.0000	3,000,000.00	4,534,431.80
Fencing	0.6818	1,100,000.00	1,133,607.65
Gymnastics	0.8163	2,800,000.00	3,454,807.17
Handball	1.0000	3,000,000.00	4,534,431.80
Horse Riding	0.2586	2,900,000.00	1,133,609.31
Hockey on Grass	0.3846	1,300,000.00	755,737.88
Judo	1.0000	3,000,000.00	4,534,431.80
Weightlifting	0.6818	1,100,000.00	1,133,607.65
Wrestling	0.4444	1,500,000.00	1,007,650.50
Modern Pentathlon	0.5325	1,300,000.00	1,046,406.60
Oar	0.3539	1,900,000.00	1,016,338.17
Rugby	1.0000	500,000.00	755,738.63
Taekwondo	0.4861	1,200,000.00	881,694.87
Tennis	0.3704	1,800,000.00	1,007,650.50
Table Tennis	0.3727	2,300,000.00	1,295,552.61
Archery	0.3846	1,300,000.00	755,737.88
Sports Shooting	0.5000	2,000,000.00	1,511,477.27
Triathlon	1.0000	2,000,000.00	3,022,954.54
Sailing	1.0000	3,000,000.00	4,534,431.80
Volleyball	1.0000	3,000,000.00	4,534,431.80
TOTAL		52,700,000.00	52,700,000.00

In analysing the second column of this table we can note that there are eight DMUs with maximum efficiency: Athletics, Water Sports, Handball, Judo, Rugby, Triathlon, Sailing and Volleyball. Despite reaching the maximum efficiency and, consequently, being in the efficiency frontier, the Confederation of Athletics is not Pareto Efficient, since there are ways of improving its situation. The Confederation of Rugby, in spite of not having won any medal, it was among the DMUs with maximum efficiency. It occurred because for the models with variable return of scale, the DMU with the smallest values of inputs has maximum efficiency, even having null outputs.

While the sports mentioned above are considered the most efficient ones, there is a group of sports needing urgent performance improvement: Badminton, Basketball, Canoeing, Cycling, Horse Riding, Hockey on Grass, Wrestling, Oar, Taekwondo, Tennis, Table Tennis and Archery. Together, these 12 confederations received approximately 42% of the total amount of funds distributed by the Agnelo/Piva Law in 2011 and they had together 98 gold medals being offered during the 2011 Guadalajara Pan American Games. However, they won only one gold medal.

The forth column of the table represents the data obtained by the financial resources reallocation made using a DEA-GSZ non radial approach. Comparing these data with data of second column, the original distribution of resources, we can reach several important conclusions.

Firstly, the DMUs that had originally received a great amount of resources and reached the maximum efficiency had more than R\$ 3,000,000.00 of resources after the reallocation. It happened for all of DMUs with maximum efficiency, except for the Rugby. This sport didn't have a big amount of money in the original distribution and it was efficient only because it has the lowest values of inputs. Therefore, it seems that the model recognizes that this sport doesn't need much more money to keep its efficiency. That is why the Rugby Confederation didn't receive much more funds in the reallocation, compared with how much it had already received originally. Still among the DMUs with maximum efficiency, those that had received exactly R\$ 3,000,000.00 were transferred the same value after the resources reallocation: R\$ 4,534,431.80. It happened because the calculation of the reallocated input for each DMU is based only on the efficiency of this DMU and on its original input.

Furthermore, all DMUs with efficiency equal to or greater than 0.6818 received a larger amount of funds after the reallocation. However, all DMUs with efficiency equal to or less than 0.6373 lost part of the original amount of funds. Therefore, according to the DEA model used, the greater is the DMU efficiency reached using the non-radial model with weight restrictions, the bigger is the amount of resources it should receive by the reallocation with a DEA-GSZ non radial approach, in order for all DMUs to reach the maximum efficiency.

5. Final Comments

This paper, using Data Envelopment Analysis, could evaluate sports performance, based on their results in the 2011 Pan American Games and on the funds coming from the Agnelo/Piva Law in 2011, and could also propose a financial resources reallocation for those funds, making all DMUS efficient. The results obtained pointed out many efficient DMUs but also many others that need urgent improvements, the ones with the lowest values of efficiency. However, those sports needing urgent improvements received less funds with the financial resources reallocation. It didn't happen because the DEA-ZSG approach used intends to be punitive, but because it intends to reward those DMUs with best performance and to serve as a warning signal for those with worst performance. The DMUs with worst performance should face the results as an indicative that they should do something to improve their performance.

The approach used in this paper is very similar to the one used in Santos et al. (2011). The major difference is that in this study the data used as outputs, representing the number of gold, silver and bronze medals conquered by each DMU, came from the 2011 Pan-American Games, while in Santos et al. (2011), they came from the 2008 Olympic Games. This chose provided more robust results for this paper compared with the other one, since data coming Olympic Games have many null results. Furthermore, Santos et al. (2011) concluded that the inclusion of the number of gold medals offered for each sport as an input in the model used in that study didn't add any value for the results. However, it could have been caused, one more time, by the countless null results in the data used for that study, inasmuch as in this paper the use of this input was fundamental for the results obtained.

Moreover, we identified a limitation in this present study, which is the non-consideration of the maintenance costs for each sport. In order to solve the problem we believe that it is necessary to take into account this cost. Thus, we expect that the efficiencies obtained with the model for each DMU will be more robust. This addition may also solve the problem of sports with high maintenance costs receiving a less amount of funds with the financial resources reallocation.

It is also interesting to highlight the importance of studies like this for Brazilian sport, due to the current situation of the country: a country in full development, host of the next FIFA World Cup and the next Olympic Games, with enormous international visibility, but also with serious problems and at the same time, basic, like a poor level of education and high rates of violence, which may find its solution with the aid of the sport. There are very few scientific studies using DEA applied to investments in sports. This is another one and it serves as an incentive for future works.

It is important to point out that, with the new values obtained for the inputs representing the funds received by each DMU, all of these DMU reached the maximum efficiency. Thereby, if the model is formulated again changing the old values for these inputs by the new ones, all DMU will be efficient.

Finally, we can say that this study validated the use of the DEA-GSZ non radial approach for financial resources reallocation on sports.

Acknowledgements

We would like to thank CNPq and FAPERJ for their financial support.

References

- ALLEN, R., ATHANASSOPOULOS, A., DYSON, R. G. & THANASSOULIS, E. 1997. Weights restrictions and value judgements in data envelopment analysis: evolution, development and future directions. *Annals of Operations Research*, 73, 13-34.
- BANKER, R. D., CHARNES, A. & COOPER, W. W. 1984. Some models for estimating technical scale inefficiencies in data envelopment analysis. *Management Science*, 30, 1078-1092.
- BANKER, R. D. & MOREY, R. 1986. Efficiency analysis for exogenously fixed inputs and outputs. *Operations Research*, 32, 513-521.
- CHARNES, A., COOPER, W. W. & RHODES, E. 1978. Measuring the efficiency of decision-making units. *European Journal of Operational Research*, 2, 429-444.
- FONSECA, A. B. D. M., SOARES DE MELLO, J. C. C. B., GOMES, E. G. & ANGULO-MEZA, L. 2010. Uniformization of frontiers in non-radial ZSG-DEA models: An application to airport revenues. *Pesquisa Operacional*, 30, 175-193.
- GOMES, E. G., SOARES DE MELLO, J. C. C. B. & ESTELLITA LINS, M. P. 2003. Busca sequencial de alvos intermediários em modelos DEA com soma de outputs constante. *Investigação Operacional*, 23, 163-178.
- LINS, M. P. E., GOMES, E. G., SOARES DE MELLO, J. C. C. B. & SOARES DE MELLO, A. J. R. 2003. Olympic ranking based on a zero sum gains DEA model. *European Journal of Operational Research*, 148, 312-322.
- SANTOS, T. P., ANGULO-MEZA, L. & SOARES DE MELLO, J. C. C. B. Allocating economic resources for Olympic sports in Brazil using a DEA-ZSG model. 3rd IMA International Conference on Mathematics in Sport, 2011 Manchester.
- SOARES DE MELLO, J. C. C. B., GOMES, E. G., ANGULO-MEZA, L. & BIONDI NETO, L. 2008. Cross Evaluation using Weight Restrictions in Unitary Input DEA Models: Theoretical Aspects and Application to Olympic Games Ranking. *WSEAS Transactions on Systems*, Forthcoming.
- VILLA, G. & LOZANO, S. A. 2004. Constant Sum of Outputs DEA model for Olympic Games target setting. *4th International Symposium on DEA*. Aston University, US.
- WU, J., LIANG, L. & CHEN, Y. 2009a. DEA game cross-efficiency approach to Olympic rankings. *Omega*, 37, 909-918.
- WU, J., LIANG, L. & YANG, F. 2009b. Achievement and benchmarking of countries at the Summer Olympics using cross efficiency evaluation method. *European Journal of Operational Research*, 197, 722-730.
- ZHANG, D., LI, X., MENG, W. & LIU, W. 2009. Measuring the performance of nations at the Olympic Games using DEA models with different preferences. *Journal of the Operational Research Society*, 60, 983-990.

A closer look at the Independence between points of the top four male players

Geoff Pollard*

*Faculty of Life and Social Sciences, Swinburne University of Technology, Melbourne, Australia,
gpollard@tennis.com.au

Abstract. The assumption made in almost all probabilistic modeling of sports such as tennis is that all points are independent and identically distributed. Pollard and Pollard (2012) suggested four specific and another four general tests to investigate this assumption and then applied these tests to Nadal's 2011 Grand Slam matches against other Top Ten players. The tests can be similarly applied when a player is serving and when a player is receiving to consider independence under both conditions.

This paper extends the analysis firstly by considering Nadal's performance against non Top Ten players, where it has been shown previously that the better player can lift, and secondly by considering the 2011 Grand Slam matches of the other Top Four players Djokovic, Federer and Murray. As with Nadal most measures are not significant, but each player has some significant results.

Thirdly this paper extends the above analysis by developing new tests achieved by treating the data as bi-points, which is particularly appropriate to tennis as each player serves successively to the first and second court during a game, but the analysis can also be applied to all points in each set. An extension to tri-points is also discussed.

1. Introduction

In the probabilistic modeling of sports such as tennis the assumption is typically made that points are independent and identically distributed. Generally it is assumed that two fixed probabilities govern a game or a set or even a whole match. One is the probability that Player A wins a point on his/her service, the other is that player B wins a point on his/her service. If the points are independent, we can then calculate a range of characteristics such as the probability of winning a game, a set or a match and the mean, variance and skewness of the number of points played.

The set was adopted as the principal component of tennis scoring systems analysis as a game is too short and a match is too long for a player to maintain a constant probability of winning a point on service. In Grand Slam Men's Singles matches, in order to win a match, a player needs to win three sets before his opponent does. Tie-breaks are excluded from the analysis.

Pollard and Pollard (2012) suggested eight measures that might be used to investigate the assumption of independence. Four of these tests looked at specific sources of potential lack of independence, namely the probability of the server winning a point (a) if he is ahead, equal or behind in a game, (b) if he won or lost the previous point, (c) if he is ahead, equal or behind in a game and won or lost the previous point, and (d) if the point is relatively important or unimportant as defined by Morris (1977). The other four tests were non-specific and compared (e) the actual and expected (assuming independence) number of games won on service, (f) the actual and expected duration of games (g) the number of runs of points won and lost, and (h) the distribution of set scores.

Pollard and Pollard applied the eight tests to Nadal's 2011 Grand Slam singles matches against other Top Ten players and although a couple of significant results were obtained, they concluded that independence between points is a good assumption for elite players against other elite players in Grand Slam tournaments, suggesting that these players are competing at full capacity on all points. They did not apply the analysis to Nadal's matches against other (lower ranked) players as it would seem that a very good player would be able to lift his game against a not-so-good player, thereby exhibiting non-independent points.

This paper looks at all the Grand Slam matches for Nadal and for the other Top Four players Djokovic, Federer and Murray, dividing these into Top Ten matches and other matches. The tests are also applied to the opponent's serving performance, which also measures each of the Top Ten players as receivers.

In tennis, players serve alternatively to the first and second court. This suggests that tennis data is eminently suitable for bi-point analysis. The data can be analysed where the first point is served to the first court and the second point to the second court or alternatively the first point is omitted and the point pair

consists of a serve to the second court followed by a serve to the first court. It is also possible to consider all points in each set regardless of service court.

2. Methods

2.1 State dependent relative frequencies.

If the probability of winning a point on service is constant and independent of the previous point, then it would make no difference if the player was ahead, equal or behind on service for the next point. Overall actual performance can be compared with expected performance on the assumption of independence and the difference tested using a chi-squared test. Alternatively a simple sign test can be applied over all the sets noting whether the number of points won on service when ahead, equal or behind is simply above, equal (split 50:50) or below the expectation for that set.

Pollard and Pollard (2012) found no significant difference for Nadal winning the next point on service whether he was ahead, equal or behind in any game against other Top Ten players. This analysis has now been extended to include Nadal's matches in earlier rounds against lower ranked players. (When ahead won 339/473; equal 227/331; behind 101/156) Again there was no significant difference. (Chi-squared with 2 d.f. is 2.87). Obviously Nadal's total Grand Slam service experience will also be not significant. (675/984; 481/744; 254/150). Whilst it was expected that Nadal could lift against weaker players and therefore the original analysis was restricted to his matches against Top Ten players, there is no evidence of lifting against the other players in Grand Slams, who it should be noted would all still be ranked in the top 100.

When considering Nadal's performance returning service, Pollard and Pollard (2012) found a significant difference against Top Ten players. This result was also obtained with the other players, giving an overall performance of (516/845; 443/785; 325/614) with Chi-squared with 2d.f. Of 9.9 which is significant. Thus when the server is ahead he outperforms and when behind he underperforms. The reverse applies to Nadal as receiver. When Nadal is behind when receiving he underperforms, but when Nadal is ahead when receiving he outperforms.

On the other hand the opposite result is obtained with Federer. When Federer is serving and ahead he performs better than when serving and behind. Overall performance was (771/1093; 491/676; 188/298) with Chi-squared 2d.f. of 10.9 which is significant. When receiving his performance was (562/960; 511/865; 340/605) which is not significant with Chi-squared with 2 d.f. of 1.6. Djokovic did not produce any significant results whether serving or receiving and whether against Top Ten or others, suggesting independence. Likewise Murray did not produce any significant results against Top Ten players but did perform significantly worse than expected when behind in a game against other players (446/656; 338/145; 157/263) giving Chi-squared 2 d.f. of 8.6

2.2 Stepwise relative frequencies.

If the probability of winning a point on service is constant and independent of the previous point, then it would make no difference if the player won or lost the previous point. Overall a player's actual performance on service can be tested against the expected under the assumption of independence using a chi-squared test. As before, a simple sign test could be applied to all the sets played.

In their amended presentation to the conference, Pollard and Pollard found no significant difference when applying the Chi-squared test to this data (after winning 368/591; after losing 253/388), but did find a significant difference when applying the t test. Similar results were obtained for matches against the other players so that overall there was no difference whether Nadal had won or lost the previous point. The overall results when serving were (727/1112; 438/657) with Chi-squared 1d.f. of 0.2 and when receiving (608/1057; 474/843) with Chi-squared 1 d.f. of 0.3. Interestingly the sign test was significant with Top Ten and with the other players, and overall in 47.5 sets Nadal won more points after losing a point and in only 27.5 sets he won more points after winning a point which is significant. ($t_{74}=2.31$). Combining the two tests suggests the difference must be small, but more often in the same direction of winning after losing.

Federer's performance when serving was (830/1219; 334/516) with Chi-squared 1 d.f. of 1.80 and when receiving was (667/1168; 463/821) with Chi-squared 1 d.f. of 0.1, neither of which is significant. Likewise Murray and Djokovic show no difference in performance after winning or losing the previous point.

2.3 Combined state and stepwise relative frequencies.

Again assuming independence and constant probability of winning a point on service throughout the set, it would make no difference to the outcome of the next point if (say) the player was ahead on the game and won the previous point or (say) was behind and had lost the previous point. There are six alternatives obtained by combining measures 2.1 and 2.2 and the actual performance for each of these can be compared with the expected using the chi-squared test.

Pollard and Pollard (2012) found that Nadal's performance on service against Top Ten players was just significant at the 5% level and it was not significant when receiving service. However against the other players it was found to be not significant when serving or receiving. Combined for all players it was also not significant when serving or receiving. Similar non significant results were obtained for Federer, Murray and Djokovic.

2.4 Importance based relative frequency.

The importance of a point within a game of tennis is defined by Morris (1977) as the probability that the server wins the game given he wins that point minus the probability that he wins the game given he loses that point. If the points are independent, then a player's performance at an important point (say 30-40 or advantage receiver) should be the same as that at a less important point (say 40-15). There are 15 different scores (30 all and deuce have the same importance, as do 30-40 and advantage receiver, and 40-30 and advantage server). A chi-squared test could be applied to all 15 scores, but it is also possible to combine them into the more important scores (say 15 all and above) and the less important scores (say 40-30 and below) and apply the chi-squared test to observed and expected frequencies.

Surprisingly no significant results were obtained for Nadal, Federer, Murray or Djokovic. Tests were applied to Top Ten and others using the above definition of important and less important points and also using very important (15-30 and above) and others (15-40 and below).

2.5 Number of games won on service.

The Probability P that the server wins his service game where p is his probability of winning a point is given by

$$P = p^4(1 - 16q^4) / (p^4 - q^4) \quad (1)$$

A sign test can then be applied to his performance indicating whether he has won more or less service games in that set to that expected under the assumption of constant p value and independence between points.

All Top Four players won significantly more service games than expected based on the t test value under the sign test. This is not surprising because they are the best four players, but an additional factor to be considered in this calculation is the number of sets where the player wins all his service games, whereas the expected must be slightly less than this, unless q=0, p=1 and P=1. Effectively the observed is discrete data, whereas the expected is nearly continuous data. Further analysis is required.

2.6 Duration of a game of tennis.

The expected duration (defined as number of points played not time duration) of a set of tennis is given by

$$\mu_1 = 4(p^4 + q^4 + 5s(p^3 + q^3) + 15s^2r^{-1} + 10s^3(3 + r)) \quad (2)$$

where $s = pq$ and $r^{-1} = 1 - 2pq$

The second non-central moment of the duration of a game of tennis is given by

$$\mu_2 = 16(p^4 + q^4) + 100s(p^3 + q^3) + 360s^2r^{-1} + 20s^3(36 + 24r + 4r^2(1 + 2s)) \quad (3)$$

and the variance of the duration of a game of tennis is given by $\mu_2 - \mu_1^2$

A sign test can be applied to his performance in each set indicating whether the set was longer or shorter than that expected assuming independence. Each Top Four player experienced a few sets with more points than expected but given the number of sets (around 80) included in each player's record this is not unusual. When the sign test was applied over all sets there was no significant difference between the number of sets with more or less points than expected.

2.7 The Wald-Wolfowitz two sample runs test.

The number of runs of wins and losses has an approximate normal distribution with mean

$$E(R) = 1 + 2n_1 n_2 / (n_1 + n_2) \quad \text{and} \quad (4)$$

$$V(R) = 2n_1 n_2 (2n_1 n_2 - n_1 - n_2) / (n_1 + n_2)^2 (n_1 + n_2 - 1) \quad (5)$$

where n_1 is the number of points won by the server and n_2 is the number of points lost.

Actual performance can be measured against expected for each set and a sign test applied over the total set-by-set results.

Each Top Four player experienced a few more or less runs than expected on one or more sets when serving and when receiving, but again this is to be expected at the usual 5% level when each Top Four player plays up to 28 matches over 4 Grand Slams and around 80 sets. However when applying the sign test to the difference between the observed and expected number of runs over all sets played, no significant results were obtained.

2.8 Distribution of set scores.

Given the p-values for each player and a knowledge of who served first in each set, and assuming points are independent, the expected distribution of set scores (6-0, 6-1, 6-2, ..., 7-6, 6-7, ..., 1-6, 0-6) can be derived. When the observed median score exceeds the expected median score then the player has effectively lifted when it matters. No evidence of lack of independence was obtained for all four players using this test.

3 Discussion

3.1 Some further thoughts related to measures of statistical independence of points in the tennis setting

The above analysis primarily looks at the independence of one point from the previous point, although there is some consideration of previous points for example when considering whether the player is ahead, equal or behind in a particular game or when considering the importance of the point being played. Also tests 2.5 to 2.8 are non-specific and may pick up other forms of lack of independence. A recent written communication and discussion with G. H. Pollard (2013) concerning further aspects of lack of independence has lead to the examples and tests in Sections 3.2 to 3.5. He believed that it was appropriate to include these details in this paper. Interestingly he felt that the test based on triplets in section 3.4 could prove to be more powerful at identifying lack of independence than several other tests.

3.2 Sporting examples with smaller variance than the binomial distribution

Example 1. Suppose the probability of winning a point having won the previous point is p_w , and the probability of winning a point having lost the previous point is p_l . Then, the steady state probability of winning a point, π_w , is given by $\pi_w = p_l / (p_l + 1 - p_w)$. Thus, in a (long) sequence of points, the probability of WW, WL, LW, and LL is $\pi_w p_w$, $\pi_w(1-p_w)$, $(1-\pi_w)p_l$ and $(1-\pi_w)(1-p_l)$ respectively. Hence, the distribution of the number of wins across two points is given in the following table, as is the corresponding binomial distribution. The table also gives values when $p_l = 0.7$ and $p_w = 0.5$, and hence $\pi_w = 7/12$.

Points won	Probability	Binomial
0	$(1-\pi_w)(1-p_l) = 0.1250$	$\pi_l^2 = 0.1733$
1	$\pi_w(1-p_w) + (1-\pi_w)p_l = 0.5833$	$2\pi_l\pi_w = 0.4861$
2	$\pi_w p_w = 0.2917$	$\pi_w^2 = 0.3403$

Thus, it can be seen that when $p_l > p_w$, the above distribution has a smaller variance than under the corresponding binomial distribution. It is clear that this is also the case if we were to consider the number of wins across three points, etcetera.

Example 2. Suppose player A has a probability of $p + \delta$ of winning a point having lost the previous point, but otherwise it is p . Then, $\pi_w = (p + \delta)/(1 + \delta)$ which equals 0.6364 when $p = 0.6$ and $\delta = 0.1$. We also have,

corresponding to above, when $p = 0.6$, the point-pair distribution in the following table. Note that, similar to example 1, there is a decrease in the probability of WW or LL when $\delta = 0.1$ relative to when $\delta = 0$.

	$\delta = 0$	$\delta = 0.1$
WW	0.36	0.3818
LW	0.24	0.2545
WL	0.24	0.2545
LL	0.16	0.1091
	1	1

Example 3. We now consider a situation with 2-step dependencies. Suppose $p(w_3/w_1 \text{ and } w_2) = 0.5$, $p(w_3/l_1 \text{ and } w_2) = 0.6$, $p(w_3/w_1 \text{ and } l_2) = 0.7$ and $p(w_3/l_1 \text{ and } l_2) = 0.8$, where, for example, the notation $p(w_3/l_1 \text{ and } w_2)$ refers to the probability that the $(n+3)$ rd point is won given that the $(n+1)$ th point was lost and the $(n+2)$ th point was won. The ‘two-state’ steady state probabilities can be shown to be given by $\pi_{ww} = (p(w_3/l_1, w_2)/p(l_3/w_1, w_2))\pi_{lw}$, $\pi_{ll} = (p(l_3/w_1, l_2)/p(w_3/l_1, l_2))\pi_{wl}$, $\pi_{wl} = \pi_{lw}$ and $\pi_{ww} + \pi_{wl} + \pi_{lw} + \pi_{ll} = 1$. The ‘one-state’ steady state probabilities are given by $\pi_w = (2\pi_{ww} + \pi_{wl} + \pi_{lw})/2$ and $\pi_l = 1 - \pi_w$. Thus, in this example, $\pi_{ww} = 48/143$, $\pi_{wl} = 40/143$, $\pi_{lw} = 40/143$, $\pi_{ll} = 15/143$, $\pi_w = (2*48+40+40)/(2*143) = 8/13$, $\pi_l = 5/13$, and the distribution of X , the number of points won in a point-pair is given by

X, Points won	Probability	Binomial
0	$15/143 = 0.1049$	$25/169 = 0.1479$
1	$80/143 =$	$80/169 = 0.4734$
2	$48/143 = 0.3357$	$64/169 = 0.3787$

Again, it is clear that the variance of X is smaller than for the associated binomial situation. For point-triplets, we have the following table, again with smaller variance than for the binomial.

Points won	Probability	Binomial
0	$3/143 = 0.0210$	$125/2197 = 0.0569$
1	$40/143 = 0.2797$	$600/2197 = 0.2731$
2	$76/143 = 0.5315$	$960/2197 = 0.4370$
3	$24/143 = 0.1678$	$512/2197 = 0.2330$

3.3 Testing for a smaller (or larger) variance than for the binomial case.

We now consider testing for a smaller variance than for the binomial, as in the above situations. We note that if X is distributed as Binomial (m, p) , $E(X) = mp$, $E(X^2) = m(m-1)p^2 + mp$, and

$$E(X^4) = m(m-1)(m-2)(m-3)p^4 + 6m(m-1)(m-2)p^3 + 7m(m-1)p^2 + mp.$$

Consider for example the first set in the final of the French Open in 2011 between Nadal and Federer. The data for this set are given in the paper by Pollard and Pollard (2012, p.114). Considering sequentially and in pairs all the points on Nadal’s service games in this set, we have

# points won, x	Frequency, f	fx	fx^2
0	3	0	0
1	12	12	12
2	7	14	28
Total	22	26	40

Firstly, we note that, for any X , $Var(X^2) = E(X^4) - (E(X^2))^2$. For the binomial with $m = 2$ and $p = 26/44$, $E(X^4) = 6.0703$, $E(X^2) = 1.8802$, and so $Var(X^2) = 2.5352$. Hence, $Var(\sum_i^n X_i^2 / n) = 2.5352 / 22 = 0.11523$, and

so the standard deviation is 0.3395. As $E(X^2)$ is 1.8802 (for the binomial hypothesis), the observed value of $40/22 = 1.8182$ has a ‘standardized Z-value’ of -0.1825.

If we repeat this analysis starting with the second point (thereby omitting the first and last points of the set from the analysis), we have the following table. Now with $m = 2$ and $p = 25/42$, $E(X^4) = 6.1508$, $E(X^2) = 1.8991$, and $\text{Var}(X^2) = 2.5442$. Thus, the observed value of $37/21 = 1.7619$ has a Z-value of -0.3941.

# points won, x	Frequency, f	fx	fx^2
0	2	0	0
1	13	13	13
2	6	12	24
Total	21	25	37

Both these Z-values are negative, indicating that the spreads of the distributions are slightly less than would be expected for a binomial distribution. Clearly, these two analyses use the same underlying data and are not independent analyses.

As a player typically has more than 20 service points in a set, n in the above analysis is typically 10 or more, so Z is approximately normally distributed (assuming the underlying binomial distribution).

This type of analysis can be repeated for triplets, quadruplets, etcetera.

3.4 A test based on triplets.

Consider 3 points in a row. As each point has two possible outcomes (W and L), there are 8 possible outcomes for the triplet. We focus on the two outcomes LWL and WLW. Under independence these two possibilities have probabilities q^2p and p^2q , which sum to pq , where $q = 1-p$. Now if $p(W/L) = p + \delta$ and $p(W/W) = p - \delta$ using an obvious notation, then $p(LWL) = (1-\pi_w)(p+\delta)(q+\delta)$ and $p(WLW) = \pi_w(q+\delta)(p+\delta)$, both increases on the above. Thus, if we consider the sequence of point outcomes in triplets, and record whether each outcome is the (binary) outcome (LWL or WLW) or something else, we can assess whether there are ‘too many’ (LWL or WLW)s than would be expected under independence.

We again consider Nadal serving in the first set against Federer in the 2011 French Open final. Starting with the first point, Nadal won 24 out of first 42 points, so $p = 24/42$, and $pq = 0.2449$. There were 2 (WLW)s and 3 (LWL)s, making 5 (WLW or LWL)s out of the 14 triplets. The ratio $5/14 = 0.3571$ is greater than 0.2449, indicating the possibility that Nadal lifts having lost the previous point or lowers having won the previous point.

Starting with the second and then third points rather than the first, the respective values for p were 25/42 and 26/42, giving pq values of 0.2409 and 0.2358 respectively. In the first case there were 4 (WLW)s and 1 (LWL), making 5 (WLW or LWL)s out of the 14 triplets, and in the second case there were 3 (WLW)s and 2 (LWL)s, again making 5 (WLW or LWL)s. These observations indicate the same possibility noted in the previous paragraph.

Of course the three analyses above are not independent. In fact they are clearly quite dependent. Nevertheless, as in some of the other tests above, it is interesting to look at all three analyses.

3.5 A binomial homogeneity test

Suppose we analyse the triplets consisting of points numbered 123, 456, 789, ...using the table below. For example, a loss on point 6, given a loss on point 4 and a win on point 5, would lead to an additional observation in the cell with row LW and column L. Then, for a player who has the capacity to increase his chance of winning a particular point, we might expect the proportion in the W column to increase as we move down the rows.

	W	L	Total
WW	N1	Diff	M1
LW	N2	Diff	M2
WL	N3	Diff	M3

LL	N4	Diff	M4
Total	N1+N2+N3+N4	Diff	M1+M2+M3+M4

We can analyse this table using a Chi-Squared analysis with 3 degrees of freedom, or we can partition the table in various ways (Lancaster (1949, 1950), Irwin (1949)), or we might combine the rows in several ways.

What we are looking for in the above table is a trend in p-values (row-wise), and Armitage (1955) gives a treatment of this.

4. Conclusion

In the probabilistic modeling of sports such as tennis the assumption is made that all points are independent and identically distributed. Four specific and four general tests were applied to test this hypothesis using the performance of the top four players Nadal, Federer, Djokovic and Murray over the four Grand Slam tournaments. The tests were applied for each player when they were serving and when they were receiving and subdivided by matches against top ten players (generally quarterfinals onwards) and against other players (who having reached the main draw of a Grand Slam were probably ranked in the top 100). Most tests confirmed the assumption of independence. There was some evidence that when he is receiving and ahead in that game Nadal can lift his play. For Federer there is some evidence that when he is serving and ahead in that game he can lift his play. For Murray there was some evidence that when he was behind in a game he performed worse on the next point than when he was ahead. No significant differences were identified for Djokovic. Overall the few differences identified suggest that for the Top Four players points the assumption that points are independent is either true or at least a reasonable approximation.

An interesting alternative approach to the above point data analysis is to consider all the points in a set of tennis as a long series which can be analysed in pairs or even triplets. The mathematics behind this approach is discussed.

References

- Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, 11, 375.
- Irwin, J. O. (1949), A note on the subdivision of Chi-Squared into components, *Biometrika*, 36, 130.
- Kemeny, J. G. and Snell J. L. (1960) Finite Markov Chains, Princeton, New Jersey, D Van Nostrand
- Klassen F. J. and Magnus J. R, (2001) Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. *Journal of the American Statistical Association*, 96, 500-509.
- Lancaster, H. O. (1949). The derivation and partition of Chi-Squared in certain discrete distributions, *Biometrika*, 36, 117.
- Lancaster, H. O. (1950). The exact partition of Chi-Squared and its application to the problem of the pooling of small expectations. *Biometrika*, 37, 267.
- Morris, C. (1977) The most important points in tennis. In Optimal strategies in sports, edited by S. P, Ladany and R.E. Machol, 131-140. Amsterdam North Holland (Vol 5 in studies in Management Science and Systems).
- Pollard, G. H. (1983) An analysis of classical and tiebreaker tennis. *Australian Journal Statistics*, 25(3), 496-505
- Pollard, G. H. (2004) Can a tennis player increase the probability of winning a point when it is important? Proceedings of the Seventh Australian Conference on Mathematics and Computers in Sport, edited by R. H. Morton and S. Ganeshalingan, Massey University, Massey, New Zealand, 253-256.
- Pollard, G. H. (2013) Some further thoughts and aspects of lack of independence in the tennis setting. (Private communication)
- Pollard, G. H., Cross, R. and Meyer, D. (2006) An analysis of ten years of the four grand slam men's singles data for lack of independence of set outcomes, *Journal of Sports Science and Medicine*, 5, 561-566.
- Pollard, G.H., Pollard, G.N., Lyle, I. and Cross, R. (2006) Bias in Sporting Match Statistics. Proceedings of the Tenth Australasian Conference on Mathematics and Computers in Sports, Darwin, July 2010, edited by A, Bedford and M Ovens, 221=228.

Pollard, G. N. and Pollard, G. H. (2012). Applying statistical tests for the independence of points in tennis, Proceedings of the 11th Australasian Conference on Mathematics and Computers in Sport, edited by Anthony Bedford and Adrian Schembri, Melbourne, Victoria, Australia, published by MathSport (ANZIAM).

Siegel, S. (1956) Non-parametric Statistics, McGraw Hill.

A comparison of the Masters scoring system and the Knock-out scoring system

Graham Pollard* and Geoff Pollard**

* Faculty of Information Sciences and Engineering, University of Canberra, Canberra, Australia,
graham@foulsham.com.au

** Faculty of Life and Social Sciences, Swinburne University of Technology, Melbourne, Australia, gpollard@tennis.com.au

Abstract. The Masters is a generic term for the year-ending event involving the top eight tennis players in the world. There is a men's Masters event and a women's event. The Masters format consists of two 'roughly equal' groups of four players in round-robin competition with the best two performing players in each group progressing to the semi-finals, which then continue on a knock-out basis. This paper compares some results for the Masters event with those for the standard knock-out event. The outcome for the Masters draw is seen to be less variable than that for the knock-out draw. This result is akin to the role of the sample size in the central limit theorem, noting that the Masters draw consists of 15 matches in total, whereas the corresponding knock-out draw consists of just 7 matches. Further, it is noted that, under reasonable assumptions, the probability that at least one of the best two players reaches the final is greater for the Masters system than it is for the knock-out system, and the probability that at least two of the best four players reaches the semi-finals is also greater for the Masters system.

1. Introduction

Scarf and Bilbao (2006) and Scarf, Yusuf and Bilbao (2009) note that there are really only two tournament designs, namely the round-robin, in which every competitor plays every other competitor and the one with the best record wins, and the knockout tournament (or single elimination tournament) in which matches are played in rounds with each winner progressing to the next round and the loser eliminated. All other designs can be considered as variations or hybrids of these two formats.

Ryvkin and Ortmann (2006, 2008) also recognize a third format called a contest, where all players compete together only once, such as in a race or in jumps or a stroke play golf tournament, and the best performer is declared the winner. The event organizer's choice of format is determined by balancing predictive power (probability better player wins) and organizational costs (time and number of matches). As costs increase the organizer will move from the round-robin to knockout and eventually to the contest format. However the contest format is not relevant to tennis.

The typical hybrid design consists of all competitors playing a round-robin in groups, with the best competitors in group play progressing to a knockout stage to determine the winner. Glenn (1960), Seals (1963), Appleton (1995), and Marchand (2002) all examined the probability that the best player wins a tournament conducted under round-robin, knockout and various hybrid formats. McGarry and Schultz (1997) measure a format's efficacy to rank all the contestants. The traditional knock-out is weaker than the round robin in this aspect, but involves fewer matches, and can be considerably improved with accurate seeding and if double elimination procedures are used.

The Masters is a generic term used to describe the year-ending tournament between the top eight tennis players in the world. Rather than a knock-out tournament as played throughout the year, the Masters format consists of two roughly equal groups of four players who play each other in round-robin format with the best two performing players in each group progressing to the semi-finals. The semi-finals and final revert to the knock-out format.

Pollard, Pollard and Meyer (2010) considered several aspects of ordering, ranking and seeding players (for the cases of 3 and 4 players). For example, they showed that, with 4 players where player A might be better than player B and player B better than player C whilst player C is better than player A, the (correct) ordering or ranking of the 4 players for a knock-out tournament can be different to that for a round-robin tournament. Given the possibility of such a situation in practice, they identified a method of ranking 4 such players in a knock-out tournament, and found a method of making the draw so as to maximize the probability that the best two players reach the final.

Given that the Masters system will be in operation for many years, it is sufficient to consider parameters that are reasonable in a very general sense. It would appear to be unnecessary to consider situations such as mentioned in the above paragraph where the principle of transitivity does not apply.

This paper compares some results for the Masters event with those for the standard knock-out event.

2. Method

2.1 The present Masters system and the knock-out system

When the draw for the Masters tournament is made, there is a given ranking of the players from 1 (highest) to 8 (lowest). The ‘rounds’ in the Masters system consist of two groups of four players, with each group of 4 players playing a round-robin. The players ranked 1 and 2 are placed in different round-robins. The third ranked player is then placed at random in one of the two groups, with the fourth ranked player placed in the other group. The fifth ranked player is placed at random in one of the groups, and the sixth ranked player in the other group. Correspondingly, the seventh and eighth ranked players are allocated at random to the two groups.

Two players from each group proceed to the semi-finals, where the player who comes first in the first group plays the player who comes second in the second group, and the player who comes first in the second group plays the player who comes second in the first group. The two winners in the semi-finals play in the final to determine the winner. Thus, the Masters consists of 15 matches with the finalists playing 5 matches, and so it is an appropriate ‘size’ for a prestige tournament of one week duration.

The knock-out draw that we consider has the following characteristics. Players 1 and 2 are in separate halves of the draw. Player 1 plays player 3 or player 4 in a semi-final (if they both get that far), and player 2 plays player 4 or 3 respectively (if they also both get that far). Player 4 plays player 5 or player 6 in the first round, whilst player 3 plays player 6 or 5 correspondingly. Player 1 plays player 8 or 7 in the first round, and player 2 plays 7 or 8 correspondingly.

It can be seen that there are 8 equally likely draws for the Masters tournament, and 8 equally likely draws for the above knock-out structure.

2.2 The underlying probability structure

The highest ranked player is denoted player $i = 1$ and the lowest ranked player is player $j = 8$. We consider the situation where the probability player i beats player j is equal to $0.5 + (j - i)*d$ for a relevant value of $d > 0$. Table 1 gives these probabilities when $d = 0.04$. We say player i is better than player j if the probability i beats j is greater than 0.5. It can be seen that players satisfying the above probability relationship have the transitivity property. That is, if player i is better than player j , and player j is better than player k , then player i must be better than player k .

Table 1. Probability player i beats player j ($d = 0.04$).

Prob(i beats j)	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$
$i = 1$	X	0.54	0.58	0.62	0.66	0.7	0.74	0.78
$i = 2$	0.46	X	0.54	0.58	0.62	0.66	0.7	0.74
$i = 3$	0.42	0.46	X	0.54	0.58	0.62	0.66	0.7
$i = 4$	0.38	0.42	0.46	X	0.54	0.58	0.62	0.66
$i = 5$	0.34	0.38	0.42	0.46	X	0.54	0.58	0.62
$i = 6$	0.3	0.34	0.38	0.42	0.46	X	0.54	0.58
$i = 7$	0.26	0.3	0.34	0.38	0.42	0.46	X	0.54
$i = 8$	0.22	0.26	0.3	0.34	0.38	0.42	0.46	X

2.3 Comparing the Masters and the knock-out systems

It can be seen that, for both the Masters and the knock-out systems, the number of players remaining in the event is reduced sequentially from 8 to 4, and then from 4 to 2. It would appear that the purpose of constructing a (well-considered) draw is to distribute the better players ‘evenly’ throughout the draw and thus reduce the probability that the best players are eliminated in the ‘early rounds’. Accordingly, one variable we might use to measure this ‘quality’ aspect of a draw or system might be X , where X is the number of the best 4 players who reach the semi-finals. Thus, given two draws A and B, it could be argued that A is a better draw than B if it has a higher value for the probability that half or more of the best 4 players reach the semi-finals (i.e. $X = 2, 3$ or 4) in the process of reducing the total number of players by one-half. Correspondingly, we might also use the variable Y , the number of the best 2 players who reach the final to measure the ‘quality’ of a draw, with system A being better than system B if it has a higher value for the probability Y equals 1 or 2.

For simplicity we consider firstly the case in which the first round-robin group of players in the Masters is the players ranked 1, 3, 5 and 7 (and the second group is players ranked 2, 4, 6 and 8). It can be seen that for these particular groupings and the values in Table 1, the ‘player/player win probabilities’ within the first group are the same as those within the second group. (Note that this is not the case for the other seven possible groupings.) Now, in general, if X_1 is the number of players ranked 1 to 4 who reach the semi-finals in the first group and X_2 is the number of players ranked 1 to 4 who reach the semi-finals in the second group, then X_1 and X_2 are independent variables, and $X = X_1 + X_2$. In our analysis we assume that in the situations in which there is a tie (in the number of matches won) between 2 or 3 players at the round-robin stage, the tie is resolved fairly (at random and with equal probabilities) across the players involved. (In practice ties are resolved by count-back methods on sets won/lost and if still equal on games won/lost.) A (somewhat complicated) spreadsheet was used to analyze the Masters system. The resultant distribution of X (for the particular groupings described above) is given in Table 2. Note that all output values in this table and throughout the paper are given to 4 decimal points.

The distribution of X for the knock-out scoring system in which player 1 plays player 5 and 3 plays 7 in one half of the draw, and 2 plays 6 and 4 plays 8 in the other half of the draw is given in the last column of Table 2. It can be seen that the variance of X is greater for the knock-out structure. Also, it can be seen that half or more of the best 4 players is more likely to reach the semi-finals under the Masters structure.

Given that the number of players in an eight-player tournament is halved by the semi-final stage, it would appear that a ‘good’ scoring system would require a ‘reasonably high’ value for the probability that at least half of the top 4 players reach the semi-finals. These probabilities are given in the last row of Table 2, with the Masters having the higher value for this probability. It can be seen that, given the two means are comparable and greater than 2, this higher probability value for the Masters corresponds to it having a smaller variance for X .

Table 2. Distribution of X , the number of the best 4 players who reach the semi-finals.

Distribution of X	Masters (1,3,5,7), (2,4,6,8)	Knock-out (1,5;3,7),(2,6;4,8)
$P(X = 0)$	0.0029	0.0134
$P(X = 1)$	0.0646	0.1038
$P(X = 2)$	0.3981	0.3021
$P(X = 3)$	0.4150	0.3910
$P(X = 4)$	0.1193	0.1897
Mean	2.5833	2.6400

Variance	0.6284	0.8976
P(X = 2, 3 or 4)	0.9325	0.8829

The spreadsheet for the Masters and knock-out draws above included results for the finals. Given that the number of players in the tournament is halved again by the final stage, it would appear that a ‘good’ scoring system would require a ‘reasonable’ value for the probability that at least half of the top 2 players reach the final. This probability is given for the same two systems in the last row of Table 3 and it can be seen that the Masters has a higher value for it. Further, given that the means are similar and close to 1, this higher probability value for the Masters corresponds to it having a smaller variance for Y.

The distributions of Y for the eight equally likely Masters draws and for the overall draw are given in Table 4. The eight individual distributions are of course very slightly different. As in the tables to follow, some of these differences may be of interest to some readers. For example, the draw (1,3,6,8), (2,4,5,7) has the highest probability that at least one of the best two players reaches the final. This particular draw also has the second largest mean and the smallest variance of Y, these two characteristics being related to the fact that it has the smallest value for P(Y = 0). Further, the draw (1,4,6,7), (2,3,5,8) has the lowest value for P(Y = 1, 2), and correspondingly the smallest mean and the second largest variance.

The corresponding results to those in Table 4 for the eight knock-out draws (and the overall knock-out situation) are given in Table 5. It can be seen that the draw (1,8)(4,6),(2,7)(3,5) has the largest value for P(Y = 1 or 2), the largest mean and smallest variance of Y. In terms of the best two players reaching the final, it could be argued that this draw is (very marginally) the ‘best’ knock-out draw. Overall, it can be seen that the distribution of Y for the knock-out draw is more variable than for the Masters draw.

Table 3. Distribution of Y, the number of the best 2 players who reach the final.

Distribution of Y	Masters (1,3,5,7), (2,4,6,8)	Knock-out (1,5:3,7),(2,6:4,8)
P(Y = 0)	0.2586	0.3379
P(Y = 1)	0.6339	0.4868
P(Y = 2)	0.1074	0.1753
Mean	0.8488	0.8374
Variance	0.3432	0.4868
P(Y = 1 or 2)	0.7414	0.6621

Table 4. Distribution of Y for the eight possible Masters draws.

Distribution of Y	Masters (1,3,5,7), (2,4,6,8)	Masters (1,3,6,8), (2,4,5,7)	Masters (1,3,5,8), (2,4,6,7)	Masters (1,3,6,7), (2,4,5,8)	Masters (1,4,5,7), (2,3,6,8)	Masters (1,4,6,8), (2,3,5,7)	Masters (1,4,5,8), (2,3,6,7)	Masters (1,4,6,7), (2,3,5,8)	Masters Overall
P(Y = 0)	0.2586	0.2578	0.2586	0.2585	0.2584	0.2588	0.2586	0.2593	0.2586
P(Y = 1)	0.6339	0.6357	0.6346	0.6346	0.6347	0.6345	0.6348	0.6341	0.6346
P(Y = 2)	0.1074	0.1065	0.1068	0.1068	0.1068	0.1066	0.1066	0.1066	0.1068
Mean	0.8488	0.8487	0.8483	0.8483	0.8484	0.8478	0.8480	0.8473	0.8482

Variance	0.3432	0.3414	0.3424	0.3423	0.3423	0.3423	0.3421	0.3426	0.3424
P(Y = 1, 2)	0.7414	0.7422	0.7414	0.7415	0.7416	0.7412	0.7414	0.7407	0.7414

Table 5. Distribution of Y for the eight possible knock-out draws.

Distribution of Y	(1,8)(3,6), (2,7)(4,5)	(1,8)(3,5), (2,7)(4,6)	(1,7)(3,6), (2,8)(4,5)	(1,7)(3,5), (2,8)(4,6)	(1,8)(4,6), (2,7)(3,5)	(1,8)(4,5), (2,7)(3,6)	(1,7)(4,5), (2,8)(3,6)	(1,7)(4,6), (2,8)(3,5)	Knock-out total
P(Y = 0)	0.2976	0.2974	0.2992	0.2980	0.2934	0.2962	0.2990	0.2972	0.2973
P(Y = 1)	0.4980	0.4970	0.4958	0.4958	0.5019	0.4996	0.4963	0.4975	0.4977
P(Y = 2)	0.2044	0.2056	0.2050	0.2062	0.2047	0.2041	0.2047	0.2053	0.2050
Mean	0.9068	0.9081	0.9058	0.9081	0.9113	0.9079	0.9058	0.9081	0.9077
Variance	0.4933	0.4946	0.4954	0.4958	0.4902	0.4919	0.4948	0.4940	0.4973
P(Y = 1, 2)	0.7024	0.7026	0.7008	0.7020	0.7066	0.7038	0.7010	0.7028	0.7027

We now consider the distribution of X, the number of the best 4 players who reach the semi-finals. The distribution of X for the various possible Masters draws is given in Table 6. Interestingly, the distribution of X for the draw (1,3,6,7), (2,4,5,8) is identical to its distribution for the draw (1,4,5,7), (2,3,6,8). Also, the distribution of X for the draw (1,3,5,8), (2,4,6,7) is identical to the distribution for draw (1,4,6,8), (2,3,5,7). The overall distribution of X is given in the last column of Table 6. Further, the draw (1,3,6,8), (2,4,5,7) which had the highest value for P(Y = 1,2), has the lowest value for P(X = 2, 3 or 4). Also, the draw (1,4,6,7), (2,3,5,8), which had the lowest value for P(Y = 1 or 2), has the highest value for P(X = 2, 3 or 4), the smallest mean and the second smallest variance for X. These results are seen to be reasonable as the means of Y are less than 1, whilst the means of X are greater than 2.5.

Table 6. Distribution of X for the eight possible Masters draws.

Distribution of X	(1,3,5,7), (2,4,6,8)	(1,3,6,8), (2,4,5,7)	(1,3,5,8), (2,4,6,7)	(1,3,6,7), (2,4,5,8)	(1,4,5,7), (2,3,6,8)	(1,4,6,8), (2,3,5,7)	(1,4,5,8), (2,3,6,7)	(1,4,6,7), (2,3,5,8)	Masters Overall
P(X = 0)	0.0029	0.0028	0.0028	0.0028	0.0028	0.0028	0.0028	0.0028	0.0028
P(X = 1)	0.0646	0.0651	0.0647	0.0643	0.0643	0.0647	0.0641	0.0641	0.0645
P(X = 2)	0.3981	0.3951	0.3987	0.3981	0.3981	0.3987	0.4002	0.4004	0.3984
P(X = 3)	0.4150	0.4191	0.4153	0.4165	0.4165	0.4153	0.4147	0.4146	0.4159
P(X = 4)	0.1193	0.1179	0.1184	0.1183	0.1183	0.1184	0.1181	0.1181	0.1184
Mean	2.5833	2.5843	2.5819	2.5830	2.5830	2.5819	2.5811	2.5810	2.5825
Variance	0.6284	0.6256	0.6264	0.6254	0.6254	0.6264	0.6247	0.6248	0.6258
P(X = 2,3,4)	0.9325	0.9321	0.9325	0.9328	0.9328	0.9325	0.9330	0.9331	0.9327

The corresponding results to the ones in Table 6 for the eight possible knock-out draws (and the overall knock-out situation) are given in Table 7. It can be seen that the means are all equal and that the distributions are ‘pairwise’ identical. Further, the draws (1,8)(3,6),(2,7)(4,5) and (1,8)(4,5),(2,7)(3,6) have the largest

value for $P(X = 2, 3 \text{ or } 4)$ and smallest variance of X . Thus, in terms of the best four players reaching the semi-final, it could be argued that these two draws are (very marginally) the ‘best’ knock-out draws. Overall, it can be seen that the distribution of X for the knock-out draw is more variable than for the Masters draw.

Table 7. Distribution of X for the eight possible knock-out draws.

Distribution of X	$(1,8)(3,6), (2,7)(4,5)$	$(1,8)(3,5), (2,7)(4,6)$	$(1,7)(3,6), (2,8)(4,5)$	$(1,7)(3,5), (2,8)(4,6)$	$(1,8)(4,6), (2,7)(3,5)$	$(1,8)(4,5), (2,7)(3,6)$	$(1,7)(4,5), (2,8)(3,6)$	$(1,7)(4,6), (2,8)(3,5)$	Knock-out total
$P(X = 0)$	0.0115	0.0117	0.0118	0.0119	0.0117	0.0115	0.0118	0.0119	0.0117
$P(X = 1)$	0.1002	0.1006	0.1004	0.1008	0.1006	0.1002	0.1004	0.1008	0.1005
$P(X = 2)$	0.3078	0.3076	0.3071	0.3068	0.3076	0.3078	0.3071	0.3068	0.3073
$P(X = 3)$	0.3977	0.3965	0.3974	0.3962	0.3965	0.3977	0.3974	0.3962	0.3970
$P(X = 4)$	0.1828	0.1837	0.1833	0.1842	0.1837	0.1828	0.1833	0.1842	0.1835
Mean	2.6400	2.6400	2.6400	2.6400	2.6400	2.6400	2.6400	2.6400	2.6400
Variance	0.8656	0.8688	0.8688	0.8720	0.8688	0.8656	0.8688	0.8720	0.8688
$P(X = 2,3,4)$	0.8883	0.8878	0.8878	0.8873	0.8878	0.8883	0.8878	0.8873	0.8878

We have noted that the ‘quality’ of a draw or tournament structure might be assessed using the variables X and Y , where X is the number of the best 4 players who reach the semi-finals and Y is the number of the best 2 players who reach the final. Using these two variables it is concluded that the Masters draw has a higher ‘quality’ of result than does the knock-out structure. The variables X and Y for the Masters draw are less variable than for the knock-out draw. This result is akin to the role of the sample size in the central limit theorem, noting that the Masters draw consists of 15 matches in total, whereas the knock-out draw consists of just 7 matches.

3. Discussion

The Masters draw has a total of 15 matches, whilst the associated knock-out draw for 8 players has only 7 matches. Midway between these two systems is a ‘Masters-like’ draw consisting of 11 matches. This system uses ‘partial round-robin’ rather than ‘full round-robin’. For example, in Table 2 we considered the Masters draw $(1,3,5,7),(2,4,6,8)$ which has the round-robins $(1,3,5,7)$ and $(2,4,6,8)$. Suppose in the first of these round-robins the matches between 1 and 3, and 5 and 7 were not played, and correspondingly in the second round-robin the matches between 2 and 4, and 6 and 8 were not played. That is, the round-robin sections of the Masters were modified so that the matches between the highest ranked 4 players, and the matches between the lowest ranked 4 players are not played. In terms of considering X , the number of the best 4 players to reach the semi-finals, the approach of deleting these matches between them, and focusing on the various matches between the highest ranked 4 and the lowest ranked 4 players would appear to make sense. It would seem that playing matches between the highest ranked 4 players would only tend to increase the probability of one of them not making the semi-finals, and playing matches between the lowest ranked 4 players would only tend to increase the probability of one of them making the semi-finals.

The ‘Masters-like’ draw with 11 matches in total and ‘partial round-robin’ as described in the paragraph above was briefly considered for the draw $(1,3,5,7),(2,4,6,8)$. The mean and variance of X were 2.6400 and 0.7552 respectively, and $P(X = 2, 3 \text{ or } 4)$ was 0.9104. The corresponding values in Table 2 were 2.5833, 0.6284 and 0.9325. Thus, the value for $P(X = 2, 3 \text{ or } 4)$ for this ‘partial case’ lay between the two values for

it in Table 2. The same was the case for the value for the variance of X, as might have been expected. Thus, this 'Masters-like' draw with 11 matches in total is of theoretical interest, although its increased likelihood of draws (given the smaller number of matches at the round-robin stage) with the need for 'count-backs' at the round-robin stage, reduces its practical relevance.

The idea of 'partial round-robbins' can be used in another way. Suppose each of the 4 players ranked 1 to 4 plays each of the players ranked 5 to 8. This involves just 16 matches, less than the 28 matches that a full round-robin between the top 8 players would involve. If this 'partial round-robin' was followed by the top 4 players proceeding to the semi-finals, there would be a total of 19 matches in such a system, and the properties of the variable X could in turn be examined. Alternatively, if the top 2 players went straight to the final, there would be a total of 17 matches.

There are other applications of 'partial round-robbins'. The following is one. Suppose we have 16 players (rather than just 8), ranked 1 to 16, and they are divided into two groups of 8 at the 'round-robin stage'. Along the lines of the present Masters tournament, the first group consists of player 1, player 3 or 4, player 5 or 6, player 7 or 8, player 9 or 10, player 11 or 12, player 13 or 14 and player 15 or 16. The second group consists of the remaining 8 players. Suppose the first group plays a 'partial round-robin' of 16 matches in which each of the top 4 players in that group plays each of the lowest ranked players in that group. The second group of 8 players plays a corresponding 'partial round-robin'. This 'round-robin stage' of a total of 32 matches could be followed by quarter-finals of 8 players (making use of the orderings that result from the two 'partial round-robbins'), or followed by semi-finals of 4 players in the same way as the Masters.

It would appear that the 'partial round-robin' is potentially a useful construct to consider in the development of tournament structures such as the Masters.

4. Conclusions

This paper compares some results for the Masters scoring system with those for the standard knock-out system. The variable defined as the number of the two highest ranked players who reach the final is proposed as a reasonable measure for comparing two such tournament structures. Under reasonable assumptions, the probability that this variable is 1 or 2 is larger for the Masters structure than for the knock-out one, and this fact is seen to be consistent with the fact that this variable has a smaller variance for the Masters structure. This smaller variance can in turn be seen as mirroring the role of the sample size in the central limit theorem, noting that the Masters draw consists of 15 matches in total, whereas the corresponding knock-out draw consists of just 7 matches.

Correspondingly, the variable defined as the number of the four highest ranked players who reach the semi-final is proposed as another reasonable way of comparing the two tournament structures. Under the same assumptions, the probability that this variable is 2, 3 or 4 is larger for the Masters structure than for the knock-out one, and this fact is also consistent with this variable having a smaller variance for the Masters structure. Again, this smaller variance mirrors the role of the sample size in the central limit theorem.

Finally, a structure with characteristics 'between' those of the Masters and the knock-out structures, and one making use of 'partial round-robbins', is outlined. It requires 11 matches in total. Such 'partial round-robbins' are shown to have uses in other situations such as where a total of 16 players are ranked and an initial 'round-robin structure' is required for the tournament.

Acknowledgement

The authors wish to thank Professor Denny Meyer for drawing attention to several earlier studies.

References

- Appleton DR. (1995) May the best man win? *The Statistician*; Vol. 44, pp. 529-538.
Glenn WA. (1960) A comparison of the effectiveness of tournaments. *Biometrika*; vol 47, pp. 253-262.
Marchand E. (2002) On the comparison between the standard and random knockout tournaments. *The Statistician*; vol. 51, pp. 169-178.
McGarry T and Schutz R. (1997) Efficacy of traditional sport tournament structures. *Journal of Operational Research Society*; vol. 48, pp. 64-74.

- Pollard GH, Pollard GN and Meyer D. (2010). Some aspects of ordering, ranking and seeding 1, Proceedings of the Tenth Australasian Conference on Mathematics and Computers in Sport, edited by A. Bedford and M. Ovens, Darwin, Australia, July, 2010, pp. 49-56.
- Ryvkin D and Ortmann A (2006) Three prominent tournament formats: predictive power and costs. CERGE-EI Working Paper Series; vol. 303, pp.1-33.
- Ryvkin D and Ortmann A (2008) The predictive power of three prominent tournament formats. Management Science; vol. 54, pp. 492-504.
- Searls DT. (1963) On the probability of winning with different tournament procedures. Journal of the American Statistical Association; vol. 58, pp. 1064-1081.
- Scarf P and Bilbao M. (2006) The optimal design of sporting contests. Salford Business School Working Paper Series. Paper no. 320/06, pp. 1-17.
- Scarf P, Yusuf MM and Bilbao (2009) A numerical study of designs for sporting contests. European Journal of Operational Research; vol. 198 pp. 190-198.

Using Forecasting to Detect Corruption in International Football¹

J. James Reade* and Sachiko Akie**

* University of Birmingham, The Johns Hopkins University, SAIS Bologna Center, j.j.reade@bham.ac.uk

** Akita International University

Abstract. Corruption is hidden action aimed at influencing the outcome of an event away from its competitive outcome. It is likely common in all walks of life yet its hidden nature makes it difficult to detect, while its distortionary influence on resource allocation ensures the importance of trying to detect it both practically and economically. This paper further develops methods to detect corrupt activity contained in Olmo et al. (2011) and Reade (2013) that make use of different forecasting methods and their information sets to detect corruption. We collect data from 63 bookmakers covering over 9,000 international football matches since 2004 and assess a claim made in early 2013 by *Europol* that the outcomes of almost 300 international matches since 2009 were fixed. Our collected data consists of match outcomes and pre-match bookmaker odds, which we use to explore the divergence between two kinds of forecasts of match outcomes: those by bookmakers, and those constructed by econometric models. We argue that in the absence of corrupt activity to fix outcomes these two forecasts should be indistinguishable as they are based on the same information sets, and hence any divergence between the two may be indicative of corrupt activity to fix matches. Such an assertion is conditional on the quality of the econometric model and in this paper we discuss the peculiarities of modelling international football match outcomes. In the absence of corroborating evidence we cannot declare any evidence procured in our manner as conclusive regarding the existence or otherwise of corruption, but nonetheless we argue that is it indicative. We conclude that there is mild evidence regarding potentially corrupt outcomes, and we also point towards yet more advanced strategies for its detection.

1 Introduction

Corruption is hidden action aimed at influencing the outcome of an event away from its competitive outcome. It likely occurs in all walks of life yet its hidden nature makes it difficult to detect, while its distortionary influence on resource allocation ensures the importance of trying to detect it both practically and economically. Practically, resources are diverted from participants in the events to those seeking to influence them. As those seeking to influence them are doing so for financial gain, this falls under the purview of fraud, since those fixing matches gain most through keeping information regarding the fix as private as possible in order to place bets on the fixed outcomes. In the context we consider, namely a sports league, the uncertainty of outcome is a particularly valued aspect of the output being produced; the *uncertainty of outcome* hypothesis of Rottenberg (1956) relates this uncertainty to the revenues generated by a sports league, and hence attempts to reduce this uncertainty must be harmful and thus there is an economic interest to ensuring corruption is detected.

This paper further develops methods to detect corrupt activity contained in Olmo et al. (2011) and Reade (2013) that make use of different forecasting methods with different information sets to detect corruption. We collect data from over 9,000 international football matches since 2004 and consider one specific recent episode of alleged match fixing. We assess a claim made in early 2013 by *Europol* that the outcomes of almost 300 matches between 2009 and 2012 were fixed.²

We note the practical relevance of our work. In response to widening concerns that corruption is harming the commercial interests of sport, various governing bodies are devoting ever increasing resources towards its detection. For example, Betfair has signed agreements with multiple governing bodies to share information on suspicious trading patterns detected on its markets, and employs a team of analysts to detect such market movements.³ It is our hope that the method of comparing information sets contained within this paper can be of use in such detection attempts.

In Section 2 we review the existing literature on forensic economics, in Section 3 we introduce our data on bookmaker odds from over 9,000 football matches, and in Section 4 we describe our econometric method, assessing in particular the forecast performance of both the econometric model developed and bookmakers recorded, and carry it out, presenting our results along the way. Section 5 concludes.

¹This is a shortened version of a longer paper which is available at <http://goo.gl/ZIWU>.

²See Harris (2013) and Hill (2013) regarding this. A press conference by *Europol* released a mix of new and old information regarding many matches known to have been fixed in recent years throughout Europe. It was later clarified that of the 700 matches mentioned, 300 were new, and 90% of these new matches were international matches.

³See ‘Anti-corruption: Technology key to catching fixers’ *Financial Times*, 16 June 2011, (last accessed 24 April 2013, <http://goo.gl/P0kTc>).

2 Using Forecasting to Detect Corruption

The field of forensic economics is expanding rapidly; as Zitzewitz (2012) notes, the aim of this field is “uncovering evidence of hidden behaviour in a variety of domains”, and already in a short number of years insights gained from economic theory regarding hidden action have facilitated empirical investigations in a wide range of areas.

Particularly relevant in the case of sports corruption are papers by Price and Wolfers (2010), Wolfers (2006) and Reade (2013). The first two consider hidden action in basketball on the part of referees and teams, and the latter considers Italian soccer and a recent match fixing scandal there. In all cases, economic theory is brought to bear to determine potentially effective channels upon which to test for the presence of corruption. Preston and Szymanski (2000) analyse the economic theory behind cheating in sport, paying particular attention to the subjective decision making process of the sports participants considering corrupt activity, borrowing from Becker (1968). They note that corrupt activity must alter the *objective probability* of any particular outcome of the sporting contest, and that the likelihood of such activity varies depending on the renumeration of participants, the importance of the individual match taking place, the likelihood of punishment and the severity of punishment.

In Reade (2013), the strategy chosen is to make use of public information available via bookmakers on football matches. Specifically, in the absence of any systematic method to influence matches (which must be private information for some subset of agents involved in a match), the forecast of a match outcome (which we can assume has true probability p_t) by bookmakers, $\hat{p}_{B,t}$ ought to be indistinguishable from that of an econometric model, \hat{p}_t , suitably specified. This assertion is based on the idea that most relevant information for predicting the outcome of a match is observed: the strengths of teams are observed via previous matches. These information sets are common to both econometric models and bookmakers in forming predictions.

In the presence of corrupt activity to fix the outcome of a football match, the objective probabilities of match outcomes are altered, say to $p_t^* = p_t + q_t$, and it may thus be that information sets differ between bookmakers and econometric methods. Hence we should expect to observe a significant difference between the bookmaker price and the econometric model price in the presence of corrupt activity due to the *difference in information sets*. From forecasting theory, basing forecasts based on larger information sets must yield an improvement, although Hendry (2011) notes this is only in the variance rather than bias. We anticipate given the nature of the forecasts we study and the insight of Preston and Szymanski (2000) that corrupt activity significantly changes probabilities of outcomes that forecasts based on subject information will be biased.

3 Data

Our dataset consists of all international football matches listed on the betting odds website www.OddsPortal.com.⁴ These matches are categorised into various regional competitions for national teams (e.g. European Championships, Asian Cup), and global events such as friendlies and the World Cup. Furthermore, most nations will have both mens and womens' teams and also youth teams (for those under the ages of 17, 19, 20 and 21 most commonly). In total, since 2004 we have 9,567 matches involving 35 different tournaments plus friendlies.⁵ Of those matches, around 32% involve youth teams and 12% involve womens' teams.⁶ Overall we have matches involving 915 teams from around 212 national teams from around the world.⁷ While it might be *a priori* anticipated that the fixed matches identified by Europol are all mens senior matches, this is not established, and hence it makes sense to consider all international matches rather than simply restrict ourselves to mens senior matches. The salaries paid to youth players are often dramatically less than for senior players, and a significant gender pay gap undoubtedly exists in football, ensuring that considering such matches provides variation along one dimension identified by Preston and Szymanski (2000) as contributing towards the corruption decision by a sports participant.⁸

⁴Information from OddsPortal.com was scraped using the Python programming language over 16–17th February 2013.

⁵See Table 6 in the Appendix of the longer paper for a breakdown of the tournaments and the number of matches in each tournament.

⁶And around 4% are womens' youth tournaments.

⁷Although there are only 193 members of the United Nations, a number of non-sovereign states have teams that participate in national championships, such as Wales, Scotland and Northern Ireland, as well as particular regions of other countries such as the Basque Country in Spain. See http://en.wikipedia.org/wiki/List_of_FIFA_country_codes for a list of FIFA members and non-member national teams..

⁸See ‘England women footballers secure central contract increase’, *BBC Sport*, 15 January 2013 (last accessed 24 April 2013, <http://goo.gl/wxJDA>) on the gender pay gap, and ‘Survey reveals footballers’ wages’, *BBC Sport*, 11 April 2006 (last accessed 24 April 2013, <http://goo.gl/99M3S>) on youth salaries.

For each match we have, on average, 24.7 bookmaker prices, with a standard deviation of 15.6 bookmakers, a maximum of 63 and a minimum of 1 bookmaker.⁹ Over all our matches around the world, we have bookmaker prices from 63 different bookmakers, all of whom are listed in the Appendix of the longer paper with the relative frequencies with which they appear in our dataset. We simply take the match outcome probabilities from www.OddsPortal.com, but usually many other types of bets exist. The choice of only match outcome prices is rather arbitrary and it is more than likely that those seeking to fix outcomes attempt to fix particular aspects of a match rather than necessarily its outcome indicating that it may be important to collect prices on other match outcomes in order to further detect corrupt activity.

4 Methodology and Results

All datafiles used in the regression models, and all codes files are available online.¹⁰

4.1 Modelling International Football Matches

International football consists of matches between national teams, rather than between club teams based on particular countries, and such matches are thus often high profile and prestigious.¹¹ As such, the allegation that such a considerable number of these matches have been fixed in recent years is important. It seems more than likely, given the insights of Preston and Szymanski (2000), that international matches are a target for corrupt activity. For example, at international level a large number of friendlies are played, upon which little rests for each team involved. Additionally, in many qualification tournaments because only the top one or two teams in a group can qualify, a large number of less meaningful matches occur between teams unlikely to qualify.

Turning to the modelling of international football match outcomes using econometric methods, while domestic football is organised into leagues of teams of similar strength, and this league structure dominates, international football has no such league distinction. National teams are composed of players qualified to play for that country (either by birth or by transferring nationality), whereas domestic football teams can be composed, in principle, of players from any country. National teams also play much more infrequently; only thirteen nations play more than 100 matches in our sample covering nine years, showing that on average national teams play at most on average 14 times per year whereas domestic football teams will play in the region of 30–60 matches per calendar year. Domestic leagues enable a simple way of assessing team strength from an econometric point of view: Each team's performance in that league. The closest in international football to a league is the qualification stages of World Cup and regional championships such as the European Championships, however even these stages are seeded such that the better teams have a greater likelihood of qualification for the latter stages meaning that teams of vastly differing qualities can meet in such mini-leagues. Indeed, McHale and Scarf (2006) make particular reference to this phenomena when studying international soccer matches relative to domestic ones. International matches also differ from club matches in the number of friendly matches that occur — something mentioned earlier in the context of match fixing.

The consequence of this lack of a common framework upon which to judge national teams (no single league, and a high variance of opposition quality) is that some other method is required to rank teams in order to construct statistical or econometric predictions. Fifa rankings could be used to attempt to approximate team quality and thus predict outcome, yet Fifa's rankings are but one attempted measure of team quality and hence have their critics, and furthermore would require additional data collection and matching with actual results.¹² An alternative is to make use of the Elo ranking system devised specifically for chess but adapted for numerous other sports. This ranking system updates for each match, affords the ability to relatively weight different types of matches differently, and provides a simple way to generate predicted outcomes for matches. We make use of Elo rankings to create a variable with which we use to help predict match outcomes. In the academic literature Hvattum and Arntzen (2010) test Elo ratings against bookmakers and econometric models as a forecast tool for English Premier League football matches, finding that bookmakers outperform Elo ratings, but Elo ratings are superior to econometric models, while Leitner et al. (2010) use Elo ratings amongst other methods when attempting to forecast outcomes from the 2008 European Championships football tournament. As Elo ratings are but one additional method for measuring team quality, it is

⁹Our dataset does include matches in which bookmakers declined to offer prices, or in which they withdrew prices.

¹⁰The workpage for our corruption research can be found at <http://goo.gl/cNPwt>.

¹¹In our sample of international friendlies, a handful of club teams do appear as occasionally higher profile club teams will play friendly matches against national teams. We do not omit these matches since they help provide information on the strength of a national team.

¹²On the criticism of Fifa rankings, see Stefani and Pollard (2007) and http://en.wikipedia.org/wiki/FIFA_World_Rankings#Criticism.

helpful to get some idea about how effectively they do this. In Appendix B in the longer paper we carry out a small simulation study to investigate the properties of the measure. We find that while some biases do exist, these are all away from the mean, implying that some Elo predictions may underestimate the true quality differences between teams. Such biases can be corrected by incorporating Elo predictions into a regression method, as we do.

Additionally, we use a variant of Elo rankings that the *World Football Elo Ratings* (WFEL) employ, which give different weights to different matches.¹³ The reason for this is to capture the idea that competitive matches reveal more about the actual quality of a team than friendly matches.

On econometric modelling, Goddard (2005) considers the two most common econometric methods for modelling and forecasting football match outcomes, notably Poisson methods to predict goal arrival, and direct limited-dependent variable models for actual match outcome, finding that the differences between the two methods are marginal. Forrest et al. (2005) carry out a direct comparison of econometric methods and bookmaker forecasts and find that bookmakers tend to forecast better, something they attribute to greater competition in the betting industry in recent years; our dataset exclusively falls in the more recent period of increased competition amongst bookmakers, as evidenced by the number of bookmakers (63 in total, on average more than 20 per match) we have prices from.

We seek to understand the outcome of a match at time t between team i and team j :

$$y_{ijt} = \begin{cases} 0 & \text{if team } j \text{ wins match at time } t, \\ 0.5 & \text{if match drawn,} \\ 1 & \text{if team } i \text{ wins.} \end{cases} \quad (1)$$

From (1), match outcome is a discrete variable with three possible outcomes. One standard way to model a variable such as y_{ijt} is to assume there exists a continuous latent variable y_{ijt}^* which, if observed in particular regions implies different outcomes for the observed match outcome variable. We thus write:

$$\mathbb{P}(y_{ijt} = 0 | \mathbf{X}) = \Phi(y_{ijt}^* < \mu_1), \quad (2)$$

$$\mathbb{P}(y_{ijt} = 0.5 | \mathbf{X}) = \Phi(\mu_1 < y_{ijt}^* < \mu_2), \quad (3)$$

$$\mathbb{P}(y_{ijt} = 1 | \mathbf{X}) = \Phi(y_{ijt}^* > \mu_2). \quad (4)$$

The parameters μ_1 and μ_2 are described as the cut-off points — the points in the distribution of y_{ijt}^* where the outcome switches from one of the possibilities to others. Hence below μ_1 , the observed outcome y_{ijt} is that team j wins the match (it is listed as an ‘away’ win), while between μ_1 and μ_2 , the match ends in a draw ($y_{ijt} = 0.5$), and above μ_2 , team i wins.

We estimate this latent variable y_{ijt}^* using an ordered probit regression model:

$$y_{ijt}^* = \beta_0 + \mathbf{X}_{ijt}\beta + e_{ijt}, \quad e_{ijt} | X_{ijt} \sim N(0, 1). \quad (5)$$

In (5) the variable \mathbf{X}_{ijt} contains explanatory variables for match outcome and in our case includes the relative difference in Elo ratings for the two national teams involved in any given match along with other variables that help describe the historical strength of the two teams involved. Such variables can be generated from each team’s historical results; information on match outcomes, ability to score goals and to prevent their concession can all be marshalled into explanatory variables for predicting match outcomes. In international matches, the venue in which the match is played can be important also, and hence we control for matches on neutral territory.

We then use our ordered probit model (5) to generate fitted probabilities for each of the possible events, which we as $\hat{p}_{E,ijt} \equiv \hat{P}_{ijt}$. We estimate (5) using data up to the end of 2009, and then use forecasts of all matches that take place from 2010 onwards in order to compare these to bookmaker forecasts.

Considering the ordered probit model for predicting match outcomes, the regression output is provided in Table 1.¹⁴ The explanatory variables are the difference between the Elo expected outcomes for the two teams competing, alongside a number of other readily calculable statistics from previous match outcomes. We include variables for recent performance (wins/draws gained, goals scored, goals conceded) and experience (in sample). While the Elo rating difference is significant, and reflects the impact the difference in quality has on outcome (negative Elo means team 2 is stronger and the dependent variable is zero if team 2 wins), it is not the most significant variable. Instead, the notional ‘points’ gained by each team per game (three points for a win, one for a draw mirroring the almost

¹³See <http://www.eloratings.net/system.html> and http://en.wikipedia.org/wiki/World_Football_Elo_Ratings for more information.

¹⁴See Table 9 in the longer paper for a description of each variable in this regression output.

universal domestic football league scoring system) is much more significant in explaining match outcomes. The difference between the Elo rating difference and the points gained by a team per game is that the latter does not adjust for opposition quality, while the former does. The significance of Elo ratings is consistent with Hvattum and Arntzen (2010) who suggest Elo ratings provide an improvement over econometric methods.

When using the output of our econometric model to compare to bookmaker prices later we use forecasts from our regression model rather than fitted values. If we compare fitted values from a regression model estimated over our entire sample from 2004 through to 2013, this would yield an inaccurate test since the econometric model would make use of information after the match in question had taken place. In that case, unusual outcomes would be already factored into the fitted values and hence we may be less likely to spot such distinct outcomes. Hence, given that our focus is the most recent three years of international matches, we estimate our model up to the end of 2009 and then forecast all matches in 2010–13 making use of data available before each match. The only notable change from using fitted values over the entire sample is that rather than our parameters be estimated on data up to 2013, they are estimated on data up to 2009; data up until the start of each match (Elo rating, recent form, etc) is still used to construct forecasts. Assuming that the true process determining the outcome of football matches is stationary, as might be expected, then estimation up to 2009 (which still allows us to estimate over 2,662 matches) should not yield significant differences from estimating up until 2013.¹⁵ In order that the Elo rating for each team is better calibrated hence more informative, we only regress on matches for which each team has already played a minimum of four matches.

4.2 Bookmaker Prices

We anticipate that bookmaker prices will reflect the presence of corrupt activity in football matches, and hence it will be important to consider the various dimensions in which this might materialise. Figure 1 gives some idea of the spread of implied probabilities for each of the three events. In our sample of 9,606 international matches since 2004, 48.4% result in wins for team1 (the first team listed, which usually is the home team apart from during tournaments played at neutral venues), 20.1% result in a draw, with 31.75% ending in a win for team2 (the second team listed, usually the away team). The distribution for the draw is much more concentrated on the lower range of the interval, with the 99th percentile falling at 30.3% compared to 89% and 87% for team1 and team2 victories respectively.

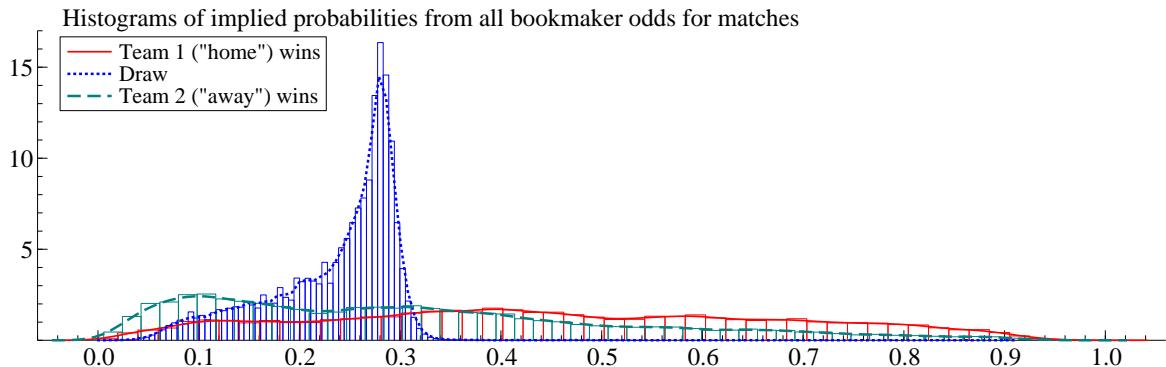


Figure 1: Histograms and estimates of empirical probability distribution of all bookmaker prices for the three events surrounding each football match. Bookmaker prices are corrected for the over-round.

Before conducting any characterisation, it is important to correct for the well-known favourite-longshot bias (FLB) in betting markets, whereby favourites win more often than their odds imply, and outsiders (longshots) win less often than their odds imply. Such bias is often corrected for using linear regression methods. Regressing the outcome, say o_{it} , on the implied probability of bookmaker i 's prices for the match at time t , $p_{B,it}$, with a constant:¹⁶

$$o_{it} = \alpha_o + \beta_o p_{B,it} + u_{it}, \quad (6)$$

¹⁵This is of course something we can check by comparing the regression model for 2004–2009 to one estimated on 2004–2013. We do this and find minimal differences.

¹⁶With a slight abuse of notation for o_{it} , since this does not vary over i .

	(1)
	outcome
outcome	
ea_team_diff	-0.970*** (-8.418)
finals	-0.260* (-2.540)
pts_last_31	-0.093*** (-4.779)
pts_last_32	0.063** (3.219)
gdiff_last31	0.042** (3.278)
gdiff_last32	-0.020 (-1.509)
experience1	0.008** (2.918)
experience2	-0.010*** (-3.420)
pts_pg1	1.674*** (19.511)
pts_pg2	-1.783*** (-20.702)
days_since_last_match1	-0.001 (-1.708)
days_since_last_match2	0.000 (0.504)
cut1	
_cons	-0.999*** (-6.239)
cut2	
_cons	-0.195 (-1.221)
<i>N</i>	2662

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 1: Ordered probit regression model output for international match outcome.

Outcome	Bookmakers	Model
Team 1	.163	.162
Draw	.159	.153
Team 2	.185	.179

Table 2: Brier scores for econometric model, calculated over the 5,695 matches in our dataset taking place after 2009, compared to bookmaker forecasts.

yields fitted values $\hat{p}_{B,it} \equiv \hat{o}_{it} = \hat{\alpha} - \hat{\beta}_o p_{B,t}$ which are the bookmaker odds $p_{B,it}$ corrected for their observed bias. The results of this correction are presented in Table 2 in the longer paper and show that indeed FLB was present in our bookmaker prices.

4.3 Forecast Comparisons

4.3.1 Brier Score

One measure of the forecast performance of a forecast model in the context of success/failure events like a win, draw or loss is the Brier score, which takes the corrected forecast probability from each forecast model and compares it to the outcome variable (see (1)):

$$\text{Brier} = \frac{1}{M} \sum_{m=1}^M (\hat{p}_{m,it} - y_{m,it})^2. \quad (7)$$

Table 2 reports Brier scores for both our econometric model and bookmakers, allowing us to compare the performance of the two methods. The scores indicate that on average, both bookmakers and our model were out by about 40 percentage points in their forecasts.

Our econometric model performs indistinguishably differently from the bookmakers when predicting either positive match outcome, and slightly better for the draw.¹⁷

4.3.2 Comparison of Forecast Differences

While the Brier score yields information on the relative quality of forecasts in general via taking averages, our main focus is on the differences between forecasts in matches we might suspect of corrupt activity, and as such we now consider methods to assess the differences between the forecasts.

Using corrected bookmaker odds $\hat{p}_{B,it}$, we can compare these to forecasts generated from our econometric model (5). Our interest is in the divergence between the two and hence we run the regression model:

$$\hat{p}_{B,it} = \alpha_p + \beta_p \hat{p}_{E,it} + \varepsilon_{it}, \quad (8)$$

and firstly consider the nature of the estimators $\hat{\alpha}_p$ and $\hat{\beta}_p$ before investigating the residuals. We theorise that in the absence of corrupt activity, $\hat{p}_{B,it} = \hat{p}_{E,it}$ and hence the residuals $\hat{\varepsilon}_{it}$ ought to be symmetrically distributed around their mean of zero, and we might anticipate $\alpha_p = \beta_p - 1 = 0$. However, the existence of minor biases in bookmaker and econometric forecasts when observed over particular dimensions (such as the draw) is such that we may observe variations in predicted probabilities, despite similar overall predictive performance, and hence to maximise the flexibility of our approach we do not require $\alpha_p = 1 - \beta_p = 0$. While the regression method ensures that $\hat{\varepsilon}_{it}$ are mean zero, nonetheless the existence of mass in either tail of the distribution may be indicative of suspicious activity, and hence we investigate the residual distribution from our models (for each team's win and the draw markets).

Table 3 presents the results of the regression model (8) for the three possible match outcomes. The regression model is carried out using every single bookmaker probability for a given match, hence we have over 160,000 observations.

¹⁷We conduct matches t-tests for the difference in the two numbers (average squared errors), and find that for the home (team 1) win, the t-statistic is 0.64 ($p = 0.64$) and for the away (team 2) win, the t-statistic is 1.4 ($p = 0.162$), while for the draw the t-statistic is 3.38 ($p = 0.001$). Nonetheless all these differences are all only at the third decimal place; see Table 2.

	(1)	(2)	(3)
	Draw	Team 1 wins	Team 2 wins
Model Probability	0.716*** (236.673)	0.694*** (348.189)	0.694*** (348.825)
_cons	0.080*** (122.860)	0.136*** (125.846)	0.085*** (108.924)
N	166902	166902	166902

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Regressions of bookmaker probabilities on model probabilities in forecast period post 2009.

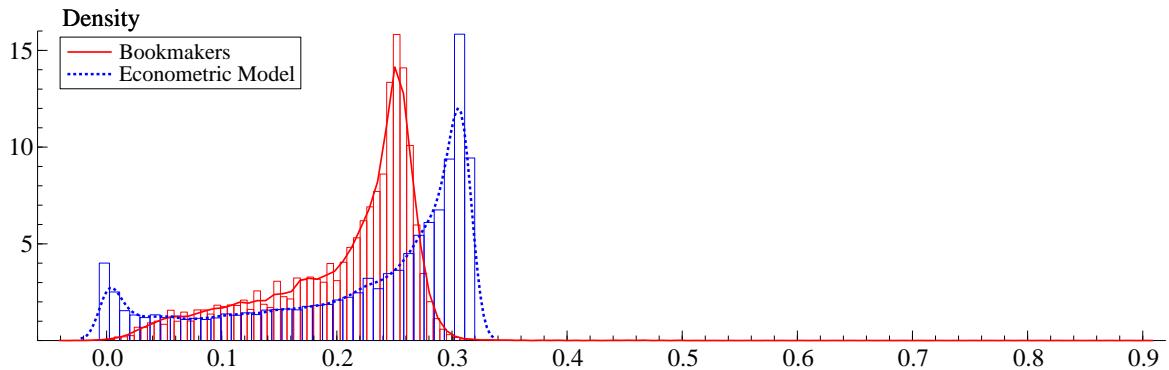


Figure 2: Plot of both bookmaker predictions for match outcomes (corrected), and econometric model predictions.

We focus on the draw, as Reade (2013) did, as opposed to either positive result. This is because often the draw is considered to be something of a ‘residual’ event, since most focus is on whether or not each team will win — and naturally, each team in a match sets out to win it in the absence of corrupt activity. As a result, the distribution of bookmaker prices (and as we will see, the predictions of our econometric model), do not venture above about a third, reflecting this residual nature. Furthermore, a draw is something more of a collaborative outcome since each team gains from it, not necessarily only in prestige terms but also in points terms in competitive matches, whereas either positive result is much more non-cooperative. As such, it seems more likely that distinct patterns in the draw market may be indicative of some kind of collusive activity to ensure such an outcome. Practically, since the implied forecasts extremely infrequently move above a third, whereas the forecasts for either positive result can and do often reach much higher levels, this also enables the spotting of unusual patterns. Thus we focus on the draw outcome.

Figure 2 provides a comparison of the distribution of forecasts for our two models for the draw outcome. The red line is the bookmaker forecasts, and the blue line is our econometric model forecasts. Figure 2 helps explain the regression results for the draw; the coefficient of 0.7 corrects for the wider dispersion of econometric model forecasts relative to bookmakers, and hence any large residual we observe in our model controls for this pattern.

In determining a large outlier indicative of a difference in information sets and hence potential corrupt activity, we use residuals larger than three standard deviations. Assuming a normal distribution, such an observation ought to be observed 0.27% of the time, and hence we might expect to see around 420 in our forecast sample of 167,916 bookmaker prices. For the draw outcome, we actually observe 1,623 such observations from 210 different matches. It should be noted that (8) includes generated regressors on both the left- and right-hand side of the regression equation, creating potential distortions for standard errors and the standard deviation of the residuals used to determine a large outlier. (Wooldridge, 2002, Ch. 6) notes that provided standard OLS assumptions hold, there is no impact on bias and consistency properties for estimators, but for standard errors the sampling variation induced by the first stage

regressions can cause problems. Lewis and Linzer (2005) suggest that when a dependent variable is generated that heteroskedasticity robust standard errors can be used to alleviate distortions. To limit the possibility that generated regressors influence the likelihood of a bookmaker price being isolated as an outlier, we report information regarding the nature of relatively large residuals based also simply on residuals, rather than a binary variable representing ‘large residuals’.

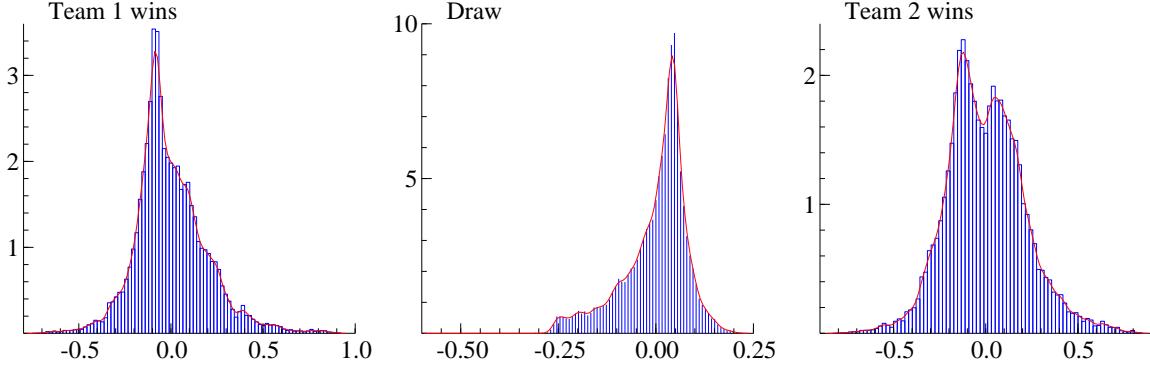


Figure 3: Histogram of residuals from the team 1, draw and team 2 win regressions of model probabilities on bookmaker probabilities.

A graphical inspection of the residuals we seek to investigate will be helpful; Figure 3 provides this for all three match outcomes. Ideally, all three distributions will be unimodal, symmetric and centred on zero — this would reflect that any departures between our econometric model and bookmakers are random. From Figure 3 however it is clear that all three distributions, but most notably the draw distribution, appear to be non-symmetric and centred away from zero. The draw distribution is particularly distinct since the mode of the distribution is positive suggesting that most often bookmakers underprice the draw relative to our econometric model, but nonetheless 40% of the distribution is negative, where bookmakers predict a draw with higher probability than our model. Indeed the largest residuals are negative and are at least three standard deviations away from the mean (and four from the mode), and this would be consistent with the observed pattern of corruption in Serie B, where matches are fixed to be draws and the implied draw probability on betting exchanges and at bookmakers reaches disproportionately high levels.

As already mentioned, the largest residuals are five standard deviations away from the mean, and hence we now investigate large residuals. It is of interest to consider the nature of the games with the largest residuals — do they happen to coincide with games that matter the least? It is somewhat more tricky to ascertain the importance of matches in the international sphere relative to those in domestic competition since the qualification process for major continental and global tournaments varies by continent with some areas (e.g. Europe) having small groups of teams competing, whilst others have much larger groups of teams competing against each other. Nonetheless, a fairly clear distinction is between friendlies and competitive matches. Of the 210 matches we observe large negative residuals for, 123 are from friendly matches (this is insignificantly larger than the frequency of friendly matches in our overall dataset). A full list of the large-residual matches is given in Table 11 of the longer paper.

Considering the teams involved in these friendly matches, many are either youth teams (under 21s, under 19s, under 17s etc) or womens teams. It might thus be hypothesised that teams for which limited data exists are those for which we find large residuals; perhaps as a result of such low numbers of observations, a low draw probability is forecast by our econometric model. However, if the match in question is the first in the sample for both teams in a match, then both teams have equal strength from their Elo score, and the draw will have a forecast of at most 31.22% since this is the largest draw probability recorded by the econometric model over our forecast sample. Hence we cannot conclude that because teams appear infrequently in our sample we might observe strange results.

Considering also the bookmakers involved, we find that a number of bookmakers appear more often in our large-residual matches relative to their occurrence in our overall sample (Table 5 in the longer version shows t-tests of differences in mean). The table reports on the bottom row (Overall frequency) the percentage of our observations that are from that bookmaker (e.g. 2.8% for 118bet), while the top row reports the increment for how often that bookmaker is observed in our large-residual matches (hence $2.8+3.1=5.9\%$ for 118bet), and the numbers beneath in parentheses are t-tests of the difference in means. Hence for all bookmakers bar bet365 in the table, they are

observed significantly more often in large-residual cases than overall in our sample. What is perhaps notable is that none of the major bookmakers appear in this list (see Table 8 in the longer paper for the frequencies with which all bookmakers are observed in our sample).

As a corollary, and also notable, is the average number of bookmakers reporting odds for these matches with large residuals. In our overall sample, on average a match has 25.1 bookmaker prices listed by *OddsPortal.com* (29.9 for matches after 2009), yet for these large-residual matches since 2009, there are on average just 7.7 bookmakers reporting prices. This in itself proves little, yet is again circumstantial as it might be expected that fewer bookmakers report prices on matches they suspect to be dubious in nature.

An additional aspect of our hypothesis is that private information becomes public through the betting markets and hence we might expect that in the cases where we identify large residuals, the majority of bookmakers for that match would report such unusual behaviour. Indeed we find that to be the case, as in 60% of matches with large residuals, more than half of the bookmakers report unusual odds, and around 30% of the time 80% or more bookmakers report large outliers. If we were picking up isolated cases, it would be expected that very infrequently would many bookmakers for the same match report odds inducing an outlier.

A final aspect of our large-residual matches that we consider are the youth and female composition of matches. As mentioned earlier, the range of pay across ages and the sexes in football is considerable, and hence it might be anticipated that youth and female matches are more likely to attract match fixers due to this. Suspect matches are significantly more likely to involve youth or female teams, which it might be argued are easier targets for fixers due to pay disparities. Specifically, almost two thirds of our sample are full international matches, yet only a third of our large outlier matches are full internationals, while only 10% of our sample are womens' matches yet they constitute 40% of our large residual matches. Both of these differences are statistically significant.¹⁸

It would be expected that the residuals for either positive outcome for the matches we identified using the draw are large, since the probabilities for all three events must sum to unity. Nonetheless, given that probabilities for either positive outcome much more readily span the unit interval (both have standard deviations more than three times that as for the draw), it seems less likely that such outcomes would necessarily attract particularly large residuals. It turns out that the residuals in either positive outcome are at least twice as large when the draw has been identified to have a large residual.

5 Conclusions

In this paper we have attempted to investigate the incidence of corrupt football matches using firstly economic reasoning based on information sets and secondly econometric methods, and specifically forecasting methods. We propose a test for suspicious patterns based on two types of forecasts built on different information sets: Those from bookmakers, and those from econometric models. In the absence of corrupt activity aimed at influencing the outcomes of football matches, the predictions of the two methods ought to be identical, conditional on the quality of the econometric model. If a match has been fixed then we might expect to see significant deviations between the two. In response to an allegation raised in early 2013 regarding international football matches over the period 2010–2013, we have investigated international football matches to see whether any suspicious patterns can be uncovered.

We find when looking at the probability of the draw in international football matches that 2,677 bookmaker prices across 210 matches have residuals of a sufficient size (more than three standard deviations away from the mean) to attract attention. We then analyse these matches, noting that there is a higher fraction of friendly matches, youth-team matches and womens matches amongst this group than the general population, and also noting that the number of bookmakers reporting prices on these matches is markedly lower than in the rest of our sample; two aspects which might attract additional attention. We also note that these large outlier bookmaker prices are clustered around a small number of matches rather than being spread amongst our sample of matches, and furthermore a statistically significantly smaller number of bookmakers report odds for these particular matches.

The evidence procured in this manner is naturally circumstantial and could be explained by legitimate factors. Nonetheless, an empirical method consistent with economic theory and hence the growing literature on forensic economics has been set out, and its potential power displayed.

¹⁸Furthermore, if we run regressions simply on residual size, we get statistically significant coefficients for women and youth matches suggesting that for these matches. However, it might be anticipated that residuals might be larger for such matches due to smaller general information sets for such matches.

References

- Becker, G. (1968), 'Crime and Punishment: An Economic Approach', *The Journal of Political Economy* **76**, 169–217.
- Forrest, D.K., J. Goddard and R. Simmons (2005), 'Odds-Setters As Forecasters: The Case of English Football', *International Journal of Forecasting* **21**, 551–564.
- Goddard, J. (2005), 'Regression Models for Forecasting Goals and Match Results in Association Football', *International Journal of Forecasting* **21**, 331–340.
- Harris, N. (2013), 'EXCLUSIVE: Top match-fix investigator reveals 'real story' about new cases', *sportingintelligence.com*.
URL: <http://www.sportingintelligence.com/2013/02/07/exclusive-top-match-fix-investigator-reveals-real-story-070201/>
- Hendry, David (2011), Unpredictability in Economic Analysis, Econometric Modelling and Forecasting, Economics Series Working Papers 551, University of Oxford, Department of Economics.
URL: <http://ideas.repec.org/p/oxf/wpaper/551.html>
- Hill, D. (2013), 'Another 'I Told You So' Moment', *Declan Hill's Blog*.
URL: <http://www.howtofixasoccergame.com/blog/?p=319>
- Hvattum, Lars Magnus and Halvard Arntzen (2010), 'Using elo ratings for match result prediction in association football', *International Journal of forecasting* **26**(3), 460–470.
- Leitner, Christoph, Achim Zeileis and Kurt Hornik (2010), 'Forecasting sports tournaments by ratings of (prob) abilities: A comparison for the euro 2008', *International Journal of Forecasting* **26**(3), 471–481.
- Lewis, J.B. and D.A. Linzer (2005), 'Estimating Regression Models in Which the Dependent Variable Is Based on Estimates', *Political Analysis* **13**(4), 345–364.
- McHale, I. and P. Scarf (2006), 'Forecasting International Soccer Match Results Using Bivariate Discrete Distributions', *Working Paper, Management and Management Sciences Research Institute, University of Salford*.
- Olmo, J., K. Pilbeam and Pouliot W. (2011), 'Detecting the Presence of Insider Trading via Structural Break Tests', *Journal of Banking and Finance* **35**, 2820–2828.
- Preston, I. and S. Szymanski (2000), 'Racial Discrimination in English Football', *Scottish Journal of Political Economy* **47**(4), 342–363.
- Price, J. and J. Wolfers (2010), 'Racial Discrimination Among NBA Referees', *Quarterly Journal of Economics* **4**, 1859–1887.
- Reade, J.J. (2013), Detecting corruption in football, in J.Goddard and P.Sloane, eds, 'Handbook on the Economics of Professional Football', Edward Elgar.
- Rottenberg, S. (1956), 'The Baseball Players' Labor Market', *The Journal of Political Economy* **64**(3), 242–258.
- Stefani, Ray and Richard Pollard (2007), 'Football Rating Systems for Top-Level Competition: a Critical Survey', *Journal of Quantitative Analysis in Sports* **3**(3), 1–20.
- Wolfers, J. (2006), 'Point Shaving: Corruption in NCAA Basketball', *The American Economic Review* **96**(2), 279–283.
- Wooldridge, J.M. (2002), *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, Cambridge, Mass.
- Zitzewitz, Eric (2012), 'Forensic economics', *Journal of Economic Literature* **50**(3), 731–69.

Integer Programming and Sports Rankings

Christian Raack*, Annie Raymond*, Thomas Schlechte*, and Axel Werner*

**Zuse-Institute Berlin (ZIB), Takustrasse 7, 14195 Berlin-Dahlem, Germany, {raack, raymond, schlechte, werner}@zib.de*

Abstract. Sports rankings are obtained by applying a system of rules to evaluate the performance of the participants in a competition. We consider rankings that result from assigning an ordinal rank to each competitor according to their performance. We develop an integer programming model for rankings that allows us to calculate the number of points needed to guarantee a team the i th position, as well as the minimum number of points that could yield the i th place. The model is very general and can thus be applied to many types of sports. We discuss examples coming from football (soccer), ice hockey, and Formula 1. We answer various questions and debunk a few myths along the way. Are 40 points enough to avoid relegation in the Bundesliga? Do 95 points guarantee the participation of a team in the NHL playoffs? Moreover, in the season restructuration currently under consideration in the NHL, will it be easier or harder to access the playoffs? Is it possible to win the Formula 1 World Championship without winning at least one race or without even climbing once on the podium? Finally, we observe that the optimal solutions of the aforementioned model are associated to extreme situations which are unlikely to happen. Thus, to get closer to realistic scenarios, we enhance the model by adding some constraints inferred from the results of the previous years.

1. Introduction

Sports fans are irrational creatures who swear eternal love to many idols and interpret the Bible of Statistics in whatever way they please. Their belief system is also filled with myths, some of which we will debunk by using a simple integer program.

For instance, it is widely believed among fans (and teams!) of the Bundesliga that a team that collects 40 points during a season is guaranteed to avoid relegation to the second league. Indeed, since the creation of the Bundesliga in 1963, no team that gathered at least 40 points (according to the modern point system which was introduced in 1995) was ever relegated. A similarly tenacious myth among fans of the NHL states that earning 95 points during the regular season guarantees a team a spot in the playoffs.

We will see that these beliefs, although well-established from previous experiences, do not hold strictly. Indeed, there have been counterexamples in minor leagues in Germany and even in the NHL itself, and it is quite easy to come up with theoretical seasons for the respective leagues in which each of these myths is thoroughly shattered. Still the question remains if there are such theoretical point bounds and, if so, what then really is the number of points necessary in the worst case. We will provide means to answer these questions using a general integer program (IP) and by offering ways to adapt it to different problems. This program also allows us to answer similar questions in other sports; we give an example of an application to Formula 1.

The most prominent sports problems are the traveling tournament problem and the referee assignment problem which were tackled by local search, integer programming, constraint programming, and tailor-made heuristic approaches. An annotated bibliography for sports scheduling over the past 40 years is provided by Kendall et al. [5]. Many articles ([2], [3], [8], [9]) also try to answer the question of whether or not, at some point in the season, some team has a chance to win (or move on) in a competition, and Gusfield and Martel [4] as well as Kern and Paulusma [6] discuss

the complexity of this problem. For example, Russell and van Beek [9] use constraint programming to determine the number of games needed to guarantee a playoff spot in the NHL at any point in the season. They do mention very briefly that for the specific schedule of 2006/2007, the teams of Toronto and Pittsburgh need 145 points at the beginning of the season to ensure a playoff spot; however, they do not discuss the maximum number of points that might be needed in general for any schedule and any team. To the best of the authors' knowledge, there are no published articles about a general ranking integer programming model that calculates the number of points needed to finish i th in any sport.

In the following Section 2, we make an introductory example from the Bundesliga precise and straighten out the 40-point rule. We also introduce the general IP model. In Section 3, we adapt this model to the setting of Formula 1 and provide some surprising results about what is possible in a racing season. Applying the model to the NHL is much more intricate and we discuss this in Section 4; here we also consider the restructuration currently under discussion in the NHL. Incorporating "experience" constraints that reflect what usually happens during a season rules out the most extreme scenarios and gives bounds that might be closer to reality. Finally, Section 5 gives a short summary of the results. All computed results mentioned throughout the paper, as well as the model files, can be obtained from a website, see Section 5.

2. The Bundesliga and the General Model

Many German football (soccer) fans believe that if a team in the Bundesliga earns 40 points, then it will not be relegated to the second league. The Bundesliga has the following structure: It is composed of 18 teams which each play against each other team twice in the season, once at home and once away. A win earns a team three points, and a tie, one point. The two worst teams of the Bundesliga are relegated to the second league, whereas the top two teams of the second league are promoted to the Bundesliga. The third worst team of the Bundesliga and the third best team of the second league affront each other in an extra game; the winner plays its next season in the Bundesliga, and the loser, in the second league. Thus the myth among the fans says that 40 points guarantee a team at least the fifteenth place. We can easily check the validity of this statement by formulating an IP.

Let $T := \{1, \dots, 18\}$ be the set of teams and G be the set of games $g = (t, t', n)$, where n denotes whether g is the first or second game between teams $t, t' \in T$. Then we can formulate the following integer program:

$$\begin{aligned} \max \quad & p_{16} \\ \text{s.t.} \quad & x_g^0 + x_g^1 + x_g^2 = 1 \quad \forall g \in G \\ & \sum_{\substack{g \in G: \\ g=(t,t',n)}} (3x_g^1 + x_g^0) + \sum_{\substack{g \in G: \\ g=(t',t,n)}} (3x_g^2 + x_g^0) = p_t \quad \forall t \in T \\ & p_{t+1} \leq p_t \quad \forall t \in T \setminus \{18\} \\ & x_g^i \in \{0, 1\} \quad \forall g \in G, i \in \{0, 1, 2\} \end{aligned}$$

Here, $x_g^i = 1$ for game $g = (t, t', n)$ if $i = 0$ and g is a tie, if $i = 1$ and t wins g or if $i = 2$ and t' wins g ; otherwise, $x_g^i = 0$. The first equation states that game $g = (t, t', n)$ ends either with the victory of t or t' or with a tie. Moreover, the second equation counts p_t , the number of points earned by team t . Note that, by the third inequality, the teams are labeled from 1 to 18 in order of

their rank at the end of the season. Finally, by maximizing p_{16} , we find out the maximum number of points that a team can earn and still be relegated.

We used the modeling language ZIMPL 3.3.0 (see [7]) and SCIP 3.0.0, a non-commercial mixed integer programming solver (see [1]) to solve the problem. The optimal solution tells us that a team may earn 57 points and be sixteenth in the league. Indeed, if the top sixteen teams win all of their home games as well as their two away games against the two bottom teams and lose their other games, then each of the first sixteen teams wins 19 games, loses 17 and thus ends the season with 57 points (tie-breakers determine which of these sixteen teams is actually sixteenth). Therefore, to guarantee a fifteenth place, a team needs to earn 58 points, not the widely believed 40.

2.1 A General Model

The preceding problem was fairly simple and could have been solved without using an integer program, however it allowed us to introduce the model we will be using. The general model is as follows:

$$\begin{aligned} \sum_{o \in O} x_g^o &= 1 \quad \forall g \in G \\ \sum_{\substack{g \in G: o \in O \\ t \in g}} s_{t,g}^o x_g^o &= p_t \quad \forall t \in T \\ p_{t+1} &\leq p_t \quad \forall t \in T \setminus \{|T|\} \\ x_g^o &\in \{0, 1\} \quad \forall g \in G, o \in O, \end{aligned}$$

where O is the set of possible outcomes o of a game, and $s_{t,g}^o$ is the number of points given to team $t \in T$ for achieving outcome o in game g . This general model applies to a variety of sports and games; it can be applied in an analogous fashion to any European football league, for instance. We can maximize or minimize the number of points earned by the i th-ranking team for any $i \in \{1, \dots, |T|\}$, answering similar questions such as what is the minimal number of points necessary to qualify for the Champions League or the Europa League in any given National League.

For other sports, depending on the structure of the teams and of the games, some adjustments may be necessary to make the model work; we present some examples in the following sections.

3. Formula 1

The general model can be applied to Formula 1 even though the structure of this sport is very different than that of the Bundesliga. In the history of Formula 1, the scoring system has changed over and over again. In 1994, when Michael Schumacher won his first Formula 1 World Championship, fourteen teams and thus 28 drivers competed in sixteen races. During that year, some drivers changed and thus the overall number of drivers was actually 46. However, for our theoretical analysis, we assume a minimal and constant set of drivers over a season. Points are awarded to the six fastest drivers of a race: 10 to the winner, 6 to the second, 4 to the third, 3 to the fourth, 2 to the fifth and 1 to the sixth.

Applying the model blindly, we let T be the set of drivers, G be the set of races, O be the $|T|!$ different ways of ordering the drivers and $s_{t,g}^o$ be the number of points awarded to driver t in race g for achieving the rank it holds in o . So the model applies, but the number of variables is very

large; even if we realize that we only need to consider the order of the six top drivers, O still has order $\binom{|T|}{6} \cdot 6!$, which is a burdensome 271252800 for $|T| = 28$.

By modifying a bit our way of thinking, we may apply the model in a different way and reduce its size. Let T be the set of drivers, G be the set of games $g = (r, i)$, which consist in finishing i th in race r for $1 \leq i \leq 6$, and $O = T$. In other words, x_g^o with $g = (r, i)$ is 1 if driver o finishes i th in race r and 0 otherwise. Thus the first equation of the model states that exactly one driver finishes i th in race r . The other two equations act as previously explained. However, the model is incomplete; nothing limits the number of positions that a driver can hold in one race. Indeed, with this model, the same driver could finish first, second, third, fourth, fifth and sixth. We must thus add the following constraint

$$\sum_{\substack{g \in G: \\ r \in g}} x_g^o \leq 1 \quad \forall o \in O, r \in R$$

where R is the set of races. This smaller model solves much faster and can be used to compute the optimal solution with respect to different objective functions. For example, $\min p_1$ will minimize the number of points needed to become World Champion, which is 15. We can also calculate

$$16 - \min \sum_{g \in G} x_g^1$$

which determines the maximum number of races in which the World Champion does not finish at least sixth, that is, the maximum number of races for which the World Champion earns no points, which is 14.

By including some additional constraints many other questions can be answered. For instance, if we optimize $\min p_1$ and add the constraint

$$\sum_{\substack{g \in G: \\ g=(r,1)}} x_g^1 = 0,$$

we find out that the minimum number of points that the World Champion can earn while winning no race is 21.

We also learn in case of 24 drivers (or less) that the World Champion must finish at least fifth in some race since $\min p_1$ with the added constraint

$$\sum_{g \in G \setminus \{6\}} x_g^1 = 0$$

is infeasible.

4. The NHL

We now modify the model to examine the National Hockey League. There is a tenacious myth among fans that earning 95 points during the regular season guarantees a team a spot in the playoffs. By making a few adjustments to the general model, we will show that 149 points are needed in the current season structure to ensure such a performance, and 156 points might be needed in the season structure currently under discussion.

The point system is and will remain as follows: two points for a victory, one point for the loss of a game that goes into overtime, and zero points otherwise.

Since 2000, the regular season is structured as follows. The thirty teams are divided equally into two conferences (West and East). Each conference is split into three divisions of five teams. Each team plays eighty-two games during the regular season: six intradivisional games against each of the teams in its own division, four interdivisional intraconference games against each team that is within its conference, but outside its division, and one interconference game against each team in the other conference. Additionally, each team plays an extra interconference game against three teams. Within each conference, eight teams continue to the playoffs: the champion of each division as well as the five teams that fared best among the rest.

Let $C := \{1, 2\}$ be the set of conferences, let $D := \{1, 2, 3\}$ be the set of divisions within a conference, let T be the set of teams $t := (i, d, c)$ which are labeled with $i = 1, \dots, 5$ according to their rank within division d in conference c , let G be the set of games $g := (t_1, t_2, n)$ where n is the number of the game played between teams t_1 and t_2 . Consider the following system:

$$\begin{aligned} x_g^{r,1} + x_g^{r,2} + x_g^{o,1} + x_g^{o,2} &= 1 & \forall g \in G \\ \sum_{\substack{g \in G: \\ g=(t,t',n)}} (2x_g^{r,1} + 2x_g^{o,1} + x_g^{o,2}) + \\ \sum_{\substack{g \in G: \\ g=(t',t,n)}} (2x_g^{r,2} + 2x_g^{o,2} + x_g^{o,1}) &= p_i^{d,c} & \forall t = (i, d, c) \in T \\ p_{i+1}^{d,c} &\leq p_i^{d,c} & \forall i \in [4], c \in C, d \in D \\ x_g^{*,i} &\in \{0, 1\} & \forall g \in G, * \in \{r, o\}, i \in \{1, 2\}, \end{aligned}$$

where $x_g^{*,j}$ is a binary variable which takes value 1 if game $g := (t_1, t_2, n)$ is won by t_j in $*$ -time (r for regular, o for overtime), and 0 otherwise.

The only difference with the general model is that the third inequality applies to the ranking within a division and not to the global ranking. Note that the extra interconference games cannot be accounted for; we do not know which teams affront each other since those games are not determined by rank. By focusing our attention on a single conference, however, say $c = 1$, we remove the problem since the two conferences access the playoffs independently. In a maximization (resp. minimization) problem, the teams of the selected conference will always win (resp. lose without overtime) all of their interconference games. Accordingly, in the following models we leave out the superscript c indicating the conference for the variables counting the points, since each model only deals with a single conference.

A second problem still arises: we know the ranking of each team in the first conference within its division, but we do not know its overall ranking in the conference. The champion of each division will automatically continue to the playoffs, but we do not know which other five teams will make the cut. We want to find the maximum number of points needed to guarantee a playoff spot, that is, we want to maximize the number of points earned by the fifth-ranked team among the teams of the first conference without the three division champions.

There are eight ranking scenarios to consider up to symmetry, see Figure 1. We represent each team by a square in a matrix-like diagram. Each column represents a division, and each row, the i th position in that division. If the team qualifies for the playoffs, its square is colored gray; otherwise, it is white. We know the top team of each division continues to the playoffs, so these squares will always be gray. We also know that if the i th-ranked team of a division makes it to the playoffs, then any j th-ranked team with $j < i$ in that division will also access the playoffs.

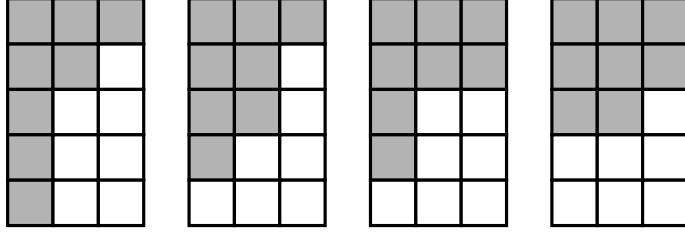


Figure 1: Possible scenarios for playoff qualification.

Note that the order of the columns is unimportant since we can decide which column is assigned to which division. Besides the top three squares, five more must be gray. There are thus only four diagrams possible.

Clearly, only the lowest gray square of a column may be the last team selected for the playoffs. So in the first diagram, either we maximize p_5^1 and add the constraint that $p_5^1 \leq p_2^2$ or we maximize p_2^2 and add the constraint $p_2^2 \leq p_5^1$. Observe how the relative order of the other teams is unimportant. Thus, for the first scenario, we get the following integer program for the selected conference:

$$\begin{aligned}
\max \quad & p_5^1 + 18 \cdot 2 \\
\text{s.t.} \quad & x_g^{r,1} + x_g^{r,2} + x_g^{o,1} + x_g^{o,2} = 1 \quad \forall g \in G \\
& \sum_{\substack{g \in G': \\ g=(t,t',n)}} (2x_g^{r,1} + 2x_g^{o,1} + x_g^{o,2}) + \\
& \sum_{\substack{g \in G': \\ g=(t',t,n)}} (2x_g^{r,2} + 2x_g^{o,2} + x_g^{o,1}) = p_i^d \quad \forall t := (i, d) \in T \\
& p_{i+1}^d \leq p_i^d \quad \forall i \in [4], d \in D \\
& p_5^1 \leq p_2^2 \\
& x_g^{*,j} \in \{0, 1\} \quad \forall g \in G, * \in \{r, o\}, j \in \{1, 2\}
\end{aligned}$$

where G' denotes the set of intraconference games and the constant term added in the objective function originates from the interconference games.

We do the same thing for the other diagrams, and we get the following results for the eight scenarios: 148, 148, 149, 149, 148, 148, 148, 148. Thus, to be guaranteed a spot in the playoffs, a team should accumulate 149 points.

Since 2000, the highest number of points collected by a team during the regular season was 124 points, and 95 points weren't enough to qualify for the playoffs only twice (Colorado in 2007 and Dallas in 2010). What can explain such a big gap between reality and theory? Two things stand out as being unrealistic in the solution we found. For one thing, teams in the NHL are mostly of similar strength; each team wins a bit more or a bit less than half of its games. A team winning 67 games is unheard of. Moreover, a relatively small percentage of games end in overtime.

We thus add some constraints to make the model more realistic. Since 2000, the top team had on average 53 wins and the bottom team, 24 wins. Since our model only includes the teams in a single conference, and assuming the probability of winning a game is the same for inter- and

intraconference games, we claim that a team wins at most 41 intraconference games and 12 interconference games. Similarly, a team wins at least 18 intraconference games and 6 interconference games:

$$18 \leq \sum_{\substack{g \in G': \\ g=(t,t',n)}} (x_g^{r,1} + x_g^{o,1}) + \sum_{\substack{g \in G: \\ g=(t',t,n)}} (x_g^{r,2} + x_g^{o,2}) \leq 41 \quad \forall t \in T$$

We also add constraints to limit the number of games that go into overtime. Since 2000, about 22% of games have gone into overtime, so we could say that between 19% and 25% of the games played by a team t , that is between 15 and 21 games, go into overtime. Once again, we change this constraint to fit our model better and suppose that a team must have between 12 and 16 intraconference games and between 3 and 5 interconference games go into overtime.

$$12 \leq \sum_{\substack{g \in G': \\ t \in g}} (x_g^{o,1} + x_g^{o,2}) \leq 16 \quad \forall t \in T$$

We can even go further and assume that the likelihood of winning or losing a game that goes into overtime is the same as winning or losing a game in sixty minutes. Thus a team should win between $\frac{24}{82}$ and $\frac{53}{82}$ of its games that go into overtime.

$$\frac{24}{82} \sum_{\substack{g \in G': \\ t \in g}} (x_g^{o,1} + x_g^{o,2}) \leq \sum_{\substack{g \in G': \\ g=(t,t',n)}} x_g^{o,1} + \sum_{\substack{g \in G': \\ g=(t',t,n)}} x_g^{o,2} \leq \frac{53}{82} \sum_{\substack{g \in G': \\ t \in g}} (x_g^{o,1} + x_g^{o,2}) \quad \forall t \in T$$

With these restrictions, the maximum number of points needed to guarantee a place in the playoffs drops to 117 points, which is still much more than 95, but already in a more realistic range.

Now let's consider the restructuration that is currently under discussion in the NHL. The plan is to have two conferences, each divided into two divisions. The two divisions in the West Conference would each be composed of eight teams, whereas the divisions in the East would be formed of seven teams. Each team in the East conference would play six games against each team in its division and forty-six interdivisional and interconference games. In the West conference, the setting would be more complicated: each team would play five games against each team in its division and an additional sixth time with three of those teams, as well as forty-four interdivisional and interconference games. The points would be assigned as before (2 for a win, 1 for a loss with overtime and 0 otherwise). The top four teams of each division would qualify for the playoffs, meaning that each division would access the playoffs independently, and so our model needs only to apply to a single division.

First, we calculate the maximum number of points needed to qualify for the playoffs for a team

in the East conference in this structure. Let G'' be the set of intradivisional games.

$$\begin{aligned}
\max \quad & p_4 + 46 \cdot 2 \\
\text{s.t.} \quad & x_g^{r,1} + x_g^{r,2} + x_g^{o,1} + x_g^{o,2} = 1 \quad \forall g \in G \\
& \sum_{\substack{g \in G'': \\ g=(t,t',n)}} (2x_g^{r,1} + 2x_g^{o,1} + x_g^{o,2}) + \\
& \quad \sum_{\substack{g \in G'': \\ g=(t',t,n)}} (2x_g^{r,2} + 2x_g^{o,2} + x_g^{o,1}) = p_i \quad \forall i \in [7] \\
& p_{i+1} \leq p_i \quad \forall i \in [6] \\
& x_g^{*,j} \in \{0, 1\} \quad \forall g \in G, * \in \{r, o\}, j \in \{1, 2\},
\end{aligned}$$

where p_i is the score of the i th-ranked team in one of the two East divisions. The solution tells us that a team might need 155 points to make the playoffs.

Now let's look at a division in the West conference. The model is exactly the same as before (except with eight teams and forty-four interdivisional and interconference games), but we cannot use the model as is since we do not know which teams will affront each other in the extra intradivisional games.

For now, suppose that every team plays only five times against each other. Then the model can be used and the maximum number of points that the fourth-ranking team can have is 150. What about the extra games? Is it possible for the fourth team to win its three extra games and still be fourth? If so, then $150 + 6 = 156$ would be the number of points needed to guarantee a playoff spot in the West conference. And indeed, this solution is feasible: Consider the extra games represented by an edge between two teams in Figure 2. The top four teams can thus win all of their extra

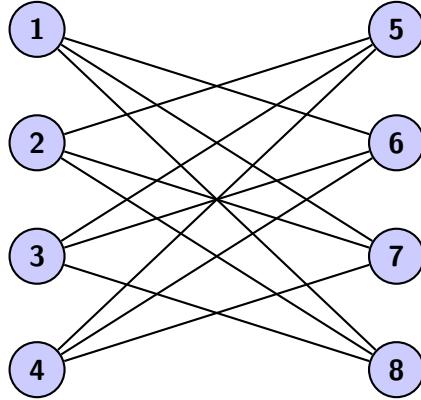


Figure 2: Bipartite graph for extra games of top 4 ranked teams.

games, meaning that the previously fourth-ranked team might have 156 points and still be fourth.

The number of points needed to guarantee a playoff spot will therefore increase from 149 to 155 or 156 (depending on the conference) if the new season structure is adopted. Will this increase be reflected in reality? Moreover, will the slight theoretical difference observed between the East and West conference make it harder for a team in the West conference to reach the playoffs?

instance	variables	constraints	nodes	time	value
bundesliga	936	341	1	0.13	57
f1_minp1	3164	599	181	19.47	15
f1_maxnpoints	3164	599	91	13.39	14
f1_minp1nowin	3164	600	718	20.10	15
f1_minp1notop5	2712	528	1	0.06	—
nhl-case3	1437	447	1	0.26	149
nhl-real-case3	1929	609	1	0.69	117
nhl-new-east	425	131	1	0.06	155
nhl-new-west	430	138	1	0.02	156

Table 1: Models sizes and running times.

5. Results and Conclusion

Table 1 lists the results of the models we discussed using ZIMPL 3.3.0 and SCIP 3.0.0, with SODEX 1.7.0 as linear programming (LP) solver, see the SCIP Optimization Suite [10]. All computations were made on a desktop machine with 4 cores (Intel Xeon Processor E3-1290 v2 @ 3.7GHz) and 16GB RAM memory. All models and solution files can be obtained from www.zib.de/schlechte/sports-ranking-ip.html under the same names as in Table 1. For each instance, we list the number of *variables* and *constraints* for the corresponding formulation and the needed branch-and-bound *nodes*, the solution *time* in seconds, and in case of feasibility, the optimal *value*.

As can be seen, the integer program we presented can be applied to many different sports to compute within a very short time the minimum or maximum number of points needed to finish *i*th in a competition. Moreover, by adding some additional constraints, the program can determine the number of points needed to ensure a certain achievement, such as avoiding relegation or qualifying for playoffs, even under unusual circumstances, as was seen in the case of Formula One or the NHL. Hence, the presented general model is very powerful and can be easily solved by free and non-commercial software.

We note that the theoretical solutions that we found are always far away from what is observed in practice. Even after adding realistic constraints to the NHL model, the number of points to guarantee a spot in the playoffs was still much higher than what has been needed in the past. It would be interesting to investigate more deeply which kind of constraints would have to be added to yield solutions even closer to reality.

References

- [1] Tobias Achterberg. Scip: Solving constraint integer programs. *Mathematical Programming Computation*, 1(1):1–41, 2009.
- [2] Ilan Adler, Alan L. Erera, Dorit S. Hochbaum, and Eli V. Olinick. Baseball, optimization and the world wide web. *Interfaces*, 32(2):12–22, 2002.
- [3] Eddie Cheng and Daniel Steffy. Clinching and elimination of playoff berth in the nhl. *International Journal of Operations Research*, 5:187–192, 2008.

- [4] Dan Gusfield and Chip Martel. The structure and complexity of sports elimination numbers. *Algorithmica*, 32:73–86, 2002.
- [5] Graham Kendall, Sigrid Knust, Celso C. Ribeiro, and Sebastián Urrutia. Scheduling in sports: An annotated bibliography. *Computers & Operations Research*, 37(1):1–19, 2010.
- [6] Walter Kern and Daniël Paulusma. The computational complexity of the elimination problem in generalized sports competitions. *Discrete Optimization*, 1:205–214, 2004.
- [7] Thorsten Koch. *Rapid Mathematical Prototyping*. PhD thesis, Technische Universität Berlin, 2004.
- [8] Celso Ribeiro and Sebastián Urrutia. An application of integer programming to playoff elimination in football championships. *International Transactions in Operational Research*, 12:375–386, 2005.
- [9] Tyrel Russell and Peter van Beek. A hybrid constraint programming and enumeration approach for solving nhl playoff qualification and elimination problems. *European Journal of Operational Research*, 218(3):819–828, 2012.
- [10] SCIP Optimization Suite. Constraint and mixed integer programming solver and modelling language, 2013. Free for academic use, download at scip.zib.de.

Dynamic opponent-dependent and position-dependent player ratings in the AFL

Jonathan Sargent* and Anthony Bedford*

* School of Mathematical and Geospatial Sciences, RMIT University, anthony.bedford@rmit.edu.au,
jonathan.sargent@student.rmit.edu.au,

Abstract. This paper explores a dynamic player rating system for the Australian Football League. The model adheres to the adjustive system whereby a player's rating can increase, decrease or remain steady in line with good, poor and mean performance, respectively. Each player's performance is measured by a linear equation that combines weighted performance indicator frequencies—achieved by each player during a single match—to arrive at a match “score”. Players are then pitched in simulated “head-to-head” contests where player i 's score and opponent j 's score are randomly generated from two independent normal score distributions from that season, prior to the impending match. Opponent j represents a player in the opposition team, playing in the same game position (for example, midfielder) as i . After 1000 simulations, a pre-match expectation (Exp) of i outscoring j is generated, where a low Exp value implies a stronger opponent and vice versa. A case study is offered in which the Geelong Football Club's midfielders are rated from the 2010 and 2011 seasons. It is anticipated that the player ratings will aid AFL and fantasy-league coaches in player performance prediction, as well as guiding betting on AFL player awards.

1. Introduction

Australian Rules football, or AFL, is an invasion game played between two teams, each with 18 on-field players (and four reserves); a regular season consists of 18 teams each playing 22 matches. The dynamics of the game are similar to world football (association football or soccer), except that AFL players can use their hands to punch (handball) the ball to the advantage of a team member. The ultimate objective is to score a goal—worth six points—by kicking a ball through two upright posts at either end of the ground. Like other invasion games, scoring is the result of a series of critical events, or performance indicators, executed between the individuals involved in the contest (Nevill et al, 2002). In modern sports media, such indicators are intensively collected and published online across an ever-increasing number of sports, prior to, during and after a match. It is common for player i 's indicators from a match to be weighted and linearly combined, resulting in a numerical performance appraisal. This methodology has become a standard for many fantasy sporting leagues—that is, to calculate players' post-match “score”, then proportionally adjust their (fantasy) market value according to their scoring fluctuations as determined by a moving average from past matches. This methodology is a simple example of a player rating model.

Ratings in sport are derived from evaluations of the performance of a team or individual player, most often with prior performances in mind. Stefani (2012) offers a succinct distinction between a rating and a ranking: “A rating is a numerical value assigned to a competitor, based on results and other factors while a ranking is the ordinal placement based on ratings.” Rating systems are beneficial to numerous parties; athletes and coaches can track form and progress and use the rating as a motivational tool, while sporting federations can publish the top-ranked (and bottom-ranked) sides for public consumption. With ratings in place, a league—and, indeed, world—ranking can be produced: teams and/or players can be compared for improved team selection; and players and teams can be compared in hypothetical situations for betting purposes.

The player rating system detailed in this paper adheres to an adjustive system where a player's rating can increase, decrease or remain steady in line with good, poor and mean performance, respectively. In general terms, a competitor's new rating is the sum of his prior rating and new information regarding his most recent performance. When calculating this new information, we noticed that each AFL player's past performance scores followed a unique, approximately normal distribution, according to his game position (defender, forward, midfielder, ruck); scores could be randomly generated for an upcoming match for the player (i) from this distribution. The same was done for his “opponent” (j), an approximately normal distribution of scores by all j in i 's position in completed matches in that season. After running 1000 simulations for each i , a probability was generated of i defeating j . A ratings increase or decrease was dependent on the difference

between the observed result (1 = win, 0.5 = draw, 0 = loss) and expected result of the player contest, weighted by the “size” of the win. The win size was i ’s actual score minus the average score of all like-positioned players from the opposition in that match. In the 2010 and 2011 season, a satisfactory relationship was achieved between all Geelong midfielders’ expected and actual performance scores, which provided confidence in the model. Furthermore, the top-five-rated midfielders finished in the top six for Geelong’s club champion award at the end of the 2011 season, suggesting the ratings correlate well with player ability.

While an abundance of literature exists on team ratings (Stefani and Clarke, 1992; Stefani, 1997; Stefani, 2012; Bedford, 2004) and individual sport player ratings (Elo, 1978; Bedford and Clarke, 2000; Barrie, 2003), there is a scarcity of published research on player ratings in team sports. In part, this reflects a lack of consensus, understanding or demand for the final product. Using factor analysis, Bracewell (2003) developed rugby player ratings where each factor represented a core trait performance across nine positional clusters. Oliver (2004) went to admirable efforts to establish a set of offensive and defensive ratings for each player in a basketball match; this was an additive system using frequencies of that player’s on-court skills, for example, effective passes and score assists. Chatterjee and Yilmaz (1999) used a covariance matrix, and a Mahalanobis distance measure to gauge the best overall performance from a list of MVP basketball winners. Although no published research could be located on AFL player ratings, the challenge in adapting and augmenting techniques applied in other sports was rewarding, adding originality to this work.

2. Rating Systems

Stefani and Clarke (1992) and Stefani (2012) define three types of rating systems across a range of internationally recognised sports: subjective, accumulative and adjustive. Subjective rating systems, as the name suggests, offer the least scientific approach to rating competitors, with a panel of experts ranking the competitors after each assessable round. Accumulative systems are the most widespread, converting performance to points which are added over a specified number of rounds to produce the rating. They are especially attractive for individual events like archery or diving because competitors’ final point totals are commensurate with their performance—accuracy in the archer’s case and execution for the diver. A limitation of accumulative ratings is the capacity to overlook absent competitors, particularly in a team sport. For example, if a rating system is concerned with the sum of performance-dependent points allocated to football players over a season, injured players who aren’t able to participate in certain matches risk being “leapfrogged” by their teammates who play more games and have the opportunity to earn more points. An injured player with fewer points—and, hence, a lower rating—is not necessarily a lesser player than one who has acquired more points. Adjustive rating systems account for leapfrogging because active contestants’ ratings can decrease at any stage, rather than increase at each stage, as with accumulative systems.

Adjustive systems have interesting and unique properties that necessitate discussion. The generalised form of the adjustive rating is:

$$R_n = R_o + k(Obs - Exp) \quad (1)$$

where R_n is the new rating for player/team i , R_o is the old rating, Obs is the observed result of a contest, Exp is the expected result and k is a multiplier that assumes different interpretations depending on the contest. A key attribute of these ratings is that Exp is predictive in nature, providing a pre-match approximation of a player’s or team’s performance. Exp is usually probabilistic, describing the likelihood of a player or team defeating his or their opponent, so is a function of opponent strength ($R_{oi} - R_{oj}$, where j is the opposition player/team). Another attractive property of the adjustive system is that a lesser participant can realise a heavy defeat against a stronger opponent without experiencing a heavy rating decrease, a powerful concept ignored by many ratings systems. Conversely, the stronger opponent may only realise a slight ratings increase after trouncing a lesser opponent. Originally developed by Arpad Elo to rate chess players, the Elo rating system (Elo, 1978) is, arguably, the most recognised adjustive system. It has since been adapted in research on individual sports such as tennis (Bedford and Clarke, 2000) and even Tiddlywinks (Barrie, 2003). The Elo system is also used to rate football teams, but should not be confused with the FIFA world rankings. Football is one of the few continuous team sports that has adapted and published official player ratings; the Castrol rankings are based on the actual performances of every player across Europe’s top five leagues (<http://www.castrolfootball.com/rankings>). Specifically, index scores are given to every player after

each match, reflecting the outcomes of every touch of the ball from different parts of the pitch, and are then converted in an adjustive points system (the algorithm is not publicly accessible). In the AFL, the only publicly accessible player rating system, with a disclosed equation, is that calculated in the AFL-sponsored fantasy football competition. Each player's rating at any stage in the season is the average of his performance scores, Y_t , to the most recent match, $t - 1$, where:

$$Y_t = \sum b_m X_m, \quad m=1, \dots, m \quad (2)$$

where X_m are AFL player performance indicators, with weights, b_m , assigned arbitrarily according to the perceived worth of one indicator unit, for example, one kick. A limitation with averaging all performances to Y_{t-1} is that the most recent performances, described as a player's "form", are not accurately portrayed. A player may have performed outstandingly in the first ten games of an AFL season, averaging 125 points (using Equation (2) for each match), but performed poorly in the last five, averaging 50. His overall average is still recorded as 100 points, which is considered excellent, but his recent form is poor. The proceeding section describes a scientific approach to player ratings in a team sport, designed to portray more accurately the form of a player with respect to his opponent.

3. Methods

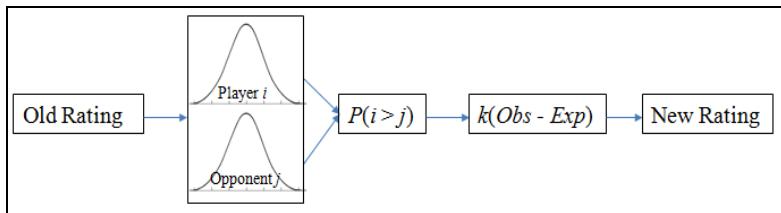


Figure 1. The adjustive AFL player rating system (APR)

While it could be argued that, in the infancy of this research, the Elo model was adopted for producing adjustive AFL player ratings (APR), it would be more correct to label it an adjustive rating system influenced by Elo. The complexities of AFL, namely the number of players on the field and the variety of roles each player assumes, made ambiguous and, at times, erroneous the application of the original Elo formula, particularly the expectation calculation. From a player perspective, AFL is not a competition where the observed player "head-to-head" result is a win, loss or draw, as in chess. Although two players will "match up" on each other during a match, the match-up rarely continues for the whole game, demanding that the data collection process account for each and every match-up on the field at every stage in the match, which is virtually impossible. Moreover, different performance appraisal methods may be required for the match-ups in order to declare the "winner" of the match-up. For example, a defender who was matched up on a forward for the entire game would need his goal prevention assessed, while the forward would need his goal kicking/assists assessed; two separate equations are required. We are now out of the bounds of a logical application of the classical Elo model, but it is possible to arrive at a logical rating system that is respectful of the Elo methodology.

A key motivation for APR was to take advantage of the predictive element of the ratings (*Exp*), that is, to arrive at an estimate of how a player might perform in an upcoming match. These estimates help determine whether a player is a sensible pick in a fantasy league given certain match conditions, namely, player position and opposition strength. Such knowledge is also beneficial when predicting the winner of player awards. APR assumed the form of Equation (1) and, like other adjustive systems, added rating points for above expectation performance and subtracted points for below expectation performance. For the sake of simplicity, each player's performance after match t was measured using an equation similar to Equation (2), but with optimised coefficients. It was important to respect Elo's "head-to-head" methodology, so the agreed approach involved calculating the probability of player i outscoring player j in an upcoming match, where player j was a randomly selected opponent from the same position as i (for example, midfield) in match t . A player's position (defender, forward, midfielder, ruck) was initially allocated as per the AFL fantasy league website (<http://dreamteam.afl.com.au/>) and assumed to be uniform for each player across the entire season, which is not realistic because players can play in different positions from week to week and, even, within a

match. Sargent and Bedford (2010) offered a player position classification system which scientifically allocates players to one of the four positions after each match, with confidence levels, using a range of AFL performance indicators. For the purposes of establishing *APR*, the positions from the fantasy league site were considered adequate.

From Equation (1), the observed and expected values for each player i are denoted as:

$$Obs = \begin{cases} 1 & \text{if } Y_i > \bar{Y}_j \\ 0.5 & \text{if } Y_i = \bar{Y}_j, \quad Exp = P(Y_i > Y_j) \\ 0 & \text{if } Y_i < \bar{Y}_j \end{cases} \quad (3)$$

where Y is a random variable describing the performance of players i and j , and \bar{Y}_j is the average observed performance of the opponents in the same position as player i in match t . Like Elo (1978), we modelled the players' performance score, Y , as an approximately normally distributed random variable, so each $Y_i \sim N(\mu_i, \sigma_i)$. It was, hence, possible to randomly generate a player's score (\hat{Y}_i) from his unique distribution in his allocated position. Figure 2 shows Geelong midfielder, Joel Selwood's performance score distribution between 2010 and 2011, fitted with a normal distribution.

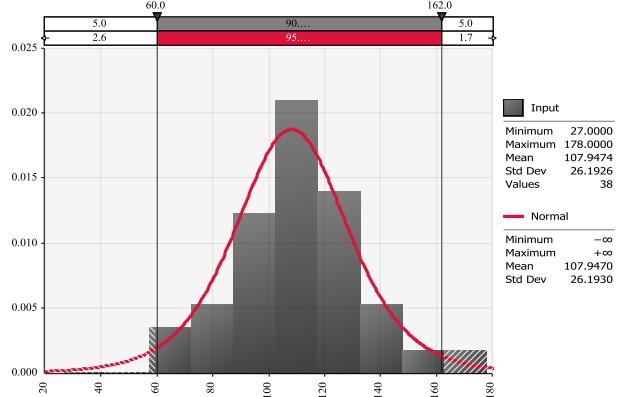


Figure 2. Joel Selwood's score frequency fitted with a normal distribution.

The estimated random opponent player score, \hat{Y}_j , was drawn from the normal distribution, $Y_j \sim N(\mu_j, \sigma_j)$, of the player j scores—in the same position as i —in matches leading up to, but not including, t in that season. By generating a unique approximately normal performance distribution for each position from each team using matches prior to t , the model was able to account for opponent strength—not just the team as a whole, but the relative strengths of each team's positions. Figure 3 offers a comparison between the midfielder score distribution of Geelong's round 23 (Sydney) and round 24 (Collingwood) opponents in the 2011 season. Collingwood played Geelong in the 2011 grand final and were considered an exceptional midfield, realising a mean score of close to 89, and 90% of scores falling between 43 and 137. Sydney finished seventh on the ladder in 2011 and was considered a weaker midfield, with a mean under 80, and 90% of scores falling between 31 and 121 (see Figure 3).

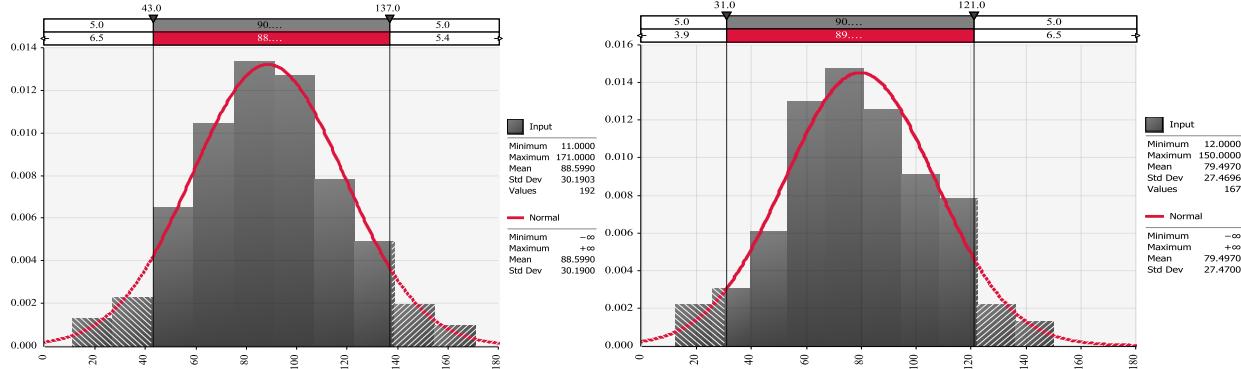


Figure 3. Collingwood (left) and Sydney (right) midfield scores fitted with normal distributions.

A Geelong midfielder, for example, would have a greater chance of outscoring a Sydney midfielder than a Collingwood midfielder. Exp was calculated by simulating 1000 different \hat{Y}_i and \hat{Y}_j from the independent player and opponent normal distributions, respectively, and recording the percentage of $\hat{Y}_i > \hat{Y}_j$. Simulations were run on every Geelong player's matches between 2010 and 2011 so that the final ratings could be adjusted on a match-by-match basis. Apart from the old and new ratings, the final value to be calculated in Equation (1) was the multiplier k , which determined the severity of the ratings fluctuation for each player. As the result of any player match-up could be quantified, k measured the size of the match-up victory or defeat for player i , or $Y_i - \bar{Y}_j$. Obs , Exp and k were then substituted into Equation (1) to develop the ratings for player I after match t . Supplementary APR system details are as follows:

- Ratings were produced from data collected over the 2010 and 2011 seasons;
- Each player started with a rating of 500 as of round 1, 2010;
- Ratings calculation was initiated after the player's third match in the rating period;
- Only Geelong midfielders were rated for this stage of the research;
- For future research, forwards will be measured against opponent forwards, and defenders against opponent defenders;
- Round 1, 2011 was excluded due to no knowledge of team form for that year;
- Injured players who were substituted out of a match and not able achieve their expected score were retained in the analysis.

4. Results

Table 1 shows Joel Selwood's ratings fluctuations for each match in the 2010 and 2011 season. Selwood was in the top three midfielders at Geelong and was consistently achieving high scores, so very rarely realised a drop in ratings. His lowest score was 27 in round 1 of 2011; we did not record this score. In this game, Selwood was injured and was substituted out of the match, which posed a question about the ethics of penalising injured players who are prevented from improving their current score. Selwood's biggest ratings decrease was -6.9 against Port Adelaide in round 4, 2010, achieving only 60. Although he fell 10 points behind Port Adelaide's midfield average, his Exp was 0.664, which is relatively low, and prevented a sharp decrease in rating. If his expectation to win a match-up in that match was 90%, his rating would have fallen a further 10 points; the system is harsher on players who are "supposed" to play better than they actually do. His sharpest increase was 27.9 against Melbourne in round 19, 2011. His expectation to win a match-up was 0.772, but his score was 178, which is outstanding. If Selwood's Exp had been set at 0.664, like in the Port Adelaide match, his rating would have increased a further 14 points. A higher Exp value suggested Selwood was expected to play well against the weaker opposition, so his rating increase was handicapped.

It was important that the APR model possessed a reliable predictive element. We ran an "internal" validation on the model by measuring the relationship between each player's Exp values and his resulting performance scores, Y_i , for every Geelong match from 2010 to 2011. Figure 4 reveals a satisfactory linear relationship between the Exp and Y_i values ($R^2 = 0.341$), suggesting performance can be predicted with modest confidence. The outlier sitting close to the x -axis near $Exp = 0.8$ is Jimmy Bartel who was concussed in Round 13, 2011 in the first quarter, having only achieved $Obs = 2$; he was taken from the field and did not return. It is likely that the accuracy of the model can be improved by additional variables to account for injury and possibly ground advantage.

A problematic area for ratings models, particularly those concerning players in team sports, is validation of the final output. How does the notational analyst know that the ratings calculated up to the prior match are reflective of observed on-field performance? We decided to compare the final 2011 ratings for the Geelong midfield with the club's "best and fairest" voting results. The award system is accumulative, with votes awarded by Geelong's internal management after each game and each player given a rating out of 10 where 10 is the highest. Firstly, we summarised every player's rating results after his last measured match. In Table 2, $f(Exp)$ is the average Exp value for player i through the rating period, $Rating$ is the final rating (the table is sorted in descending order of $Rating$), $Movement$ is the average rating movement per round and $Rank$ is where each player finished in the best and fairest voting. The winner of the award for 2011 was Corey Enright—he achieved 150 points (ranked one)—who played as a defender for the year, but more through the

midfield than a key defender who tends to stay closer to the opposition goals. The interesting result was that our top-five-rated Geelong midfielders finished ranked between second and sixth place in the award voting. This impressive correlation between the final ratings and perceived player ability gave us confidence regarding the application of the final ratings in performance measurement and prediction.

Table 1. Joel Selwood's ratings between 2010 and 2011

Year	Opponent	Round	Y_i	\bar{Y}_j	Exp	k	Rating
2010	ESS	1	107	-	-	-	500.0
2010	HAW	2	90	-	-	-	500.0
2010	FRE	3	87	63	0.768	24	505.6
2010	POR	4	60	70	0.660	10	498.7
2010	CAR	5	86	78	0.504	9	502.9
2010	RIC	6	126	84	0.562	42	521.4
2010	SYD	7	107	66	0.696	41	533.8
2010	BRI	8	110	70	0.712	40	545.3
2010	COL	9	96	75	0.566	21	554.4
2010	MEL	10	110	73	0.652	37	567.4
2010	WCE	11	103	73	0.676	30	577.0
2010	ESS	12	90	59	0.766	31	584.4
2010	STK	13	120	107	0.612	13	589.5
2010	NOR	14	109	73	0.782	36	597.3
2010	HAW	15	104	79	0.826	25	601.6
2010	ADE	16	118	85	0.760	33	609.5
2010	BRI	17	162	79	0.780	83	627.7
2010	SYD	18	111	60	0.804	51	637.7
2010	COL	19	96	96	0.674	0	637.8
2010	WBG	20	139	74	0.664	65	659.6
2010	CAR	21	113	79	0.702	34	669.6
2011	STK	1	27	-	-	-	669.6
2011	POR	3	129	66	0.814	63	681.3
2011	SYD	4	92	67	0.746	25	687.8
2011	HAW	5	134	77	0.702	57	704.9
2011	NOR	7	124	70	0.836	54	713.8
2011	COL	8	121	77	0.612	44	730.7
2011	CAR	9	122	82	0.690	40	743.1
2011	GST	10	116	67	0.756	49	755.2
2011	WBG	11	92	73	0.740	19	760.1
2011	HAW	12	108	85	0.756	23	765.7
2011	BRI	17	70	71	0.700	1	764.7
2011	RIC	18	108	80	0.778	28	770.8
2011	MEL	19	178	56	0.782	122	797.5
2011	GST	20	135	63	0.804	72	811.6
2011	ADE	21	84	89	0.780	5	807.7
2011	SYD	23	100	76	0.750	24	813.8
2011	COL	24	118	81	0.716	37	824.4

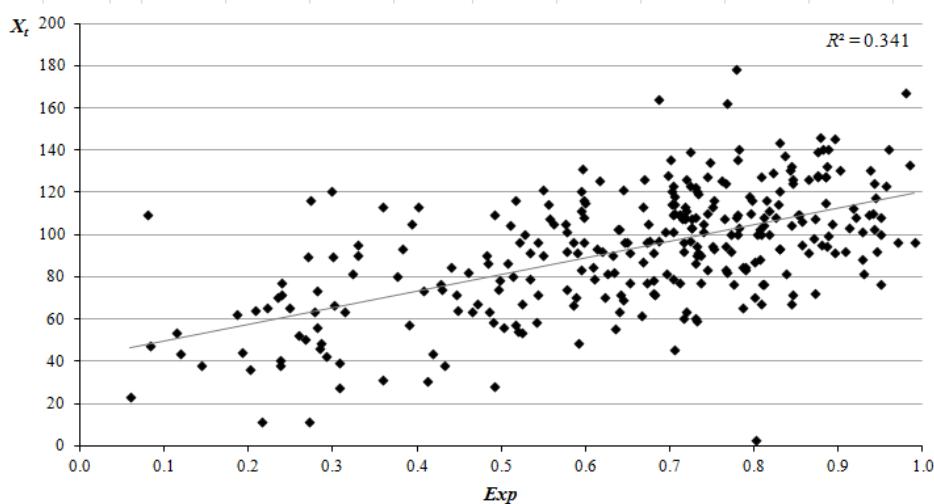


Figure 4. Relationship between Exp and Y_t for all Geelong midfield matches (2010 - 2011).

Table 2. Final ratings results for Geelong midfielders in 2011.

(* Gary Ablett was only rated for 2010 as he was drafted by another club for the 2011 season)

Player	Matches	$f(Exp)$	Rating	Movement	Rank
Joel Selwood	38	0.661	823.10	8.54	6
Jimmy Bartel	41	0.726	678.95	4.36	3
Cameron Ling	36	0.487	670.37	4.73	4
Joel Corey	30	0.580	648.27	4.94	2
James Kelly	21	0.593	623.51	5.88	5
Paul Chapman	20	0.672	608.37	5.42	-
Allen Christensen	16	0.206	589.01	5.56	-
Gary Ablett *	21	0.824	560.09	2.86	-
Mitchell Duncan	18	0.427	499.75	-0.01	-
Josh Cowan	3	0.049	493.52	-2.16	-
Simon Hogan	12	0.223	465.92	-2.84	-

5. Conclusion

An adjustive player rating system was shown to adequately describe and predict player performance in the AFL, with player form and opponent strength the key determinants. This was proven using the Geelong Football Club's midfielders' expected performances between 2010 and 2011 as an estimator of impending match performance and comparing their final ratings to 2011 club champion voting. While arguments exist that rating systems in team sports are too player-centric—ignoring the important concepts of teamwork and cooperation (Gould & Gatrell, 1979/80)—the adjustive rating model has proven to be an interesting tool, particularly for consumption by parties interested in a particular player's upcoming performance. Simulation of player performance proved to be a pragmatic methodology, mainly because the player and opponent score distributions were approximately normally distributed. It is expected that the predictive power of this model will be improved by additional predictors such as ground advantage, travel effects and injury compensation. Moreover, a player scoring formula that is more reflective of a player's teamwork within any match has been flagged as an important development.

6. References

- Barrie, P.J. (2003) A new sports ratings system: the tiddlywinks world ratings. *Journal of Applied Statistics*, 30, 4, pp. 361-372.
- Bedford, A. (2004) Predicting Women's World Cup Handball using Optimised Ratings Models, in proceedings of 7th Australasian Conference on Mathematics & Computers in Sport, H. Morton and S. Ganeshalingam (Eds). pp. 66-74.

- Bedford, A. & Clarke, S.R. (2000) A Comparison of the ATP Ratings with a Smoothing Method for Match Prediction, in proceedings of 5th Australasian Conference on Mathematics & Computers in Sport, G. Cohen and T. Langtry (Eds), pp. 43-48.
- Bracewell, P.J. (2003) Monitoring meaningful rugby ratings. *Journal of Sports Sciences*, 21, pp. 611-620.
- Chatterjee, S. & Yilmaz, M. R. (1999) The NBA as an evolving multivariate system. *The American Statistician*, 53, 257-262.
- Elo, A.E. (1978) *The Rating of Chess Players, Past and Present*, London, Batsford.
- Gould, P. & Gatrell, A. (1979/80). A structural analysis of a game: the Liverpool v Manchester United Cup final. *Social Networks*, 2, 253-273.
- Nevill, A.M., Atkinson, G., Hughes, M. D. and Cooper, S-M. (2002) Statistical methods for analysing discrete and categorical data recorded in performance analysis. *Journal of Sports Sciences*, 20, pp. 829-844.
- Oliver, D. (2004). *Basketball on Paper*, Washington, D.C.: Brassey's Inc.
- Stefani, R.T. & Clarke, S.R. (1992) Predictions and home advantage for Australian rules football. *Journal of Applied Statistics*, 19, 2, pp.251-261.
- Stefani, R.T. (1997) Survey of the major world sports rating systems. *Journal of Applied Statistics*, 24, 6, pp. 635-646.
- Stefani, R.T. (2012) Predictive success of official sports rating in international competition, in proceedings of 10th Australasian Conference on Mathematics & Computers in Sport, A. Bedford and A. Schembri (Eds), pp. 35-40.
- Sargent, J. & Bedford, A. (2010) Long-distance relationships: positional classification of Australian Football League players, in proceedings of 10th Australasian Conference on Mathematics & Computers in Sport, A. Bedford and M. Ovens (Eds), pp. 97-102.

Modelling and optimisation of the sport and exercise training process.

P. Scarf*, M. Shrahili**, S. Jobson*** and L. Passfield****

**Salford Business School, University of Salford, Salford M5 4WT, email: p.a.scarf@salford.ac.uk*

***Salford Business School, University of Salford, Salford M5 4WT, email: m.m.shrahili@edu.salford.ac.uk*

****Department of Sports Studies, University of Winchester, UK, email: simon.jobson@winchester.ac.uk*

*****School of Sport and Exercise, University of Kent, UK, email: l.passfield@kent.ac.uk*

Abstract. In this paper we present an analysis of a mathematical model of the training process that allows the relationship between training inputs and performance output to be quantified. When this relationship is known, training can then be scheduled in order to maximise performance at a future time. We propose a training input measure and a performance output measure that can each be calculated from training data but that depend on a number of unknown athlete specific parameters. The challenge is the estimation of these parameters. We describe some preliminary results for two competitive cyclists.

1. Introduction

This paper is concerned with modelling the training process in sport and exercise and in cycling in particular. Our aim is to provide a quantitative model that can be used to optimize training in advance of a major competition. Training is the method by which an athlete improves his/her specific performance and develops individual character according to the requirements of a specific competition. This paper uses a statistical model to relate training to performance. To do this, both training and performance must be measured, and we do so using field data relating to power output and heart rate. The measure of training we use is an established one based on the concept of accumulated training load; this is broadly an exponentially weighted moving average of the total load on the cardio-vascular system during training of an athlete over all time. However, this accumulated training load measure depends on a number of unknown parameters that must be specified; only then can training be optimized for a specific athlete.

The measure of performance of an athlete that we use is the estimated heart-rate required by the athlete in order to produce power output at a defined, high level. Such a level corresponds to an upper percentile (e.g. 90%) of the athlete's power output distribution, considered over all time. We then determine those values of the parameters of the accumulated training load measure such that this measure is most closely related to the performance measure. At this stage, we do not use the model to optimize training; we merely imply through the results that we obtain that this is now possible in principle.

Training data from a number of competitive athletes were available to us. To illustrate our methodology, we chose two of these athletes with reasonably complete training records. These cyclists gave written, informed consent to participate in our study, providing us with data from their training. The study received local ethical committee approval and was carried out according to the principles of the Declaration of Helsinki (World Medical Association, 2013). For each athlete, for a number of training sessions typically extending over a 300 day period between December 2006 and September 2007, power output and heart-rate were recorded every five seconds, e.g. figure 1. Power output data were collected using SRM cranks (SRM, 2012).

For each session, we calculate a training load and a performance measure. These measures are defined in sections 3 and 5 respectively. The relationship between training load and performance is investigated in section 6. In the following section, we briefly review the literature on training and performance.

2. The relationship between training and performance

The relationship between training and performance is very important for coaches who look to determine a training program for their athletes. Research that has investigated this relationship by using quantitative data can be traced back to the seminal work of Banister et al. (1975). However, in spite of the time that has elapsed since these early ideas were described, predicting the results of a particular training program is difficult, and in particular predicting performance output from training input remains an unsolved problem (Jobson et al., 2009). The relationship between training and performance is highly individualised (Avalos et

al., 2003), with many features (e.g. genetic factors, individual training background, psychological factors, technical factors and speciality) that are very difficult to quantify (Hellard et al., 2006).

Qualitative predictions and descriptions of the effect of training have been made. For example, one can observe a rapid improvement in performance when the initial performance is low, but as an athlete becomes fitter and better trained, it becomes more difficult to observe further improvement in performance by continued training. Banister and Calvert (1980) then point out that it is important for an athlete to avoid overtraining and injury that may decrease performance. Such arguments have been reinforced e.g. Borresen and Lambert (2009). A quantitative description of the relationship between training and performance remains illusive.

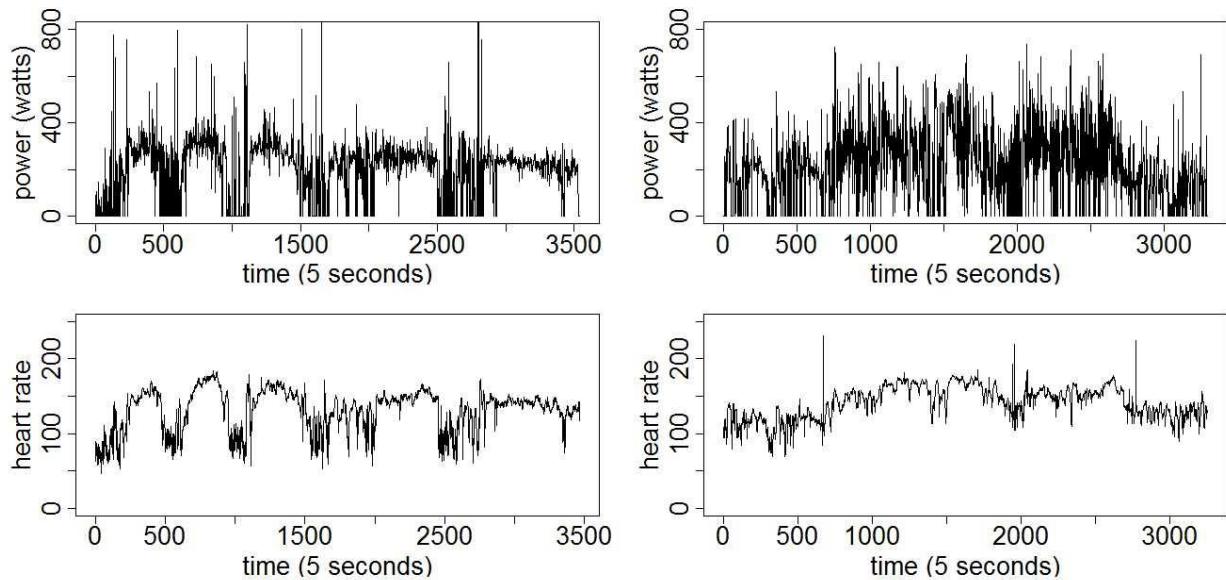


Figure 1. An example of power output (Watts) and heart-rate (beats per minute, bpm) traces from one session for athletes A (left) and B (right).

3. The Banister model

Banister et al. (1975) proposed a model that quantifies training load and its effect on performance. This model describes the progress of an athlete in terms of training and fatigue. The model was developed through the study of the training and performance profiles of a top class swimmer over 105 days of training. The original model considered four components: skills, psychology, cardiovascular and strength. Calvert et al. (1976) simplified this model to two components, fitness and fatigue. The model has been criticised, for example, with regard to: the accuracy of the model to predict future performance; the difference between estimated and actual changes in performance; and the poor corroboration of the model with physiological mechanism (Hellard et al., 2006; Hayes and Quinn, 2009). However, the model remains the basis of the theory of progressive training.

The Banister model defines the accumulated training effect at time t of training sessions occurring up to time t as

$$W(t) = w_0 + k_a \sum_{i=1}^{n_t} w_{s_i} e^{-(t-s_i)/\tau_a} - k_f \sum_{i=1}^{n_t} w_{s_i} e^{-(t-s_i)/\tau_f}. \quad (1)$$

$W(t)$, the accumulated training effect (ATE) at time t , can then be interpreted as the fitness at time t and hence represents the potential performance at time t . In equation (1), w_{s_i} is the known training load during session i and it is defined as a function of \mathbf{h}_i , the heart rate history for session i alone. n_t is the number of sessions up to time t . One possible candidate of training load is training impulse or TRIMP (Borresen and Lambert, 2009). We define TRIMP in the next section. w_0 corresponds to net training effect at time $t = 0$ of sessions in $(-\infty, 0]$.

We will call $w_{s_i} e^{-(t-s_i)/\tau_a}$ the training benefit at time t of session i that took place at time $s_i < t$ and $w_{s_i} e^{-(t-s_i)/\tau_f}$ the training detriment at time t of session i that took place at time s_i . Thus the fitness and fatigue associated with a particular session decay at different rates depending on the parameters τ_a and τ_f , the fitness and fatigue decay time constants, respectively. The decay in both fitness and fatigue is assumed to be exponential and in principle, the decay of fitness is slower than the decay of fatigue: $\tau_a < \tau_f$. k_a and k_f are scale constants that control the relative size of the immediate training benefit with respect to the immediate training detriment. Strictly, one or other of these parameters is redundant as the scale of $W(t)$ is arbitrary. Therefore, without loss of generality we will set $k_a = 1$ throughout.

Thus $W(t)$ is the resultant accumulation of decaying benefits and detriments over time, and the level of potential performance is the difference between the total training effect and total fatigue effect.

The key to quantitatively predicting performance output from training input is the estimation of the parameters τ_a , τ_f and k_f of the Banister model for a specific athlete. This requires performance to be measured reliably and sufficiently often during the training of an athlete. Such performance measurement is the crux of the problem. Arguably, performance can only be measured in a very specific way (e.g. by performing a specific test); consequently it can be difficult to do this sufficiently often; also, in spite of the specificity of the test, the athlete may not perform “maximally”. To circumvent these issues, in this paper, we describe how performance can be measured from training data. The measurement of training load is somewhat more straightforward and is described next.

4. Training impulse (TRIMP)

In its simplest form, the training impulse measure is defined as $TRIMP = T \times \bar{H}$ where T is the duration of a session) and \bar{H} is the average heart rate of the session in (beats per minute). Thus, here, TRIMP is the total number of heart beats during a session. The original formula of Banister and Calvert (1980) was modified by Morton et al. (1990) to include a multiplicative factor that gave greater weight to high-intensity training: $TRIMP = T \times a \times H_{ratio} e^{bH_{ratio}}$, where $H_{ratio} = (H_{ex} - h_0)/(h_{max} - h_0)$, H_{ex} is the average heart rate during the exercise and h_0 is the athlete specific resting heart rate (the number of heart beats per minute). h_0 should be calculated upon waking and while still lying in bed. h_{max} is the athlete specific maximum heart rate. The constant a is taken to be 0.64 for males and 0.86 for females ; the constant b is based on blood lactate and it is equal 1.92 in males and 1.67 in females (Borresen and Lambert, 2009). To calculate the accumulated training effect (1) we use this latter measure of TRIMP.

5. Measuring performance from training data

Our performance measure is determined using the power output, heart-rate relationship. First we describe this relationship.

The relationship between power output and heart-rate excess (the difference between heart rate and resting heart rate) is considered to be proportional. For example, Grazzi et al. (1999) explored the power output, heart-rate relationship in a study involving 500 tests conducted with 290 participants, and observed a correlation of 0.98 or above for many athletes tested. There exists also a delay or time lag between the change in power output and the heart-rate response. The value of this delay or lag is less clear from literature: Jeukendrup and van Diemen (1998) argue its existence for periods of exercise of short duration, as the circulatory system is not able to fully adapt to change in exercise intensity, but do not indicate the size of the lag; The study of Stirling et al (2008) suggests that large changes in heart-rate, say from 80 to 160 (beats per minute), occur over approximately 30-60 seconds, for both increases and decreases in heart-rate. For the data in our study, short term changes in heart-rate tend to be smaller than in the Stirling et al. study. We speculate that for sessions where intensity changes gradually power output will be best explained by a heart-rate lag towards the bottom end of the 30-60 second range, or indeed less. We investigate different lags between power output and heart rate (0, 10, 20 and 30 seconds) and find the strongest relationship when the lag is 10 seconds for almost sessions. Figure 2 illustrates this relationship for a single session.

Our performance measure is now defined as follows. For a specific athlete of interest, first determine a high percentile (e.g. the 90th) of power output using the entire training history of the athlete e.g. see figure 3. Then the performance measure for a session is defined as the expected heart rate (given a linear model that

relates power output to heart rate excess) at this power output percentile, and denoted h_{Pq} in general or h_{P90} in particular, say, e.g. figure 3. This performance measure is then calculated for all sessions.

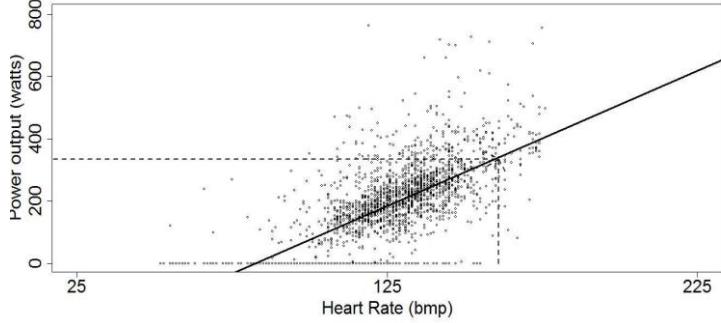


Figure 2. Power output versus heart rate for a typical session, with a lag of 10 seconds.

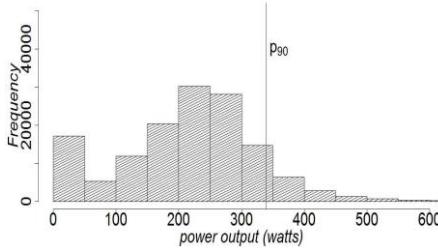


Figure 3. The histogram of power output pooling all sessions for a specific athlete.
 $(P_{50} = 227, P_{90} = 339, P_{95} = 383, P_{99} = 497)$.

6. Estimating the Banister model parameters

We now present findings on the estimation of the Banister model parameters: k_f, τ_a, τ_f , with $k_a = 1$. These are preliminary findings, obtained using a crude search, for two athletes. Table 1 shows the correlation between the performance measure and accumulated training effect for various values of the parameters. On the basis of this search, we now set $k_f = 4, \tau_a = 16, \tau_f = 2$ for rider A and $k_f = 2, \tau_a = 7, \tau_f = 4$ for rider B. The performance measures and the accumulated training effect (ATE) over time are presented in Figure 4 for these two riders. Figure 5 shows the performance measure versus the accumulated training effect for the “best” parameter values for each athlete.

We make the following observations. We would expect h_{P90} to reduce as the athlete becomes fitter and ATE to increase (and these to be negatively correlated). These are indeed the case for athlete B. Matters are inconclusive for rider A, although there is a period during which training data is not recorded for reasons unknown to us. Therefore our procedure has worked reasonably for athlete B but not for athlete A.

7. Discussion

With fully specified parameters of the Banister model, training can be scheduled to maximise performance at a particular time T in the future. We do not determine how to do this optimisation. We merely estimate the parameters in order to do so. Many issues and questions remain, not least investigation of: more rigorous methods of estimation of the Banister model parameters; the usefulness our procedure given the potentially weak relationship between performance output (e.g. h_{P90}) and training input (ATE); the choice of an appropriate percentile for the specification of the performance measure; the consequences of training data that is not recorded; the determination of appropriate values for the maximum heart-rate, resting heart-rate and heart-rate lag; the role of weight loss for performance improvement; the applicability to other sports; and the specification of an algorithm to optimise training . However, in our view, this paper makes an important contribution to the quantification of training and a significant step towards optimising training.

Table 1. The correlation between the performance measure h_{P90} and the accumulated training effect (ATE) for all sessions for two athletes, A and B, for various values of the ATE parameters.

case	k_a	k_f	τ_a	τ_f	$corr(ATE, h_{P90})$	case	k_a	k_f	τ_a	τ_f	$corr(ATE, h_{P90})$		
					A B						A B		
1	1	2	2	1	-0.11	0.08	19	1	2	3	1	-0.08	0.20
2	1	2	4	1	-0.09	0.20	20	1	2	5	1	-0.09	0.19
3	1	2	8	1	-0.12	0.15	21	1	2	7	1	-0.11	0.16
4	1	2	4	2	-0.18	-0.14	22	1	2	3	2	-0.10	-0.30
5	1	2	8	2	-0.17	0.06	23	1	2	5	2	-0.18	-0.01
6	1	2	16	2	-0.16	0.03	24	1	2	7	2	-0.17	0.05
7	1	2	8	4	-0.13	-0.30	25	1	2	5	3	-0.11	-0.33
8	1	2	16	4	-0.17	-0.07	26	1	2	7	3	-0.18	-0.15
9	1	2	32	4	-0.10	-0.09	27	1	2	7	4	-0.09	-0.35
10	1	4	2	1	-0.11	-0.24	28	1	4	3	1	-0.16	-0.13
11	1	4	4	1	-0.17	-0.03	29	1	4	5	1	-0.17	0.03
12	1	4	8	1	-0.16	0.07	30	1	4	7	1	-0.16	0.06
13	1	4	4	2	-0.08	-0.32	31	1	4	3	2	-0.06	-0.32
14	1	4	8	2	-0.17	-0.23	32	1	4	5	2	-0.11	-0.31
15	1	4	16	2	-0.19	-0.09	33	1	4	7	2	-0.15	-0.27
16	1	4	8	4	-0.02	-0.33	34	1	4	5	3	-0.03	-0.32
17	1	4	16	4	-0.10	-0.33	35	1	4	7	3	-0.07	-0.34
18	1	4	32	4	-0.09	-0.21	36	1	4	7	4	-0.01	-0.31

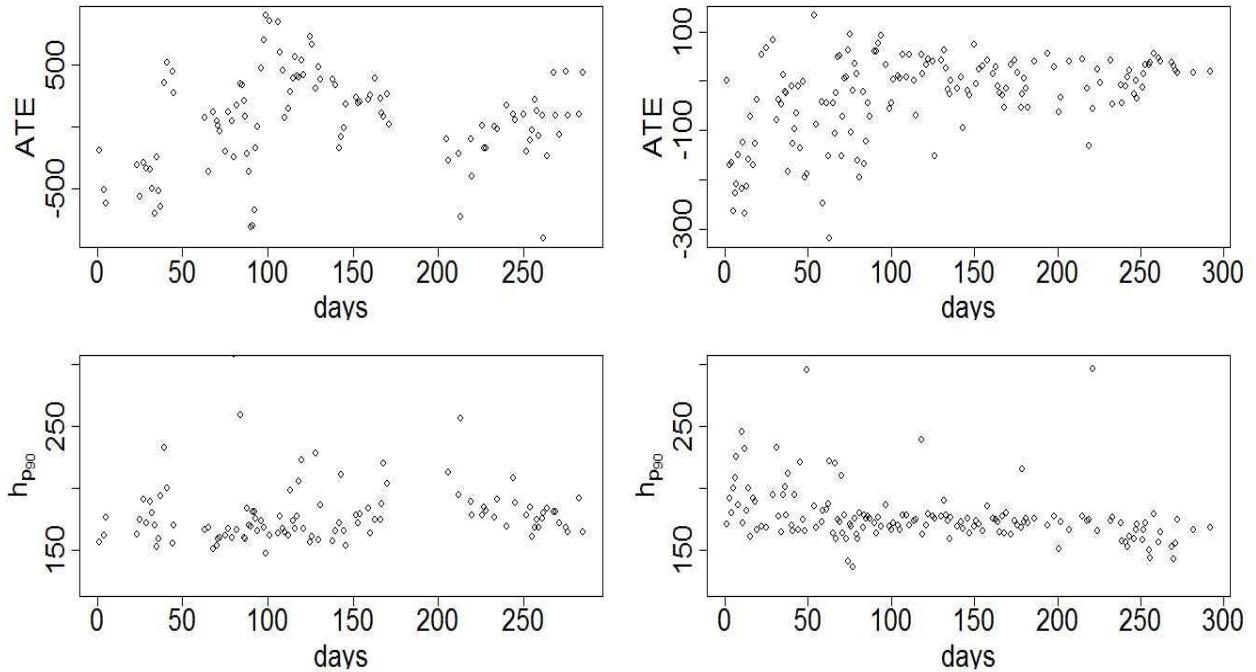


Figure 4. The performance measure h_{P90} and the accumulated training effect (ATE) over time for two riders. Parameter values: $k_a = 1, k_f = 4, \tau_a = 16, \tau_f = 2$ for rider A (left) and $k_a = 1, k_f = 2, \tau_a = 7, \tau_f = 4$ for rider B (right).

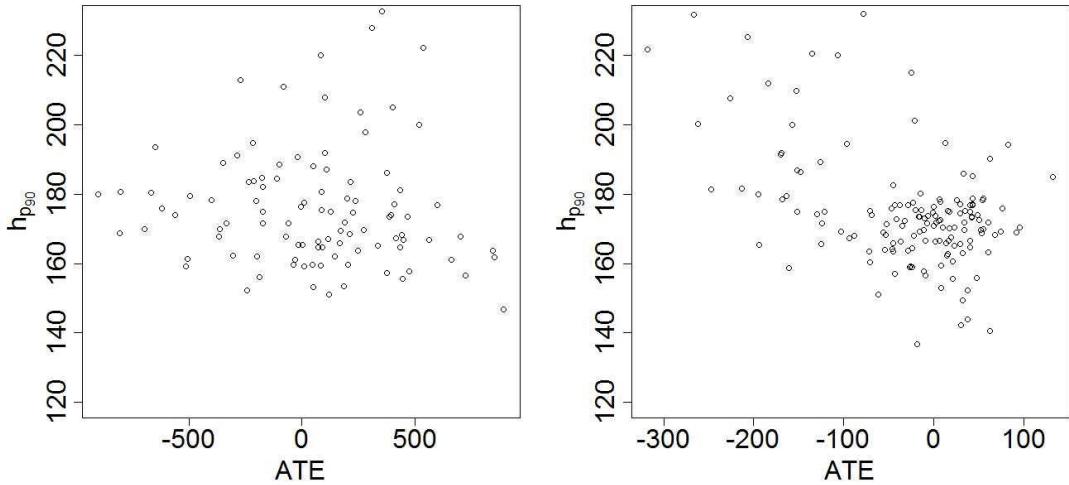


Figure 5. The performance measure h_{P90} versus the accumulated training effect (ATE) for two athletes.

Parameter values: $k_a = 1, k_f = 4, \tau_a = 16, \tau_f = 2$ for rider A (left) and
 $k_a = 1, k_f = 2, \tau_a = 7, \tau_f = 4$ for rider B (right).

References

- Avalos, M., Hellard, P., & Chatard, J. (2003). Modelling the training-performance relationship using a mixed model in elite swimmers. *Med. Sci. Sports Exercise*, 35, 838-846.
- Banister, E. W., Calvert, T. W., Savage, M. V., & Bach, T. M. (1975). A system model of training for athletic performance. *Australian Journal of Sports Medicine*, 7(3): 57-61.
- Banister, E. W., & Calvert, T. W. (1980). Planning for future performance: Implications for long term training. *Can. J. Appl. Spt. Sci.*, 5:3, 170-176.
- Borresen, J., & Lambert, M. I. (2009). The quantification of training load, the training response and the effect on performance. *Sports Med*, 39(9), 779-795.
- Calvert, T. W., Banister, E. W., Savage, M. V., & Bach, T. (1976). A system model of the effects of training on physical performance. *IEEE Transactions on System*, 6, 94-102.
- Grazzi, G., Alfieri, N., Borsetto, C., Casoni, I., Manfredini, F., Mazzoni, G. and Conconi, F. (1999). The power output/heart-rate relationship in cycling: test standardization and repeatability. *Medicine and Science in Sports and Exercise*, 31: 1478-1483.
- Hayes, P. R., Quinn, M. D. (2009). A mathematical model for quantifying training. *Eur J Appl Physiol*, 106, 839-847.
- Hellard, P., Avalos, M., Lacoste, L., Barale, F., Chatard, J., & Millet, G. (2006). Assessing the limitations of the banister model in monitoring training. *Journal of Sports Sciences*, 24(5), 509-520.
- Jeukendrup, A. and van Diemen, A. (1998). Heart-rate monitoring during training and competition in cyclists. *Journal of Sports Sciences*, 16: S91-S99.
- Jobson, S. A., Passfield, L., Atkinson, G., Barton, G., & Scarf, P. (2009). The analysis and utilization of cycling training data. *Sport Medicine*, 39(10), 833-844.
- Morton, R. H., Fitz-Clarke, J. R., & Banister, E. W. (1990). Modelling human performance in running. *J. Appl. Physiol.*, 69(3), 1171-1177.
- Morton, R. H. (1997). Modelling training and overtraining. *Journal of Sports Sciences*, 15, 335-340.
- SRM (2012) SRM power meter. <http://www.srm.de/index.php?lang=en> (last accessed 1.4.2012).
- World Medical Association (2013) Declaration of Helsinki—ethical principles for medical research involving human subjects. <http://www.wma.net/en/30publications/10policies/b3/> (last accessed 11.4.2013).
- Stirling, J. R., Zakynthinaki, M., Refoyo, I. and Sampredo, J. (2008). A Model of Heart-rate Kinetics in Response to Exercise. *Journal of Nonlinear Mathematical Physics*, 15: 426 - 436.

Measuring competitive balance: the case of European premiership rugby union

P. Scarf*, P. Williams** and M.M. Yusof***

*Salford Business School, University of Salford, M5 4WT, UK, p.a.scarf@salford.ac.uk

**Salford Business School, University of Salford, M5 4WT, UK, p.williams@salford.ac.uk

***Universiti Utara Malaysia, 06010 UUM Sintok, Kedah Darul Aman, Malaysia , mmy@uum.edu.my

Abstract. Competitive balance in the rugby union premierships in England, France and Italy is measured using a Bradley-Terry type model. Intra-season competitive balance is measured by the distribution of team strengths; inter-season competitive balance by the correlation of points scored by teams in successive seasons. We find that intra-season competitive balance is similar in each nations' premiership, while inter-season competitive balance is, among the three nations, greatest in the England and least in Italy.

1. Introduction

Tournaments and the institutions that underlie them have evolved to promote competitiveness among participants, as competition is the purpose of their existence (Szymanski, 2003). Measures of competitive balance are therefore required for evidence-based decision-making by tournament administrators. Many such measures exist, all broadly measuring variability of outcome under repeated observation of a tournament, and three dimensions to competitive balance can be distinguished (Lenten, 2009). The first is concerned with the uncertainty of outcome of a single contest, match or tie; the second with the distribution of outcomes across competitors in a tournament, intra-tournament competitive balance; the third with the distribution of outcomes across tournaments, inter-tournament competitive balance. Given the generality of Lenten's classification, it is useful to illustrate these three dimensions in the context of a home and away round-robin tournament that is played annually e.g. the FA Premier League in England and Wales. The first dimension considers uncertainty in relation to a match between team i , say, and team j , and can be considered for all teams i and j in the tournament. In the second dimension, and defining outcome as the number of wins, or win percentage, or points achieved by each competitor in the completed tournament, the variability of such an outcome across competitors in any given season is of interest. In the third dimension, the outcome of interest is typically the label of the tournament winner, or the labels of the top m teams, and variability in the labels from season to season. Arguments about the domination of particular teams in the long run might then be supported or otherwise by considering this third dimension of competitive balance. That a particular team dominated or is dominating in a particular season would be evidenced in the second dimension. The original "uncertainty of outcome" hypothesis (e.g. Forrest and Simmons, 2002; Buraimo and Simmons, 2008) is a matter for the first dimension. All three dimensions of competitive balance are of concern in European sports tournaments played in a league format. Revenue generation underlies these concerns in all dimensions, and simply stated the dimensions relate to consumption of sport in the short, medium, and long term.

Measures of competitive balance were developed mostly for the North American market (Scully, 1989; Quirk and Fort, 1992) where early studies focused on the second dimension. The draft system that is typical of major North American sports leagues ensures that long-term domination by particular teams cannot occur, and this, and that their specification is not obvious (Szymanski and Smith, 2002), is perhaps why inter-season competitive balance measures are less well developed. Since European leagues rely on promotion and relegation to regulate competitiveness in the long-term, and the effectiveness of such regulation is debatable, it is more interesting to measure inter-tournament competitive balance for European sports. We discuss such a measure in this paper, and apply it to premiership rugby union in England, France and Italy.

We are concerned with a tournament played in a league format over a season although our ideas are not confined to such a format. Thus, in tennis for example, in a grand slam tournament, although the tournament is a single elimination standard seeded design (McGarry, 1998) competitive balance in relation to each of the dimensions (individual matches, intra-tournament and inter-tournament) can still be defined in principle although they may be of little interest. For example, in the second dimension, a win-percentage type of measure that is typically used for a league will, for a single elimination tournament, be fixed over repetitions of the tournament. The third dimension may not be an issue because while particular players may dominate

in the medium term, turn-over of winners in the long-term is guaranteed. In hybrid tournaments that are a mix of league and knock-out formats (e.g. UEFA Champions League) each of the three dimensions may be of interest to the tournament administrators: dull group-round matches (first dimension) should be avoided; uncertainty of outcome of qualifiers for the knockout stage (second dimension) should be high; uncertainty of outcome of the tournament winner should remain high throughout the tournament (second dimension). Long term domination has not been an issue to date, perhaps a result of the fact that the latter stages are knockout thus inducing a high uncertainty of outcome with respect to the tournament winner.

When considering competitive balance in relation to different tournament designs, it becomes clear that competitive balance depends not only on variability of the strengths of teams but also on the design of the tournament. Thus, uncertainty of outcome in the second and third dimensions is considerably higher for a knockout tournament than for a round-robin—and Scarf and Yusof (2011) argue this point at length. At face value one might think that in the first dimension (match outcome) competitive balance does not depend on the tournament design. However, within-match competitive balance must depend on the size of the tournament (number of teams) and the schedule of ties. For example, seeding in a tennis grand-slam tournament ensures that within-match competitive balance is high in later stages of the tournament because seeding ensures the likely progression of high seeded players to the latter stages—it is not unusual for the each of the four top seeds to reach the semi-finals. Thus match outcome uncertainty at the semi-final stage will depend not only on the strengths distribution of players but also on whether a seeded or unseeded design is used. First round match uncertainty of outcome will depend very much on the size of the tournament, for obvious reasons. Thus, the three dimensions of competitive balance are related. Underlying this relationship is the strength distribution of competitors and the design of the tournament. In this paper, the design is considered fixed: a home-away round robin played over a “season”.

Central to our development is the existence of a model of the tournament outcome, that competitive balance is a property of this model and that competitive balance is a function (possibly non-scalar) of model parameters. One then attempts to estimate competitive balance given a model. Consequently, due to uncertainty of outcome, one can never have complete information about competitive balance, so that analysis of competitive balance should account for imprecision in its estimation.

2. Elite club rugby in England, France and Italy

2.1. Background

The three leagues we discuss have evolved in different domestic rugby environments, but are, perhaps, more similar now than at any previous point. While some external influences on competition, like the move to professionalism (1995) and the Kolpak ruling on player eligibility (2003), have applied to all three leagues, initiatives available to league administrators seeking to promote competitiveness have not been adopted in any uniform fashion. In England, prior to the introduction of a league structure in 1987, clubs enjoyed “first class” status by virtue of a loosely-defined system of historical playing strength. Powerful clubs fought hard to maintain their status. The meritocracy of leagues sharply increased competitiveness. Like its French equivalent, the Rugby Football Union dithered in response to the advent of professionalism. The clubs, to an extent financed by wealthy benefactors, took matters in their own hands; prices paid for players escalated; and ultimately a number of big-named clubs dropped out of the professional tier. To stabilise finances, league administrators introduced the salary cap in 1999. Two years later, play-offs and bonus points were introduced in an explicit attempt to improve competitiveness. The size of the top division has varied from ten to fourteen clubs, but since season 1999-2000 has settled at twelve, and clubs must win the second tier division to gain promotion. The league is still considered too reliant on rich backers, with few clubs able to sustain themselves financially.

In France, rugby enjoys higher status than in either England or Italy, and it is considered the national sport in southern France. Since professionalism, the size of the top tier of clubs has reduced from 32 clubs (in four pools) to 14, with some pressing for a further reduction to 12. The use of a play-off system to determine the champion club is well-established. However, one of the most popular features of professional sports leagues where improvements in competitiveness are explicitly sought, a salary cap, was not employed until the 2010-11 season. It is likely this was a reaction to the entry of rich club backers, most notably at Toulon

and at the two Paris clubs of Stade Francais and Racing Metro. Generously pitched at almost twice that applicable in England, the cap has not prevented some French clubs from amassing large squads of international stars from around the globe. Less wealthy clubs have struggled to compete and some have recently been at or near bankruptcy. Recognising that teams dominated by foreign imports can distort competitiveness and reduce the pool of players available for international selection, the Ligue Nationale de Rugby recently announced a quota system requiring clubs to field a minimum percentage of French players: 40% in 2010-11; 50% in 2011-12; and 60% in 2012-13. Of note in French rugby is the strong tradition of managing playing resources heavily in favour of home victories.

In Italy, as in France, all but two of the top teams (from 2001, the Super 10) are located in one half of the country, the North. Until the 1980s, Italian club rugby was small scale with relatively low playing standards. From the 1980s, financial incentives brought investment (perhaps most notably from the Benetton fashion company) and foreign stars into the domestic league. A liberal approach to player payment, tacitly ignored by the International Rugby Board, supported this development, such that David Campese, an Australian test star with Italian ancestry playing for Milan in this period, famously declared himself “rugby’s first millionaire” in a period when rugby was still officially amateur. Standards have risen steadily, and recent international successes reflect this. No salary cap is in operation, and this has contributed to frequent changes in the structure and membership of the “Super 10”, with weaker teams either folding, merging with neighbouring clubs or returning to amateur status in lower divisions. The top Italian clubs have struggled to be competitive since their inclusion in European competitions. In an attempt to raise standards, in season 2010-11 two representative teams from a restructured Super 10 were admitted to the Magners League, the elite competition for Celtic regional sides from Wales, Scotland and Ireland.

2.2. *Intra-season competitive balance*

Many competitive balance measures for sports leagues are based on the win percentages and more precisely win proportions. The most well known is the intra-tournament measure the ASD/ISD ratio. The numerator is the standard deviation of the observed win proportions, p_i , of each team: $\{\sum_{i=1}^N (p_i - \frac{1}{2})^2 / N\}^{1/2}$, assuming no ties and each team plays an equal number $M(N-1)$ of matches in M rounds. The denominator is its expected value when outcomes are purely random, $\frac{1}{2} / \sqrt{M(N-1)}$ (this follows because $p_i = w_i / \{M(N-1)\}$ and $w_i \sim \text{bin}(M(N-1), \frac{1}{2})$ if all strengths are equal whence outcomes are purely random), interpreted as the standard deviation of win proportions of competitors in an idealized tournament. This and other similar measures can be regarded as somewhat ad-hoc.

Koning (2000, 2009) instead uses a statistical model and considers competitive balance as a property of such a model. Specifying an explicit model allows us to study the statistical properties of a competitive balance measure, and secondly facilitates criticism of model assumptions. Inevitably the model is an approximation to the reality of a tournament. For a tournament with n teams labelled $i=1,\dots,n$ that play paired matches, the Bradley-Terry model (Bradley and Terry, 1952) is the simplest. This model associates a (fixed) positive quantity α_i with team i (strength) and states that the probability that team i beats team j is given by

$$\Pr(i \text{ beats } j) = \alpha_i / (\alpha_i + \alpha_j). \quad (1)$$

subject to a constraint (e.g. $\alpha_1 = 1$) to ensure a unique set $\{\alpha_i; i=1,\dots,n\}$ determines the individual match outcome probabilities. This model assumes: strengths are fixed over a set of matches (a tournament); the outcomes of matches in a tournament are statistically independent. A model that relaxes the latter assumption would be difficult to build, and to our knowledge has not been attempted, although “carry-over” effects are acknowledged in tournament scheduling (Kendall et al., 2010). There has been recent development of models that allow strengths to vary over time (e.g. Crowder et al., 2002; Owen, 2011), thus relaxing the former assumption; in this context an instantaneous competitive balance measure might be defined.

More general paired comparison models exist. The general linear paired comparison model (Glickman, 2008) assumes that the probability that team i beats team j in match k is given by $\Pr(i \succ j) = F(\theta_{ik} - \theta_{jk})$ where θ_{ik} is the strength of team i in match k and θ_{jk} is the strength of team j in match k , and F is a specified probability distribution function. With F the standard logistic distribution function and $\alpha_i = \exp(\theta_{ik})$ gives the Bradley Terry model. With F the standard normal distribution the Thurstone-

Mosteller model is obtained; a variation of this model was used by Koning (2000) to look at the development of within-season competitive balance in Dutch soccer over time. Stern (1990) proposed a paired comparison model that contains these models as special cases, and has argued that all linear paired comparison models are broadly equivalent (Stern, 1992).

Home advantage can be handled by introducing a (non team-specific) home advantage parameter δ such that when i plays at home against j , $\Pr(i \text{ beats } j) = \delta\alpha_i / (\delta\alpha_i + \alpha_j)$. Home advantage may also be modelled to be team specific.

When outcomes may be tied, the general linear paired comparison model can be extended so that in match k the probability that team i beats team j is $F(\theta_{ik} + \phi_{ijk} - \chi_{ijk} - \theta_{jk})$, and the probability that team j beats team i is $1 - F(\theta_{ik} + \phi_{ijk} + \chi_{ijk} - \theta_{jk})$. The parameter ϕ_{ijk} is the home advantage of team i when they play team j at home in match k . The parameter $\chi_{ijk} \geq 0$ represents the tendency for a draw to occur when team i play team j in match k . With both the home advantage parameter and the draw parameter non-team and non-match specific, with $\delta = \exp(\phi_{ijk})$ and $c = \exp(\chi_{ijk})$ ($c \geq 1$), and F the standard logistic distribution, the extended Bradley-Terry model for the trinomial outcome $Y_{ij}=1$ if i wins, $Y_{ij}=0$ if match tied, $Y_{ij}=-1$ if j wins, is $\Pr(Y_{ij}=1) = \delta\alpha_i / (\delta\alpha_i + c\alpha_j)$, $\Pr(Y_{ij}=0) = (c^2 - 1)\delta\alpha_i\alpha_j / \{(\delta\alpha_i + c\alpha_j)(\delta\alpha_i + \alpha_j)\}$, and $\Pr(Y_{ij}=-1) = \alpha_j / (\delta\alpha_i + \alpha_j)$. Similarly the extended Thurstone-Mosteller model is $\Pr(Y_{ij}=1) = 1 - \Phi(-\phi + \chi - \theta_i + \theta_j)$, $\Pr(Y_{ij}=0) = \Phi(-\phi + \chi - \theta_i + \theta_j) - \Phi(-\phi - \chi - \theta_i + \theta_j)$, and $\Pr(Y_{ij}=-1) = \Phi(-\phi - \chi - \theta_i + \theta_j)$ where ϕ is the common home advantage effect, and $\chi \geq 0$ is the additional parameter (the cut parameter in the ordered probit) that represents the tendency of matches in the tournament to be drawn: when $\chi=0$, $\Pr(Y_{ij}=0)=0$. Ties can of course also be modelled through the indirect approach that models the scores for each competitor in a match (e.g. Maher, 1982 for soccer; Lee, 1999, rugby league; Stefani, 2009, rugby union; Baker and McHale, 2013, NFL). A team's strength is then vector-valued, so that immediate strength comparisons are not facilitated. However, as suggested by Scarf and Yusof (2011), strength estimates may be constructed using the estimated attacking and defensive parameters by simulating a very large, repeated round robin.

Extensions to multiple comparisons are possible to handle "matches" with >2 competitors: the Plackett-Luce model (Farmer, 2003) generalizes the Bradley-Terry model so that the probability of the outcome "competitor r_1 beats competitor r_2 , r_2 beats r_3, \dots , r_{n-1} beats r_n " is $\prod_{k=1,\dots,n} \alpha_{r_k} / (\alpha_{r_k} + \alpha_{r_{k+1}} + \dots + \alpha_{r_n})$. This model implies the existence of a general linear multiple comparison model (Joe, 2001, p.197). The Plackett-Luce model has been used by Baker and McHale (2012) to rate professional golfers. Home advantage can be handled in a standard way. Ties are more difficult to handle in the multivariate case.

Estimation for these models can proceed via maximum likelihood. Interestingly, it is not always necessary to know the results of individual matches to estimate the strength parameters. In fact, for the rugby data that we analyse in this paper, we only have the final league table for each season, showing for each team the number of matches played and the numbers of wins, draws, and losses. If n teams play a double round-robin of $n(n-1)$ matches in total, then for the model (1) the log-likelihood is

$$l = \sum_{i=1}^n w_i \ln \alpha_i - \sum_{i=1}^n \sum_{j \neq i} \ln(\alpha_i + \alpha_j) + \ln c(w_1, \dots, w_n) \quad (2)$$

where w_i is the number of wins for team i ($0 \leq w_i \leq 2(n-1)$, $i=1,\dots,n$, and $c(w_1, \dots, w_n)$ is a combinatorial number being the number of ways that the n competitors could win w_1, \dots, w_n matches, respectively, the exact form of which is not required. The maximum likelihood estimates are then the solutions of the system of equations $\partial l / \partial \alpha_i = 0$, $i=1,\dots,n$, that is of

$$w_i / \hat{\alpha}_i - \sum_{j \neq i} 2 / (\hat{\alpha}_i + \hat{\alpha}_j) = 0. \quad (3)$$

$\hat{\alpha}_i$ is also the moment estimator of α_i . This fact and the symmetry of (3) imply that the estimated strengths are in one to one correspondence with the win percentages if the league is balanced. This suggests there exists an F with a parameter θ in the general linear paired comparison model such that for a balanced tournament the proportion of wins for each team is the maximum likelihood estimate of θ . Extending the likelihood to handle home advantage is possible; with a common home advantage parameter (2) becomes $l = \sum_{i=1}^n w_i \ln \alpha_i - \sum_{i=1}^n \sum_{j \neq i} \ln(\delta\alpha_i + \alpha_j) + \ln c(w_1, \dots, w_n) + z \ln \delta$ where z is the total number of home wins among the $n(n-1)$ matches. Team specific home advantage can also be considered in principle provided the numbers of home wins and away wins is known for each team.

In general, (w_1, \dots, w_n) is a sufficient statistic for $(\alpha_1, \dots, \alpha_n)$ provided outcomes are restricted to win and loss, and the tournament is balanced—that is, if every team plays every other exactly t times. Then (ignoring the constant term) $l = \sum_{i=1}^n w_i \ln \alpha_i - \sum_{i=1}^n \sum_{j < i} t \ln(\alpha_i + \alpha_j)$. Where drawn matches occur, provided they are not too frequent a pragmatic solution is to regard a drawn match as “halved”, so that we count the number of “wins” as $w'_i = w_i + d_i / 2$ where d_i is the number of draws for team i . As (w_1, \dots, w_n) is sufficient for $(\alpha_1, \dots, \alpha_n)$ when the design is balanced, knowing individual match outcomes will not increase the precision of the maximum likelihood estimates. When the design is unbalanced, the likelihood (2) is not valid, and $\{(x_{ij}, m_{ij}), \text{ all } i < j\}$, where m_{ij} is the number of times i beats j in the m_{ij} matches between i and j , is then a sufficient statistic for $(\alpha_1, \dots, \alpha_n)$ (Feinberg and Larntz, 1976). If $\{(x_{ij}, m_{ij}), \text{ all } i < j\}$ is partially known, e.g. we may have $\{(x_{i\bullet}, m_{ij}), \text{ all } i < j\}$ where $x_{i\bullet}$ is the total number of wins for team i , then a missing-value procedure such as the EM algorithm is required to estimate $(\alpha_1, \dots, \alpha_n)$ and the precision of estimation is necessarily reduced. Where only the numbers of wins is observed then a parsimonious model is advantageous; the team specific home advantage model has twice as many parameters as the simplest model. One could imagine that match specific home advantage effects might apply i.e. parameters $\delta_{ij} \quad i \neq j = 1, \dots, n$ may be introduced; multiple seasons would be required to fit such a model, perhaps with individual strengths varying from season to season and individual home advantages fixed over seasons.

The collection of the estimated strengths, e.g. $\hat{\alpha}_i \quad i=1, \dots, n$ for the Bradley-Terry model (1), measured at the season-end is then in this paper the intra-season competitive balance. The estimated strengths might be conveniently represented graphically, such as in a box and whisker plot, or a dot plot. Such a graphical representation would then provide more information about the strength distribution than a single scalar measure such as the ASD/ISD measure which is itself sensitive to outliers (Lenten, 2009, p. 414) or the total deviation from average strength, $\sum_i (\hat{\alpha}_i - \hat{\alpha}_\bullet / n)^2$, where $\hat{\alpha}_\bullet = \sum_i \hat{\alpha}_i$ (Koning, 2000). An advantage of the single scalar measure is that the precision of this measure can also be determined (e.g. using the delta method, Casella and Berger, 2002, p.240) and then variation of the measure and hence variation in competitive balance over time can be interpreted while accounting for statistical uncertainty in the measure.

2.4. Inter-season competitive balance

When considered over multiple seasons, the variability in win proportions can be decomposed into within season and between season components, the latter then used to measure inter-tournament competitive balance (Eckard, 2001a&b, 2003; Humphreys, 2002; Szymanski and Smith, 2002). Other popular measures are based on concentration indices (e.g. Utt & Fort, 2002) which can be used both in a within and between season sense. Lenten (2009) and Williams (2012) attempt to capture the mobility of finishing position of teams from one season to the next. The season to season correlation idea of Koning (2009) aims to do likewise.

From a statistical perspective, a natural measure of persistence from one season to the next is the correlation of performance in season t with performance in season $t+1$ measured across the teams, that is, the correlation ρ_t of $w_{i,t}$ with $w_{i,t+1}$ ($i=1, \dots, n$), where $w_{i,t}$ is the number of wins for team i in season t , with drawn matches halved. The correlation of the estimated strengths might equally be used although it is not equivalent. This is because in general the correlation coefficient is not invariant under a monotonic increasing transformation of the margins. Rank correlation measures (correlations based on ranks) are invariant under such transformations. However the correlation of ranks from one season to next has in our view a significant drawback: in a persistence measure one would wish to put greater weight on persistence in the tails of the rank distribution, that is, at the top and bottom of the table. It is our view that this is natural; as the “mid-table” tends to be congested, the articulation of persistence then focuses on the top and the bottom: do the leaders continue to lead and are the promoted subsequently relegated? In other words, the number of wins is a much better measure of strength than position in the final league table.

If correlation is high (close to 1) then strengths persist, the strong remain strong, and the weak remain weak. In the absence of a draft system we would not expect the correlation to be negative. Strange effects that a correlation measure would fail to pick up (e.g. where both those who do well and those who do poorly in season t might do poorly in season $t+1$) could be observed by representing evolution of strength of each

team over time as a line graph. Longer term persistence could be measured by considering the correlation of $w_{i,t}$ with $w_{i,t+k}$ ($i=1,\dots,n$), for $k>1$.

Uncertainty of the correlation measure of inter-season competitive balance can be quantified by using Monte Carlo simulation. With strengths estimated for season i , the correlation between the number of wins in two simulated seasons can be calculated; this provides a simulated inter-season correlation, assuming strength is persistent (unchanged). Repeating this simulation a large number of times and calculating the correlation each time, percentiles of the simulated correlation distribution under the hypothesis H of persistence can be determined. A Monte Carlo test of H then compares the actual observed between season correlation with the Monte Carlo lower percentile, $\tilde{\rho}_{t,\alpha}$, rejecting H at the 5% level, say, if $\rho_t < \tilde{\rho}_{t,0.05}$.

2.5. Results

Strength estimates, for the Bradley-Terry model (1) fitted using the likelihood (2) with drawn matches halved, for each of the England, France, and Italy rugby union premierships are shown in figure 1. In the France, 4 team-playoff results have been not been included. Also in 01/02—03/04 in France a design was used in which the top 8 of 16 after 14 matches played-off in a hybrid tournament—these seasons are omitted. The strength of the strongest team has been set to unity. From these graphs, one can roughly calculate probabilities of a win when the strongest plays the median strongest, and the strongest best plays the lower quartile strongest. Broadly, intra-season competitive balance is the same in each league, with no particular trends over time in any league, although in Italy over the seasons 2001/2 to 2004/5 the tournaments are unbalanced to the greatest extent and competitive balance in England has appeared to increase from 2000/1 through to 2009/10. The former observation is apparent in the line graphs of team strengths, figure 2. Williams (2012) provides a qualitative discussion of these matters and their underlying causes.

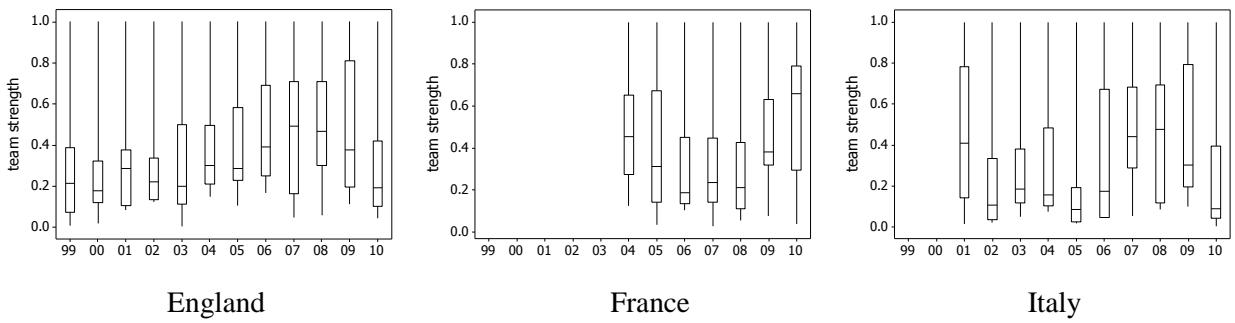


Figure 1. Boxplots of maximum likelihood estimates of strengths for seasons 1999/2000 through to 2010/11.

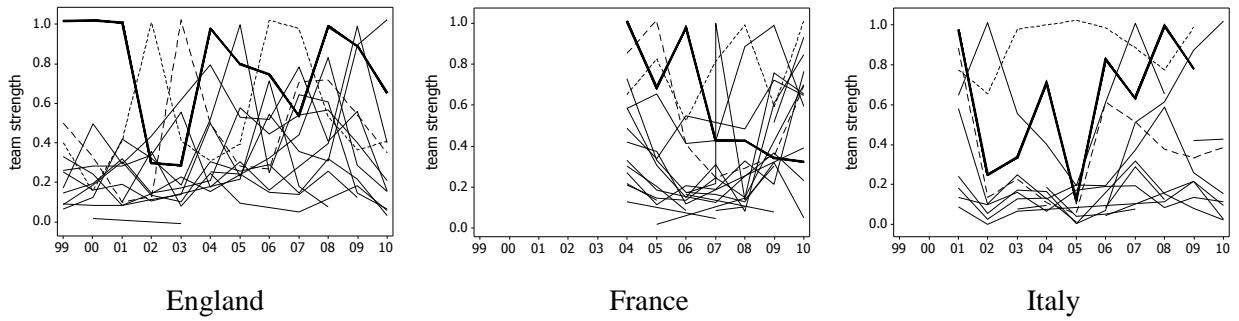


Figure 2. Line graphs of maximum likelihood estimates of strengths by club team: top 3 in 1999/2000 season indicated differently (e.g. England: 1st, bold, Leicester; 2nd, dashed, Bath; 3rd, dotted, Gloucester).

The inter-season competitive balance measure, the between season correlation, is shown for each league in figure 3. Simulated median, and lower 5% and 1% percentiles under the hypothesis H of persistence are also shown and these indicate that there is some persistence in France and in Italy but markedly none in the England premiership. It is perhaps noticeable that the domination of a league by a small of teams (suggested by the line graph for England, figure 2) is not captured by the between-season correlation (figure 3). It would

have been interesting to extend these plots further back in time in order to address the questions regarding the effects of significant policy changes. However, changes in the tournament formats make this difficult.

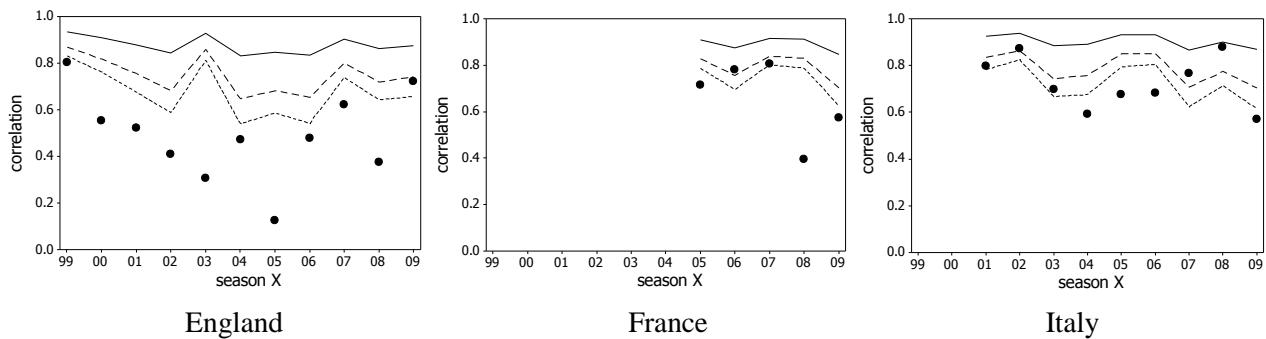


Figure 3. Pearson correlation of number of wins in season X with number of wins in season $X+1$ (solid symbols) with simulated median correlation (solid) and lower 5% (dashed) and lower 1% (dotted) percentiles when the strengths of teams are unchanged from season X to season $X+1$.

3. Discussion

This paper reviews some measures of competitive balance for league tournaments, for within season and between seasons. Two measures are adopted: the distribution of team strengths for measuring within or intra-season competitive balance; and the season to season correlation of the number of wins for between or inter-season competitive balance. These measures are illustrated for the England, France and Italy rugby union premierships. They are calculated using the Bradley Terry paired comparison model that is fitted to data contained in the end-of season league tables rather than data on individual match results. The intra-season competitive balance for each league appears to be broadly the same. Inter-season competitive balance appears to be greater in England than in France and Italy although there is some indication of persistent domination of the league in England by one or two teams in particular.

This paper has concentrated on two dimensions of competitive balance: inter-season and intra-season. The other dimension, that of individual match competitive balance, is not addressed. As a final aside, if one wanted to make individual matches in rugby more competitive, administrators might simply reduce the number of scoring events. The evidence for this is discussed in Koning (2000): “soccer results are more random...as only a few goals are scored each game and chance may be quite influential in determining outcome”. However a systematic study of the relationship between uncertainty of direct outcome and the scoring systems across sports has not been carried out, so that the extent to which competitive balance is influenced by the scoring system is an open question. It is our contention however that if in rugby union one were to award 1 point for a try as the only means of scoring then match outcome uncertainty would increase significantly.

Acknowledgements

The authors are grateful for the discussion of this work with Professor Rose Baker and for her contribution to the development of the likelihood for the tournament outcome when only the number of wins and draws for each team is known.

References

- Baker R.D. and McHale I. (2012) How good is Tiger Woods? *Mathematics Today*, June 2012, 124-126.
- Baker R.D. and McHale I. (2013) Forecasting exact scores in National Football League games: a birth process model of exact scores. *International J. Forecasting* 29, 122-130.
- Bradley R.A. and Terry M.E. (1952) The rank analysis of incomplete block designs 1. The method of paired comparisons. *Biometrika* 39, 324-345.
- Buraimo B. and Simmons R. (2008) Do Sports fans really value uncertainty of outcome? evidence from the English Premier League. *International Journal of Sport Finance* 3, 146-155.
- Casella and Berger (2002) *Statistical Inference*, 2nd ed. Duxbury, USA.

- Crowder M., Dixon M.J., Ledford L. and Robinson M. (2002). Dynamic modelling and prediction of English Football League matches for betting. *The Statistician* 51, 157-168.
- Eckard E. (2001a) Baseball's blue ribbon economic report: Solutions in search of a problem. *Journal of Sports Economics* 2, 213-227.
- Eckard E. (2001b) Free Agency, competitive balance, and diminishing returns to pennant contention. *Economic Enquiry* 39, 430-443.
- Eckard E (2003) The ANOVA-based competitive balance measure: A defense. *Journal of Sports Economics* 4, 74-80.
- Farmer C.J. (2003) Probabilistic modelling in multi-competitor games. MSc Thesis, University of Edinburgh.
- Feinberg S.E. and Larntz K. (1976) Log linear representation of paired and multiple comparisons models. *Biometrika* 63, 245-54.
- Forrest D. and Simmons R. (2002) Outcome uncertainty and attendance demand in sport: the case of English soccer. *The Statistician* 51, 229-241.
- Glickman M.E. (2008) Bayesian locally optimum design of a knockout tournament. *Journal of Statistical Planning and Inference* 21, 2117–2127.
- Humphreys B.R. (2002) Alternative measures of competitive balance in sports leagues. *Journal of Sports Economics* 3, 133-148.
- Kendall G., Knust S., Ribeiro C.C. and Urrutia S. (2010) Scheduling in sports: An annotated bibliography, *Computers & Operations Research* 37, 1-19.
- Joe H. (2001) *Multivariate Models and Dependence Concepts*. Chapman and Hall, London.
- Koning R.H. (2000) Balance and competition in Dutch soccer. *The Statistician* 49, 419-431.
- Koning R.H. (2009) Sport and measurement of competition. *De Economist* 157, 229-249.
- Lee A. (1999) Modelling rugby league data via bivariate negative binomial regression. *Australian & New Zealand Journal of Statistics* 41, 141-152.
- Lenten L.J.A. (2009) Towards a new dynamic measure of competitive balance: a study applied to Australia's two major professional "football" leagues. *Economic Analysis & Policy* 39, 407-428.
- Maher M.J. (1982) Modelling association football scores. *Statistica Neerlandica* 36, 109-118.
- McGarry T (1998) On the design of sport tournaments. In: *Statistics in Sport* (J. Bennett, ed.), Edward Arnold, London, pp. 199–217.
- Owen A.J. (2011) Dynamic Bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. *IMA J Management Mathematics* 22, 99-113.
- Quirk J. and Fort R. (1992) *Pay Dirt: The Business of Professional Team Sports*. Princeton: Princeton University Press.
- Scarf P.A. and Yusof M.M. (2011) A numerical study of tournament structure and seeding policy for the soccer World Cup Finals. *Statistica Neerlandica* 65, 43-57.
- Scully G.W. (1989) *The Business of Major League Baseball*. University of Chicago Press, Chicago.
- Stefani R.T. (2009) Predicting score difference versus score total in rugby and soccer. *IMA J Management Mathematics* 20, 147-158.
- Stern H. (1990) A continuum of paired comparisons models. *Biometrika* 77, 265-73.
- Stern H. (1992) Are all linear paired comparison models empirically equivalent? *Mathematical Social Sciences* 23, 103-117
- Szymanski S. (2003) The economic design of sporting contests. *Journal of Economic Literature* 41, 1137-1187.
- Szymanski S. and Smith R. (2002) Equality of opportunity and equality of outcome: static and dynamic competitive balance in European and North American sports leagues. In *Transatlantic Sports: the Comparative Economics of North American and European Sports* (Barros, Ibrahim, and Szymanski, eds.), Edward Elgar, USA, pp. 109-124.
- Utt J. and Fort R. (2002) Pitfalls to measuring competitive balance with Gini coefficients. *Journal of Sports Economics* 3, 367-373.

Williams P. (2012) Any given Saturday: competitive balance in elite English rugby union. *Managing Leisure* 17, 88-105.

It is Harder, not Easier, to Predict the Winner of the Champions League

Jeroen Schokkaert* and Johan Swinnen

LICOS Centre for Institutions and Economic Performance, KU Leuven, Jeroen.schokkaert@kuleuven.be

Abstract. European Cup football has experienced a major change in format with the introduction of the Champions League in 1992 and a major change in admission rules with direct qualification for multiple teams from the highest ranked leagues in 1999. We show that, in line with popular press reports and other studies, qualification in lower rounds has become more predictable in the Champions League. At the same time, however, outcomes at later stages have become less predictable. We provide evidence and an explanation.

1. Introduction

The major European competition among football (soccer) teams, the Champions League, is the most prestigious football club competition in the world. The Champions League is organized as a successor of the European Champion Clubs' Cup – often simply referred to as the “European Cup” – in which national league winners competed in a knockout tournament from its inception in 1955.

Both the popular press and academic studies argue that the change from the European Cup (EC) to the Champions League (CL) has made the CL less exciting and its outcomes more predictable (see e.g. Haan et al., 2002, 2012; Milanovic, 2005; Cross, 2009; Fisher, 2012; Gall, 2012). For example, Milanovic (2005) finds that in the CL the same teams are more likely to qualify for the quarterfinals.

However, the argument that the CL's outcomes are more predictable seems inconsistent with ad hoc observations on the history of EC and CL winners (see Table 1). Several teams were able to win the EC several times in a row. In fact, in 13 out of 37 seasons, the winner of the EC was the winner of the previous season. In contrast, not a single team has been able to win the CL twice in a row.¹ To put it differently, if one would bet on the winner of the championship and put his money on the winner of the previous championship, the chances were 35% that one would win in the EC and 0% in the CL. This observation seems inconsistent with the argument that the CL rules have made the major European competition less exciting and more predictable.

In this paper we first confirm – consistent with Milanovic's (2005) findings – that it is indeed easier to predict who will qualify for lower knockout rounds, such as the round of 16 and the quarterfinals, under the current CL rules. However we also show that the uncertainty of who wins in the competition increases beyond these stages with the CL. So it is harder, not easier, to predict the winner.

We provide an explanation for both empirical findings. Our argument is that the group-round and the new qualification rules of the CL, allowing multiple teams from the highest ranked leagues to directly participate, make it more likely that these teams qualify for the round of 16 – and less likely that teams from lower ranked leagues progress. This is consistent with the earlier argument of less excitement and more predictability. However, we argue further that the same rules also cause smaller quality differences between teams in later rounds of the tournament, making the outcomes of the final rounds less certain and less predictable.

¹ Admittedly, whereas 37 EC editions were organized, only 20 CL editions have been organized up to now. We return to this issue when discussing our results.

The paper is organized as follows. Section 2 discusses the main differences between the CL and the EC. Section 3 provides the theoretical predictions. Section 4 comments the data, discusses the empirical indicator we use and presents the results. Section 5 concludes.

2. Main Differences between the Champions League and the European Cup

The CL and the EC differ in two aspects: the tournament format and the admission rules (as summarized in Table 2). The EC was established by the UEFA (Union of European Football Associations) in 1955 as a pure knockout tournament between the champions of European national leagues (Granville, 1991).² However, due to complaints from the teams from higher ranked leagues that the knockout format of the EC favored the teams from lower ranked leagues while the value of this competition was dependent on the teams from higher ranked leagues, which attract larger (television) audiences, UEFA introduced a mini-league system into the format of the EC in the 1991/92 season (see e.g. King, 2004; Holt, 2007). At the start of the 1992/93 season, UEFA ratified this change of format and renamed the tournament the CL.

The CL initially implied a change from an (unseeded) knockout tournament to a hybrid tournament combining a (seeded) round robin tournament in which groups of four teams compete to determine qualifiers and a (seeded) knockout tournament between the qualified teams (Scarf et al., 2009).³

Later, UEFA also changed the admission rules. As in the EC, the first editions of the CL included only national league champions (and the titleholder). From 1999 onwards multiple teams from the highest ranked leagues were admitted. Since the 1999/2000 season, the runners-up from the six highest ranked leagues, assigned according to the official UEFA coefficients, also directly qualify for the group-round.⁴ Since the 2009/10 season also the third-placed teams of the three highest ranked countries qualify directly.

3. Theoretical Predictions

The operational research literature uses simulation methods to compare match or tournament outcome uncertainty under different tournament formats (see e.g. Appleton, 1995; McGarry and Schutz, 1997; Scarf et al., 2009; Scarf and Yusof, 2011; Koning and McHale, 2012). For this purpose prediction models are used to simulate the expected outcomes of games between two teams. This can then be extended to a complete tournament to obtain estimates for different indicators such as a team's probability to qualify for certain tournament rounds, the probability that the team with the highest ranking before the tournament wins the tournament, the correlation between a team's pre- and post-tournament ranking, etc.

The tournament format influences these tournament outcomes. Scarf et al. (2009) simulate CL outcomes for 11 different possible tournament formats. The simulated formats include the “unseeded 2 leg knockout” structure (which was the format of the EC) and the “seeded 1 group-round and 2 leg seeded knockout in later rounds” structure (the current format of the CL). The simulation estimates show that the

² Except for the first edition, where the organizers decided on the playing schedule, the schedule was determined by random draw (UEFA, 2004).

³ If tournaments are seeded, highly ranked teams play lowly ranked teams in earlier rounds, which maximizes the probability that the highly ranked teams qualify for later rounds (see e.g. Noll, 2003; Monks and Husch, 2009).

⁴ The UEFA coefficients are calculated based on the performance of teams from each country in the main European club competitions, the CL and the Europa League. In general, each participating team gets two points for a win, one point for a draw and some bonus points for qualifying for later tournament rounds. The UEFA coefficient assigned to a country is the sum of points obtained by all the participating teams from that country divided by the number of those teams.

average ranking of the teams that qualify for the different knockout rounds in later rounds is higher for the current CL format than for the EC format. This suggests that on average the highly ranked teams who participate in the tournament are more likely to qualify for the round of 16 in the CL than in the EC, and vice versa for lowly ranked teams.

These simulation results do not take into account the effect of UEFA's decision to allow multiple teams from the highest ranked leagues to qualify directly for the CL (compared to only one team per league for the EC). Because of this decision it is more likely that a specific highly ranked team from one of the highest ranked leagues qualifies year after year since they qualify even if they do not win the title in their national league but end second (or third). Additionally, it is more difficult for a team from a lower ranked league to qualify since only the champions qualify for their national leagues (as in the EC) and since fewer spots remain to directly qualify for the group-round.

These two effects from UEFA's decision to allow multiple teams from the highest ranked leagues to qualify directly for the CL suggests that a specific highly ranked team from one of the highest ranked leagues is more likely to qualify for the round of 16 in the CL than in the EC, and vice versa for a team from a lower ranked league.

However, another effect of UEFA's decision to allow multiple teams from the highest ranked leagues to qualify directly for the CL (compared to only one team per league for the EC) and of the knockout format versus the hybrid format is that the average ranking of the qualified teams is higher, and that the quality difference between the teams is smaller at later stages in the tournament. This implies that while it is less likely that a team from a lower ranked league progresses to the round of 16, it is also more difficult for a specific highly ranked team from one of the highest ranked leagues to progress at later stages in the tournament.

In sum, the introduction of the group-round and the rule change towards direct qualification for multiple teams from the highest ranked leagues has made it more likely for a specific highly ranked team from one of the highest ranked leagues to qualify for the round of 16 in the CL than in the EC and – at the same time – less likely to qualify for later stages in the CL than in the EC. Hence, we predict that on average it is easier to predict who will qualify for the lower rounds in the CL than in the EC and harder to predict who will qualify for later stages in the CL than in the EC.

4. Empirical Evidence

Table 3 and Table 4 present summary statistics on the number of teams that reached various stages in the EC and the CL. More detailed data on all teams and their performance in the EC and the CL can be found in Appendix.

192 teams reached the round of 16 in the EC at least once over the 1955-1991 period. For the CL this was 84 teams over the 1992-2011 period. What matters, of course, is the average number. Comparing the number of teams that reached the round of 16 and the quarterfinals over five-year periods in the EC with the CL, we observe that these numbers are substantially lower in the CL than in the EC. Over five-year periods the number of round of 16 participants was almost always at least 50 in the EC. Except for the first CL editions, this number decreased to less than 40 in the CL. Similarly, while over five-year periods on average 29 teams have reached the quarterfinals in the EC, except for the first CL editions not more than 22 teams have been able to reach this round in the CL.

The difference in the number of qualified teams between the CL and the EC is smaller and even reversed at later stages in the tournament. On average 15 teams have reached the semifinals in a five-year period in the EC, compared to 12 in the CL. Over five-year periods, on average seven teams have reached the finals in the EC and six teams in the CL. Finally, on average more different teams (4) won the CL than the EC (3).

To provide better evidence for our theoretical predictions, we calculate an indicator which is commonly used in the literature, i.e. “uncertainty of outcome”. The literature identifies three different levels of uncertainty of outcome (UO): for a match, a season and a championship (see e.g. Szymanski, 2003; Buzzacchi et al., 2003; Goossens, 2006). Most studies have focused on uncertainty of outcome for a particular match or season.

The focus of our study is on UO in the CL and the EC, which is a “dynamic” measure, i.e. it measures particular teams’ dominance of a championship and the predictability of the winners *across* seasons. Following Hadley et al. (2005) and Pawlowski et al. (2010), we use a Markov model to compare UO between two periods.⁵ In a Markov process the outcome at time $t + 1$ is determined by the state at time t (Krautmann and Hadley, 2006). The probability that a particular entity i will transition from one state at time t , $S_{i,t}$, to another state at time $t + 1$, $S_{i,t+1}$, is called the “transitional probability”. In our case the entity is a football team and the two states refer to qualification for a particular tournament round: either team i qualified or team i did not qualify. For each knockout round, we assign a dummy variable equal to 1 if team i qualified and equal to 0 if team i did not qualify. For each of the dummy variables, four transitional probabilities can be calculated (see e.g. Koop, 2003):

$$p_{00} = \Pr(S_{i,t+1} = 0/S_{i,t} = 0); \quad (1)$$

$$p_{01} = \Pr(S_{i,t+1} = 1/S_{i,t} = 0); \quad (2)$$

$$p_{10} = \Pr(S_{i,t+1} = 0/S_{i,t} = 1); \quad (3)$$

$$p_{11} = \Pr(S_{i,t+1} = 1/S_{i,t} = 1), \quad (4)$$

where p_{00} is the probability that team i repeats in not qualifying across two seasons, p_{01} the probability that team i goes from not qualifying in one season to qualifying in the next season, p_{10} the probability that team i goes from qualifying in one season to not qualifying in the next season and p_{11} the probability that team i repeats in qualifying across two seasons.

Table 5 presents the UO indicators and Figure 1 shows the difference in the UO indicator between the CL and the EC for the various knockout rounds and for winning the tournament. Figure 1 shows that the difference in the UO indicator between the CL and the EC is negative for the round of 16, the quarterfinals and the semifinals. On the other hand, the difference in the UO indicator between the CL and the EC is positive for the finals and for winning the tournament. These empirical indicators are consistent with our hypotheses that it is easier to predict who will qualify for lower rounds in the CL than in the EC and that, at the same time, it is harder to predict who will qualify for later stages in the CL than in the EC.⁶

Next, we can try to see which of the rule changes under the CL was crucial in causing the effects. As we explained in Section 2, the current CL implied two key changes compared to the EC: the tournament format and the admission rules. The second change (direct qualification for multiple teams from the highest ranked leagues) was implemented only in 1999 when the CL was already in place for several years. Figure 2 shows the difference in UO indicators between the CL after 1999 and before 1999. There is a strong relationship. The difference in the UO indicator is negative for the round of 16 and the quarterfinals. There is almost no difference in the semifinals and the difference in the UO indicator is

⁵ Hadley et al. (2005) use the Markov model to compare a team’s probability of qualifying for postseason play in Major League Baseball (MLB), the highest level of professional American baseball in the United States, before and after the players’ strike in 1994. Pawlowski et al. (2010) use the Markov model to compare a top team’s probability of qualifying for the CL group-round before and after the change of the distribution system of CL revenues in 1999.

⁶ UEFA introduced the group-round in the 1991/92 season but officially renamed the tournament the CL in the 1992/93 season. Our results are robust to switching the data from the 1991/92 season between the EC era and the CL era.

positive for the finals. This suggests that it is in particular the change in the admission rules that makes it easier to predict who will qualify for lower rounds in the CL and harder to predict who will qualify for later stages in the CL.

Finally, as Eckard (1998, 2001) and Pawlowski et al. (2010), we also compare our empirical indicator over periods of equal length to check whether our results are not driven by the longer sample period of the EC. We split the EC sample into two periods (practically) equal to the CL period: the 1955-1973 period and the 1973-1991 period. Table 5 shows that there are only small differences in the UO indicator between the two EC periods for the round of 16, the quarterfinals and the semifinals. There is a larger difference in the UO indicator for the finals and for winning the tournament, with the UO indicator lower for the early period. However, regardless of the EC period under comparison, uncertainty of the final outcome of the tournament is considerably higher in the CL than in the EC.⁷

5. Conclusion

In this paper we empirically examine how the change in format with the introduction of the Champions League in 1992 and the change in admission rules with direct qualification for multiple teams from the highest ranked leagues in 1999 affected uncertainty of outcome in European Cup football.

We first explain how the introduction of the group-round and the new admission rules has made it more likely for a specific highly ranked team from one of the highest ranked leagues to qualify for the round of 16 in the Champions League than in the European Cup. This is consistent with the existing argument of less excitement and more predictability in the Champions League. However, we also argue that the same rules cause smaller quality differences between teams in later rounds of the tournament. This has made it less likely for a specific highly ranked team from one of the highest ranked leagues to qualify for later stages in the Champions League than in the European Cup, making the outcomes of the later rounds in the Champions League less certain and less predictable.

Through a comparison of differences in uncertainty of outcome between the Champions League and the European Cup and between the Champions League after 1999 and before 1999, we confirm that it indeed is easier to predict who will qualify for lower rounds such as the round of 16 and the quarterfinals in the Champions League. However, we also show that it is harder to predict who will qualify for later stages such as the finals and who will win the tournament. So it is harder, not easier, to predict the winner of the Champions League.

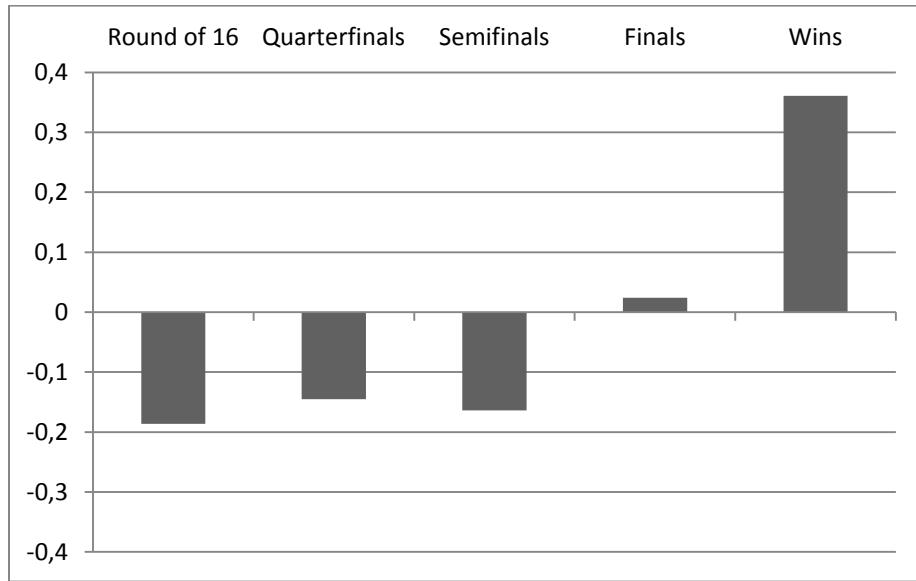
References

- Appleton, D.R. (1995). May the Best Man Win?. *Journal of the Royal Statistical Society: Series D*, 44(4), 529-538.
- Buzzacchi, L., Szymanski, S., & Valletti, T.M. (2003). Equality of Opportunity and Equality of Outcome: Open Leagues, Closed Leagues and Competitive Balance. *Journal of Industry, Competition and Trade*, 3(3), 167-186.
- Cross, J. (2009). Why Successful English Teams Have Made the Champions League Predictable and Boring. Article Available at <http://www.mirrorfootball.co.uk>, 12 March 2009.
- Eckard, W.E. (1998). The NCAA Cartel and Competitive Balance in College Football. *Review of Industrial Organization*, 13(3), 347-369.
- Eckard, W.E. (2001). Free Agency, Competitive Balance, and Diminishing Returns to Pennant Contention. *Economic Inquiry*, 39(3), 430-443.

⁷ Also the results of the difference in uncertainty of outcome between the CL after and before 1999 are robust to comparing our empirical indicator over periods of equal length. Results are available upon request.

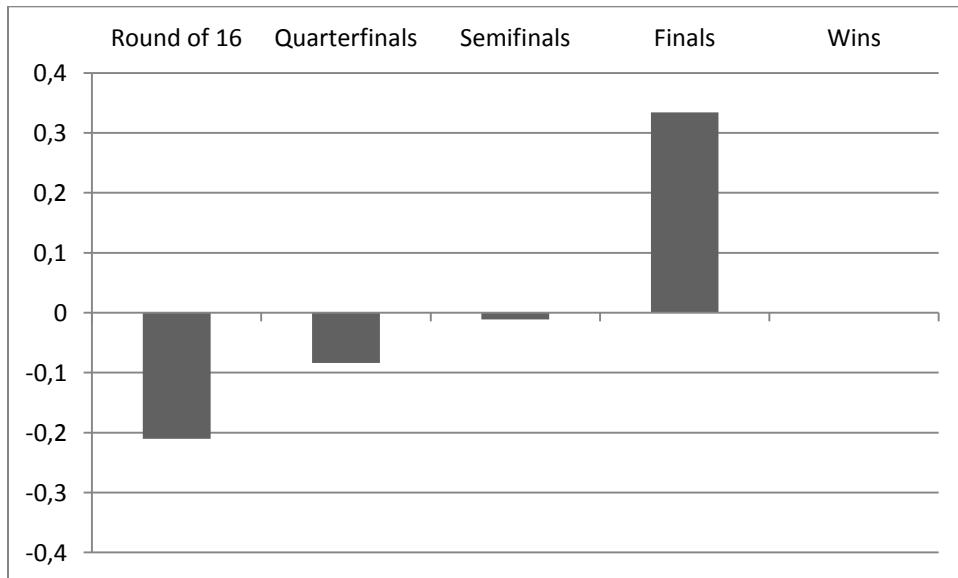
- Fisher, S. (2012). What Would an Expanded Champions League Mean for Scottish Clubs. Article Available at <http://www.heraldscotland.com>, 29 November 2012.
- Goosens, K. (2006). Competitive Balance in European Football: Comparison by Adapting Measures: National Measure of Seasonal Imbalance and Top 3. *Rivista Di Diritto Ed Economia Dello Sport*, 2(2), 77-122.
- Granville, B. (1991). *Champions of Europe*. Enfield: Guinness Publishing.
- Haan, M.A., Koning, R.H., & van Witteloostuijn, A. (2002). Market Forces in European Soccer. *SOM Research Report*, Faculty of Economics, Groningen.
- Haan, M.A., Koning, R.H., & van Witteloostuijn, A. (2012). The Effect of Institutional Change in European Soccer. *Journal of Economics and Statistics*, 232(3), 318-335.
- Hadley, L., Ciecka, J., & Krautmann, A.C. (2005). Competitive Balance in the Aftermath of the 1994 Players' Strike. *Journal of Sports Economics*, 6(4), 379-389.
- Holt, M. (2007). The Ownership and Control of Elite Club Competition in European Football. *Soccer & Society*, 8(1), 50-67.
- Gall, J. (2012). The Problem With Football Today: Europe. Article Available at <http://www.dailysensible.co.uk>, 6 March 2012.
- King, A. (2004). The New Symbols of European Football. *International Review for the Sociology of Sport*, 39(3), 323-336.
- Koning, R.H., & McHale, I. (2012). Estimating Match and World Cup Winning Probabilities. In W. Maennig & A. Zimbalist (Eds.), *International Handbook on the Economics of Mega Sporting Events* (pp. 177-193). Cheltenham and Northampton: Edward Elgar.
- Koop, G. (2003). Modelling the Evolution of Distributions: An Application to Major League Baseball. *Journal of the Royal Statistical Society: Series A*, 167(4), 639-655.
- Krautmann, A.C., & Hadley, L. (2006). Dynasties versus Pennant Races: Competitive Balance in Major League Baseball. *Managerial and Decision Economics*, 27(4), 287-292.
- McGarry, T., & Schutz, R.W. (1997). Efficacy of Traditional Sport Tournament Structures. *The Journal of the Operational Research Society*, 48(1), 65-74.
- Milanovic, B. (2005). Globalization and Goals: Does Soccer Show the Way?. *Review of International Political Economy*, 12(5), 829-850.
- Monks, J., & Husch, J. (2009). The Impact of Seeding, Home Continent and Hosting on FIFA World Cup Results. *Journal of Sports Economics*, 10(4), 391-408.
- Noll, R. (2003). The Organization of Sports Leagues. *Oxford Review of Economic Policy*, 19(4), 530-551.
- Pawlowski, T., Breuer, C., & Hovemann, A. (2010). Top Clubs' Performance and the Competitive Situation in European Domestic Football Competitions. *Journal of Sports Economics*, 11(2), 186-202.
- Scarf, P.A., & Yusof, M.M. (2011). A Numerical Study of Tournament Structures and Seeding Policy for the Soccer World Cup Finals. *Statistica Neerlandica*, 65(1), 43-57.
- Scarf, P., Yusof, M.M., & Bilbao, M. (2009). A Numerical Study of Format for Sporting Contests. *European Journal of Operational Research*, 198(1), 190-198.
- Szymanski, S. (2003). The Economic Design of Sporting Contests. *Journal of Economic Literature*, 41(4), 1137-1187.
- UEFA (2004). *50 Years of the European Cup*. Nyon: UEFA.
- UEFA (2012). *Regulations of the UEFA Champions League 2012-15 Cycle*. Nyon: UEFA.

Figure 1: Differences in uncertainty of outcome between the Champions League (1992-2011) and the European Cup (1955-1991)



Notes: (i) The Figure shows the difference in the teams' average probability of qualifying for the different knockout rounds (or winning) in one season and not qualifying (or not winning) in the next season between the Champions League and the European Cup. A positive difference is associated with higher uncertainty of outcome in the Champions League and vice versa.

Figure 2: Differences in uncertainty of outcome between the Champions League after 1999 (1999-2011) and before 1999 (1992-1998)



Notes: (i) The Figure shows the difference in the teams' average probability of qualifying for the different knockout rounds (or winning) in one season and not qualifying (or not winning) in the next season between the Champions League after and before 1999. A positive difference is associated with higher uncertainty of outcome after 1999 and vice versa. (ii) No club has won the Champions League twice in a row, such that the difference in uncertainty of outcome between the Champions League after 1999 and before 1999 equals 0 for this stage.

Table 1: European Cup and Champions League winners

European Cup		Champions League	
Season	Team	Season	Team
1955/56	Real Madrid	1992/93	Olympique Marseille
1956/57	Real Madrid	1993/94	AC Milan
1957/58	Real Madrid	1994/95	Ajax
1958/59	Real Madrid	1995/96	Juventus
1959/60	Real Madrid	1996/97	Borussia Dortmund
1960/61	Benfica	1997/98	Real Madrid
1961/62	Benfica	1998/99	Manchester United
1962/63	AC Milan	1999/2000	Real Madrid
1963/64	Internazionale	2000/01	Bayern München
1964/65	Internazionale	2001/02	Real Madrid
1965/66	Real Madrid	2002/03	AC Milan
1966/67	Celtic	2003/04	FC Porto
1967/68	Manchester United	2004/05	Liverpool
1968/69	AC Milan	2005/06	FC Barcelona
1969/70	Feyenoord	2006/07	AC Milan
1970/71	Ajax	2007/08	Manchester United
1971/72	Ajax	2008/09	FC Barcelona
1972/73	Ajax	2009/10	Internazionale
1973/74	Bayern München	2010/11	FC Barcelona
1974/75	Bayern München	2011/12	Chelsea
1975/76	Bayern München		
1976/77	Liverpool		
1977/78	Liverpool		
1978/79	Nottingham Forest		
1979/80	Nottingham Forest		
1980/81	Liverpool		
1981/82	Aston Villa		
1982/83	Hamburger SV		
1983/84	Liverpool		
1984/85	Juventus		
1985/86	Steaua Bucuresti		
1986/87	FC Porto		
1987/88	PSV Eindhoven		
1988/89	AC Milan		
1989/90	AC Milan		
1990/91	Red Star Belgrade		
1991/92	FC Barcelona		

Notes: (i) Data are gathered from UEFA.

Table 2: Main changes in tournament format and admission rules

Period	Tournament format	Directly qualified teams
1955-90	knockout	1 st from each country
1991-93	knockout, 1 group-round	1 st from the majority of countries
1994-1998	1 group-round, knockout	1 st from country ranked 1-8
1999-2002	2 group-round, knockout	1 st from country ranked 1-10, 2 nd from 1-6
2003-08	1 group-round, knockout	1 st from country ranked 1-10, 2 nd from 1-6
From 2009	1 group-round, knockout	1 st from country ranked from 1-12, 2 nd from 1-6, 3 rd from 1-3

Notes: (i) Tournament format refers to the tournament excluding qualifying or preliminary rounds. (ii) Before the 1966/67 season, preliminary rounds coincided with first rounds. (iii) Except for the final, games have always been played in two legs – one home game and one away game for each team. (iv) The group-round in the 1993/94 season was followed by a knockout round between semifinalists. (v) From the inception of the EC, the titleholder also qualifies directly for next season's campaign. See UEFA (2012) for implications for other teams from the same country if the titleholder does not qualify through its national competition. (vi) The number of UEFA countries increased substantially during the 1991-93 period because of the dissolution of socialist political entities like the Soviet Union, Yugoslavia and Czechoslovakia.

Table 3: Number of teams that reached various stages in the European Cup

Round	Period	1955-91	1955-91 5-year average	1955-	1960-	1965-	1970-	1975-	1980-	1985-
Round of 16	91	192	54	56	56	60	55	46	52	50
Quarterfinals		112	29	30	31	29	29	27	30	26
Semifinals		62	15	13	16	15	15	16	16	13
Finals		33	7	5	6	9	7	8	7	7
Wins		19	3	1	3	5	2	3	4	4

Notes: (i) The theoretical maximum number of teams reaching the various stages respectively equals 80, 40, 20, 10 and 5. (ii) Data from the 1990/91 season and from the 1991/92 season are excluded from the last column for comparability reasons.

Table 4: Number of teams that reached various stages in the Champions League

Round	Period	1992-2011	1992-2011 5-year average	1992-	1997-	2002-	2007-
Round of 16	2011	84	39	50	38	33	37
Quarterfinals		50	23	29	22	20	22
Semifinals		27	12	13	11	14	10
Finals		12	6	6	6	7	5
Wins		13	4	5	3	4	4

Notes: (i) The theoretical maximum number of teams reaching the various stages respectively equals 80, 40, 20, 10 and 5. (ii) There was no semifinal during the 1992/93 season and no round of 16 during the 1997/98 season and the 1998/99 season.

Table 5: Uncertainty of outcome in the Champions League and the European Cup

Round	Period	Champions League			European Cup		
		1992-2011	1992-98	1999-2011	1955-91	1955-73	1973-91
Round of 16		0,535	0,648	0,438	0,721	0,712	0,725
Quarterfinals		0,612	0,667	0,583	0,757	0,771	0,743
Semifinals		0,635	0,636	0,625	0,799	0,806	0,792
Finals		0,816	0,583	0,917	0,792	0,75	0,833
Wins		1	1	1	0,639	0,556	0,722

Notes: (i) The Table shows the teams' average probability of qualifying for the different knockout rounds (or winning) in one season and not qualifying (or not winning) in the next season for various periods of the Champions League and the European Cup. A higher average probability is associated with higher uncertainty of outcome and vice versa.

Appendix

Table A.1: Performance of teams in the European Cup (for teams which qualified at least once for the round of 16 in the EC)

Team	Round of 16	Quarterfinals	Semifinals	Finals	Wins
SK Tirana	88;89				
Arat Yerevan		74			
Austria Wien	61;62;81;85;86	84	78		
Gwardia Warsaw	55				
Rapid Wien	56;64;67;82;87	55;68;83	60		
Wacker Innsbruck	89;90	77			
Wiener Sportklub		58;59			
Anderlecht	55;64;66;67;68;72	62;65;74;86;87;91	81;85		
Club Brugge	73;88;90	76		77	
KV Mechelen		89			
Royal Antwerp	57				
RWD Molenbeek	75				
SK Beveren	84				
Standard Liège	70;82;83	58;69;71	61		
Dinamo Minsk		83			
FK Sarajevo	67				
CDNA Sofia	58;60;62;71;72;82;83;90	56;73;80;89	66;81		
Levski Sofia	65;77;84				
Lokomotiv Sofia	64;78				
Spartak Plovdiv	63				
Hajduk Split	74	75;79			
Apoel Nicosia	86				
Apollon Limassol	91				
Omonia Nicosia	72;79;85;87				
Banik Ostrava	76;81	80			
Bohemians Praha	83				
Dukla Praha	57;58;64;79	61;62;63	66		
Hradec Kralove		60			
Sparta Praha	87;89	65;67;84;91			
TJ Vitkovice	86				
Zbrojovka Brno	78				
AB København	68				
AGF Aarhus	55;57;87	60			
B1903 København	77				
B1909 Odense	59				
B1913 Odense	61				
Brøndby IF	91	86			
Esbjerg fB	62;80				
Hvidovre IF	67				
KB København	81				
Lyngby BK	84				
Vejle BK	73;79				
Arsenal	91	71			
Aston Villa		82			81
Burnley		60			
Derby County	75		72		
Everton		70			
Ipswich Town	62				
Leeds United			69	74	76;77;80;83
Liverpool	66;73	81;82	64	84	67 78;79
Manchester United			56;57;65;68		
Nottingham Forest					
Tottenham Hotspur			61		
Wolverhampton Wanderers	58	59			
Haka Valkeakoski	61				
HJK Helsinki	74;82				
HPS Helsinki	58				
Kuusysi Lahti		85			
Reipas Lahti	68				
TPS Turku	76				
AS Monaco	63	88			
AS Saint-Etienne	67;69	76	74	75	

Team	Round of 16	Quarterfinals	Semifinals	Finals	Wins
FC Nantes	66;77;80				
Girondins Bordeaux		87	84		
OGC Nice		56;59			
Olympique Marseille	71;91		89	90	
RC Strasbourg		79			
Stade de Reims	60	62			55;58
Dinamo Tbilisi	79				
1. FC Magdeburg	72;74				
1. FC Kaiserslautern	91				
1. FC Köln		64	78		
1. FC Nürnberg		61			
1. FC Saarbrücken	55				
Bayern München		72;76;85;87	80;89;90	81;86	73;74;75
BFC Dynamo Berlin	80;81;84;86	79;83			
Borussia Dortmund	56	57	63		
Borussia Mönchengladbach	70;71	75	77	76	
Carl Zeiss Jena		70			
Dynamo Dresden	73;77	76;78;90			
Eintracht Braunschweig		67			
Eintracht Frankfurt				59	
Hamburger SV	83		60	79	82
Rot-Weiß Essen	55				
Schalke 04		58			
TSV 1860 München	66				
Vorwärts Berlin	61;65	69			
Werder Bremen	65	88			
Wismut Karl-Marx-Stadt	57;60	58			
AEK Athens	78;89	68			
Olympiakos Piraeus	74;82;83				
PAOK Thessaloniki	76				
Panathinaikos	60;64;65;77	91	84	70	
ETO Györ	83		64		
Ferencváros	69;76	65			
Honvéd Budapest	56;80;85;89;91				
MTK Budapest	58	55			
Ujpest Dozsá	60;74;75	71;72	73		
Vasas Budapest	62;66	67	57		
IA Akranes	75				
Valur Reykjavík	67				
Bohemians Dublin	78				
Cork Celtic	74				
Derry City	65				
Dundalk	79				
Waterford United	70				
AC Milan	59;69	63;90	55	57	62;68;88;89
AS Roma				83	
Cagliari	70				
Fiorentina		69		56	
Hellas Verona	85				
Internazionale			65;80	66;71	63;64
Juventus	75;81;86	61;85	67;77	72;82	84
Napoli	90				
Sampdoria				91	
Torino	76				
Jeunesse d'Esch	59;63				
Sliema Wanderers	71				
AZ Alkmaar	81				
Ajax	73;80	57;66;77	79	68	70;71;72
DWS Amsterdam		64			
Feyenoord	61;74	71	62		69
PSV Eindhoven	55;76;78;91	63;88;89	75		87
Rapid JC Heerlen	56				
Sparta Rotterdam		59			
Glentoran	77;81				
Linfield Belfast	84	66			
Fredrikstad FK	60				
Lillestrøm SK	78;87				
Lyn Oslo	64				
Rosenborg BK	86				

Team	Round of 16	Quarterfinals	Semifinals	Finals	Wins
Vålerenga IF	66				
Górnik Zabrze	63;65;66;72;87;88	67			
Lech Poznań	90		70	69	
Legia Warsaw					
Polonia Bytom	62				
Ruch Chorzów	75	74			
Szombierki Bytom	80				
Widzew Łódź			82		
Wisła Kraków		78			
Benfica	63;69;72;73;81;84	65;68;75;77;83;91	71	62;64;67;87;89	60;61
FC Porto	79;85;87;88	90			86
Sporting CP Lisbon	55;58;62;70	82			
Arges Pitești	72;79				
CCA București	57;86;89		87	88	85
Dinamo București	56;63;64;65;71;73;82;84;90		83		
Rapid București	67				
Universitatea Craiova		81			
UT Arad	70				
CSKA Moscow	71				
Spartak Moscow	88	80	90		
Zenit Leningrad	85				
Aberdeen	80	85			
Celtic	72;77;82;86;88	68;70;79	71;73	69	66
Dundee FC			62		
Dundee United			83		
Glasgow Rangers	56;57;75;90	61;64;78;87	59		
Hibernian			55		
Kilmarnock	65				
Red Star Bratislava	59				
Slovan Bratislava	56;70				
Spartak Trnava	69	72;73	68		
Athletic Bilbao	83	56			
Atlético Madrid	66	77	58;70	73	
FC Barcelona			59;74	60;85	91
Real Madrid	60;68;69;76;78;89	64;66;90	67;72;75;79;86;87;88	61;63;80	55-59;65
Real Sociedad			82		
Sevilla		57			
Valencia	71				
Atvidabergs FF		74			
Djurgårdens IF		55			
IFK Göteborg	58;59;91	84;88	85		
IFK Malmö		60			
IFK Norrköping	56;57;62;63				
Malmö FF	75;89;90			78	
FC Basel	70;80	73			
FC Zürich			63;76		
Grasshoppers Zürich	71;84	56;78			
La Chaux-de-Fonds	64				
Servette FC Genève	55;61;79;85				
Xamax Neuchâtel	87;88				
Young Boys	57;59;60		58		
Besiktas	58	86			
Fenerbahce	59;61;68;74;85				
Galatasaray	63	62;69	88		
Trabzonspor	76				
Dinamo Kiev	67;69;78	72;75;81;82;91	76;86		
Dnipro Dnipropetrovsk		84;89			
Zaria Voroshilovgrad	73				
Partizan Belgrade	61;83	55;63		65	
Red Star Belgrade	59;68;69;77;88	57;73;80;81;86;91	56;70		90
Vojvodina Novi Sad		66			

Notes: (i) Data are gathered from UEFA. (ii) The order of the teams is by nationality (with alphabetic ranking of the countries). (iii) In case a team merged during the EC period, the Table shows the name of the team during their first EC round of 16 participation. (iv) 2 teams did not have to play the round of 16 during the 1968/69 season and one team did not have to play the round of 16 during the 1982/83 season, such that the number of round of 16 observations equals 293 rather than 296. (v) There was no semifinal during the 1991/92 season such that the number of quarterfinals observations equals 150 rather than 148 and the number of semifinals observations equals 72 rather than 74.

Table A.2: Performance of teams in the Champions League (for teams which qualified at least once for the round of 16 in the CL)

Team	Round of 16	Quarterfinals	Semifinals	Finals	Wins
Austria Salzburg	94				
Austria Wien	92;93				
Rapid Wien	96				
Sturm Graz	00				
Anderlecht	94;00	93			
Club Brugge		92			
Levski Sofia	93				
Hajduk Split		94			
Apoel Nicosia		11			
Sparta Praha	93;99;01;03				
AaB Aalborg	95				
FC København	93;10				
Arsenal	01;02;04;06;10;11	00;03;07;09	08	05	
Blackburn Rovers	95				
Chelsea	05;09	99;10	03;04;06;08	07	11
Leeds United	92		00		
Liverpool	05	01;08	07	06	04
Manchester United	93;94;03;04	97;99;00;02;09	96;01;06	08;10	98;07
Newcastle United	02				
Tottenham Hotspur		10			
AJ Auxerre		96			
AS Monaco	04		93;97	03	
FC Nantes	01		95		
Girondins Bordeaux	99	09			
Lille OSC	06				
Olympique Lyon	00;06;07;08;10;11	03;04;05	09		
Olympique Marseille	99;10	11			92
Paris Saint-Germain	00		94		
1. FC Kaiserslautern		98			
Bayer Leverkusen	02;04;11	97		01	
Bayern München	03;05;10	97;01;04;06;08	94;99	98;09;11	00
Borussia Dortmund	02	95	97		96
Hertha BSC	99				
Schalke 04		07	10		
VfB Stuttgart	03;09				
Werder Bremen	04;05	93			
AEK Athens	92;94				
Olympiakos Piraeus	07;09	98			
Panathinaikos	00;08	01	95		
Ferencváros	95				
AC Milan	96;00;07;09;10	03;11	05	92;94;04	93;02;06
AS Roma	01;02;08;10	06;07			
Fiorentina	99;09				
Internazionale	06;07;08;11	98;04;05;10	02		09
Juventus	01;03;08	04;05	98	96;97;02	95
Lazio Roma	00	99			
Napoli	11				
Ajax	05	02	96	95	94
Feyenoord	93;99				
PSV Eindhoven	05	92;06	04		
Rosenborg BK	95;99	96			
Lech Poznan	92;93				
Legia Warsaw		95			
Widzew Łódź	96				
Benfica		94;05;11			
Boavista	01				
FC Porto	95;01;04;06;07;09	92;96;99;08	93		03
Sporting CP Lisbon	08				
Dinamo Bucuresti	92				
Steaua Bucuresti	93;94;95;96				
CSKA Moscow	11	92;09			
Lokomotiv Moscow	02;03				
Spartak Moscow	94;00	93;95			
Zenit St. Petersburg	11				
Celtic	06;07				

Team	Round of 16	Quarterfinals	Semifinals	Finals	Wins
Glasgow Rangers	95;96;05	92			
Slovan Bratislava	92				
Atlético Madrid	08	96			
Celta de Vigo	03				
Deportivo de la Coruña	02	00;01	03		
FC Barcelona	92;04;06	94;02	99;01;07;09;11	93	05;08;10
Real Madrid	04;05;06;07;08;09	95;98;03	00;02;10;11		97;99;01
Real Sociedad	03				
Sevilla	07;09				
Valencia	10	02;06		99;00	
Villarreal		08	05		
IFK Göteborg	96	92;94			
FC Basel	02;11				
FC Sion	92				
Grasshoppers Zürich	95;96				
Fenerbahce	96	07			
Galatasaray	94;01	93;00			
Dinamo Kiev	94;99	97		98	
Shakhtar Donetsk		10			

Notes: (i) Data are gathered from UEFA. (ii) The order of the teams is by nationality (with alphabetic ranking of the countries). (iii) In case a team merged during the CL period, the Table shows the name of the team during their first CL round of 16 participation. (iv) There was no round of 16 during the 1997/98 season and during the 1998/99 season, such that the number of round of 16 observations equals 144 rather than 160. (v) There was no semifinal during the 1992/93 season, such that the number of quarterfinals observations 82 rather than 80 and the number of semifinals observations 38 rather than 40.

The Macbeth Method for Ranking Olympic Sports: a Complementary Analysis for the DEA Efficiency

J. C. C. B. Soares de Mello*; J. Benício**; L. Bragança*** and V. Guimarães****

* Universidade Federal Fluminense: joaocsmello@gmail.com

** UNILASALLE-RJ: juliana.benicio@hotmail.com

*** Universidade Federal Fluminense: livia_braganca@yahoo.com.br

**** Universidade Federal Fluminense: viniciusguimaraes@yahoo.com.br

Abstract The main objective of this article is to propose a ranking of Brazilian performance in different Olympic modalities in which Brazil participated in the Beijing 2008 Olympic Games. We investigate the way each sport earns medals in view of the numbers of athletes in the team and the investment made. That analysis is justified because investors need to focus their incentives in the most efficient sports. However, the efficiency analysis made with DEA is not adequate for rankings. For this reason we use the MACBETH method to realize the ranking as a complementary analysis to DEA efficiency results. This article analyses if both methods converge or diverge in their results. After application and analyses of results generated by both methods, it is possible to conclude that, for this specific case, there is no convergence between MACBETH and DEA.

1. Introduction

Public policies that support sports activities are important tools to aid the development of a country. According to the Ministério dos Esportes (2011), the practice of sports has the power to develop fully man as an autonomous, democratic and participatory citizen. Furthermore, sport is an important tool for leisure and reaffirmation of national identity. The sport must be presented for the students in their initiation, and must be developed in a broad, involving hygiene practices and cultural education (Betti, 1998). Seen this, public policies cannot neglects the potential of sports and must assist the development of national sport. Although, you need parameters that lead the investments choices in order to implement the best actions, since resources are scarce and must be allocated efficiently.

The results presented in competition for a particular sport is an important indicator. The performance of each sport in different types of competition can be an important indicator of how sports policy is being conducted nationwide. In this case, it is known that the better results obtained in larger competitions are positive externalities arising from this practice, because sports can raise the victorious national self-esteem and increase the number of future practitioners of the sport.

Among the sports competitions, that which encompasses the largest number of sports and has more visibility worldwide is the Olympics. For this reason, the result coming from this competition should gain prominence among studies in the area. As an example, we can cite Li, Liang, Chen and Morita (2008) and Lins, Gomes Soares de Mello et al (2003) who have been interested in the investigation of the results of the Olympics Games based in mathematician and economic tools.

The verification the better sports results, given certain amount of resources allocated initially, can be made by an analysis of the results obtained by applying the model DEA. This analysis was done in the study of Bragança and Lima (2011). The results obtained in this study provided an analysis of how Olympic sports "produced" medals given amount of financial investment and competition athlete participating in the Beijing Olympic Games and their results will be presented throughout this article. This analysis not only considers the importance of results to encouraging the successful sport in the Olympics, but it also reflects about the capacity of the sports confederation uses fewer resources to produce the medals.

However, the efficiency of generating winners in a sport should not just be the only bias analysis for decision making for the public manager. The ranking of the teams is also important decision tool for public managers in which sports should receive public incentives. As noted by Talluri (2000) DEA is not a suitable model for ranking, as the efficiency obtained can be engaged with a pattern of irrational weights. Thus, the advancement of research will propose the use of Macbeth Multicriteria Model, such as tool to produce a

ranking based on the evaluated criteria weighted by specialist. Therefore, it is expected that the results found in Macbeth will add value to the results found by DEA.

2. Objectives and Methodology of the Study

The central objective is to achieve a ranking of the performance of different Olympic sports that Brazil had participated in the Beijing Olympics Games in 2008, in order to provide inputs for investors allocate more efficiently its incentives in sports. So, this paper uses Method MACBETH Multicriteria Decision to provide a ranking, from the point of view of the performance of different Olympic sports and, additionally, compares the results of the relative efficiencies obtained by the DEA modeling. For both models, the same variables were used.

The combination of the two models allow the qualitative valuation of convergence or divergence between them, from the comparison of the results obtained, for providing decision makers information more consistent performance achieved by the sports

3. Literature Review

3.1. Multicriteria Decision Method – MACBETH

The multicriteria analysis methods should be used to assist in the selection process. In the present article, the search method is related to the necessity of ranking of the alternatives available criteria to be considered.

Despite the diversity of approaches, methods and techniques Multicriteria Decision Support, the basic elements are related to the way that people make a decision, taking into account the multiplicity of criteria, the corresponding need for consistent evaluation and structuring of complex situations (Pinheiro, et Souza Castro, 2008).

The MACBETH modelling used in the present study aims to build a value function intra-criteria given the decision maker's judgment, which assign weights to the criteria proposed. Through the determination of the weights, the function will allow aggregate several intra-critérias into single criteria.

According Bana e Costa & Vansnick (1995) MACBETH method determines a value scale that represents the cardinal value judgments of the decision maker from the verification of possible inconsistencies. This scale is obtained from the comparison of alternatives, pairwise, so as to evaluate the difference between pairs of attractiveness. That is, given two alternatives, the decision maker must evaluate which is the most attractive and what is the extent of attractiveness. The degree of attraction is given on a scale that has a semantic correspondence with an ordinal scale, with $C_k = 1, 2, 3, 4, 5, 6$ (Table 1). The scale obtained is normalized and generates the values of the weights for the alternatives under evaluation.

Table 1: Degrees of attractiveness between alternatives

Level of Attractiveness	Difference in Attractiveness	Semantic scale
C_1	Very weak	$C_1 = [s_1, s_2] \text{ e } s_1=0$
C_2	Weak	$C_2 =]s_2, s_3]$
C_3	Moderate	$C_3 =]s_3, s_4]$
C_4	Strong	$C_4 =]s_4, s_5]$
C_5	Very Strong	$C_5 =]s_5, s_6]$
C_6	Extreme strong	$C_6 =] s_6, +[$

The Macbeth consists of Linear Programming Problems (CPPs) sequential. The software that implements the computational method makes it the consistency analysis and generation of a cardinal cardinal scale, indicating possible inconsistencies and alternative ways to solve it. The PPL suggests a cardinal scale of values for alternatives and a range of values in this range may vary without making the problem infeasible (no solution).

As the scale can vary within a range defined by linear programming, it is possible that the decision maker, for a sensitivity analysis, adjust the scale values graphically suggested, respecting the defined range. As Bana

e Costa & Vansnick (1997), these adjustments only after the decision-maker made based on expert knowledge, which is the cardinal scale of values is defined.

An alternative method widely used in the literature for generating weights is AHP (Analytical Hierarchy Process (Saaty, 1980)), but this method was rejected since MACBETH allows flexibility in the result of the weights found. The AHP provides fixed weights as a result, in contrast, has MACBETH as possible results upper and lower limits of weights that can be adjusted according to the preference of decision maker. Thus, the decision maker can adjust the result aimed at rapprochement with his expert opinion.

3.2. Data Envelopment Analysis – DEA

The purpose of modeling is to assess the DEA relative efficiency of different production units decision makers, called DMU (Decision Making Unit). Second Angle Meza, and Gomes Neto (2007), efficiency is a relative concept which compares, through the determination of a production frontier, the productivity of a given DMU relative to the maximum yield that could achieve this DMU, and the productivity ratio Products generated (outputs) and allocated resources (inputs).

The evaluation of the efficiency considering multiple inputs and outputs is one of the main advantages of DEA modeling. Another relevant feature is the flexibility of DEA for orientation model to evaluate the efficiency of each DMU. According to the problem to be solved, the modeling may have two (2) possible orientations: input orientation, seeking efficiency of DMU from the minimization of its inputs to produce the same outputs, or output orientation that seeks maximize outputs, applying the same inputs, used in this study.

To determine the production frontier there are two classic multidimensional models: CRS, originally introduced by Charnes et al. (1978), and model VRS proposed by Banker et al. (1984). The CRS model is characterized by defining a production frontier piecewise linear and consider constant returns to scale. The BCC model, used in this article, does not consider the proportionality between inputs and outputs, considering the variable returns to scale.

For the CRS and BCC models, the efficiencies of the DMUs can be calculated by means of two different methods: (1) Method of Multipliers, which allows calculation of the relative efficiencies and optimal weights for the variables, and (2) Envelope method, the dual methods of multipliers, which allows calculation of the relative efficiencies, DMUs benchmarking, slacks and goals to be achieved by DMUs.

According to Soares de Mello, Angulo Meza, and Gomes Neto et all (2005), modeling the DEA has the advantage of defining ordinations without the opinion makers, different methods of Multicriteria Decision Support, but its evaluation is extremely benevolent, because the model allows each DMU to choose the weights of each variable (inputs and outputs) to maximize its efficiency.

As mentioned earlier, this article uses to calculate the efficiencies, the CRS model (Method Envelope) with output orientation. Therefore, the following LPP (Linear Programming Problem) is solved:

$$\begin{aligned}
 & \text{Max } h_0 \\
 & \text{Subject to} \\
 & x_{io} - \sum x_{ik}\lambda_k \geq 0, \forall i \\
 & -h_0 y_{jo} + \sum y_{jk}\lambda_k \geq 0, \forall j \\
 & \sum \lambda_k = 1 \\
 & \lambda_k \geq 0, \forall k
 \end{aligned} \tag{1}$$

Where:

0 = observed DMU

h_0 = inverted efficiency ($Eff = 1/h_0$)

x_{io} = input i of observed DMU

x_{ik} = inputs i of DMU k , where $i = 1, \dots, r$

y_{jk} = outputs j da DMU k , where $j = 1, \dots, s$

λ_k = represents the contribution of DMU $_k$ for the projection of DMU $_0$ in the frontier

4. Modelling

4.1. Application of DEA

The DEA results presented in this article were based on modeling initially proposed by Lima and Bragança (2011) in "Evaluation Methodology for the DEA results of Brazil at the Beijing Olympics, 2008." The aim of this study was to evaluate the efficiency of each Olympic sport and the conversion of investments into results, considering each of the 23 (twenty three) Olympic sports in which Brazil had participation as a DMU. In this universe of 23 (twenty three) Olympic disciplines, the authors considered for the calculation of efficiencies only eleven (11) arrangements because the others did not come to compete medals.

The modeling by Lima and Bragança (2011) was the VRS Method. The output orientation was used, since the goal of a country in the Olympics is to maximize your winnings. The variables used were:

A_Inputs:

- Number of athletes / teams for sport
- Public investment in Brazil for sport, in the 02 years preceding the 2008 Games
- Number of medals available for sport

B_Outputs:

- Number of gold medals
- No. of silver medals
- No. of bronze medals
- No. disputes medal for sport

However, this paper adapts the original model proposed by Lima and Bragança (2011), in view of the need to conduct a comparative analysis of the results of DEA and MACBETH methods before more related parameters.

First, the input variable "Number of medals available for sport" was not considered in modeling DEA, since it would not be applicable as a variable to the method MACBETH, because it is not a decision variable for analysis of preferences but an intrinsic factor competition.

Second, CRS modeling is used instead of the VRS modeling. As told before, this choice is based on the characteristic more discretionary of the CRS, facilitating the identification of the sources of efficiency (or inefficiency). To enable the use of CRS instead of VRS, the input variables were normalized. This normalization aimed to reduce the influence of the scale of operation of the units in the efficiency measure. Also included are restraints on weights, such that:

- Gold medal is of equal or greater importance than the silver medal, where $vo \geq vp$, $vo - vp \geq 0$.
- Silver medal is of equal or greater importance than the bronze medal, where $vp \geq vb$, $vp - vb \geq 0$.
- Bronze medal is of equal or greater importance than play a game worth medal, where $vd \geq vb$, $vb - vd \geq 0$.
- The difference between gold and silver is greater than that between silver and bronze, in turn, is greater than the difference between brass and contention for medal where $vo - 2Vp + 2 vb - vd \geq 0$.

The data for the input variables and outputs for each DMU are shown in Table 2.

Table 2: Beijing Olympic Games Datas (2008)

DMU (Olympic sport)	INPUT		OUTPUT			
	Nº of athletes or teams	Investments in R\$ MI (*10 ⁶)	Nº of disputes for medals	Nº of Gold medals	Nº of Silver medals	Nº of Bronze medals
Athletics	45	4,17	16	1	0	0
Handball*	2	4,32	0	0	0	0
Water sports	32	4,64	9	1	0	1

Volleyball	6	5,61	5	1	2	1
Equestrianism	13	3,54	7	0	0	0
Judo	13	3,78	4	0	0	3
Basketball*	1	4,69	0	0	0	0
Sailing	12	6,16	7	0	1	1
Gymnastics	8	4,13	6	0	0	0
Rowing*	4	3,57	0	0	0	0
Boxing*	6	2,20	0	0	0	0
Cycling	5	2,64	5	0	0	0
Tennis*	3	2,49	0	0	0	0
Table tennis*	4	2,90	0	0	0	0
Taekwondo	3	1,37	2	0	0	1
Triathlon	3	2,10	3	0	0	0
Shooting sports*	2	2,74	0	0	0	0
Canoeing*	2	2,75	0	0	0	0
Fencing*	2	1,22	0	0	0	0
Archery*	1	1,30	0	0	0	0
Weight lifting*	1	1,19	0	0	0	0
Fights*	1	1,14	0	0	0	0
Modern Pentathlon	1	1,29	1	0	0	0

* Sports that do not enter to the calculation of efficiencies

4.2. Modeling the problem - Using MACBETH

This paper aims to conduct a comparative analysis and complement the results found by the DEA model, as described earlier. Therefore, to allow comparison between studies, we attempted to follow the same line and the same analytical modeling of the problem. Based on this, the criteria to be analyzed are the same inputs and outputs considered in the DEA model with a few caveats to be highlighted below. It is worth noting that the models use the axioms of Roy (Roy et Bouyssou, 1993) as the basic criterion of selecting criteria, namely exhaustion, not redundancy and cohesion.

The input model of the DEA number of medals available will not be a criterion, but your data will be used as the denominator of indexes representing criteria: number of disputes, the number of gold medals, silver and bronze. That is, the four criteria cited represent a ratio of their gross value and the number of medals.

Public investment will not be a criterion. These data considered in a cost-benefit assessment, to be held in the last step of the analysis, in which we share the results found by the method Macbeth by investment value corresponding to each sport. Thus, the final result will be compared with the DEA index generated by the cost-benefit ratio of the result Macbeth and total public investment in 2 years.

After those considerations, the criteria to be assessed are:

- Number dispute medal / Number of medals available - DM.
- Number of gold medals / Number of medals available - MO.
- Number of silver medals / Number of medals available - MP.
- Number of bronze medals / Number of medals available - MB.

Thus, the country to win a gold medal in football given the value 1 in the criterion, given that first gold medal was won and 1 was available. Finally, it is noteworthy that the study done on DEA considered as inputs the number of athletes / team per sport and numbers of medals for sports. Such inputs will not be

considered as criteria in this study. The first will not be used, because little help in a multicriteria analysis. The second ended up being used indirectly in the indices of medals won.

A matrix of judgments that expresses the preference relation pairwise can be verified by the decision maker in the Table 3. The SQ column represents the status quo:

Table 3: Matrix judgments of difference in attractiveness to the criteria

	MO	MP	MB	DM	SQ
MO		Strong	Very Strong	Very Strong	Very Strong
MP			Moderate	Strong	Very Strong
MB				Moderate	Strong
DP					Strong
SQ					

The above matrix was considered consistent and rational decision maker. The results by application of the method are presented in Table 4:

Table 4: Weights generated by application of MACBETH

Criterias	Gold Medals	Silver Medals	Bronze Medals	Disputes for medals
Weights	33.33	26.98	22.22	17.46

As can be noted the structure of weights privileges winning medals rather than medals disputes. That is, values the achievements, although considering some importance to the dispute.

5. Results

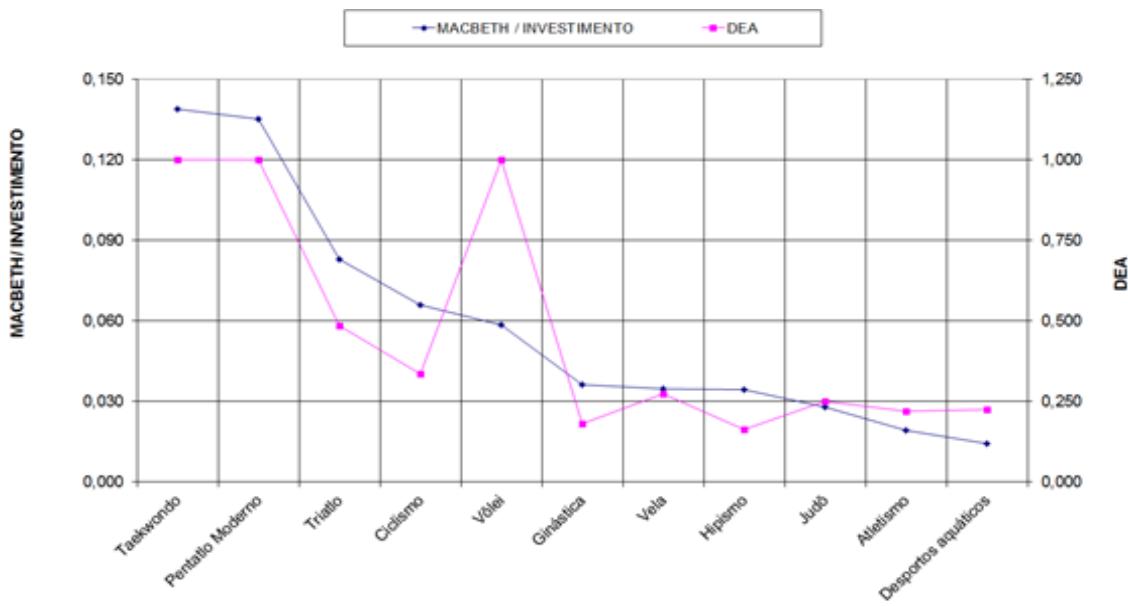
Based in the outcomes, shown in Table 4, we can compare the results between the application of two methods, MACBETH and DEA CRS, as can be seen from the ordering generated by each of these in Table 5.

Table 5: Comparison of the results generated by the DEA and MACBETH

Olympic sport	Macbeth Result	Macbeth / Investment	DEA Result	Ranking Macbeth / Investment	DEA ordination
Taekwondo	0,190	0,139	1,000	1	1
Modern Pentathlon	0,175	0,135	1,000	2	1
Triathlon	0,175	0,083	0,486	3	2
Cycling	0,175	0,066	0,337	4	3
Volleyball	0,328	0,058	1,000	5	1
Gymnastics	0,150	0,036	0,182	6	8
Sailing	0,214	0,035	0,275	7	4
Equestrianism	0,122	0,035	0,163	8	9
Judo	0,105	0,028	0,251	9	5
Athletics	0,080	0,019	0,219	10	7
Water sports	0,066	0,014	0,224	11	6

Basketball *	0,000	0,000	NA	12	NA
Handball *	0,000	0,000	NA	12	NA
Rowing*	0,000	0,000	NA	12	NA
Table tennis *	0,000	0,000	NA	12	NA
Canoeing*	0,000	0,000	NA	12	NA
Shooting sports*	0,000	0,000	NA	12	NA
Tennis*	0,000	0,000	NA	12	NA
Boxing*	0,000	0,000	NA	12	NA
Archery *	0,000	0,000	NA	12	NA
Fencing*	0,000	0,000	NA	12	NA
Weight lifting *	0,000	0,000	NA	12	NA
Fights*	0,000	0,000	NA	12	NA

Graphs 1 and 2 reflects the values shown in Table 4 and confirms that there is convergence between the methods applied in the present study.

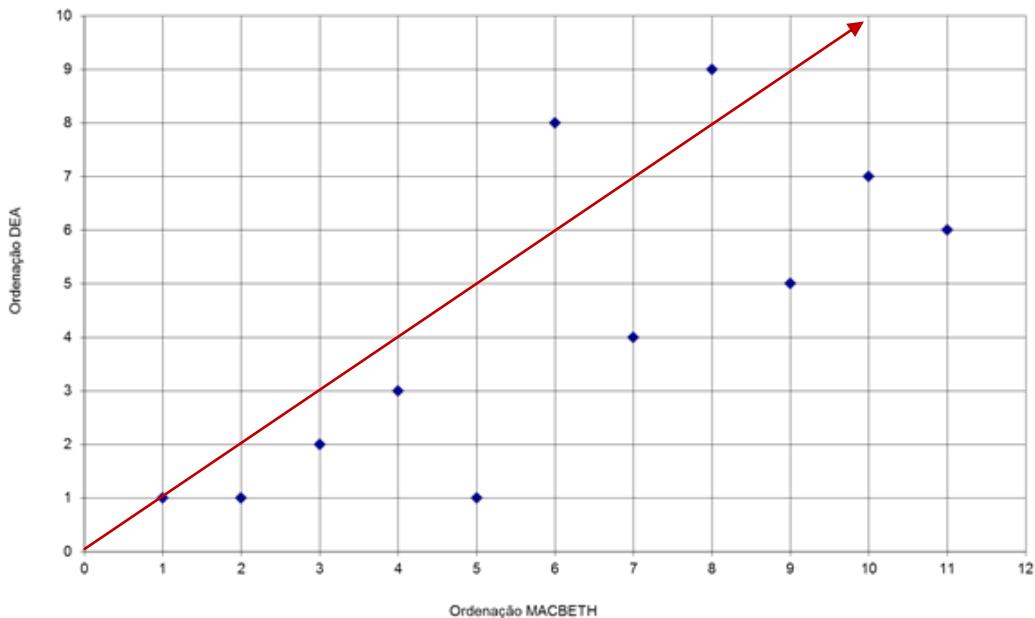


Graphic 1: Comparison of the results generated by the DEA and MACBETH

Graphic 1 displays a tendency to confluent results between the 2 methods. It is noticed that the areas that separate the results have hardly changed, except for the result of volleyball; sport that most of distancing result was presented.

It is noteworthy that the purpose of the chart is not one to relate the efficiencies found in DEA with the values of additive function of MACBETH. The purpose of this is to show divergence of results.

Graph 2 shows the difference between the rankings generated by the 2 methods. The red line indicates the points where the 2 methods would show the same result in the ordering.



Graphic 2: Comparison between Macbeth and ordering generated by DEA

Graphic 2 shows that, generally speaking, the results are close to the red arrow indicating a closeness of the results presented.

6. Conclusions

From the results, it was possible to verify the existence of a correlation between the results of DEA and MACBETH methods for assessing the performance of sports in the Olympic Games. This is explained by the fact that when experts ponder relevant criteria to analyze the results of the various Olympic Games Olympic sports in Brazil, they generate a ranking among the sports which resembles the result of efficiencies generated by DEA. That is, the value of each criterion made in MACBETH method generates results that volume weighted invested; goes against the ability of DMUs are efficient in producing medals, given the volume of investment and the number of athletes available to the dispute.

The rationality of the decision maker, who produced a ranking for Olympic sports that were supported by efficiency analysis made by DEA, is a positive point to be added to the possible applications of the method MACBETH, since the ranking produced by it is more assertive in order to DEA.

Thus, the central aim of the article was reached, once the relation between the results analyzed actually presented important tools for investment decisions.

As a future study intends to check the convergence of results will give when the same methodology is applied to other countries or for the results of other years.

7. References

- Bana e Costa, C.A. & Vansnick, J.C. (1995). Uma nova abordagem ao problema da construção de uma função de valor cardinal: MACBETH. *Investigação Operacional*, 15, 15-35.
- Bana e Costa, C.A. & Vansnick, J.C. (1997). Thoughts on a theoretical framework for measuring attractiveness by categorical based evaluation technique (MACBETH). In: *Multicriteria Analysis* [edited by Clímaco, J.J], Springer-Verlag.

- Betti, M. A janela de vidro: Esporte, televisão e educação física. Campinas, SP: Papirus, 1998 (Coleção Fazer / lazer).
- Li, Y.; Liang, L.; Chen, Y.; Morita, H. (2008). Models for measuring and benchmarking Olympics achievements. *Omega – International Journal of Management Science*. 36 (6), 933-940.
- Lima, A.; Bragança, L. (2011). Avaliação pela metodologia DEA dos resultados do Brasil nas Olimpíadas de Pequim, 2008.
- Lins, M. P.; Gomes, E.; Soares de Mello, J. C.; Soares de Mello, A. J. (2003). Olympic ranking based on a zero sum gains DEA model. *European Journal of Operational Research*. 148 (2), 312-322.
- Ministério dos Esportes. Disponível em <http://www.esporte.gov.br/>. Acesso em 12 de julho de 2011.
- Pinheiro, P.; Souza, G.; Castro. A.K. (2008). Estruturação do problema multicritério para produção de jornal. *Pesquisa Operacional*, vol.28, no.2, Maio/Agosto. Rio de Janeiro.
- Roy, B. (1985). *Méthodologie multicritère d'aide à la décision*. Economica, Paris.
- Silva Filho, S.; Tejerina, L.; Malheiros, N. (2009). Metodologia de rankeamento de tecnologias desenvolvidas em laboratórios científicos: Segundo sua viabilidade de transferência do meio acadêmico ao mercado através de spin-offs. *XIX Seminário Nacional de Parques Tecnológicos e Incubadoras de Empresas*, Florianópolis (SC).
- Soares de Mello, J. C.; Lins, M. P.; Gomes, E. (2001). O Uso de Análise Envoltória dos Dados e Auxílio Multicritério à decisão na análise dos resultados das Olimpíadas de 2000. Disponível em http://www.abepro.org.br/biblioteca/ENEGEP2001_TR62_0711.pdf. Acesso em 13 de julho de 2011.
- Soares de Mello, J.C.; Gomes, E.; Mangabeira, J.A. (2008). Índice Multicritério de Bem Estar Social Rural em um Município da Região Amazônica. *Pesquisa Operacional*, vol. 28, no.1, pg. 141-160.
- Talluri, Srinivas (2000). Data Envelopment Analysis: Models and Extensions. *Production Operational Management*, Decision Line, pg 8-11.

Tennis Player Ranking using Quantitative Models

A. Demetris Spanias* and B. William Knottenbelt**

* Department of Computing, Imperial College London, South Kensington Campus, London, SW7 2AZ
d.spanias10@imperial.ac.uk

** Department of Computing, Imperial College London, South Kensington Campus, London, SW7 2AZ
wjk@doc.ic.ac.uk

Abstract. The Association of Tennis Professionals (ATP) and the Women's Tennis Association (WTA) generate weekly rankings for professional tennis players by awarding points to each player depending on how far the player has advanced in a countable tournament. Since tournaments are designed such that top players face the lower-ranked players in the earlier rounds, a bias is introduced which favours the top players. In this paper we demonstrate two new algorithms, SortRank and LadderRank, which rank professional tennis players. Both ideas make use of a quantitative tennis model to assess the performance of individual players and then compare them with each other. SortRank uses traditional sorting algorithms to rank the players using the result of a simulated match between the two players as the comparison criterion. LadderRank ranks players using a "sports-ladder" style iterative algorithm, which also compares players based on the result of a simulated match between them. Both algorithms are flexible as they can be implemented using any underlying quantitative model. The ranking systems are demonstrated and assessed based on their ability to predict the outcome of matches played within the period used to rank the players.

1. Introduction

Professional tennis rankings are at the centre of attention of the tennis world. Both the Association of Tennis Professionals (ATP) and the Women's Tennis Association (WTA) rank professional tennis players and use their rankings to decide both the participation of players in tournaments, as well as the ultimate champion of the year. Being a top ranked player generates a great deal of prestige and popularity. In fact, most professional tennis players have, in one way or the other, mentioned their passion to reach the top of the rankings.

One may argue that any absolute ranking system is by definition flawed when applied to such a complex sport in which there is an unknown degree of transitivity and a multitude of parameters to take into account. Nonetheless, an overall ranking is a simplistic method of determining who is performing better at the sport and captivates both the public and media. Simplistic as it may be, there is a general desire for the overall ranking system to be "fair". Unfortunately rankings, as they are currently calculated, provide the top players with an unfair advantage as seeded tournaments make it increasingly difficult for lower ranked players to climb the rankings.

This bias which favours the top players of the rankings, does not only affect the lower ranked players but also researchers who have used these rankings as a tool for prediction of match outcomes. Clarke & Dye (2000) propose an approach based on regression which uses ATP ranking points to simulate professional tennis matches. Additionally, del Corral & Prieto-Rodriguez (2010) attempt to assess the degree to which the difference in ranking points are good indicators of the outcome of Grand Slam matches.

Some research has also been directed towards the invention of different ranking systems. Clarke (1994) proposed a ranking system which uses a player rating which is adjusted after each match played by the player. The adjustment is calculated using exponential smoothing on the difference between an expected result suggested from the previous ranking difference, and the actual match result. A more recent method was proposed by Radicchi (2011) which makes use of an algorithm similar to Google PageRank by Brin & Page (1998). Radicchi uses PageRank by assigning prestige values for all professional players and adjusts them relative to the number of victories they achieved against other players. This ranking system can be used to rank all players regardless of the time period they were active and thus contributes to an investigation on the best player of all time. A few years later, Dingle, et al. (2013) presented further evidence of PageRank's usefulness as a ranking tool for both female and male professional tennis players and also showed that a ranking generated using PageRank is a better predictor of match results than the official ATP rankings.

In this paper, we provide evidence towards the forementioned bias inherent in the current ATP ranking system and we attempt to introduce a new, flexible concept of ranking systems. We compare our ranking systems with PageRank for tennis and the official ATP Rankings by quantifying the extend to which they reflect the outcome of the set of matches used to generate them and show up-to-date (March 2013) results.

2. ATP Ranking System

The Emirates ATP Rankings is the official ranking system ATP used for 2013. It is “a historical objective merit-based method used for determining entry and seeding in all tournaments” as the official ATP World Tour website states. The ranking is generated using a summation of points players acquire while proceeding within seeded tournaments. Tournaments themselves are split into categories with some tournaments awarding more points than others (see Table 1).

Table 1 – ATP Ranking points awarded for main ATP tournament categories. (Additional points are awarded from the Barclays ATP World Tour Finals, Olympic Games, Challenger and Futures tournaments that are not included in this table. Numbers in brackets are dependent on the tournament draw size. Points for qualification are also dependent on draw size.)

	W	F	SF	QF	R16	R32	R64	R128	Qual.
Grand Slams	2000	1200	720	360	180	90	45	10	25
ATP World Tour Masters 1000	1000	600	360	180	90	45	10(25)	(10)	25
ATP 500	500	300	180	90	45	(20)	-	-	20
ATP 250	250	150	90	45	20	(5)	-	-	12

The summation of ranking points is over a maximum of 18 tournaments played within the previous 52 weeks, out of which four are the Grand Slam tournaments, eight are the compulsory ATP World Tour Masters 1000, and the rest are the best six results from the ATP 500, 250 and other tournaments (given a minimum of 4 ATP 500 tournament participations). Additionally, players who have finished within the top eight rankings at the end of the ATP tennis season, qualify to play at the Barclays World Tour Finals to earn points that count towards crowning the final champion of the year. In those years where the Olympics occur, the players also win extra points for the position they get in the Olympics.

While the Emirates ATP Rankings provide accurate rankings for the top 32 players, by design players ranked lower than the top 32 are at a disadvantage. The seeded tournament system makes it increasingly difficult for lower ranked players to proceed into the latter rounds of tournament and thus earn the necessary points to climb the rankings.

Almost all countable tournaments have seeded players, i.e. the top 16 or 32 players who are participating in the tournament have a seeded position. The tournament draw is set up in a way such that the seeded players do not face any other seeded players in the first round. The reasoning behind this is to avoid situations where top ranked players face off in the earlier rounds and get knocked out earning fewer points. This non-random selection of draws creates a bias towards the top 32 players as any players ranked lower than that have a much higher chance to face the top players in the early rounds of the tournament and therefore have much higher chance of being knocked out without earning the points they deserve. This can make it difficult to rank the true performance of these players especially when compared to one another. Evidence of this will be presented in the results section later on.

On the other hand, seeded tournaments together with the Emirates ATP Ranking system, create a much more accurate ranking of the top 32 players when compared to each other. The reasoning behind this is that these players get to face each other in higher frequency as they are more likely to proceed in the latter rounds and therefore there is a more data on which the rankings are based upon. Additionally, the difference in points earned for each victory is higher in the latter rounds. This higher difference in points boosts players

who achieve victories against other high ranked players and thus enhances the subtle differences in their performance.

3. Background of Rankings and Tennis Models

In this section we briefly introduce some tennis models which have been presented in past literature and are used in combination with our ranking algorithms, SortRank and LadderRank. We first introduce how one can construct a hierarchical Markov model to estimate the probability of a player winning a match against another player using only the probabilities of the two players winning points while serving. We then describe a Markov model that can be used to calculate the probability of a player winning a point against another player while serving. Finally, we briefly present the PageRank ranking system for tennis that is used as a comparison system in the results section.

3.1 Hierarchical Markov Model for Tennis

A study performed by Klaassen & Magnus (2001), shows that even though points in tennis are not independent and identically distributed (i.i.d.), one may assume that they are for the purpose of modelling a tennis match because the deviation from independency is small. This means that one can estimate the probability of a player winning a game while serving or even the probability of a player winning a tiebreaker by constructing a Markov Chain of the game/tiebreaker which uses only two parameters, the probabilities of the two players winning a point while serving. This in turn allows one to calculate the probability of a player winning a set using only the probabilities of each player winning a service game and the probability of a player winning a tiebreaker. Finally, one can hierarchically calculate the probability of a player winning a match using only the probabilities of the two players winning sets in which they served first. Barnett & Clarke (2002) demonstrate this idea using a simple spreadsheet application which recursively calculates the probability of a player winning a match from every score-line.

3.2 Low-Level Point Markov Model

Having discussed how to hierarchically model a tennis match, all that remains is a method to estimate the probability of a player winning a point on serve. Spanias & Knottenbelt (2012) present a Markov chain in an attempt to model a tennis point and show two techniques of parameterising the model using historical player statistics. The first technique, named the “*Uncombined*” model, estimates the probability of a player winning a point while serving against the “average” professional player. The second technique, named the “*Combined*” model, estimates the probability of a player winning a point while serving against a specific player by combining serving statistics of the server with return statistics of the receiver. These two models will be used in conjunction with the hierarchical Markov model introduced earlier to generate rankings using SortRank and LadderRank.

3.3 PageRank Tennis Ranking

The PageRank tennis ranking system was first introduced by Radicchi (2011) and further investigated by Dingle, et al. (2013). It is an effective ranking system for tennis players which we use in this paper as a good comparison for the ranking systems introduced by this paper. The system is based on Google’s PageRank algorithm summarized in Brin & Page (1998) which is used for ranking websites.

In order to explain how PageRank can be applied in tennis we need to define a few variables. Let w_{ji} be the amount of tennis matches player j has lost against player i, and s_j^{out} be the total defeats suffered by player j. Also let α be a weight factor between 0 and 1 and N be the total number of players being ranked. The prestige of player i is then described by the following equation.

$$P_i = (1 - \alpha) \sum_j P_j \frac{w_{ji}}{s_j^{out}} + \frac{\alpha}{N} + \frac{(1 - \alpha)}{N} \sum_j P_j \delta(s_j^{out})$$

P_i , the prestige value assigned to player i, is calculated as the summation of three parts. The first part is the amount of prestige that is transferred from player j to player i, the second part is a constant redistribution

of prestige and the third part is used as a constant value for players with no outward links (no defeats). An algorithm can be designed which will iteratively calculate the prestige of all players until they converge. Players are then ranked according to the amount of prestige they hold.

4. SortRank and LadderRank

SortRank and LadderRank are two similar approaches to ranking professional tennis players. They both use an underlying tennis model which estimates the outcome of a simulated match between players and rank players based on that outcome. SortRank is a faster algorithm but has the requirement that the underlying model is absolutely transitive. LadderRank is a slower algorithm which expands the idea of SortRank taking into account the non-transitive nature tennis models may have. This is done by sorting the same players over and over even after a regular sorting algorithm would have finalized their position. Also the algorithm has the ability to compare players with other players who are not immediately next them.

4.1 SortRank

The concept behind SortRank is very simple: take any tennis model, convert it into a binary model and then use it as the comparison criterion of a sorting algorithm. For example: let's assume that we have a list of players to rank. A sorting algorithm such as QuickSort, as described by Hoare (1961), can be used to sort this list of players by using a binary model which outputs a comparison criterion between players.

A limitation of any sorting algorithm is that it assumes absolute transitivity. This means that if Player A can beat Player B and Player B can beat Player C then it must hold that Player A can beat Player C. As a consequence, any model that is used as the comparison criterion should also be absolutely transitive.

An example of a fully transitive model is the “*Uncombined*” model mentioned in section 3.2. This model is transitive by definition as the opponent is not taken into consideration when estimating the parameter of a player winning a point while serving. Therefore, the output of any probability from the model is always compared against the constant “average” player. This “*Uncombined*” model can be converted into a binary model by using the resulting probability of Player A winning a match against Player B. If this probability is greater than 0.5 then the binary model returns “true”, otherwise it returns “false”.

This binary model can be joined with any sorting algorithm to generate a ranking. For this to happen, the sorting algorithm, when comparing two players, A and B, should use the binary model as the comparison criterion. That is, if the binary model returns “true” for Player A winning a match against Player B, the sorting algorithm places Player A above Player B in the rankings. By completing the algorithm for the entire list of players, the end result is a sorted list of players based on their performance, with the best player at the top of the list, thus a ranking.

4.2 LadderRank

To overcome the limitation of absolute transitivity, we constructed a new algorithm that does not assume the comparison criterion is absolutely transitive. This algorithm is inspired by normal “sports-ladders”. In a “sports-ladder” there is an initial ranked list of players, and each of those players is allowed to challenge another player that is ranked up to X positions higher. If the challenger is victorious in the challenge, then he/she overtakes the player challenged and pushes everyone in-between one position down. The resulting algorithm is described by the pseudo-code below.

For this algorithm to function correctly it must be provided with these crucial variables: the *number_of_iterations*, the *positions_above_allowed_to_challenge* and the *ranking_list*. To ensure complete ranking of the players the *number_of_iterations* must always be larger than the number of players being ranked. The *positions_above_allowed_to_challenge* defines the number of positions in the ranking list that any player is allowed to jump after any challenge. Finally the *ranking_list* is the list of players ranked in an initial order.

```

for (int i =0; i < number_of_iterations; i++) {
    foreach (current_player in ranking_list) {
        if (current_player.ranking > 0) {
            x = positions_above_allowed_to_challenge
            if (x > current_player.ranking) { x = current_player.ranking }
            for (int position = x; position > 0; position--) {
                PlayerA = PlayerWithRanking(current_player.ranking - position)
                PlayerB = current_player
                if (Compare(PlayerA, PlayerB) == false) {
                    //if player A loses the match-up move player B
                    //above A and push all players inbetween 1 spot down
                    MovePlayerToRanking(PlayerB, PlayerA.ranking)
                    position = 0 //stop challenging
                }
            }
        }
    }
}

```

The function `PlayerWithRank(integer)` which appears in the algorithm retrieves the player which has the ranking provided as the integer parameter. The function `MovePlayerToRanking(player, integer)` changes the ranking of the player to the integer value provided and shifts all rankings of players which were between the player and the new ranking by 1 position towards the direction of the player's current ranking. For example in a list of three players, A, B and C ranked as 1, 2 and 3 respectively, the function `MovePlayerToRanking(C, 1)` will change the rankings of A, B and C to 2, 3, 1 respectively.

5. Evaluation and Results

In this section we will present and discuss the results of our implementation of the LadderRank ranking system when using the “*Combined*” model as the comparison criterion. Figure 1 illustrates the top 100 players in the ATP Official Rankings on the 18th of March 2013 and corresponding LadderRank ranking generated over the same period for x=3. It can be observed that players ranked by the ATP in positions 1-32 are positioned very close to the y=x line. This means that the LadderRank system ranks them in a similar position to the ATP ranking system. The two systems start to deviate in the rankings a lot more for players ranked in positions greater than 32 by the ATP. This appears to support the theory that seeded tournaments deteriorate the accuracy of rankings of players ranked greater than 32 by the current ATP system. In fact similar results have been produced using the PageRank ranking system and are also evident in other periods (see Dingle, et al. (2013)).

In Figure 1, any players that appear above the y=x line are players which according to LadderRank are ranked higher than they should be by the ATP. Similarly players that appear below the line are players that are ranked lower than they should be.

A striking case is Mardy Fish who has dropped to ATP position 33 on the 18th of March 2013, from being number 9 in the world on the 19th of Match 2012. Mardy Fish on the other hand is still ranked in the top 10 players on the LadderRank system. The reason for this is the underlying model, which uses the average statistics of the player over the past year. Therefore the “*Combined*” model itself does not adapt fast to changes in performance of players and as such, the LadderRank ranking did not adapt quickly and is still showing Mardy Fish as one of the top players. This can be fixed by using a heavier weighting to more recent statistics when calculating the probabilities of winning the point on serve in the underlying model. Therefore this is not a problem of the LadderRank algorithm but a problem with the “*Combined*” model.

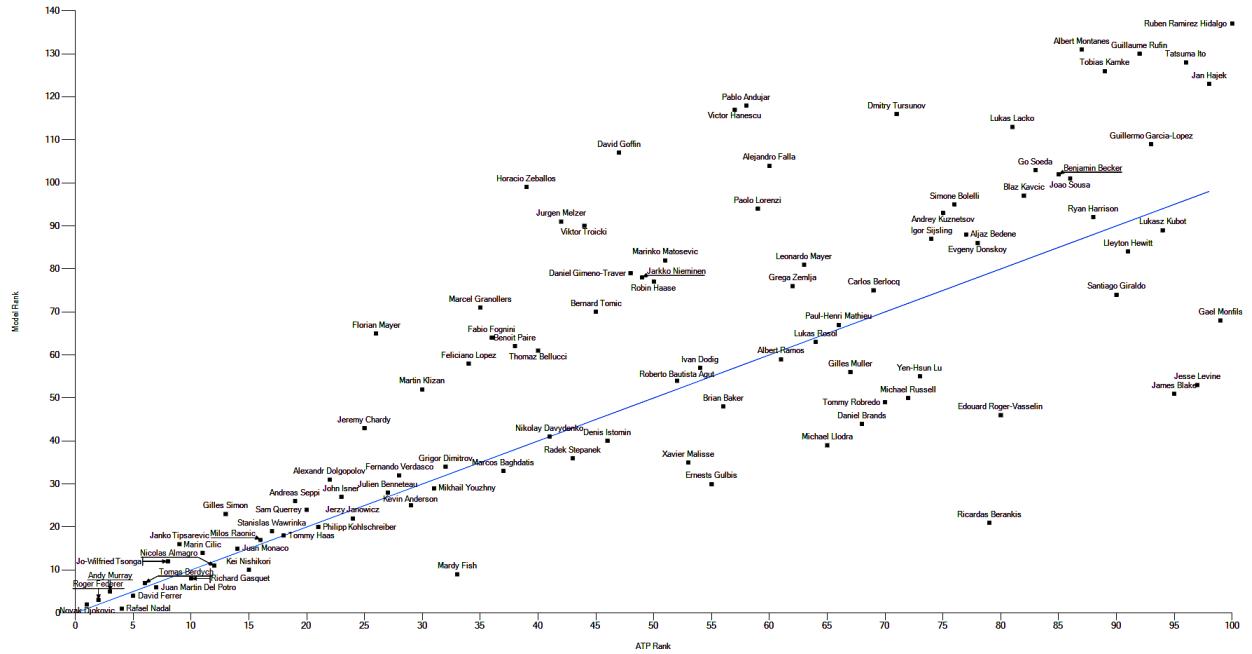


Figure 1 – Comparison between LadderRank-Combined model with $x=3$ and the ATP Rankings of the Top 100 players over the period 18/03/2012-18/03/2013.

To analyse the performance of our ranking systems further and get a metric to compare them with other systems such as the PageRank and the ATP, we used the rankings generated by each system and tested whether those rankings “predict” the outcomes of the matches within the period that was used to generate them. To put it simply, we calculate the percentage of matches in which the winner of the match is better ranked than the loser.

Using 2552 matches that were played in the period 18th of March 2012 to 18th of March 2013, we generated these percentages for 5 ranking systems – the ATP Official rankings, the PageRank system using Match Victories as weights, the SortRank-Uncombined system and the LadderRank-Combined system with $x=1$, 3 and 5). Table 2 shows these results in detail.

Table 2 – Comparison of predictive power of ranking models over matches played in the period used to generate them. These rankings were generated for 298 players who competed in 2552 matches over the period 18/03/2012 to 18/03/2013. In SortRank and LadderRank prediction results, 381 matches were not attempted as there were insufficient statistics (less than 10 matches) to model one or both the players who took part in those matches.

ATP	Match PageRank ($\alpha=0.15$)	SortRank Uncombined (381 Skipped)	LadderRank Combined (381 Skipped) ($x=1$)	LadderRank Combined (381 Skipped) ($x=3$)	LadderRank Combined (381 Skipped) ($x=5$)
69.83354%	71.15987%	66.97374%	70.65868%	70.70474%	70.65868%

Observing the results presented in Table 2, it is evident that the PageRank system appears to describe the matches which were used to generate it better than the rest of the systems. Also the SortRank-Uncombined system seems to perform worse than the rest of the systems – something which is expected as the “*Uncombined*” model which used to generate the rankings also performs poorly. The LadderRank-Combined system appears to be in second place with marginally better performance when the allowed challenge positions, $x=3$. Both the LadderRank-Combined system and the PageRank system outperform the ATP Official Rankings.

In an attempt to provide further evidence that the ATP ranking system is inaccurate at ranking players with rankings greater than 32, we selected a subset of the matches played only in-between players ranked 32-80 according to the ATP Official Rankings on the 18th of March 2013. This subset was comprised of 275 matches that were played within the period 18/03/2012-18/03/2013. The small size of this subset also hints to the problem of the seeded tournament system as it is only 275 matches out of a total of 2552 that were played in the period. This means that this group of players play a much smaller number of matches between them and as a result there are not enough matches to compare the performance of these players against one another.

Table 3 presents how the ranking systems perform at predicting the outcomes of this subset of matches.

Table 3 – Comparison of predictive power of ranking models over 275 matches played between players ranked in positions 32 to 80 by ATP rankings in the period 18/03/2012 to 18/03/2013.

ATP	Match PageRank	SortRank Uncombined	LadderRank Combined (x=1)	LadderRank Combined (x=3)	LadderRank Combined (x=5)
55.27273%	58.54545%	54.54545%	56.00000%	56.3634%	56.00000%

The ATP Official rankings perform much more poorly in this subset of matches with a success rate as low as 55.27273%. The other models also perform much worse than when using the full range of players but still outperform the ATP Rankings. This generic drop in the success rate of the ranking systems to reflect the outcomes of matches played by players ranked in the range 32-80, could occur for a number of reasons. It could be because the players of this range are more unstable in their performance which adds to the uncertainty of the outcome. Also, since we are comparing a group of players which are more similar to each other, the outcomes in the matches played between players in this group would also have increased uncertainty. Additionally, the small number of matches played between these players also affects the quality of the models: the PageRank model uses match victories to rank players and the “*Combined*” model uses average statistics from these matches. In other words, the seeded tournament system affects all these ranking systems as players ranked 32-80 face each other a lot less, thus reducing the quality of statistics available for these players.

6. Conclusion

We introduced a new, flexible idea for ranking professional tennis players by simulating a “*sports-ladder*” driven by a tennis model in the background. We demonstrated this idea by using existing models from the literature and comparing the rankings that they generate with the official ATP Rankings. We identified problems such as the slow adaptation of the LadderRank-Combined system and discussed how they could be solved by changing the underlying model to account for them.

Despite the slow-adapting underlying model that we used, comparing the LadderRank-Combined system’s performance against the ATP rankings in terms of how well the rankings represent the set of matches used to generate them, the LadderRank algorithm outperformed the ATP rankings.

We also detected the bias created by seeded tournaments which is inherent in the official ranking systems and we provided evidence which support this. By simply comparing the differences in the rankings assigned to players by the various ranking systems we provided evidence towards the bias by detecting an explosion of disagreement with the ATP for players ranked greater than 32. By testing the performance of the ATP Ranking system on a subset of matches that were played between players ranked in the range 32-80, we found further evidence of the poor representation those same players have in the official rankings.

To sum up, even though the “*LadderRank*” ranking system joined with the “*Combined*” model does not perform as well as the PageRank ranking system, it still outperforms the ATP Official Rankings. This proves that it works as an idea. Also, since the quality of the rankings generated by the “*LadderRank*” system is directly dependent on the quality of the model that drives the comparisons of players, by using a more

sophisticated model one can improve the performance further and this is exactly what makes “*LadderRank*” so flexible.

References

- Barnett, T. J. & Clarke, S. R., (2002) Using Microsoft Excel to model a tennis match. *6th Australian Conference on Mathematics and Computers in Sport* (G. Cohen ed.), pp. 63-68.
- Brin, S. & Page, L., (1998) The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, Volume 30, pp. 107-117.
- Clarke, S. R., (1994) *An adjustive rating system for tennis and squash players*. Queensland, Australia, Bond University, pp. 43-50.
- Clarke, S. R. & Dyte, D., (2000) Using official ratings to simulate major tennis tournaments. *International Transactions in Operational Research*, 7(6), p. 585.
- del Corral, J. & Prieto-Rodriguez, J., (2010) Are differences in ranks good predictors for Grand Slam Tennis matches?. *International Journal of Forecasting*, Volume 26, pp. 551-563.
- Dingle, N., Knottenbelt, W. & Spanias, D., (2013) On the (Page)Ranking of Professional Tennis Players. In: M. Tribastone & S. Gilmore, eds. *Computer Performance Engineering*. s.l.:Springer Berlin Heidelberg, pp. 237-247.
- Hoare, C. A. R., (1961) Algorithm 64: Quicksort. *Commun. ACM*, 4(7), p. 321.
- Klaassen, F. J. G. M. & Magnus, J. R., (2001) Are Points in Tennis Independent and Identically Distributed? Evidence from a Dynamic Binary Panel Data Model. *Journal of the American Statistical Association*, 96(454), pp. 500-509.
- Radicchi, F., (2011) Who Is the Best Player Ever? A Complex Network Analysis of the History of Professional Tennis. *PLoS ONE*, Volume 6, p. e17249.
- Spanias, D. & Knottenbelt, W., (2012) Predicting the outcomes of tennis matches using a low-level point model. *IMA Journal of Management Mathematics*.

The London Olympics in Perspective: Athletics, Swimming and Home Nation Medal Advantage

Raymond Stefani*

* California State University, Long Beach, USA. Raystefani@aol.com

Abstract. A mention of London 2012 brings to mind images of the charismatic Usain Bolt in athletics, Michael Phelps with his record haul of swimming medals and Jessica Ennis striking gold in the women's heptathlon, with all of Great Britain's weight on her shoulders. There is much more to learn about athletics, swimming and home nation medal advantage. Athletics winners at London were worse than those in 1988, after which a crackdown began on performance enhancing drugs. The fraction of 1988 winners who would still win dropped from 68% in 2000 to 50% in 2008 only to rise to 57% in 2012. One possible cause of that recent deterioration was an increase in proactive anti-doping surveillance. Following the 2008 Olympics, when men's swimming times improved 1.69% over one Olympiad, FINA concluded that high tech suits were contributing to that improvement (even though times had improved 1.63% per Olympiad from 1956-1988). Effective January 1, 2010, lower tech suits had to be used. If FINA was correct, times should have especially deteriorated for men. That did not happen. There had been a 0.82% average four-year improvement for the last World Short Course, World Long Course and Olympic competition using the high tech suits. In fact, there was a higher average four-year improvement of 0.99% for those same three competitions when the lower tech suits were first worn. Over the last 13 fully attended Games, the home nation won 13 more medals at home than four years before and then won seven fewer medals, four years after being host. For London, Great Britain was projected to win $47+13=60$ medals (65 were won). China was projected to win $100-7=93$ medals (88 were won). The average error was zero. For Rio in 2016, Brazil is projected to win $17+13=30$ medals while Great Britain should win $65-7=58$ medals.

1. Introduction

A mention of London 2012 brings to mind images of the charismatic Usain Bolt in athletics, Michael Phelps with his record haul of swimming medals and Jessica Ennis striking gold in the women's heptathlon, with all of Great Britain's weight on her shoulders. There is much more to learn about athletics, swimming and home nation medal advantage.

Athletics at the London Games inherited the legacy of past indiscretions regarding the use of performance enhancing drugs. The infamous 100 m run at the 1988 Games was "won" by Canada's Ben Johnson in "world record time", only to have those accomplishments wiped forever from the record books when he was disqualified for using steroids. The subsequent Dubin (1990) inquiry initiated by the Canadian government determined that the use of performance-enhancing drugs was rampant in athletics and weightlifting. Performances in athletics at London will be compared to those from 1988 to demonstrate the affect of reduced use of such drugs post-1988 and another apparent affect of introducing pro-active drug testing (the biological passport) post 2008. Some 57% of 1988 gold medal winners would still have won in London including some iconic names from the past.

As Michael Phelps won more gold in swimming, he did so in a lower tech swim suit mandated for use starting in 2010 to counteract high tech suits that had been supposed by FINA to grant unreasonable advantages to the swimmers. Three sets of recent winning performances in swimming will be compared before and after introduction of those new lower tech suits to determine whether performances were degraded as one would suppose, had those high tech suits given advantage. Some direct laboratory test of various high tech suits will be summarized to be compared with the results from actual competition.

Gold medal-winning athletes such as Jessica Ennis have made the home nation proud. Just how much advantage does the home nation gain in medal count? Comparative data from the last 13 fully-attended Games are examined. The accuracy of predicted 2012 medal count for Great Britain (home nation) and China (four years after being home nation) are revealed and predictions are made for the medal count in 2016 for the home nation Brazil and for Great Britain's legacy of hosting the Games four years before.

2. Athletics

Many of the winners at London might not have won at all had the 1988 winners still been in competition (removing age as a barrier, of course). We would have seen the 1988 winner Florence Griffith-Joyner (USA, with her amazingly long fingernails) become a double winner for women beating 2012's Shelly Fraser-Pryce (JAM) at the 100 m final (:10.54 to :10.75) and Allison Felix (USA) at the 200 m final (:21.34 to :21.88). Moving to the long jump, 1988's Jackie Joyner-Kersee (USA) would have been the women's winner over 2012's Brittney Reese (USA) (7.40 m to 7.12 m) while iconic 1988 winner Carl Lewis (USA) would have won again for the men over Greg Sunderland (GBR) (8.72 m to 8.31 m).

Some 34 in-stadium events were contested in both 1988 and 2012. The marathons and walks that left the stadium and were contested over varying terrain are not included nor are scored events such as the decathlon for which the table has changed. Table 1 includes the 19 winners from 1988 that would have still won in 2012 as well as one 1988/2012 tie in the men's high jump. Thus 19.5 of 34 or 57% of the 1988 winners would still have reigned supreme at London 24 years later. Note that ten of the twenty 1988 winners in Table 1 are from nations that no longer exist, the USSR (6) and East Germany (4).

Table 1 19 Events Where the 1988 Athletics Olympic Winning Performances are better than the 2012 Winning Performances and the Men's High Jump where Performances are Equal

<i>Event</i>	<i>1988 Winner</i>	<i>1988 Performance</i>	<i>2012 Winner</i>	<i>2012 Performance</i>
<i>M = Men</i>				
<i>W = Women</i>				
M 400 m	Steve Lewis (USA)	43.87	James (GRN)	43.94
M 5 km	Ngugi (KEN)	13:11.70	Farah (GNR)	13:41.66
M 10 km	Boutalb (MOR)	27:21.46	Farah (GNR)	27:30.42
M 3000 m Steeple.	Kariuki (KEN)	8:05.51	Kemboi (KEN)	8:18.56
M 400m Hurdles	Phillips (USA)	47.19	Sanchez (DOM)	47.63
M 4x400m Relay	USA	2:56.16	JAM	2:56.72
M Discus	Schult (E Ger)	68.82 m	Harting (GER)	68.27 m
M Shot Put	Timmerann (E Ger)	22.47 m	Matewski (POL)	21.89 m
M Hammer Th	Litvinov (USSR)	84.80 m	Krisztian (HUN)	80.59 m
M High Jump	Avdeyenko (USSR)	2.38 m	Ukhov (RUS)	2.38 m
M Long Jump	Carl Lewis (USA)	8.72 m	Rutherford (GBR)	8.31m
W 100 m	Griffith-Joyner (USA)	10.54 w	Fraser-Price (JAM)	10.75
W 200 m	Griffith-Joyner (USA)	21.34	Felix (USA)	21.88
W 400 m	Byrzgina (USSR)	48.65	Richard-Ross (USA)	49.55
W 800 m	Wodors (E Ger)	1:56.10	Savinova (RUS)	1:56.19
W 1500 m	Ivan (USSR)	3:53.96	Alptekin (TUR)	4:10.23
W 4x400 m	USSR	3:15.18	USA	3:16.87
W Discus	Hellmann (E Ger)	72.30 m	Perkovic (CRO)	69.11 m
W Shot Put	Lisovskaya (USSR)	22.24 m	Adams (NZL)	20.70 m
W Long Jump	Joyner-Kersee (USA)	7.40 m	Reese (USA)	7.12 m

In order to put into perspective the information from Table 1 covering 1988-2012, Table 2 covers that period and also a comparable earlier period from 1968-1992. None of the 1968 winners would still have won 20 and 24 years later (compared to 50% and 57% respectively for the 1988 winners). Just four years later, only 24% of the 1968 winners could have won compared to 79% superiority by the 1988 winners in 1992. Conversely, in 2000, 12 years after 1988, 68% of the 1988 winners would still have won versus only 7% of the 1968 winners who would have won in 1980.

Notice that progress was made by post-1988 winners in athletics in 2004 and 2008 in that the percent of still-dominant 1988 winners dropped to 56% and then to 50%. By 2008, parity was achieved by those winners under more intense scrutiny for using performance enhancing drugs compared to those of 1988. In fairness, the winners from 1988 that were not disqualified must be assumed innocent until proven guilty. Given the Dubin Inquiry (1990) conclusions regarding the rampant use of performance enhancing drugs in 1988, the honest athletes would have had to train with unprecedented effort to keep up (the bar was raised). Following action by the International Olympic Committee to create and empower the World Anti-Doping Agency, the cheaters would have been less able to benefit from those drugs and the bar would have been lowered for the rest. The WADA website <http://www.wada-ama.org/en/About-WADA/History/A-Brief-History-of-Anti-Doping/>, claims that "in the 1990s, there was an evident connection between more effective

test methods and a remarkable drop in the level of top results in some sports. the downturn in performances in some events was due to anti-doping scrutiny".

Table 2 Percent of Olympic Athletics Winners from 1968 and from 1988 that Would Still Win Subsequently

Years Later	1968 Winners Would Still Win		1988 Winners Would Still Win	
	Olympics	% That Would Win	Olympics	% That Would Win
4	1972	24	1992	79
8	1976	7	1996	41
12	1980	7	2000	68
16	1984	14	2004	56
20	1988	0	2008	50
24	1992	0	2012	57

The landscape of athletics performance in the years after 1988 was as different from that just after 1968 as the terrain of Mars differs from that on Earth. Two issues will now be considered. First, the amount of improvement though 1988 will be quantified and contrasted with that of the post 1988 period. Second, why might performances have regressed in 2012 compared to 2008? Table 3 is aimed at the former. The latter will be considered shortly.

Table 3 Percent Improvement Per Olympiad (%I/O) For Athletics Winning Performances in Indicated Games

Period	Olympics	%I/O Running	%I/O Jumping	%I/O Throwing	%I/O Overall
WW1 and Recovery	1900-1924	1.91	2.41	5.19	2.76
WW2 and Recovery	1928-1952	0.92	1.57	4.27	1.82
Cold War	1956-1976	0.79	1.93	4.39	1.88
Boycotts and Recovery	1980-1988	0.64	1.90	1.69	1.08
Post 1988	1992-2012	-0.07	-0.08	-0.30	-0.12
All	1900-2012	0.63	1.35	2.76	1.24
	1992	-0.76	-0.79	-3.14	-1.20
	1996	0.26	0.50	-0.08	0.20
	2000	-0.42	-1.16	-0.45	-0.60
	2004	0.41	1.65	0.33	0.65
	2008	0.33	-0.11	0.82	0.34
	2012	-0.17	-0.75	-0.03	-0.26

No results were calculated for the men's javelin in 1988 and women's javelin in 2000 (javelins were rebalanced), for walks and marathons (terrain varied), and for the heptathlon and decathlon (scoring tables varied).

Five periods of Olympic history are chosen for Table 3, each period representing one cycle of Games as affected by outside influences. The percent improvement per Olympiad for Olympic winning performances, %I/O, is first found for each athletics event of the Games of 1900 compared to 1896. Subsequently, improvement is similarly found for each event of a given Games compared to the immediately preceding Games and then averaged separately for running, jumping and throwing events. The first group of six Games worth of improvement, 1900-1924, spans WW1 and ends with the second post-WW1 Games, when competition had returned to near normal. The second period of five Games 1928-1952 similarly spans WW2. Women first competed in athletics in 1928; hence, improvements were first calculated for 1932. The third period covers the six Cold-War Games 1956-1976 where East and West competition spawned consistently high improvement. The fourth period includes the two boycotted Games of 1980 and 1984, ending with the fully-attended Games of 1988. The fifth period covers the six post-1988 Games 1992-2012. The values for each of the six recent Games are shown in the bottom part of Table 3. The highest improvements for running, jumping and throwing were in the first period. Improvement in running was about

the same for the second, third and fourth periods as was true of jumping. Improvement in throwing was high for the second and third periods and then it declined during the fourth period. Improvement is negative for running, jumping and throwing as evidenced by the 1992-2012 Games, post anti doping crackdown.

Tables 2 and 3 are consistent in that negative improvement in Table 3 for any of the 1992-2012 Games coincides with an increase in the fraction of 1988 winners who would still win in Table 2; while a positive improvement in Table 3 coincides with a lower fraction of 1988 winners who would still win in Table 2.

The relationship between two performances in running depends on relative power/weight times relative efficiency while the relationship between two performances in jumping depends on the square of relative power/weight times relative efficiency. See Stefani (2006, 2008b) for that and the following analyses. In turn, power/weight depends on (lean body mass)/(total body mass), LBM/m, and training. Efficiency depends on technique, coaching and equipment. A one percent increase in power/weight times efficiency implies a one percent improvement in running and a two percent improvement in jumping. If the percent improvements in running and jumping are known, the percent increase in power/weight times efficiency follows. The average %I/O over all past Games was 0.63% for 405 running events, 1.35% for 146 jumping events and 2.76% for 154 throwing events. Assuming that LBM/m was relatively unchanged over a typical Olympiad, then training times efficiency improved 0.63% per Olympiad.

The amount of elicit advantage due to steroid use in 1988 can be estimated versus 2008. The average %I/O in 2004 and 2008 was 0.37% for running and 0.77% for jumping (about twice as much as for running, consistent with the above data). If we assume that those improvements represent the “new normal” as of 2008 of efficiency improvement, given constant power/weight for a cleaner athlete, then over the 5 Olympiads after 1988 there should have been an improvement of 1.85% in running and 3.85% in jumping instead of zero improvement. Then, power/weight (LBM/m times training) for the athlete of 2008 must have declined a total of 1.85% versus their counterparts in 1988, the “smoking gun” of steroid benefit.

This brings us to the question of why the 2012 results were worse than in 2008, when parity had been achieved versus 1988 under the “new normal”. What happened after 2008? One known change was the introduction of the biological passport by the World Anti-Doping Agency, <http://www.wada-ama.org/en/Science-Medicine/Athlete-Biological-Passport/>. Post 2008, each athlete was then required to submit blood and urine tests on a regular basis (the so-called biological passport). Those stored vials could be tested at any time to detect subtle changes consistent with use of performance enhancing drugs or masking agents. It was no longer necessary to actually detect those agents, only the affect of such use. A stranglehold was thus placed on the ability of an athlete to use such a drug and wait until traces were gone. Beside the loss of efficacy of performance enhancing drugs, there are two other possible explanations for the downturn in performance in 2012 versus 2008. Perhaps the weather in London or the track conditions were negative influences. Prior to 1992, athletics performances regressed only three times: after each world war and due to the boycotting of the 1984 Games. Had weather and track conditions been a sensitive factor, downturns should have been noted before. We are left with the conclusion that the downturn in performance from 2008 to 2012 was due to the implementation of the biological passport. The performance downturn in 2012 compared to 2008 was about equal to a reduction in power/weight of 0.65%, about 1/3 of the 1.85% noted above from 1988 to 2008. The total benefit of the use of performance enhancing drugs in 1988, versus the “new-new normal” of 2012, with the biological passport in place, is estimated to be 1.85% plus 0.65% or 2.5%. The honest athlete can now work on LBM/m enhanced by nutrition while training and efficiency can be enhanced by dedicated effort in preparation for the Games of 2016, while worrying much less about others cutting corners. We hope to see something of the order of the improvements of 2004 and 2008 in Rio.

3. Swimming

For over 80 years, Speedo and other manufacturers have touted the ability of newly designed suits to improve swimming performances. The prevalent men’s suit changed from full upper-body coverage in the 1920s, to minimalist suits of the 1970s, back to full upper-body coverage again in recent years. The full upper-body coverage of the 1920s seems to have been reinvented as of 2008. See Stefani (2012a) for an evocative photo from 1924 and discussion of swim suit trends over the years. Ironically, the same Michael Phelps who has served as model/spokesperson for high tech suits has contributed to showing that the

swimmer and not the suit has caused performances to improve, using his and other winning performances from the 2012 Olympics.

Following the 2008 Olympics, when men's swimming times improved 1.69% over one Olympiad, the world swimming federation, FINA, concluded that high tech suits were contributing to that improvement (even though times had improved 1.63% per Olympiad from 1956-1988). We could argue that it was a matter of remarkably poor memory and not remarkably high improvement. As of January 1, 2010, lower tech suits had to be used. If FINA was correct, times should have especially deteriorated for men. That did not happen. Table 4 summarizes the four-year improvement for the last three major competitions using the former high-tech suit versus the same three competitions when the lower tech suits were first used, based on the winning performances for men.

Table 4 The Last use Of High-Tech Suits Versus the First Use of Lower-Tech Suits

<i>Competition</i>	<i>Years</i>	<i>%Improvement Last High-Tech (Compared to High-Tech Four Years Before)</i>	<i>%Improvement First Lower-Tech (Compared to High-Tech Four Years Before)</i>
World Short Course	2002-2006	0.41	
World Long Course	2003-2007	0.37	
Olympics	2004-2008	1.69	
World Short Course	2006-2010		2.45(2.10)
World Long Course	2007-2011		0.91
Olympics	2008-2012		-0.05
Average		0.82	0.99

When the high-tech suits were last used, the average improvement was 0.82% over four years. When the lower-tech suits were first used, the average improvement was 0.99%, a bit higher than before. Had the high-tech suits been a major boost to performance, the lower-tech suits should have engendered significant loss of performance versus those high tech suits used four years before. That did not happen. Note that for the last three uses of the high-tech suit there was one high improvement (at the 2008 Olympics, sparking the FINA reaction) and two lower improvements and similarly for the first three uses of the lower-tech suit there was one high improvement (at the 2010 World Short Course championships) and two lower improvements. It would have been better for FINA to have taken a more comprehensive view than just focusing on the 2008 Olympic swimming competition. The value used for the World Short Course %Improvement of 2012 could have been higher; but the value was shrunk from 2.45% to 2.10%, a four-year equivalent. The competition occurred four years and eight months from that in 2004 to switch to a winter schedule. Further, if there was one competition that would display the disadvantage of a non drag-reducing suit compared to a drag-reducing suit, it would be a short course competition. In such competition, the swimmer negotiates more than twice as many turns as in a long-course race at the same distance. That is, the swimmer glides more through the same type of smooth water used by manufacturers to measure drag. (At 200 m there are three turns in a long course race and 7 in short course race, a ratio of 7/3 or more than twice as many).

Direct laboratory measurements have been made of the most recent two high tech designs by Speedo. Sanders et al. (2001) and Toussaint (2003) used a flume of water with the result that the Fastskin provided no significant drag reduction. In fact, the shark skin-scale feature which is fine for sharks, does not reduce drag for the much slower velocity of human swimmers. Moria et al. (2010) used a wind tunnel to determine that the surface design of the LZR Race did not produce significant drag reduction. Halvorson (2011) measured the compression and lift of the LZR Races but no significant drag reduction was noted.

Table 4 suggests that a one percent improvement is quite possible for the swimming events at Rio in 2016, low tech suit or not. Swimmers should realize that it is the preparedness of the swimmer in the suit that will pave the way for swimming gold, not the suit on the swimmer.

4. Home Nation Medal Advantage

In London, Jessica Ennis was among the 65 medalists for Great Britain (Team GB as it was affectionately called), striking gold in the women's heptathlon, with all of her nation's weight on her shoulders. In team

sports, home teams outperform the visiting team due to three factors: physiology (the home team players are more rested, having traveled less far), psychology (home team supporters are loud and supportive) and tactics (home team players are familiar with the venue), Stefani (2008a). In Olympic competition, one additional helpful factor is present: the home nation automatically qualifies more athletes than it would have had some four years before. Jessica Ennis was able to maintain grace and composure under fire, making her supporters proud, as had many such past Olympic home-nation medalists. Table 5 summarizes the medal advantage that home nations have enjoyed during the last 13 fully attended Games. The first year in Table 5 is 1956, with home nation medal count compared with 1952 when Olympic competition had returned to reasonable normality, post-WW2. Also, 1952 was the year the then Soviet Union reentered competition. The average home nation in Table 5 gained 13 medals versus four years before being host.

Only one host nation earned fewer medals than four years prior to being host: the USA earned seven fewer medal in 1996 than in 1992. That seemingly counter-intuitive result is likely due to international events. The Barcelona Games of 1992 were the last Games for the Soviet Union, called the Unified Team in 1992. By 1996, the constituent republics had their own teams. Further, many international athletes post-1992 probably anticipated a greater chance for success, with the Soviet Union out of the way. Competitive balance improved in 1996 versus 1992. The top three medal-winning nations (medals won) in 1992 were the Soviet Union/ Unified Team (111), the USA (108) and Germany, unified after 1988, (82) for an average of 100 medals. The top three medal-winning nations (medals won) in 1996 were the USA (101), Germany (65) and Russia (63) for an average of 76 medals, 24 fewer than in 1992. Since the USA won 7 fewer medals in 1996, that host nation could be viewed as having had a 17 medal home nation advantage, a much more intuitive result.

Table 5 Home Nation Medal Advantage for the Last 13 Fully-Attended Olympics

Year, Location	13 Completed Games			11 Follow-On Games	
	Host Nation Medals	Four Years Before	Host Nation Medal Increase	4 Years Later	Medal Decrease from Being Host Nation
1956 Melbourne	35	11	24	22	-13
1960 Rome	36	25	11	27	- 9
1964 Tokyo	29	18	11	25	- 4
1968 Mexico City	9	1	8	1	- 8
1972 Munich	40	26	14	39	- 1
1976 Montreal	11	5	6		
1988 Seoul	33	19	14	29	- 4
1992 Barcelona	22	4	18	17	- 5
1996 Atlanta	101	108	- 7	97	- 4
2000 Sydney	58	41	17	49	- 9
2004 Athens	16	13	3	4	-12
2008 Beijing	100	63	37	88	-12
2012 London	65	47	18	58 (est)	
2016 Rio	30 (est)	17			
Average	42	29	13		- 7

Table 5 follows 11 of those 13 host nations to the Games four years later (Canada did not compete in Moscow in 1980 and Great Britain has not yet competed in Rio). The average host nation won seven fewer medals four years later, leaving a six medal gain compared to four years before being host: the residual advantage of new sport infrastructure and a new cadre of athletes.

The average advantages including the London Games (13 and -7) were the same values as after the Beijing Games. Prior to the London Games, in Stefani (2012b), we predicted that Team GB would add 13 medals to their count of 47 in 2008 for a predicted total of 60 medals. In fact, Team GB won five more. Also,

China was predicted to win seven less than in Beijing, amounting to 93 medals in London. The actual value was 88, five fewer. The total predicted medal count for Team GB and for China was exactly correct.

For Rio, following Table 5, Brazil should add 13 medals to their count of 17 in London for a total of 30 medals. Team GB should leave Rio with seven fewer medals than in London, 58 medals.

5. Conclusions

Beclouded by media coverage of the athletics winners at London was the fact that the winning performances were generally worse than in 2008, and in fact were worse than in 1988, some 24 years earlier. Iconic athletes from 1998 would still have won, such as Florence Griffith-Joyner, Jackie Joyner-Kersee and Carl Lewis. In sharp contrast, none of the athletics winners from 1968 would have won 24 or 28 years later. Some 68% of the 1988 winners would still have been supreme in 2000, compared to 56% in 2004 and 50% in 2008 when parity had been achieved versus 1988, the year that Canada's Ben Johnson, apparent winner at 100 m in apparent world record time, was disqualified for use of performance enhancing drugs. The fact that it took until 2008 to achieve parity with 1998 suggests that the 1988 era athletics competitor had about a 1.85% higher power/weight ratio compared to 2008. Post 2008, the biological passport was introduced. That proactive approach to anti-doping appears to have caused the downturn in winning performances in 2012 versus 2008. Total power/weight ratio benefit of performance enhancing drugs 1988 versus 2012 totals about 2.5%. The honest athlete can now work on lean body mass/total body mass enhanced by nutrition while training and efficiency can be enhanced by dedicated effort in preparation for the Games of 2016, while worrying much less about others cutting corners. We can hope to see something of the order of the improvements of 2004 and 2008 in Rio.

Following the 2008 Olympics, when men's swimming times improved 1.69% over one Olympiad, the world swimming federation, FINA, concluded that high tech suits were contributing to that improvement (even though times had improved 1.63% per Olympiad from 1956-1988). When the high-tech suits were last used, the average improvement was 0.82% over four years. When the lower-tech suits were first used, the average improvement was 0.99%, a bit higher than before. Had the high-tech suits been a major boost to performance, the lower-tech suits should have engendered significant loss of performance versus those high tech suits used four years before. That did not happen. Swimmers should realize that it is the preparedness of the swimmer in the suit that will pave the way for swimming gold, not the suit on the swimmer. A one percent average improvement is quite possible for the swimming events at Rio in 2016, low tech suit or not.

Based on the last 13 fully-attended Games, the host nation won 13 more medals than four years before. Based on 11 follow-on Games, the host nation won seven fewer medals four years after being host. For London, Great Britain had been predicted to win 60 medals (65 were won) while China was to win 93 medals (88 were won). For Rio, Brazil should add 13 medals to their count of 17 in London for a total of 30 medals. Great Britain should leave Rio with seven fewer medals than in London, for a 58 medal total.

References

- Dubin, C.L. (1990) *Commission of Inquiry into the use of drugs and banned practices intended to increase athletic performance*. Ottawa: Canadian Government Printing Office.
- Halvorson, L.M. (2011) Analyzing Compression and Buoyancy in Technical Swimming Suits, *Proceedings of the ASME International Mechanical Engineering Congress and Exposition*, Denver, Colorado, USA.
- Moria, H. et al. (2010) Contribution of Swimsuits to Swimmer's Performance, *8th Conference of the International Engineering Associations, Science Direct, Procedia Engineering*, 2(2010), 2505-2510.
- Sanders, R., Rushall, B., Toussaint, H., Stager, J. and Takagi, H. (2001) Bodysuit yourself: but first think about it, *American Swimming Magazine*, 5, 23-32.
- Stefani, R. (2006) The relative power output and the relative lean body mass of World and Olympic male and female champions with implications for gender equity. *Journal of Sports Sciences*, 24(12), 1329-1339.
- Stefani, R. (2008a) Measurement and Interpretation of Home Advantage, *Statistical Thinking in Sports*, Chapman and Hall/CRC Press

- Stefani, R. (2008b) The physics and evolution of Olympic winning performances, *Statistical Thinking in Sports*, Chapman and Hall/CRC Press.
- Stefani, R. (2012a) Olympic swimming gold: The suit or the swimmer in the suit, *Significance*, 9(1), 13-17
- Stefani, R. (2012b) How Many Medals Will Great Britain Win in the 2012 Olympics
<http://www.significancemagazine.org/details/topic/868421/Sports.html>, Published March 05, 2012
- Toussaint, H.M. (2002) The FAST_SKIN body suit: Hip, hype, but does it reduce drag during front crawl swimming. *Proceedings of the XXV Congress on Swimming*, Porto, Portugal.

Performance inequality at the Olympic Games

E. Sterken*

**Faculty of Economics and Business, University of Groningen, The Netherlands, e.sterken@rug.nl*

Abstract. We measure inequality of performances in final events of both the Olympic Summer and Winter Games using the Deaton specification of the Gini coefficient. We observe a decrease of performance inequality over time, which could be attributed to an increase in competition. For some of the events real income per capita increases lead to lower inequality of performances.

1. Introduction

The analysis of sports performances often focuses on the best results. In this paper we analyze the dispersion of top sports performances; we take the Olympic Games final events of some individual as example cases. We first describe dispersion of the range of top individual performances with a focus on inequality of results. And next we try to explain the historical development of performance inequality since the start of the modern Olympic Games in 1896.

In team sports inequality of results often leads to organizational measures to restore competitive balance. In this paper we do not focus on team but on individual performances. Individual sports heroes are often commemorated for their extreme outperformances, but sometimes also have to be unmasked as violators of good sportsmanship after some years. An early diagnosis of sports results could perhaps be helpful in detection of excellence or fraud. Sometimes sports commentators describe a certain event to be more competitive than the other without having an absolute measure. Some measure of inequality would give quantitative support to these subjective statements. And finally it is valuable to analyze whether competitiveness in sports is driven by economic conditions. If so, spreading of economic welfare will increase competition and ultimately lead to better performances.

In this paper we use top-8 results of various Olympic Summer – and Winter Games final events to measure the times series development of sports performance inequality. We use data of track and field, swimming, speedskating, and cross-country events from 1896 up to and including 2012. We construct the Deaton (1997) specification of the Gini coefficient, which is well known in statistics to measure income inequality. We describe the historical development of the Gini coefficients and explain their historical development using data on real gross domestic product per capita.

We find that the Gini coefficient is a valuable measure to describe the intensity of competition. Moreover we do find some evidence for the impact of economic conditions on inequality of sports results. We first describe the ideas to investigate inequality of performances in Sections 2 and 3. Next we introduce the data in Section 4 and present the estimation results in Section 5. We summarize and conclude in Section 6.

2. Measuring inequality of sports performances

In sports statistics one is generally interested in measuring the performance of the best athlete. For instance Kuper and Sterken (2003) present an example of an historical analysis of world record performances in speedskating. Usually there is less interest in the dispersion of the results. In sports competitions the inequality of the results is interesting, in particular if one wants to measure competitiveness. How is the top performance to be considered knowing the rest of the results? Considering inequality or asymmetry of sports results it is appropriate to make a distinction between team and individual sports. In general sports competition is more interesting if the final result is rather unpredictable. A strong dispersion of sports results therefore might increase public interest. In particular, if the dispersion is an unbiased reflection of differences in quality, spectators are willing to accept and applaud clear performance differences. But inequality of results can also become predictable and especially in team sports this might lead to a lower attention. There are various initiatives to stimulate competitive balance principles (see Sanderson and Siegfried, 2003). For instance sports organizations are interested in creating the optimal level playing fields. Sometimes, equalizing principles like salary caps, first drafting rights (like in the NBA) are used to create more equality among competitors.

These principles can be applied to team sports, but are less applicable to individual sports. There are exceptions though and in this paper we present an example for the 50k cross-country skiing event, where the organization changed the model of an individual start procedure into a mass start. Sometimes, organizational incentives given to athletes can lead to uneven performances. Frick and Prinz (2007) present empirical evidence for professional marathon athletes, who are sensitive with respect to changes in prize money. But illegal stimuli can lead to unequal performances, like in recent times in professional cycling. So finding an interpretation of unequally distributed performances can find its origin in many aspects, ranging from differences in talent, in incentives, in training background, in (lack) of competition, or even illegal stimuli.

To measure inequality the Gini coefficient (Gini, 1912) is an interesting measure. Usually the Gini coefficient is used to measure inequality of income and or wealth. For instance, for the US the wealth inequality exceeds the income inequality. The Gini coefficient is usually based on the Lorenz curve, which plots the proportion of variable A that is represented by the bottom y percent of the sample population. If the Gini coefficient measures 1 the distribution is unequal; for a Gini coefficient of 0 the distribution is fully equal. Deaton (1997) proposes to use the ratio to the mean of half of the pairs of the absolute deviations between the population elements. In a population $x(i)$, (with mean xm) of size N there are $N(N-1)/2$ distinct pairs, so the Gini coefficient is:

$$G = \frac{1}{xm} N(N-1) \sum_{i>j} \sum_j |x(i) - x(j)|. \quad (1)$$

It is more convenient to write the Gini coefficient in the equivalent form GD , see Deaton (1997):

$$GD = \frac{N+1}{N-1} - \frac{2}{N(N-1)xm} \sum_{i=1}^N \rho(i)x(i). \quad (2)$$

Where $\rho(i)$ is the rank of observation i in the x -distribution. If the Gini coefficient applies to income data $xm(1-GD)$ is a measure of welfare. The Deaton interpretation of the Gini coefficient satisfies the so-called transfer principle. Suppose that without changing the order of the distribution of the elements in the population, a fraction of a larger observation is transferred to a smaller observation, the measure should decrease (or not increase in a weaker form of the transfer principle).

Alternative measures of variation are the coefficient of variation, the standard deviation divided by the mean, and for instance Theil's entropy measure:

$$T = \frac{1}{N} \sum_{i=1}^N \frac{x(i)}{xm} \ln\left(\frac{x(i)}{xm}\right). \quad (3)$$

We do not use the Theil measure in this paper, since it is quite sensitive to outliers with few observations.

3. Description of data

We use data of the Olympic final events in track and field, swimming, cross-country skiing, and speedskating. For track and field we include the 100, 200, and 400 meters events, the high- and long jump, discus throw and for men the marathon event. For swimming we use the 100 meters and the 1500 meters events for men. For cross-country skiing we use data of the 50k marathon event, and for speedskating the 500, 1000, 1500 meters, and 5k events for men and women, and the 3k meter for women and 10k for men.

The progress of the performances in these events is summarized for the 1948-2010/2012 timeframe in Table 1. For the winter events we use data before and including the 2010 Vancouver Games data and for the Summer events the 2012 London Games. One can observe a wide spread in the rate of progress across the events. In general the progress is larger for women than for men.

Table 1 – Percentual progress rates of ‘golden’ performances 1948-2010/2012

	Track and Field	
	Men	Women
100m	6.50	9.66
200m	8.44	10.33
400m	4.89	4.73
marathon	17.33	
discus	29.35	64.86
long jump	6.27	25.02
high jump	20.20	22.02
	Swimming	
100m	17.07	20.06
1500m	24.81	
	Speedskating	
500m	19.00	17.11
1000m	13.09	18.64
1500m	23.28	19.50
3000m		22.83
5000m	26.46	5.35

We use data for the best 8 performances when available. We use the winner's performance as an indicator of the progress rate. Next we compute the Deaton measure of the Gini coefficient GD . We weight the best time and the largest and winning distance in high- and long jump and discus throw events with the largest weight. In Tables 2 and 3 we present descriptive statistics of the time series of the GD for different events for men and women respectively.

Table 2 – Descriptive statistics: Gini coefficients men’s events

	Mean	Median	Maximum	Minimum	Observations
Track and Field					
100 meter	0.0066	0.0060	0.0129	0.0026	27
200 meter	0.0082	0.0080	0.0146	0.0036	26
400 meter	0.0100	0.0069	0.0575	0.0019	26
Marathon	0.0122	0.0073	0.0869	0.0019	27
High Jump	0.0117	0.0090	0.0404	0.0032	27
Long Jump	0.0177	0.0175	0.0349	0.0073	27
Discus	0.0197	0.0187	0.0391	0.0110	27
Swimming					
100 meter	0.0075	0.0070	0.0140	0.0035	25
1500 meter	0.0109	0.0097	0.0251	0.0019	24
Speedskating					
500 meter	0.0039	0.0033	0.0088	0.0013	20
1000 meter	0.0035	0.0036	0.0057	0.0017	10
1500 meter	0.0053	0.0049	0.0155	0.0023	20
5k	0.0062	0.0061	0.0091	0.0031	20
10k	0.0085	0.0071	0.0296	0.0048	20
Cross Country					
50k	0.0090	0.0070	0.0254	0.0001	21

Table 3 - Descriptive statistics: Gini coefficients women's events

	Mean	Median	Maximum	Minimum	Observations
Track and Field					
100 meter	0.0086	0.0076	0.0180	0.0041	20
200 meter	0.0078	0.0082	0.0101	0.0037	17
400 meter	0.0082	0.0071	0.0152	0.0043	13
High Jump	0.0127	0.0120	0.0205	0.0068	20
Long Jump	0.0149	0.0141	0.0280	0.0105	17
Discus	0.0268	0.0201	0.0726	0.0141	20
Swimming					
100 meter	0.0112	0.0076	0.0384	0.0033	23
Speedskating					
500 meter	0.0058	0.0056	0.0095	0.0022	14
1000meter	0.0056	0.0051	0.0109	0.0019	14
1500 meter	0.0060	0.0050	0.0100	0.0029	14
3k	0.0075	0.0078	0.0108	0.0033	14
5k	0.0078	0.0078	0.0113	0.0036	7

Tables 2 and 3 show that speedskating events are competitive, because they have low average and median values of the Gini coefficients. Overall, we find that the inequality measure decreases over time. In Figure 1 we plot the averages across the events for men and women. The numbering starts at the 1896 Athens first edition of the Olympic Summer Games and ends with the London 2012 Summer Games and 2010 Vancouver Winter Games (number 27). Two things can be noted. The first observation is that before World War II the inequality of results was by far larger than after World War II. Secondly, one can see for the events for women an increase in inequality at the end of the 1970's and early 1980's. If any, this could point at the influence of the use of illegal drugs or doping. Still there is a slightly larger equality of results for men in the last editions.

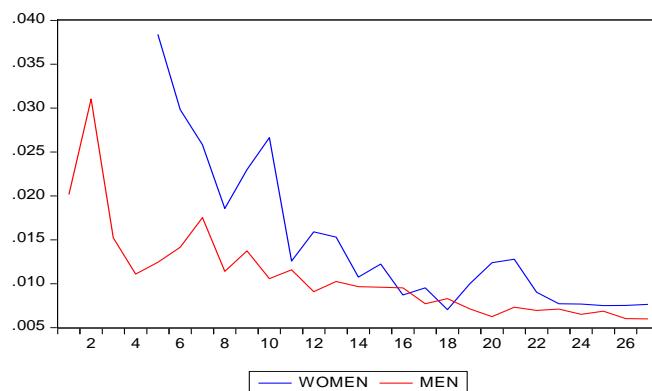


Figure 1 – Average Gini coefficients for men's and women events

Next we show four special cases to illustrate the type of data we use. We present the graphs for the 100 meters dash for men (TFM100MGINI), the marathon for men (MARMGINI), the cross-country 50k for men (CC50KMGINI) and the speedskating 3k event for women (SKW3000MGINI). The 100 meters dash figure shows that on average inequality is rather stable after World War II. For the marathon event we see that before World War II there have been events with largely unequal results, but inequality is rather stable afterwards. For the cross-country event one observes the impact of the mass start in recent years, leading to

the most competitive event in 2006, where the number 8 finisher was only 3.3 seconds after the winner (with a time of 2 hours 6 minutes 11.8 seconds). Finally the plot for the 3k speedskating event for women illustrates again a rather constant inequality, but with a peak period in the early 1980's like observed before.

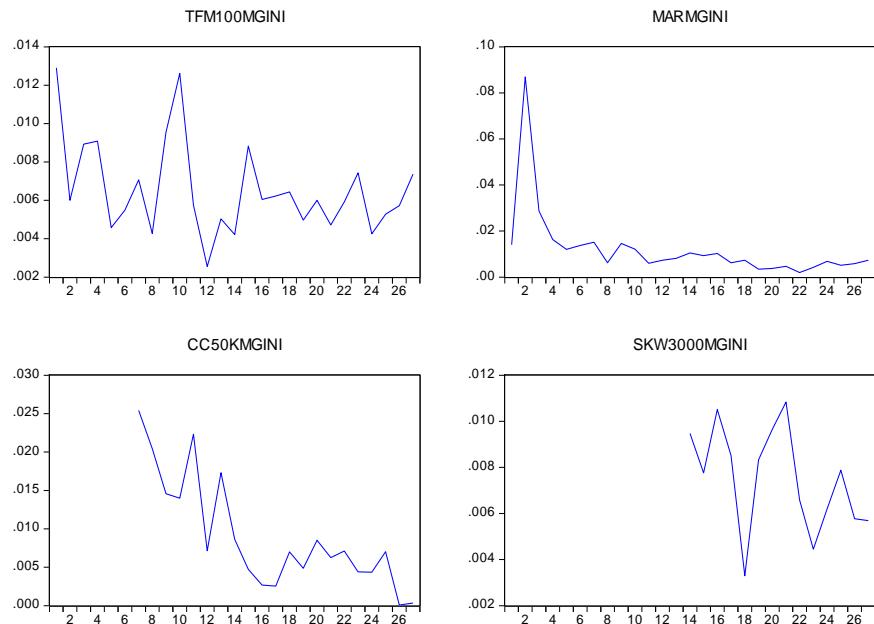


Figure 2 –Gini coefficients for 100 meter dash for men (TFM100MGINI), the marathon for men (MARMGINI), the cross country 50k for men (CC50KMGINI) and the speedskating 3k event for women (SKW3000MGINI).

5. Results

In this section we explore a simple model to explain the time series development of the Gini coefficients of the Olympic final event results. We use three determinants. First an indicator of progress, here measured by a variable labeled real gross domestic product per capita, as provided by the Growth and Development Center of the University of Groningen (www.ggdc.net). This variable also picks up the role of a time trend, explaining technological progress (see Haake, 2009), progress in training facilities, sports knowledge, etc. Next we include the performance of the winner of the gold medal. This variable represents progress, but also indicates the occurrence of an exceptional individual athlete (for instance Jesse Owens or Usain Bolt in track and field). And finally we include for the track and field events the temperature, as indicated by Peiser and Reilly (2004). This variable is an indicator for the circumstances under which the athletes had to perform their performances in competition.

We use a simple loglinear model in the performance of the winner and the real GDP per capita, allowing for an interpretation of the estimated parameters as elasticities. We use the White-correction for heteroskedasticity and indicate significance of the parameters by indication in bolds. We present the estimated standard errors in the line below the parameter estimates.

Table 4 illustrates that the golden performance generally increases inequality, especially in the track and field running events. Generally an increase in real world income per capita decreases inequality, indicating that income increases allow for more athletes to compete and therefore increase competition. On average we find that a 1%-increase in real per capita income leads to a 1%-decrease of the Gini coefficient. The temperature only had an impact on the men's track and field results: higher temperatures led to more inequality in the sprints events and the high jump.

Table 5 shows that the model fits the shorter distances for men's speedskating events and the women's events in general. Real income increases increase equality of results. And here we clearly find that the gold medal results generally shift the overall performance, leading to more equality.

Table 4 – Estimation results Summer Games events; White-corrected standard errors; significance at the 5%-confidence interval indicated by figures in bold.

	Gold	GDPCAP	Temperature	Intercept
Track and Field - Men				
100 meter	11.90 2.961	-0.726 0.150	0.011 0.011	29.133 8.134
200 meter	14.30 2.826	-0.661 0.207	0.041 0.008	43.60 10.38
400 meter	15.18 4.440	-1.130 0.256	0.064 0.023	62.19 18.86
Marathon	1.652 1.065	-0.538 0.256	0.032 0.027	-15.49 11.47
High jump	5.795 3.479	-1.109 0.459	0.027 0.020	0.669 1.868
Long jump	2.884 1.452	-0.635 0.152	0.002 0.013	-4.230 2.009
Discus	0.432 0.828	-0.110 0.231	-0.011 0.011	-0.972 1.715
Track and Field - Women				
100 meter	16.13 3.030	-1.139 0.187	-0.016 0.025	45.58 8.785
200 meter	3.609 2.828	-0.590 0.400	0.043 0.023	11.12 12.33
400 meter	13.81 5.546	-1.165 0.374	-0.011 0.029	60.92 22.44
High jump	3.659 2.588	-0.696 0.357	-0.025 0.026	0.573 2.184
Long jump	3.031 2.028	-0.326 0.263	-0.004 0.019	-6.676 1.977
Discus	0.708 0.803	-0.794 0.310	-0.009 0.021	1.240 1.382
Swimming - Men				
100 meter	2.251 3.710	-0.095 0.604		-13.01 20.49
1500 meter	0.682 3.066	-0.463 0.496		4.452 25.75
Swimming - Women				
100 meter	1.490 3.470	-0.513 0.574		-5.942 19.57

Table 5 – Estimation results Winter Games events; White-corrected standard errors; significance at the 5%-confidence interval indicated by figures in bold.

	Gold	GDPCAP	Intercept
Speedskating - Men			
500 meter	6.212	-1.555	32.09
	3.966	0.569	19.90
1000 meter	11.39	-3.604	79.49
	5.880	1.363	38.66
1500 meter	3.528	-1.224	23.40
	3.179	0.713	22.06
5k	2.781	-0.713	18.68
	2.812	0.557	22.40
10k	4.475	0.736	-42.08
	5.650	1.117	49.08
Speedskating - Women			
500 meter	-15.91	-4.032	93.91
	3.010	0.774	18.67
1000 meter	-12.52	-3.886	88.63
	4.780	1.308	33.82
1500 meter	-9.351	-3.129	71.29
	2.122	0.572	15.60
3k	-10.16	-3.198	83.64
	2.274	0.558	18.08
5k	-16.64	-4.681	143.61
	4.861	1.638	45.33
Cross-country - Men			
50k	-2.086	-2.457	37.57
	2.363	1.361	34.31

6. Summary and conclusions

Using historical data for both Summer and Winter Games final events we analyze inequality of performance results. We show that over time generally results have become more competitive, especially for women. We construct a simple model to explain inequality of results using the golden performances, real income per capita and in some instances temperature as a measure of circumstances as explanatory variables. We find that especially the results of the track and field events and of the women speedskating events can be explained by ‘golden’ performances and real income per capita increases have led to lower inequality of results, maybe hinting at increases in competitiveness.

We conclude that a simple measure like the Gini coefficient is helpful in analyzing athletic performances. One is able to quantify in a single figure the degree of competitiveness of events. Moreover, the Gini coefficient might be helpful in diagnosing ‘unnatural’ performances in an historical perspective. The other main message in this paper is that economic conditions matter to sports performances: increases in income are a prerequisite to maintain competitiveness and to create a level playing field.

References

- Frick, B. and Prinz J. (2007) Pay and performance in professional road racing: the case of city marathons, *International Journal of Sport Performance* 2, 25-35.

- Gini, C. (1912), *Variabilità e mutabilità, Contributo allo Studio delle Distribuzioni e delle Relazioni Sistiche*, C. Cuppini, Bologna; Reprinted in: *Memorie di metodologica statistica* (Pizetti and Salvemini Eds.), Libreria Eredi Virgilio Veschi, Rome.
- Haake, S.J. (2009), The impact of technology on sporting performance in Olympic sports, *Journal of Sports Sciences*, **27**(13), 1421-1431.
- Kuper, G.H. and Sterken E. (2003) Endurance in speed skating: development of world records, *European Journal of Operational Research* **148**, 2, 65-73.
- Peiser, B. and Reilly T. (2004) Environmental factors in the summer Olympics in historical perspective, *Journal of Sports Sciences* **22**, 981-1002.
- Sanderson, A.R. and Siegfried J.J. (2003) Thinking about competitive balance, *Journal of Sports Economics* **4**, 255-277.
- Wallechinsky, D. and Loucky J. (2012) *The Complete Book of the Olympics: 2012 Edition*, Aurum.

Evaluating Regional Balance in the NCAA Men's Basketball Tournament using the Tournament Selection Ratio

John A. Trono*

*Saint Michael's College, Colchester, VT 05439 (USA), jtrono@smcvt.edu

Abstract. The placement of teams, or players, into a tournament's brackets can have a significant impact on who will emerge as the champion. At Wimbledon, 32 players are seeded among the 128 singles entrants in that particular tennis tournament. UEFA coefficients are used to select those European soccer clubs that will compete for the Champions League and Europa League titles in the following season. Teams that survive qualifying are seeded into groups for the finals of World Cup competition in soccer, basketball and ice hockey. An important question to answer is whether or not the criteria used to determine such seeds will create an equitable tournament bracket. The Tournament Selection Ratio (TSR) was designed specifically to quantitatively address this question with regards to the annual NCAA men's basketball tournament, where 64 teams are selected from over 340 teams located across the entire United States. The TSR metric will be used to evaluate: if the four regions (16 seeded teams in each) in recent NCAA Men's basketball tournament have been evenly balanced; if the appointed selection committee, who creates the tournament bracket, has invited the best teams to compete in this sport's culminating event; and if those invited teams have been assigned the most accurate seeds. The results of current NCAA practices and procedures, as applied by the selection committee for this tournament, will also be compared against previous invitational guidelines.

1. Introduction

The National Collegiate Athletic Association (NCAA) is the official body that oversees every intercollegiate sport in the United States. In almost every one of these sports, a championship tournament is held – after the regular season, and after all postseason, conference tournaments (in most sports) have finished – to allow direct competition to determine who will have earned the annual, prestigious title of national champion. With regards to basketball, the NCAA men's tournament is a roughly month long event where every eligible team has the opportunity to earn an invitation to participate in this extravaganza.

Any team that can win its conference's postseason tournament is automatically invited, and almost half of the teams competing in the NCAA basketball tournament receive these automatic bids; the remaining tournaments spots are filled by the worthiest teams as chosen by the NCAA's selection committee (for this tournament). Starting in 1985, 64 teams have battled to become the national champion in this tournament, with 16 teams (seed #1 to #16) competing in four separate regions. The four regional champions continue on to what is known as the Final Four, where three more games will produce the winner of this single elimination tournament; that team, who has outlasted the entire field of 64, then receives the national champion's trophy.

The number of conferences receiving automatic bids has varied between 29 and 32 since 1985, and 31 tournament positions have been allocated each year since 2001. In 2001, the field was also expanded to 65 teams, with the two teams from the weakest conferences (that typically have received automatic bids) competing in a 'play-in' game. The field was expanded again to 68 teams in 2011, now with four play-in games. These additional invitations have increased the selection committee's chances of not excluding qualified teams when selecting who will receive the remaining at-large bids. However, there is always

plenty of discussion concerning did the committee invite the best teams, from those remaining, after the automatic bids have been awarded?

The rest of this paper will describe: the Tournament Selection Ratio (TSR), and its performance predicting the teams that will be invited to the NCAA tournament by said tournament selection committee as well as how accurate the assigned seeds have been; the model used to determine the probability that a team will reach the Final Four; the TSR strength function and its predicted effectiveness regarding the teams reaching the Elite Eight (the teams still left in this tournament in the round before the Final Four begins); and finally, how balanced the four NCAA tournament regions have been since 1985 in comparison to the regional balance before the field expanded to 64 teams that year.

2. Essentials of the Tournament Selection Ratio

The NCAA does not hold a tournament to determine who the national champion is with regards to the sport of football. Instead, the results of all the games that season inform the Bowl Championship Series (BCS) methodology, which has evolved since its inception in 1998. The TSR mimics the BCS formula in several ways. Normalized results from the two polls (sports writers in one, and coaches in the other) comprise two thirds of the BCS formula, while six computer ranking models contribute the other third. Using the trimmed Borda method, the lowest and highest computer rankings are dropped, and the other four values are averaged (and normalized) before being added into the final BCS ranking. The top two teams in the final BCS standings compete for college football's national championship.

The BCS methodology was modified in 2002 to only employ ranking models exclusively after the NCAA deciding to have all computer systems ignore margin of victory, to lessen the motivation for teams to run up large wins over weaker teams (in hopes that the computer models would consider their team as one of the best teams that year). The TSR does not solely rely on ranking models, which limits the winning point differential in any contest to be at most one point, because limiting margin of victory has been proven to decrease prediction models when forecasting future outcomes (Berry, 2003). The TSR consists of four rating and four ranking systems, utilizing the trimmed Borda method, with this average rank contributing 50% to the final TSR. The two polls, once normalized by the maximum voting points a team can receive, each add another 25% to this ratio (which ranges from zero up to a maximum of one, since the trimmed Borda mean is normalized as well). Therefore, the TSR weights equally the objective results, provided by those eight quantitative systems, with the subjective expertise of those selected to participate in the two major polls. Of the eight systems, four have been widely recognized as valuable, quantitative measures of team performance; the other four (two rating and two ranking systems apiece) have been devised by the author over the past twenty years. These eight systems will now be briefly described; more detailed explanations can be found in the references.

The Rating Percentage Index (RPI) was been a mainstay of the NCAA men's basketball tournament selection committee since the early 1980s, helping them to objectively select (and seed) the teams that have best demonstrated their worthiness to compete for this title. The RPI is also a weighted formula, like the TSR, with a team's won-loss percentage contributing 25% to the aggregate. Another 50% comes from the average won-loss percentage of each team's opponents, with the last 25% being derived from each team's opponents' opponent's won-loss percentage. (This formula was modified in 2004 to weight road wins more than wins earned at a neutral site, which are weighted more highly than wins in front of one's home fans. All references below are with respect to the original RPI formula, which I refer to as RP.)

The power rating system (PW, as described in Carroll, et al, 1988) has been verified to be a very accurate predictor of games not yet played (Trono, 2010). This system iterates over all games that season until each team's strength of schedule component, which directly impacts the overall rating a team is

assigned when applying this technique, stabilizes to within a specified tolerance. If the scores are modified so that the largest margin of victory is at most one, then this rating system becomes a ranking system (P1).

Another ranking system, one that was designed for possible inclusion in the BCS formula, is the Rewards system (Trono, 2007). In this system (RW), an average win value is computed by weighting each win in an exponential fashion, once those wins are sorted from best to worst (using the aforementioned P1 system). The team appearing at top of the P1 ranking will have a rating of roughly one, and the rating for the team at the bottom will be roughly zero, with the rest of the team's ratings being normalized in accordance to their P1 rating (versus the best and worst teams). The normalized P1 rating of the best opponent a team defeats is weighted roughly 40% more than the second strongest beaten opponent, which is roughly 40% more than the third best, and so on. Each team's average win value is then multiplied by the number of wins by that team that year, and losses lessen this amount, which produces the team's final Rewards rating, generating the ranking according to this system.

The fourth ranking system included in the TSR (MD, the modified percentage stabilizer) re-computes a team's "winning percentage" in an iterative manner as well to determine how much a win over an opponent should contribute to its modified percentage. Starting with the actual percentage, a new percentage is computed by adding the opponent's percentage in the previous iteration for each win, and subtracting $(1 - \text{opponent's percentage})$ for each loss. Then, each team's new percentage is just this accumulated sum divided by the total number of games they played. After a normalization step, once every game has been considered, this process continues until the new set of percentages converges to be within a specified tolerance of the percentages determined in the previous iteration. Therefore, wins over weaker teams will probably lower a team's next computed percentage, and losses to strong teams only slightly penalize the losing team.

The Sagarin (SG) ratings, for every NCAA football and basketball team, have been included in the *USA Today* daily newspaper since the mid-1980s. Jeff Sagarin devised his rating system to be an undisclosed combination of his pure rating system (that appears to have strong similarities to PW), with his pure ranking system (that is modeled after the ELO chess rating system). With regards to recent NCAA men's basketball tournament games, the opening Las Vegas betting lines have closely matched the differences between the competing team's Sagarin ratings.

The PW system is the sum of a team's average score differential (offensive average – defensive average) and its computed strength of schedule. The expected difference (ED) system relies solely on the average score differential to generate each team's rating. If team A has an expected score of X ($(A\text{'s offensive average} + B\text{'s defensive average})/2$), and its opponent's expected score is Y ($(B\text{'s offensive average} + A\text{'s defensive average})/2$), then the actual game differential above (or below) $X-Y$ is added to A's rating, and likewise with $Y-X$ and B's rating. Good teams typically perform better than the expected game score; therefore, their ratings increase, especially when playing more games against strong teams than weaker ones.

The eighth system is a modification of the discrete rating system (DIS) which is essentially integer-based, as opposed to the seven previously described systems – all of which are based on continuous mathematics. In the DIS (Trono, 2010), each team has an integer rating, and this is multiplied by a specified, constant point value; the difference between the two team's ratings can therefore be used to generate a predicted point spread for a game between them. If that game's actual point spread differential is within a certain range, i.e. plus or minus a specified threshold, then there is no need to update either team's rating. However, if the favorite (let's say team A) wins by more than the predicted differential plus the threshold, its rating is incremented and its opponent's (team B) rating is decremented; likewise, if A

loses and/or doesn't cover the predicted spread minus the threshold, its rating is decremented and B's is incremented.

All teams start the season with a rating of zero, and the ratings earned by the end of the season then becomes a team's initial rating for the next time that same season's games are used to update the ratings; all team's ratings are recomputed in this fashion until every set of ratings matches a set of ratings that have been previously generated, most likely the previous iteration – though some cycles do appear as this stabilization process unfolds. Many teams will finalize to the same integer rating, so to break these ties, penalties for losses are then applied to the integer rating (times the constant point value) almost guaranteeing unique teams ratings for the stabilized DIS (SD). (As reported in Trono, 2010, a threshold value of 1, and a rating point equaling 1 as well, had the best retrodictive prediction accuracy of all combinations when examining performance after the season is over.)

The eight systems used to compute the TSR were chosen to include a wide variety of approaches, and the ratings/rankings were either easy to obtain (SG), or readily computable by software implementation (the other seven). A breakdown of the top ten teams in 2011, according to the TSR, is provided in Table 1. (AP and COA represent the normalized poll quantities, sports writers and coaches respectively; the specific values listed beneath the 8 system names, will be explained more fully a little later on.)

Table 1 – Full breakdown for top 10 teams in 2011 (according to the TSR)

#	TSR	W	L	AP	COA	PW	P1	RP	RW	EX	MD	SD	SG	Team Name
1	0.99787	32	2	0.998	0.999	70	70	69	70	70	69	70	70	Kentucky
2	0.95889	31	2	0.951	0.942	65	69	70	69	66	70	68	66	Syracuse
3	0.91820	30	4	0.898	0.923	64	63	59	68	64	66	69	64	Missouri
4	0.90474	29	5	0.863	0.841	66	68	68	67	68	68	64	65	NorthCarolina
5	0.89756	27	7	0.853	0.852	68	66	66	62	67	64	65	68	MichiganSt
6	0.85912	27	6	0.818	0.756	67	65	65	64	65	65	63	67	Kansas
7	0.84220	27	7	0.756	0.746	69	64	64	63	69	63	62	69	OhioSt
8	0.80545	27	6	0.713	0.719	60	67	67	65	57	67	50	60	Duke
9	0.74028	27	7	0.655	0.621	57	61	63	60	56	61	43	59	Baylor
10	0.69683	25	7	0.561	0.583	55	62	62	59	54	59	46	56	Marquette

3. TSR's Invitation Performance

Now that the basic elements composing the TSR have been described, it is appropriate to evaluate its efficacy. A model known simply as the “Dance Card” (DC) has been developed (Coleman and Lynch, 2001); it is a statistical model that was intended to capture the behavior of the NCAA selection committee, regarding the teams they have invited to the men's basketball tournament. The Dance Card was trained using previous NCAA tournaments (1994-1999), and its performance has been measured ever since. Table 2 illustrates that the TSR has averaged over 90% accuracy when predicting what teams will receive the at large tournament invitations, which is only slightly less than the DC model – and the TSR had no training whatsoever!

The DC model incorporates the RPI, team records against top 50 and top 100 teams, inter-conference records, post season tournament performance, etc., so there is some overlap with what is employed by the TSR. However, the TSR does seem to capture a team's merits, as witnessed by the strong correlation with the selection committee. (More specifics concerning years when certain teams may have been “wrongly” excluded from the tournament can be found in one of the appendices in the lengthy, comprehensive report: Trono, 2013.) The ranking produced via the TSR also correlates nicely with the teams' seeds in

this tournament, as decided by the selection committee. Considering only the invited teams, and assigning the teams with the four highest TSR values a #1 seed, and the next four as #2 seeds, and so on, the average seed difference (between what TSR would have assigned, versus the selection committee) was only 1.056: on average, for the NCAA tournaments from 1985-2011, 24 teams received the same seed in both; 24 more teams had a plus or minus one difference between the two seeding methods; 9 were plus or minus two; and 7 (out of the 64 invited teams) had a seeding difference greater than two.

This examination validates that the TRS formula devised does indeed produce a quantitative measure that matches what the qualified experts have also observed, with regards to the set of teams under scrutiny for tournament inclusion. Table 2 lists how effect each objective system has been, as compared to the DC.

Table 2 - TSR performance predicting NCAA tournament at-large bids.

Span	TSR	PW	P1	RP	RW	EX	MD	SD	SG	DC	All
01-11	340	307	346	338	324	302	332	288	337	351	377
94-00	218	185	215	212	207	187	216	179	205	224	240
85-93	277	251	275	270	260	238	269	230	265	-----	306
Total	835	743	836	820	791	727	817	697	807	575	923
Pct.	91	81	91	89	86	79	89	76	88	93	100

4. Probability Models

Many models have been published that estimate the probability that teams will advance in the NCAA tournament (Brown and Sokol, 2010, and Brady, 2008). These employ quantitative measurements of each team; however, a very simple approach was proposed (Breiter and Carlin, 1997) that only relied on a team's seed to accurately compute how likely it was that a team would reach the Final Four. Their formula for the probability that a team with seed n would defeat a team with seed k is $k / (k+n)$. Using this formula, they then calculated the probability that each seed would reach the Final Four. This strategy was modified (Berry, 2000) to use a seed's strength, all sixteen of which were empirically determined to minimize the sum of the squared error terms when compared against the observed results in the NCAA tournaments from 1985 to 2000. Berry's formula to determine the probability that seed n defeats seed k is $\text{Strength}_n / (\text{Strength}_n + \text{Strength}_k)$, and those seed strengths have performed just as well in all of the years after the specified training period.

To incorporate Berry's strategy, a derived TSR strength value (for each team) was required; the ordering generated by the TSR is converted into a value from 5000 (for the #1 team) down to 101 (for the 70th team in the TSR ranking). As can be seen in Table 1, the highest value listed, beneath the eight computer system names, is 70, and those values decrease indicating where that team was placed by that system, with 70 being at the top of that system's ordering. (Regarding the trimmed Borda value in the TSR, this average is divided by 70, before being multiplied by 0.5, when the TSR rating is calculated.) Because both polls included in the TSR typically only include 35 to 50 teams (each of which receives at least one vote), it seemed reasonable not to rank order all teams (from N down to 1) according to the eight objective systems. Seventy seemed like a good choice because it was only slightly larger than the number of teams invited (64), so most teams competing in the NCAA tournament should appear in some system's top 70 teams. It was also chosen because $70^2 + 100$ is a nice 'round value'. (One hundred is added to reduce the ratio between the strength value of the top team and the #70 team when using the strength value formula: $(71 - \text{TSR rank})^2 + 100$. This ratio is roughly 50 (5000/101), whereas it would 4900 if that

constant (100) wasn't added in the strength value formula. Since no #16 seed has ever defeated a #1 seed, a 2% chance of them winning seemed acceptable. Before 2012, only four #15 seeds have defeated a #2 seed, out of the 108 games played since 1985, so that percentage seemed to be in line with actual results, and is only slightly larger than the #1 vs. #16 percentage. Other similar ratio comparisons were also favorable – with regards to this strength formula.)

Using the updated formula, where the probability that team n defeats team k is $\text{TSRstrength}_n / (\text{TSRstrength}_n + \text{TSRstrength}_k)$, the probability of each team reaching the Final Four can be calculated (Berry, 2000). From 1985-2011, #1 seeds have had, on average, a 28% chance to reach the Final Four, +/- 3%, using this TSR-based model, and the #2 seeds had a 23% chance (+/- 3% as well). Teams with a high TSR ranking, who are also in a weak region, will have a higher expected likelihood of reaching the Final Four than this average.

Table 3 compares the likelihood of each seed making its way through the first four rounds (ignoring all 'play-in' games when more than 64 teams are invited) for the actual TSR strength values from 1985-2011 versus the strength values that would be assigned if the top 64 teams were invited, and the first four were #1 seeds, the next four #2 seeds, and so on down to the last four invited being #16 seeds.

Table 3 – TSR related probability of teams to reach the Final Four.

	Actual Avg.	Prob.	Top 64 Avg.	Prob.
1	4768	28.10	4793	26.69
2	4185	22.94	4261	21.71
3	3678	17.26	3761	16.75
4	3212	12.60	3293	12.26
5	2679	7.27	2857	8.59
6	2310	5.08	2453	5.80
7	1831	2.60	2081	3.66
8	1655	1.85	1741	2.14
9	1367	1.03	1433	1.21
10	1189	0.71	1157	0.66
11	892	0.31	913	0.33
12	829	0.26	701	0.15
13	348	0.02	521	0.06
14	209	0.00	373	0.01
15	59	0.00	257	0.01
16	21	0.00	173	0.00

Table 4 compares the number of teams predicted to reach the Elite Eight versus the actual counts. (The Elite Eight round has twice as many teams still competing in the tournament as the Final Four round, and so the former provides more non-zero counts for the lower seeds – for comparison purposes in Table 4.) The seed strengths, as determined by Berry empirically, are listed as well, but as can be easily observed, these values are quite uneven in their distribution whereas the TSR strength values are smoother when transitioning from the high seeds down to the lower seeds. The expected counts decrease more regularly when the TSR strength values are applied than when utilizing Berry's trained, seed strength values.

Table 4 - Expected number of teams to reach the Elite Eight: 1985-2011

	Actual	Top 64	TSR Str.	Berry	B (Str)
1	78	49	52	76	100
2	53	43	45	51	43
3	25	35	36	24	25
4	14	27	28	14	25
5	7	20	18	9	21
6	12	15	13	15	21
7	6	10	8	8	17
8	7	7	6	4	17
9	1	4	4	4	17
10	7	3	3	5	14
11	5	2	2	3	10
12	1	1	1	2	10
13	0	0	0	0	6
14	0	0	0	1	5
15	0	0	0	0	2
16	0	0	0	0	1

5. Estimating Regional Balance: 1985-2011

Given how well the untrained TSR strength values generated predictions, matching fairly closely those observed totals in Table 4, it appears that the TSR: closely matches where the selection committee places teams in the NCAA tournament bracket, and, the corresponding quadratic, TSR strength value formula succeeds reasonably well in matching the prediction of how far teams will advance in this tournament. The TSR strength value can also be used to evaluate regional balance in the following manner. By summing the TSR strength values of all teams in a region, this will produce a quantitative measure of the relative, overall quality of teams placed into each region. Using a worst case analysis, assigning the #1, #5, #9 ... and #61 teams into one region (i.e. the best #1, the best #2, etc.) and the #4, #8, ... #64 team into another region (i.e. the worst #1, the worst #2, and so on), this produces an acceptable variation in what the TSR strength value sums could be – per region. (After performing said associated calculations, this acceptable regional sum variation is 3696.)

Table 5 lists the TSR strength value sums from 2000, along with the maximum difference between any two regions for each year. Nine of these twelve years have acceptable variations between the regions, i.e. 4300 or less, and the largest difference (7095) occurred in 2007. After careful review, that difference shrinks to 1927 if the two #6 seeds in the specific regions (East and Midwest) are interchanged. The average sum per region in Table 5 is 28980.5, and the average difference between each region's sum, and that value, is 1293.4, which is roughly equal to one quarter of the acceptable variation that would constitute an equitable allocation of the invited teams across the regions.

Table 5 – Aggregate TSR strength values (across the regions)

Year	East	South	Midwest	West	Avg. Diff.	Max. Diff.
2000	31851	29479	25723	29025	29019.50	6128
2001	27173	31423	30173	28397	29291.50	4250
2002	26184	29563	28230	29084	28265.25	3379
2003	27974	27983	29820	31475	29313.00	3501
2004	30445	31552	25856	28882	29183.75	5696
2005	27187	30787	29894	30053	29480.25	3600
2006	27611	30229	28793	28197	28707.50	2618
2007	24103	28531	31198	29282	28278.50	7095
2008	29384	26585	29413	30057	28859.75	3472
2009	28186	29930	26906	29231	28563.25	3024
2010	28678	29306	28762	30748	29373.50	2070
2011	31212	28797	29921	27790	29430.00	3422

For the entire twenty seven year period (1985-2011), over half of those years (fourteen to be exact) the maximum, regional TSR sum difference was than 4000, with another eight years between 4000 and 6000, four more (including 2007) were between 6000 and 8000, and 1988 was the largest difference (at 11095). Even with the latter, swapping two or three judiciously chosen, same-seeded teams (between the regions with the smallest and largest TSR sums) brings even **that** bracket back into line.

Now certainly any regional imbalance can be corrected by exchanging teams between the outlying regions in question. However, when such exchanges are with teams who have been recognized as being roughly equivalent by the tournament selection committee, since those teams have been assigned the same regional seeds, this implies that the correction is fairly minor – especially when compared to how teams were placed into regions before seeds were assigned (as described in the next section).

6. Regional Balance Before 1985

The previous sections illustrated that since 1985, the NCAA tournament selection committee has been quite equitable when placing teams across the four regions, maintaining a reasonably balanced bracket in almost all cases. However, the current methodology in place has gone through quite an evolution since the first NCAA tournament in 1939. From 1951 until 1974, a tournament bracket was published before the season even began, placing designated conference champions into specified slots in geographically determined regions. This template format, known as the ‘tournament draw’, had openings for at large teams that were independent of any conference; six of the sixteen bracket slots in 1951 were reserved for these at large invitations. Prior to 1974, a serious invitation restriction was in effect then: only one team per conference could be invited to compete in this tournament. Many of the conferences automatically rewarded the team who won their conference’s postseason tournament with that conference’s automatic place in the NCAA tournament, which allowed for some really strong teams to be excluded from the NCAA tournament (when they did not win their own conference’s post season tournament).

It seems obvious that predetermining each team’s placement in the bracket (without any consideration to a team’s qualities) could be quite arbitrary, and possibly even unfair to certain teams (or regions). Teams were locked into specific regions by geographical location and/or conference affiliation, and so the

top teams might meet before the Final Four, since teams were not seeded when applying the tournament draw, bracket strategy.

From 1953 to 1974, 22 to 25 teams competed in the NCAA men's basketball tournament, with four to seven teams per region (and even that quantity varied from year to year as well). Given these circumstances, the sum including every team's TSR strength value could be misleading, so two other values will be examined to illustrate how unbalanced the regions were before seeds were introduced into this tournament's bracket formation in 1979.

Table 6 – Sum of the TSR strength values for the top four teams in each region: 1953-1974

	East	Mideast	Midwest	West
Minimum	10350	10806	5836	7385
Maximum	17598	17485	15851	17094
< 10,000	0	0	3	5
10,000-12,999	7	7	10	10
13,000-14,999	5	9	6	4
15,000-16,999	7	4	3	2
> 17,000	3	3	0	1
Average	14414.2	14245.6	12428.9	11993.4
Top four ratio	32.51%	33.50%	37.20%	39.34%
# > 40%	1	2	6	11

The first metric is simply the sum of the four largest TSR strength values associated with teams in each region (since four is the fewest number of teams in any region during this time period). Table 6 contains the results when applying this metric to the 22 years in question: from 1953 to 1974 (since only 16 teams were invited in 1951 and 1952). One obvious observation is that two regions (the Midwest and West) appear to be weaker than the other two during the tournament draw era. (The top four ratio in Table 6 is simply the largest TSR strength value, in each region, divided by the sum of that region's four largest TSR strengths values.) However, Table 6 would not expose weak regions if several very strong teams were to confuse this metric – when all the other teams in that region were weak ones.

Table 7 – Region ratios: teams with highest TSR strength / second highest TSR strength

	East	Mideast	Midwest	West
1985-2011 (Avg)	1.1502	1.1104	1.1436	1.1198
1953-74 (Avg)	1.1752	1.1152	1.2798	2.3769
>1.35	4	2	6	11
>1.5	1	1	2	6
>2.0	0	0	1	2

Therefore, the second metric is simply the ratio of the highest TSR strength value in a region over the second highest one. Table 7 highlights (once again) that when compared to previous practices, the current methodology in place to select teams, and place them into different regions, is vastly superior to prior procedures. Only two times after 1984 has this ratio been greater than 1.3, yet in the Midwest and West regions, where apparently there are not as many strong teams/conferences as the other two, this ratio is was larger than 1.3 seventeen times.

The final table (#8) is a composite snapshot that also magnifies how the tournament draw created mis-scheduling decisions once teams were placed into their designated, geographically aligned bracket slots. First round games were scheduled only when more than 32 teams were invited to this tournament, and each subsequent round has half the number of teams competing in the tournament as the previous round. By the fourth round, only eight teams are left, and the fifth round features the opening games in the Final Four. Table 8 lists the number of games where teams were scheduled to play **one round** before a correctly seeded tournament bracket would pair them against each other. (Table 8 does **not** include the eight games where contests were played two rounds too early.)

Table 8 - Number of games where tournament pairings were one round too early

Span	1 st	2 nd	3 rd	4 th	5 th
1951-74	-----	3.01	15.10	10.42	8.33
1975-78	-----	17.19	6.25	0.39	0.00
1979-84	0.00	6.25	2.50	0.00	0.00
2001-2011	5.21	3.70	5.56	0.93	0.00

Such premature matchups would not occur at Wimbledon, for example, where a properly seeded bracket is employed. When examining such a canonical bracket of 32 players, the #1 player would be scheduled to play the player seeded as #32 in the opening round, and that winner would play the victor of the match between the #16 and #17 seeds. A match between the #3 seed and the #8 seed should not occur until the third or fourth rounds, yet such contests did occur several times, during the early transition period (1975-78) when the NCAA tournament selection process was still evolving, in the opening round!

This NCAA tournament was expanded to include 32 teams for four years (1975-78). The maximum number of teams allowed from each and every conference was also increased to two in 1975, and teams receiving at large invitations could be placed into any region. (A significant increase in the number of premature matchups occurred in this brief period than before – or after – this four year span.) The tournament held in 1979 was the first one where teams were assigned seeds, in an effort to balance the four regions; the cap, pertaining to the number of invitations allowed per conference, was also completely abolished that year. There were 40 teams invited in 1979, and this increased to 48, then 52 and eventually 64 (in 1985).

7. Summary

The Tournament Selection Ratio (TSR) was introduced as one possible, quantitative method to evaluate how well a team has performed over the entire season. The TSR was then used to determine if the most deserving teams have been invited to the NCAA men's basketball tournament as well as generating individual strength values (per team) that can be used to predict the likelihood of a game's outcome in this

tournament. Continuing such probabilistic analysis, these calculations can be extended to cover several rounds of this tournament to see how accurate these predictions/probabilities are.

In both cases, the TSR was seen to be a very accurate predictor regarding which teams would be invited to this tournament as well as deriving the expected number of rounds each team would advance in the tournament. Several other measurements (which rely only on the TSR) were taken, and these illustrated how much more evenly balanced the four tournament regions have been since 1985, when compared to the previous tournament team selection strategies that were employed (prior to 1985).

References

- Berry S. (2000) A Statistician Reads the Sports Page. *Chance* 13(3), 56-61.
- Berry S. (2003) College Football Rankings: The BCS and the CLT. *Chance* 16(2), 46-49.
- Breiter, D.J. and Carlin, B.P. (1997) How to Play the Office Pools if You Must. *Chance* 10(1), 5-11.
- Brown, M. and Sokol, J. (2010) An Improved LRMC Method for NCAA Basketball Prediction. *Journal of Quantitative Analysis in Sports* 6(3), article 4.
- Carroll, B., Palmer, P. and Thorn, J. (1988) *The Hidden Game of Football*. Warner Press.
- Coleman, B.J. and Lynch, A.K. (2001) Identifying the NCAA Tournament ‘Dance Card’. *Interfaces* 31(3), 76-86.
- Trono J. (2007) An Effective Nonlinear Rewards-Based Ranking System. *Journal of Quantitative Analysis in Sports* 3(2), article 3.
- Trono J. (2010) Rating/Ranking Systems, Post-Season Bowl Games, and 'The Spread'. *Journal of Quantitative Analysis in Sports* 6(3), article 6.
- Trono J. (2013) A Longitudinal Study of Regional Bracket Equality in the NCAA Men’s Basketball Tournament. Saint Michael’s College Technical Report: SMC-2013-CS-001. <http://academics.smcvt.edu/jtrono/Papers/BracketStudy.doc>.
- West B. (2008) A Simple and Flexible Rating Method for Predicting Success in the NCAA Basketball Tournament: Updated Results from 2007. *Journal of Quantitative Analysis in Sports* 4(2), article 8.

Metaheuristic Optimisation of Parameterised Betting Exchange Strategies

P.Tsirimpas* and W.J. Knottenbelt**

* Department of Computing, Imperial College London, London, United Kingdom: pt307@doc.ic.ac.uk

** Department of Computing, Imperial College London, London, United Kingdom: wjk@doc.ic.ac.uk

Abstract. Stochastic optimisation algorithms have been growing rapidly in popularity over the last three decades, with a number of methods now playing a significant role in the analysis, design, and execution of betting strategies. This paper presents the optimisation platform of SPORTSBET, an event-driven tool for the quantitative evaluation of generic betting exchange trading strategies. Strategy parameters are automatically refined using a stochastic search heuristic in order to improve strategy performance. Walk Forward Analysis is employed to avoid overfitting. To demonstrate the applicability and effectiveness of the platform, case studies are presented for betting strategies in horse racing.

1. Introduction

Gambling and mathematics have a long mutual history. Though the ability to analyse and act on high-frequency real time betting markets is relatively new and very challenging. Since their introduction in June 2000, betting exchanges have revolutionised the nature and practice of betting. Betting exchange markets share some similarities with financial markets in terms of their operation. However, in stark contrast to financial markets, there are very few quantitative analysis tools available to support the development of automated betting exchange trading strategies.

SPORTSBET (Specification and Performance Optimisation of Real-time Trading Strategies for Betting Exchange platforms) (Tsirimpas & Knottenbelt, 2011) is a toolset developed to specify, execute, back-test and optimise parameterised betting strategies for a wide range of sports. SPORTSBET allows the definition of betting strategies in a novel generic betting strategy specification language (UBEL) as sets of concurrent processes which make use of event-calculus-like operators. Strategy performance is quantified by synchronizing multiple real time or historical data streams with a dynamic market reconstruction. The development of a trading strategy is a complex process consisting of a number of different stages such as: formulation, specification in a computer testable form, back-testing, optimisation, evaluation, real time trading, monitoring trading performance and finally refinement and evolution.

This paper confronts challenges related to the back-testing, optimisation and execution of parameterised automated trading strategies for betting exchange markets, and presents the optimisation platform of SPORTSBET. The organisation of the rest of this paper is as follows. Section 2 explains the back-testing and optimisation process in general. Section 3 explores the practical impacts the different types of search and evaluation methods have upon the outcome and quality of the historical simulation and on the optimisation process and presents the optimisation platform of SPORTSBET. Section 4 presents a case study in horse racing and finally, Section 5 summarises this paper and gives directions to future works.

2. Back-testing and Optimisation preliminaries

Back-testing is a specific type of historical testing that calculates how a strategy would have performed if it had actually been applied in the past. This requires the back-test to replicate the market conditions of the time in question in order to get an accurate result. While back-testing does not allow one to predict how a strategy will perform under future conditions, its primary benefit lies in understanding the vulnerabilities of a strategy as it encountered real world conditions of the past (Wikipedia, 2013). The more accurate the back-testing, the better the real-time trading results are likely to be.

To optimise a trading strategy is to obtain its peak trading performance. Most trading strategies have a set of parameters that highly affect their performance. Any strategy that can accept different values for these parameters is eligible for optimisation. An optimisation algorithm is an algorithmic method that can be applied to solve optimisation problems. Numerous optimisation algorithms are available but choosing one for solving a given optimisation problem depends much on the characteristics of the optimisation problem at hand. Many optimisation methods are especially designed for specific types of search spaces, objective and

constraint functions. This work focuses on optimisation methods that are not dependent on any knowledge about the system or model of the optimisation problem. For example: imagine if you are trying to find an optimal set of parameters for a betting exchange strategy. You have a simulator for the betting strategy and can test any given set of parameters and assign it a quality. And you have come up with a definition for what a strategy parameter sets look like in general. But you have no idea what the optimal parameter set is, not even how to go about finding it. These optimisation problems are known as black box optimisation problems (see Figure 1).

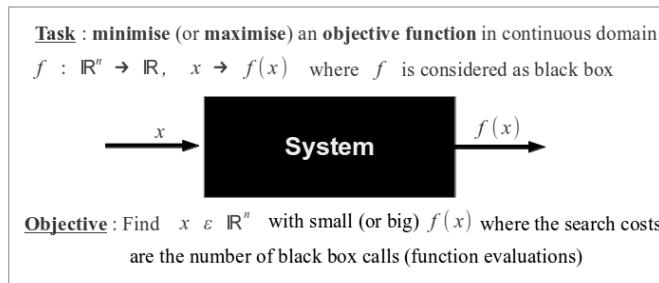


Figure 1 Black box optimisation

Optimisation has many drawbacks and this is because there are many ways that it can be done incorrectly. Usually an optimisation done incorrectly is overfitted. Overfitting occurs when the optimisation process identifies parameters that produce good trading performance on historical data but produces poor trading performance on unseen data. This is because someone can always find a combination of rules and trading parameters that fits perfectly to the available historical data, resulting in exceptional trading results based on those tests. However, when those rules are tested on a live market, they can fail and lose money very quickly. There are also degrees of overfitting. A trading strategy, where the degree of overfitting is not extreme, can still produce real-time profit but it will certainly underperform its optimisation results. On the other hand, a highly overfitted trading strategy will produce disastrous real-time trading losses. Thus to avoid overfitting, it is essential to further test any strategy with the optimized parameters on a set of historical data that is distinct from that used in the optimisation process. Some of the most well known and used techniques are the K-fold Cross Validation (Rodriguez, et al., 2009), Regularization (Gencay & Qi, 2001) and Walk-Forward Analysis (Pardo, 2008). In SPORTSBET optimisation platform Walk-Forward Analysis is employed to avoid overfitting.

3. SPORTSBET Optimisation Platform

During the optimisation process, a historical simulation (back-testing) will be calculated for a large number of different values of the key strategy parameters. In order to do that, all optimisation processes use some type of search method. The methods adopted in stochastic optimisation attempt to model the uncertainty in data by assuming that the input is specified in terms of a probability distribution. Metaheuristics are the most general of these kinds of algorithms, and can be applied to a wide range of problems. The search method will determine the number of back-tests to be performed and therefore the amount of processing time required to complete the process. Moreover, the search method will guide the search in productive directions. However, the directed search methods have some drawbacks. Since a direct search method does not evaluate every possible candidate solution, there is a potential for a certain lack of thoroughness. This can be minimised by selection of the appropriate search method.

In order to retrieve from the optimisation process the trading strategy parameters that are most likely going to produce real-time and long term trading profits, we need to understand the impact of the objective function. The objective function is used during the optimisation process to assign a score in each candidate solution.

3.1 SPORTSBET search method

Evolution strategies (ESs) are robust stochastic search algorithms designed to minimize objective functions f that map a continuous search space \Re^n into \Re . An Evolution Strategy is broadly based on the principle of biological evolution. In each generation (iteration) new candidate solutions (denoted as x) are generated by variation, usually in a stochastic way, and then some individuals are selected for the next generation based on their objective function value $f(x)$. Like this, over the generation sequence, individuals with better and better $f(x)$ are generated. An Evolution Strategy follows these steps:

- **Initialisation:** The initial population can be based on known good solutions or can be generated randomly.
- **Evaluation:** The evaluation uses the objective and constraint functions of the optimisation problem to assign a quality score to each individual. Individuals with a higher score will have a higher probability of surviving and passing on their genetic material (i.e., the candidate solution) to future generations.
- **Selection:** There are two types of selection, the parental selection and the survivor selection. Parental selection is a stochastic selection type that selects the parents that are used for the recombination of a new offspring. In this selection type, the fitter individuals have a higher probability to be selected as parent for recombination. Survivor selection is a deterministic selection that selects the μ fittest individuals either out of the λ offspring (elitist selection) or out of the λ offspring and the μ old parents (non-elitist). This selection type is commonly referred as $(\mu + \lambda)$ when denoting elitist selection, and as (μ, λ) when denoting non-elitist selection.
- **Mutation and recombination:** Mutation operators add small perturbations to the individuals in the population. Recombination operators recombine two or more individuals in the population into a new individual.
- **Termination:** The termination condition can depend on the available computation time the available number of evaluations/generations, or on convergence criteria such as a predefined target fitness that is to be reached. After termination, the best solution(s) is found throughout the evolution cycle.

The simplest Evolution Strategy is the (1+1)-ES (one parent, one offspring) (Hoffmeister & Back, 1991). SPORTSBET optimisation platform uses the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) (Hansen, 2006). The CMA-ES is a $(\mu/\mu_w, \lambda)$ -ES in which all offspring are generated from the same recombinant, computed as the weighted centre of mass of the μ selected individuals. For more details see (Beyer & Sendho, 2008).

3.2 Objective Function

A search method is continually accepting or rejecting trading strategies in the process of seeking the best parameter set in the least time possible. Thus, it is critical to use a proper objective function, which correctly characterises the quality of a trading strategy. As an example if the optimisation process is using as objective function the highest net profit, caution must be exercised in the event that a large proportion of the profit arises from a single large and likely unrepeatable trade with a favourable outcome. Furthermore, using the highest net profit in isolation completely ignores the question of risk. The strategy with the highest net profit could also have a very large and unacceptable drawdown. Or, the strategy parameters selected may have a very small number of trades, which brings into question the statistical validity of these parameters. All of these criteria are very crucial and cannot be ignored when building the objective function.

SPORTSBET's back-testing process returns as its result an array of values representing the evaluation of several objective functions, which the user may use as they are, or combine them to produce a new one. The array contains:

- **Profit:** the difference between the winning and the losing trades.
- **Profit after commission:** the difference between the winning and the losing trades including commission.

- **Maximum Drawdown (MDD):** measures the largest single drop from peak to bottom in an account balance during the life of a strategy.

$$MDD = PV - LV, \quad MDD(%) = MDD/PV \quad (1)$$

where PV is the peak value before the largest drop and LV is the lowest value before new high established.

- **Maximum Run Up (MRU):** measures the largest single increased from bottom to peak in an account balance during the life of a strategy.

$$MRU = PV - LV, \quad MRU(%) = MRU/LV \quad (2)$$

where LV is the lowest account before the largest increased and PV is the highest value before new lowest established.

- **Return On Investment (ROI):** evaluate the efficiency of an investment. Formula:

$$ROI(%) = \text{Profit after commission} / \text{Total Investment} \quad (3)$$

- **Risk-Adjusted Rate of Return (RAR):** measures the amount of risk involved in an investment's return.

$$RAR(%) = \text{Net profit after commission} / (\text{Risk} + \text{Initial Account}), \quad \text{where Risk} = 2 \times MDD \quad (4)$$

- **Reward to Risk Ratio (RRR):** provides an easy comparison of reward to risk.

$$RRR(%) = \text{Net profit after commission} / MDD \quad (5)$$

- **Perfect Profit:** it is the total profit produced if the strategy was winning the highest net profit in each market during the historical period.
- **Number of winning trades:** the number of winning trades. A winning trade is the one where the net profit in a market is positive.
- **Number of losing trades:** the number of losing trades. A losing trade is the one where the net profit in a market is negative.
- **Pessimistic return on margin (PROM):** a measure that pessimistically assumes that a trading strategy will win less and lose more in real-time trading than it did in its historical simulation.

$$PROM(%) = ([\bar{W} \times (WT - \sqrt{WT})] - [\bar{L} \times (LT - \sqrt{LT})]) / \text{margin} \quad (6)$$

where \bar{W} is the average winnings, \bar{L} is the average losses, WT the number of winning trades, LT the number of losing trades, and margin the initial account balance. $PROM$ is a robust measure because it takes in account a number of significant performance statistics as the ones mentioned above. Moreover, $PROM$ penalises the small trade samples because of the adjustment of gross profit and loss by the square root of their respective number.

- **Strategy Efficiency (SE):** measures how efficiently a trading strategy converts the perfect potential profit into realised trading profits.

$$SE(%) = \text{Net profit after commission} / \text{Perfect profit} \quad (7)$$

3.3 Walk-Forward Analysis

The primary purpose of a Walk-Forward Analysis is to determine the consistency of a trading strategy's performance. This can be achieved by judging the performance of a trading system exclusively on data which were never part of the optimisation process which is a far more reliable measure than performance

based only on in-sample simulation. The automatic Walk-Forward Analysis is a system design and validation technique in which you optimise the strategy parameter values on a past segment of market data (“in-sample”), then verify the performance of the strategy by testing it forward in time on data following the optimisation segment (“out-of-sample”). The evaluation of the trading strategy is based on how well it performs on the test data (“out-of-sample”), not the data it was optimised on. The process is repeated over subsequent time segments (see Figure 2).

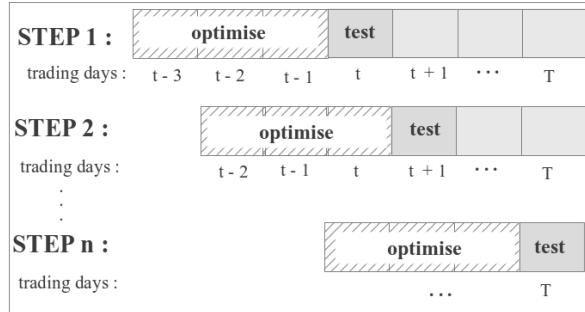


Figure 2 Walk-Forward Analysis

In each step the *Walk-Forward Efficiency (WFE)* is calculated and saved.

$$WFE(\%) = \text{Annualised net profit from testing} / \text{Annualised net profit from optimisation} \quad (8)$$

The average WFE over the Walk-Forward Analysis can be used to provide some estimation of the rate of profit to be earned during real-time trading. Research has clearly demonstrated that robust trading strategies have WFEs greater than 50-60 percent. Finally another thing to take in consideration is the consistency of the trading strategy. Example:

- **Consistency on profits:** 70 percent of the Walk-Forward windows were profitable.
- **Distribution of profits:** no individual time window contributes more than 50 percent.
- **Maximum Drawdown:** no individual time window had a drawdown of more than 40 percent of initial capital.

4. A Horse Racing Case Study

A horse racing case study was used as a means of demonstrating and evaluating the results yielded by the SPORTSBET optimisation platform. The trading strategy is: 1 minute before a race start, back a range of horses $[X_2, X_3]$ where $X_2 \leq X_3$, $X_2 \geq 0$, $X_3 \leq 5$, $X_2, X_3 \in \mathbb{Z}$ for £40 if the odds of the favourite are more than X_1 where $X_1 \in \mathfrak{N}$, and spread the profit across all the backed horses. So we want to find the best combination of X_1 , X_2 , X_3 . The strategy was tested in 990 markets using 10 time windows. The chosen objective function was the PROM and each time window consisted 99 markets. The step size of the Walk-Forward Analysis was three time windows. Table 1 shows an optimisation sample of the time windows 4, 5 and 6. Table 2 shows the Walk-Forward Efficiency of the strategy in each time window, where W is the current optimisation window, I the iteration of the searching algorithm, Ev the number of evaluation (how many times you performed a back-test), CW is the maximum number of consecutive winnings and CL is the maximum number of consecutive losses.

Table 1 Optimisation sample of time windows 4-5-6

<i>W</i>	<i>I.</i>	<i>Ev.</i>	<i>X</i>	<i>Fitness</i>	<i>Prof.After.</i>	<i>MDD</i>	<i>ROI</i>	<i>Profit</i>	<i>MDD(%)</i>	<i>MRU</i>	<i>MRU(%)</i>	<i>Perf.Profit</i>	<i>PROM</i>	<i>RAR</i>	<i>RRR</i>	<i>SE</i>	<i>CW</i>	<i>CL</i>	<i>W</i>	<i>L</i>
6	1	1	[4.6, 3, 5]	25.66	829.58	280	48.23	934.29	28.46	909.58	757.99	5910.23	25.66	32.40	296.28	14.0364	2	6	14	29
6	1	2	[4.6, 1, 5]	14.63	316.85	218.03	18.42	367.21	35.416	471.09	325.95	1595.83	14.63	13.01	145.33	19.855	4	4	27	16
6	1	3	[5.6, 4, 4]	14.82	853.6	280	118.55	928	56.49	1053.6	0	6338.4	14.82	33.34	304.86	13.4671	2	6	4	14
6	1	4	[5.2, 2, 5]	20.98	552.25	120	55.22	604.48	28.01	592.26	370.16	1846.6	20.98	24.65	460.21	29.9066	3	3	14	11
6	1	5	[4.6, 3, 4]	-21.44	7.6	738.8	0.44	88	78.06	826.4	688.66	12288	-21.44	0.21	1.02	0.061849	2	13	5	38
6	1	6	[5, 2, 4]	-1.63	110.34	317.54	9.85	158.25	77.21	376.61	401.77	3339.08	-1.63	4.187	34.74	3.30455	3	6	8	20
6	1	7	[5.7, 4, 5]	5.27	536.6	200	95.82	588	56.56	696.6	1741.5	4681.6	5.27	22.35	268.3	11.4619	1	4	3	11
6	1	8	[6.4, 3, 4]	-1.17	201.2	80	167.67	216	16.62	321.2	200.75	889.2	-1.17	9.31	251.5	22.6271	1	2	1	2
6	1	9	[4.8, 2, 5]	23.49	607.79	159.99	47.48	671.35	34.86	647.79	404.87	2354.48	23.49	26.19	379.89	25.8142	3	4	17	15
6	1	10	[4.9, 3, 4]	-7.42	271.2	440	24.21	336	57.20	609.2	380.75	8126.96	-7.42	9.42	61.63	3.33704	1	11	4	24
6	2	11	[3.7, 2, 5]	-7.58	-64.79	388.21	-1.88	47.58	109.15	339.73	1143.95	6045.35	-7.58	-2.33	-16.69	-1.07184	3	6	31	55
6	2	12	[6, 4, 4]	4.76	503	160	179.64	540	34.63	623	778.75	2428.2	4.76	21.68	314.37	20.7149	1	3	2	5
6	2	13	[4.9, 2, 5]	25.57	657.34	120	58.69	717.2	28.01	697.34	435.84	2054.65	25.57	29.34	547.78	31.9929	3	3	16	12
6	2	14	[5, 2, 4]	-1.63	110.34	317.54	9.85	158.25	77.21	376.61	401.77	3339.08	-1.63	4.19	34.74	3.30455	3	6	8	20
6	2	15	[5, 1, 3, 5]	36.13	1050.62	200	93.8	1139.6	19.81	1090.62	681.64	3870	36.13	43.77	525.31	27.1477	3	5	12	16
6	2	16	[5.8, 3, 4]	7.91	582.4	240	121.33	632	35.27	662.4	414	3526.4	7.91	23.48	242.67	16.5154	1	6	3	9
6	2	17	[5, 1, 3, 5]	36.13	1050.62	200	93.8	1139.6	19.8	1090.62	681.636	3870	36.13	43.78	525.31	27.1477	3	5	12	16
6	2	18	[5, 1, 2, 5]	25.57	657.34	120	58.69	717.2	28	697.34	435.84	2054.65	25.57	29.35	547.78	31.9929	3	3	16	12
6	2	19	[4.7, 2, 4]	-2.02	110.92	330.02	7.92	169.39	65.43	390.92	244.33	4113.06	-2.02	4.17	33.612	2.69688	3	7	10	25
6	2	20	[5.8, 4, 5]	-3.38	303	280	63.12	340	140	583	828.75	3990	-3.38	11.83	108.21	7.59398	1	6	2	10
6	3	21	[5.9, 3, 5]	12.26	478.59	160	99.70	518.52	45.34	518.59	324.12	1645.13	12.26	20.62	299.12	29.0916	2	4	5	7
6	3	22	[5.4, 4, 5]	52.15	1776.46	200	211.48	1899.44	32.01	1856.46	1547.05	7261.26	52.15	74.01	888.23	24.4649	2	4	7	14
6	3	23	[5.1, 2, 5]	25.57	657.34	120	58.69	717.2	28.01	697.34	435.84	2054.65	25.57	29.34	547.78	31.9929	3	3	16	12

Table 2 Walk-Forward Efficiency

Time Window	X	Profit After Commission	MDD	Winning Percentage	WFE
4	[3.95, 1, 5]	-116.51	236.11	50	-196.39
5	[5.4, 4, 5]	758	80	66.66	128.01
6	[3.4, 3, 5]	-33.85	373.43	22.58	-8.18
7	[3.5, 0, 3]	30.35	200	63.33	13.7
8	[2.6, 0, 4]	442.38	239.6	44	33.3
9	[2.8, 0, 3]	123.29	265.39	39.5	19.4
10	[3.1, 0, 4]	218.43	178.6	42.43	35.36

So we have $\overline{WFE} = 3.6\%$, which means the strategy didn't pass the Walk-Forward Analysis test. A robust strategy must have \overline{WFE} more than 50%. In case the \overline{WFE} is more than 50% it means we have a robust trading strategy. In that case, in order to choose which set of parameters will be used for the real time trading, we use another metric called the Sharpe ratio (Wikipedia, 2013).

4. Conclusion – Future Work

In this paper we have presented a way to find near-optimal parameterisation of betting strategies for betting exchange markets. As illustrated in the case study, the SPORTSBET optimisation platform implements Walk-Forward Analysis for the robust parameterisation of betting exchange trading strategies without overfitting. Future work includes the usage and comparison of different metaheuristics inside the SPORTSBET platform and the automated evolution of entirely new betting strategies.

References

- Beyer, H. & Sendho, B., 2008. Covariance Matrix Adaptation Revisited. The CMSA Evolution Strategy. In: *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 123-132.
- Gencay, R. & Qi, M., 2001. Pricing and hedging derivative securities with neural networks: Bayesian regularization, early stopping, and bagging. *Neural Networks, IEEE Transactions*, 12(4), pp. 726-734.
- Hansen, N., 2006. The CMA Evolution Strategy: A Comparing Review. In: *Towards a New Evolutionary Computation*. s.l.:Springer Berlin Heidelberg, pp. 75-102.
- Hoffmeister, F. & Back, T., 1991. Lecture Notes in Computer Science. In: Heidelberg, ed. *Genetic Algorithms and evolution strategies: Similarities and differences*. Springer Berlin, pp. 455-469.
- Pardo, R., 2008. *The Evaluation and Optimization of Trading Strategies*. 2 ed. s.l.:Wiley.
- Rodriguez, J., Perez, A. & Lozano, J., 2009. Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. *Pattern Analysis and Machine Intelligence, IEEE Transactions*, 32(3), pp. 569-575.
- Tsirimpas, P. & Knottenbelt, W. J., 2011. *SPORTSBET: A Tool for the Quantitative Evaluation and Execution of Betting Exchange Trading Strategies*. Aachen, pp. 155-156.
- Wikipedia, 2013. *Backtesting*. [Online] Available at: <http://en.wikipedia.org/wiki/Backtesting>
- Wikipedia, 2013. *Sharp ratio*. [Online] Available at: http://en.wikipedia.org/wiki/Sharpe_ratio

Sensitivity of court-side in tennis?

Michelle Viney*, Anthony Bedford* and Elsuida Kondo*

*School of Mathematical and Geospatial Sciences, RMIT University, Melbourne, Australia

Corresponding author: Anthony Bedford@rmit.edu.au

Abstract. Several professional tennis players in the ATP circuit serve better on a particular side of the tennis court than the other. Generally left hander's have an advantage due to their wide serve to a typically a right hander's weaker side, the backhand. In this paper, we empirically investigated whether players win more points on a particular side of the tennis court. Analysis of the top 50 ATP male player's last twenty matches at the end of the calendar year was analysed. To explore whether left handers do have an advantage on the advantage side of the court, analysis of the top 100 left handers were performed. A set simulator was implemented to determine whether altering a player's probability of winning a point on serve in respect to what side of the court was played on, was more effective than keeping a player's probability of winning a point in serve constant.

1. Introduction

When serving in tennis you serve on two sides of the court, the deuce and the advantage side. Every service game begins on the deuce side where the server serves diagonally in the left direction. The advantage side involves serving diagonally to the right side of the court. Several professional tennis players in the ATP circuit have a more effective serve on a particular side of the court than the other. In particular, a left hander generally has an advantage due to their wide serve to a typically right hander's weaker side, the backhand.

A broad range of research has been performed on the effect and the advantage of being a left hander in tennis (Chappell, 2003; Holtzen, 2000 and Wood & Aggleton, 1989). It is still unknown whether the advantage is primarily due to a biological or a tactical aspect. The biological aspect is based on the hypothesis that left handed people have a greater developed right hemisphere and therefore have a more developed motor, attention and spatial functions. The tactical advantage relates to left hander's generally play right hander's in their tennis career, therefore they have developed their game to face a right hander whereas right handers face a left hander a minority of the time. Therefore a left hander represents a rare and awkward opponent, where generally a left hander's best stroke goes diagonal to a generally weaker backhand of the right hander.

Pollard (2008) mathematically investigated whether players who have a more effective serve on one side of the court has an advantage on the outcome of the service game. Pollard concluded that if the server has a more effective serve on one side on the court they have a higher probability of winning the game on serve. Pollard outlined that any player regardless of their dominant hand can be equally rewarded if they have a superior serve to a particular side of the court, regardless of which side is more superior.

The aim of this paper is to empirically validate the statements that Pollard proposed. By having the knowledge of which players serve better on a particular side of the court can be a good wagering tool to have when wagering in tennis in-play.

2. Methods

To determine whether ATP players win more points on one particular side of the court, player analysis was performed. The top 50 ATP players and the top 100 left handers in the ATP tour were analysed. Using a large database provided by KAN-soft (www.oncourt.info), OnCourt provides various match facts and statistics on most tennis matches since the 2003 French Open. It's important to note that not all match statistics are given for all ATP events.

Point by point data was collected for the last twenty matches played in the calendar year 2012, regardless of the outcome of the match. In order to perform analysis on the point by point data, match statistics such as who won the point, what side of the court was it played on and whether the player was serving was established.

For a deeper analysis to determine whether court side has an effect in the game of tennis, a set simulator was applied. A set simulator was built using the add-on @Risk for Microsoft Excel. This simulator has the ability to modify the probabilities as the simulation is occurring, and the set can commence at any set score for any server. Once a simulation is complete the output displays the number of times each player has won to a particular set score.

Before the simulation can commence, the following input parameters are entered into the simulator: the probability of winning on serve for both players, the game and set score and the server of the current game. To determine the winner of the point, a uniform distributed random number is generated to compare against the server's probabilities. This process of generating a uniformly random number is repeated many times until the set is complete.

Two models were implemented into the simulator to analyse whether adjusting the probability of winning a point on serve in relations to the side of the court that was played on was more effective. The two models were *Court-side* and *Normal*.

The *Normal* approach follows the Markov Chain model, where the probability of winning a point on serve remains constant for the entire set, due to the assumption of independence and identical distribution (IID). The Markov Chain model is typically still used to predict outcomes of tennis matches before and during the match. Barnett, Brown & Clarke (2006) applied the properties of the Markov Chain to derive a recursive formula to calculate the probability of winning from any state within a game, set and match.

In terms of a game, the probability of Player A winning the game at point score (a, b) is given by:

$$P(a, b) = pP(a + 1, b) + (1 - p)P(a, b + 1) \quad (1)$$

with boundary conditions:

$$P(a, b) = 1 \text{ if } a = 4, b \leq 2$$

$$P(a, b) = 0 \text{ if } b = 4, a \leq 2$$

$$P(3,3) = \frac{p^2}{p^2 + (1-p)^2},$$

where p is the probability of Player A winning a point on serve which remains constant for the entire match.

In similar fashion, the probability of either player winning a tiebreak set can be calculated using a Markov chain. Let $P_A^{GST}(c, d)$ represent the conditional probability of Player A winning a tiebreak set from game score (c, d) when Player A is serving. It is expressed as followed:

$$P_A^{GST}(c, d) = p_A^g P_B^{GST}(c + 1, d) + (1 - p_A^g) P_B^{GST}(c, d + 1) \quad (2)$$

with boundary conditions

$$P_A^{GST}(c, d) = 1 \text{ if } c = 6, 0 \leq d \leq 4, c = 7, d = 5$$

$$P_A^{GST}(c, d) = 0 \text{ if } d = 6, 0 \leq c \leq 4, c = 5, d = 7$$

$$P_A^{GST}(6,6) = p_A^{gT},$$

where p_A^g represents the probability of Player A winning a game on serve and p_A^{gT} represent the probability of player A winning a tiebreak game.

For a detailed explanation, see Barnett, Brown & Clarke (2006)

The *Court-side* approach takes into consideration the side of the court the server has won or lost the point, and updating the probability of serve for the next time the server serves on that particular side of the court. In tennis there are two sides of the court to serve from, deuce and advantage court. For example if the server lost the point on the advantage court, then the next time the server serves on the advantage court, their probability is decreased by a weighting parameter, theta (θ).

For Player A serving at point, p

$$p_{a,p[cs]}^* = p_{a,p-2[cs]}^* + 1_{\{A \text{ wins}\}} - 1_{\{A \text{ loses}\}} \frac{1}{\theta} \quad (3)$$

where θ = the weighting parameter, $p_{a,0[cs]}^* = p_a$ and $p_{b,p[cs]}^* = p_{b,p-2[cs]}^*$.

In order to compare the different models, all models were linked to each other to ensure all methods have the same random variable value.

3. Results

Analysis of point by point data of player's last twenty matches in the calendar year, 2012 was implemented. When data collection commenced, there were fourteen left handers in the ATP top 100. The results found that twelve out of fourteen players won a greater average of points on the advantage side of the court overall and whilst serving. Martin Klizan and Horacio Zeballos were the two players who won a greater average of points on the deuce side, overall and whilst serving. The average difference for the fourteen players whilst serving was 0.014. Therefore 1.4% more points were won on the advantage side of the court.

Table 1. The top 50 left hander player's average probability of winning on the deuce and advantage side (Avg Diff= Advantage-Deuce).

Player's	<u>Avg win point on</u>		Avg Diff	<u>Avg win point serving on</u>		Avg Diff
	Deuce	Adv		Deuce	Adv	
Rafeal Nadal	0.574	0.588	0.014	0.689	0.704	0.016
Fernando Verdasco	0.493	0.517	0.024	0.641	0.671	0.030
Jurgen Melzer	0.474	0.491	0.017	0.617	0.626	0.009
Martin Klizan	0.541	0.534	-0.007	0.638	0.630	-0.008
Thomas Bellucci	0.504	0.507	0.003	0.622	0.656	0.035
Feliciano Lopez	0.470	0.502	0.032	0.634	0.676	0.042

Referring to Table 1, it displays the average probability of winning on both sides of the tennis court. As shown in Table 1, all players except Klizan recorded a higher average probability of winning on the advantage side. Feliciano Lopez recorded the highest average difference of 0.042 between the two sides of the court. Therefore Lopez won 4.2% more points on the advantage side of the court compared to the deuce side.

The next process involved analysing and interpreting the data of the top 50 players in the ATP to determine whether court-side sensitivity exists in these samples of players. The same analysis was applied as for the left handed analysis. The results found that 38.6 percent of right hander's in the Top 50 won more points on serve on the advantage side of the tennis court.

Table 2. Summary of the top five players in the top 50 ATP who had the largest average difference between the two sides of the court (Average difference=Advantage-Deuce).

	Deuce Side			Advantage Side			Average Difference
	Avg	Max	Min	Avg	Max	Min	
Overall							
Feliciano Lopez	0.470	0.607	0.293	0.502	0.660	0.286	0.032
John Isner	0.492	0.575	0.403	0.523	0.598	0.413	0.031
Benoit Paire	0.474	0.588	0.340	0.501	0.648	0.379	0.027
Fernando Verdasco	0.493	0.615	0.346	0.517	0.641	0.354	0.024
Jeremy Chardy	0.494	0.604	0.353	0.514	0.627	0.383	0.020
On Serve							

John Isner	0.691	0.800	0.547	0.737	0.880	0.595	0.046
Benoit Paire	0.573	0.733	0.313	0.618	0.794	0.464	0.045
Feliciano Lopez	0.634	0.862	0.424	0.676	0.818	0.375	0.042
Jeremy Chardy	0.636	0.795	0.448	0.674	0.861	0.457	0.038
Thomaz Bellucci	0.622	0.755	0.455	0.656	0.882	0.438	0.035

Table 2 displays the top five ATP players in the top 50, that recorded the highest average difference between the two sides of the court for overall and whilst serving. Table 2 also displays the average, maximum and minimum probability of the two sides of the court and the average difference is advantage minus deuce side. John Isner recorded the highest average probability difference on serve, of a value of 0.046. Therefore 4.6% of the time Isner will win a point on serve on the advantage side of the court.

To compare and contrast which method was more effective, the simulator was performed on all theta values using 10,000 simulations. Theta ranged from 200 to 400 with increasing increments of a value of ten. To determine whether there were any effects on court side, a “balance situation” was applied. A “balance situation” is where in both approaches both players are given a starting probability of winning on serve a value of 0.5. Once the set simulator has commenced the probabilities alter in accordance to what is occurring in the set.

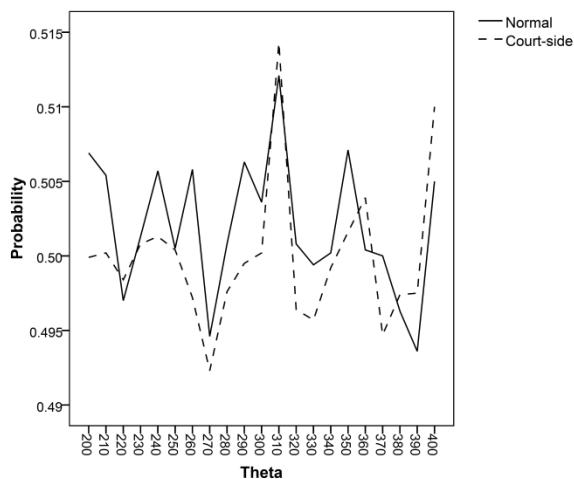


Figure 1. A graphical representation of the two approaches on the average probability of Player A winning a set when Player A commenced serving first.

In the figure above, it represents the probability of Player A winning the set where Player A begins serving at the start of the set. Referring to the Figure 1 the highest probability of winning the set occurred at theta value 310, where the lowest probability is at theta value 270 for *Court-side* and 390 for the *Normal* approach. The largest average probability difference between the two methods was at theta value 260, with the absolute difference of 0.0086. Figure 1 shows that *Court-side* records a lower probability of winning the set than the *Normal* approach for most theta values, when Player A serves at the beginning of the set.

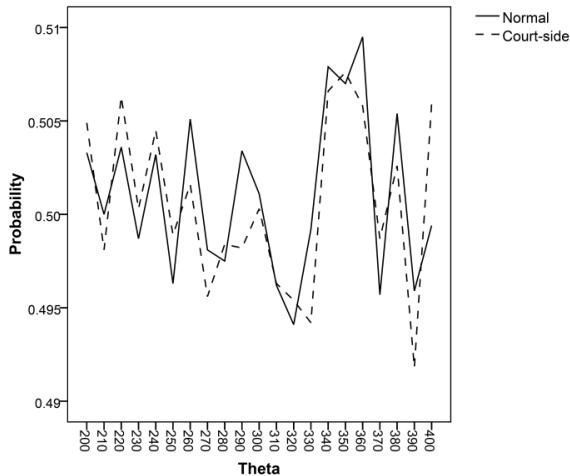


Figure 2. A graphical representation of the two approaches on the probability of Player A winning a set when Player B commenced serving first.

In the figure above, it represents the probability of Player A winning the set with Player B serving first in the set. Referring to Figure 2, the highest probability of winning the set occurred at theta value 350 for *Court-side* and 360 for the *Normal* method, where the lowest probability is at theta value 390 for *Court-side* and 320 for the *Normal* approach. The largest average probability difference occurred at theta value 400, with the absolute difference of 0.0065.

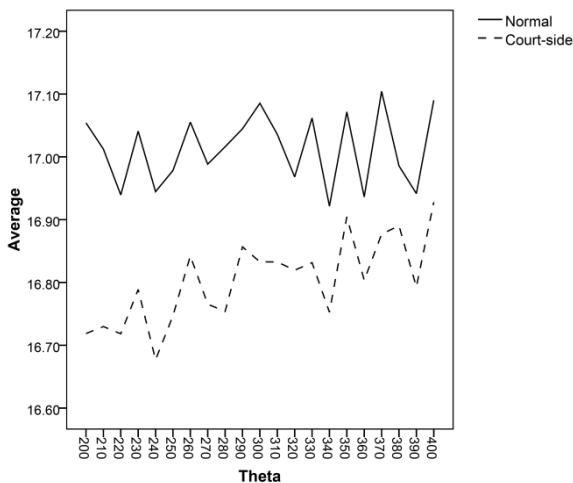


Figure 3. A graphical representation of the two approaches on the average amount of points Player A wins on serve.

Figure 3 represents the average quantity of points Player A wins on serve when Player A serves first in the set. It shows that the *Normal* approach in all theta values, have a higher average of Player A winning a point whilst serving than the *Court-side* approach when Player A is serving first. The average difference between the two methods is 0.210, where the maximum difference occurs at theta value 200 at a value of 0.335 and a minimum difference of 0.095 occurred at theta value 380. This pattern exists for both player's holding and breaking on serve, when either player commence serving at the beginning of the set.

A Pearson's chi square test was performed comparing the two approaches. The results found that when Player A commenced serving in the set the following theta values were not statistically significant at 0.05

level, theta values: 320, 340, 350, 360, 380 and 390. When Player B commenced serving, all theta values above 340 were not statistically significant at a 0.05 level.

From the Pearson's chi square testing, most theta values above 300 were not statistically significant at a 0.05 level. Although no optimal theta value can be determined from this research, the theta value of 250 was chosen to display a graphical representation of how the set was won, comparing the two methods.

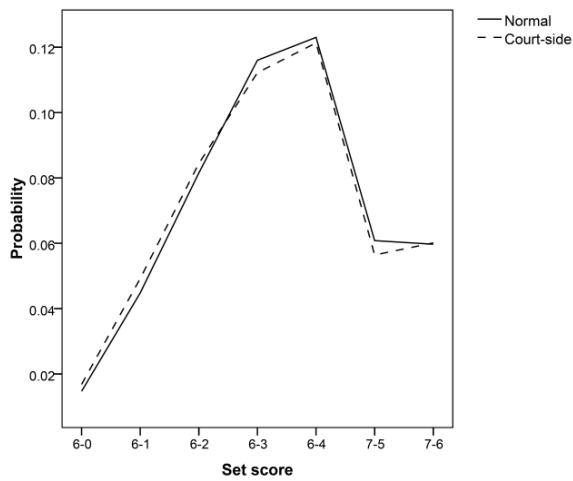


Figure 4. A graphical representation of the outcome of the set with theta value 250 with Player A serving first in the set.

Referring to Figure 4 it shows that the two methods mimic each other's pattern. The highest peak was at 6-4 with a denoted probability of 0.12. Therefore 12% of the time, 6-4 will be the set score when the starting probability of winning a point on serve is 0.5. The same pattern occurred when Player A is serving first in the set.

4. Discussion

Looking into whether sensitivity of court side exists in tennis, we note the following trends. Firstly empirically, most left hander's and some right hander's performed and therefore succeed better on the advantage side of the court. Having this knowledge, may be useful to increase the probability of winning a point on serve to that particular side when a particular player has a court side difference. Therefore it may increase the accuracy and validity of the outcome of the match. These findings may be useful when wagering in-play. If knowledge on a server's preferred court side is known, then the individual can wager to win the point on that particular side of the court.

Secondly, a set simulator was applied to determine whether altering the probability of winning a point on serve is more effective than not. The results found that in comparison of the two models, when a player is serving first in the set their likelihood of winning the set is lower for the *Court-side* approach than the *Normal* approach. Analysing the average number of points won and broken on serve, there was a large difference between the two methods. Future research is required in applying a simulator in tennis, as the optimal theta value is yet to be determined.

Although the results can be concluded and applied for the future, it's important to note that the database, OnCourt sometimes contain errors or missing data. Therefore it may not give an exact representation of the match. For further research in applying empirically evidence of ATP players, it may be useful to analyse point by point data in relations to the surface of the court. By undertaking this research it may give a more accurate representation of how each player performs on different surfaces.

5. Conclusion

In this research we empirically investigate whether players in the ATP circuit win more points on a particular side of the court. It is shown that most left handers and a handful of right handers win a greater amount of points on the advantage side of the court. By having this knowledge, it can be applied to a set simulator and alter the probability of winning a point on serve in relation to court-side. Therefore it may increase the accuracy and validity in determining the outcome of the tennis match. These findings could be applied as a wagering tool to wager in-play.

References

- Barnett, T.J, Brown, A. & Clarke, S.R. (2006) Developing a model that reflects the outcomes of tennis matches. *Proceedings of the 8th Australasian Conference on Mathematics and Computers in Sport*, 178-188.
- Chappell A.S. (2003) Left handedness: A blessing or a curse? *Med Sci Tennis*, **8**(1).
- Holtzen D.W. (2000) Handedness and professional tennis. *Intern J Neurosci*. **5**, 101-19.
- Pollard, G. (2008) An advantage of serving left handed in tennis. *Med Sci Tennis*, **13**, 34-36.
- Wood, C.J. and Aggleton, J.P. (1989) Handedness in 'fast ball' sports: Do left handers have an innate advantage? *Br J Psychol*. **80**(2):227-40.

Intelligent computational optimisation of sport skills

Joe Wright* and I. Jordanov**

* School of Computing, University of Portsmouth, Lion Terrace, Portsmouth, PO1 3HE, UK, jonathan.wright@port.ac.uk

** School of Computing, University of Portsmouth, Lion Terrace, Portsmouth, PO1 3HE, UK, ivan.jordanov@port.ac.uk

Abstract. This paper presents a proof of concept for the use of simulation and optimisation techniques in sport contexts. After presenting the potential benefit of these techniques for sport, an approach is presented incorporating physical simulation, control algorithms, optimisation methods and skill specific fitness functions. For a vertical jump sport skill, movement patterns were generated and optimised for a bipedal system. Comparisons are made between two different control algorithms – central pattern generators and recurrent neural networks, and two different optimisation techniques – genetic algorithms (GA) and particle swarm optimisation (PSO). The results, in terms of best solutions, convergence speed and accuracy are critically analysed and compared for the different methods. The choice of fitness function is discussed, considering how correctly framing the problem is crucial to produce optimal results and to avoid inappropriate movement patterns. A conclusion summarising the results of this investigation with implications for sport training and suggestions for future work is finally given.

1. Introduction

Optimisation in sport typically involves constructing very simplified models of the target system (Glazier and Davids, 2009) (the system will include the participant(s), environment and any props such as projectiles). This simplification is done in order to make the optimisation a straight-forward task. Often quite basic mathematical analysis can give answers to the problems set.

It was noted in (Glazier and Davids, 2009) that this simplification results in three problems: **reliability** of the results – movement patterns derived in the simplified model may not be (even approximately) optimal in the real world system; the simplified model may not contain the **detail** required. Some sport skills have significant contributions from large and small muscle groups simultaneously (Hore and Watts, 2005), but it is difficult to incorporate both scales in a simple model; it is difficult to express differences between **individuals** in a simplified model.

To address these problems techniques need to be developed for optimising more detailed models. Highly accurate models have many degrees of freedom (Damsgaard et al., 2006) and, therefore, generalised control methods applied to these models will have many parameters. So it is necessary to develop control methods capable of producing the desired sport skills, and techniques capable to optimise a large parameter set usually required by these control methods.

Inspired by bipedal movement control optimisation in the robotics literature (Kim and Lee, 2007) this paper presents a comparison of movement control algorithms, optimisation techniques, and fitness functions for a vertical jump sport skill.

In the next section we introduce the design of the physical model and control algorithms; section 4 investigates heuristic optimisation approaches; in section 5 we present results from the optimisation and finally in section 6 we give concluding remarks and suggestions for future work.

2. Design

The optimisation procedure adopted for this research has four components: a physical simulator: this is where the participants, environment and props are modelled and are simulated to experience forces generated externally (e.g., gravity) and internally (e.g., muscle forces); a control algorithm that specifies control joint torques (to approximate muscle forces) for each time iteration of the physical system; an optimiser that specifies the parameters for the control algorithm; a fitness function that drives the optimisation towards a desired sport skill goal.

2.1. Physical simulator

This paper uses OpenHRP3 (Open Architecture Human-centered Robotics Platform version 3) (Kanehiro et al., 2004) which is a software platform for robot simulations. OpenHRP3 was chosen as it has often been used successfully in movement control ((Kanehiro et al., 2004), (Guan et al., 2006)). It allows external automation so can readily be used for the multiple simulations required by the optimisation procedures outlined in section 4. It can be used for bipedal

simulations and includes a humanoid model that was used for this paper. However, for sport skill research, a biomechanical software package may be preferred (Arnold et al., 2010).

The main environment is typically run as a java programme hosted in Eclipse (Figure 1). This provides a user interface for visualising the simulation, loading and creating physical models, specifying control scripts, as well as other functions.

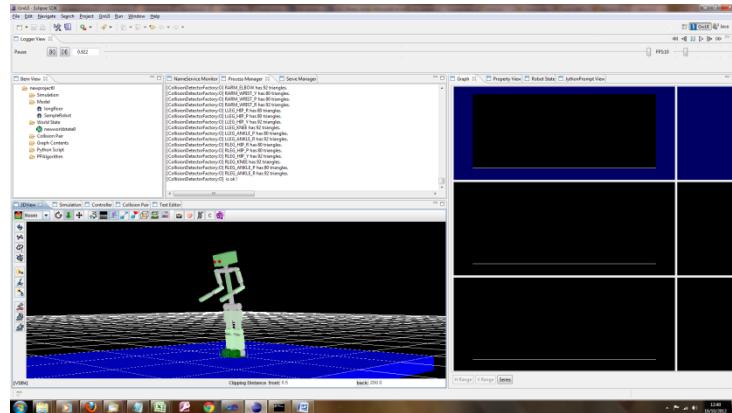


Figure 1. OpenHRP3 Main user interface hosted in Eclipse

2.2. Control algorithm

The control algorithms investigated here output desired joint angles at each time iteration. They are then converted into joint torque values which are fed into the physical simulation.

This was done using a proportional–integral–derivative controller (PID controller) (Kiam Heong et al., 2005). A PID controller outputs control values based on the error between the current measured position and desired position (here of the joint angles). It is given in discrete time form by equation 1:

$$u(t_k) = K_p e(t_k) + K_i \sum_{i=1}^k e(t_i) \Delta t + K_d \left(\frac{e(t_k) - e(t_{k-1})}{\Delta t} \right), \quad (1)$$

where t_k is the time at iteration k , $u(t_k)$ the output (joint torque), $e(t_k)$ the error (difference between target joint angle and measured joint angle), and K_p , K_i , and K_d are tuning parameters representing proportional, integral and derivative gains respectively.

Additionally, in order to cope with situations where the control algorithms changed their output too rapidly the output joint torque was limited to a maximum value.

Drawing from other movement control research the authors have chosen to compare central pattern generators (Watanabe et al., 2008) and recurrent neural networks (Reil and Husbands, 2002) as control algorithms.

2.1.2. Central Pattern Generator (CPG)

A CPG is essentially a system of interconnected oscillators. For this study, Van der Pol oscillators were used (Watanabe et al., 2008). Each interconnected oscillator targets a joint and is governed by:

$$\begin{cases} \dot{x}_i(t) = \dot{x}_i(t-1) + \Delta t \left(\alpha_i (p_i^2 - x_i(t)^2) \dot{x}_i(t-1) - \omega_i^2 \left(x_i(t) + \sum_{j=1}^n \lambda_{ij} x_j(t) \right) - k_i \right). \\ x_i(t+1) = x_i(t) + \dot{x}_i(t) \Delta t \end{cases} \quad (2)$$

In equation 2 $x_i(t)$ is the i 'th oscillator output at iteration t , α_i , p_i , ω_i and k_i are constants to be tuned by optimisation, λ is the interconnection matrix (with $\lambda_{ii}=0$), Δt is the length of time between iterations, and n is the number of oscillators. α_i , p_i , ω_i and k_i control the shape, amplitude, frequency and amplitude of oscillation respectively for the i 'th oscillator.

The parameter ranges used in this investigation are given in table 1 and the encoding scheme used in the optimisation algorithms presented in section 4 is given in table 2.

Table 1. CPG parameter ranges

Initial oscillator phase x	Set equal to the initial joint positions in the standing pose
Initial oscillator speed \dot{x}	$[-4.0, 4.0]$
α	$[0.0, 4.0]$
p^2	$[0.0, 8.0]$
k	0
ω^2	$[0.0, 40.0]$
λ_{ij}	$[-1.0, 1.0]$

Table 2. Optimisation encoding (represents one chromosome or particle)

n constants α_i	n constants p_i^2	n constants ω^2	$n(n-1)$ matrix λ_{ij}
--------------------------	-----------------------	--------------------------	--------------------------------

2.2.2. Recurrent Neural Network (RNN)

A neural network processes information through a network of connected units (neurons). This can be done in a feed-forward network where information is passed through layers from input to output, or in a recurrent network where information is fed back around the network.

Typically, neural networks are used to process inputs but in this case recurrent neural networks can self-generate patterns by cycling information around the network. The network used here has a fully connected structure (all neurons are connected to each other), (Reil and Husbands, 2002) shown in figure 2.

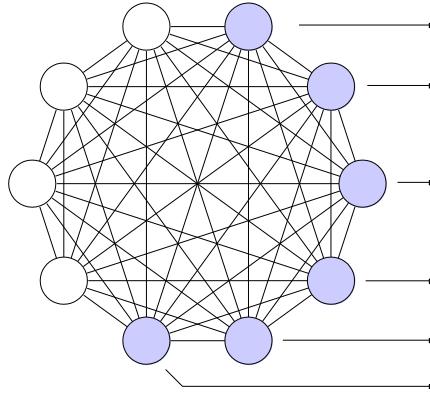


Figure 2. A 10 neuron fully connected recurrent neural network with 6 outputs (shaded).

$$\begin{aligned}\Delta A_j &= \Delta t \frac{\left(-A_j + \sum_i w_{ij} O_i \right)}{\tau_j} \\ O_j &= \left(1 + e^{(\alpha_j - A_j)} \right)^{-1}\end{aligned}. \quad (3)$$

The neurons are controlled by equation 3 where A_j is the activation of neuron j , τ_j is time constant of neuron j , w_{ij} is weight between neurons i and j , O_j is output of neuron j , α_j is bias of neuron j , and Δt is sampling period. Table 3 lists the parameter ranges for the network.

Table 3. Parameter ranges for RNN.

Initial oscillator phase x	Set equal to the initial joint positions in the standing pose
τ_j	[0.001, 5.0]
w_{ij}	[-16.0, 16.0]
α_j	[-4.0, 4.0]

3. Heuristic optimisation

In this paper, two different heuristic optimisation techniques are compared: genetic algorithms (GA) and particle swarm optimisation (PSO). The former was chosen as it has been often studied in the context of movement control optimisation (Kim and Lee, 2007), (Vundavilli and Pratihar, 2010), whereas the latter was chosen because the authors could not find examples of its use in this domain in the literature. We believe its use will contribute new results to the field and it is also useful for comparison purposes.

Both techniques are used to configure the parameters of the control algorithms, provide a set of solutions which are evaluated by the physical simulator, according to a fitness function. Based on those evaluations, a new set of configurations is produced, searching for better population fitness.

3.1. Genetic algorithms (GA)

A GA (Goldberg and Holland, 1988) is a search heuristic inspired by the biological process of evolution. Candidate solutions are encoded as a string of genes called a chromosome. Several chromosomes (initially randomly generated) are contained in a pool and each chromosome is assessed for its fitness. In the experiments presented here, a chromosome encodes the parameter set of each control algorithm, with one gene representing a parameter. The algorithm is then used in the physical simulation to test the candidate solutions. The fitness is derived accordingly to the sport skill under evaluation.

A new generation of the chromosome pool is produced at each iteration. The prevalence of chromosomes and genes carried through to the next generation is related to how good their fitness was – more information from fitter chromosomes will be present, than such from less fit chromosomes, in general. Candidate solutions are combined, inspired by breeding in nature, using cross-over reproduction operators in an attempt to build better solutions by finding combinations of good elements of other solutions. Furthermore, mutation is used to randomly alter individual genes in an attempt to move through and explore the search space.

The genes of each chromosome consisted of real values in the interval [0.0, 1.0]. These specified the parameters for the control algorithms, being linearly mapped onto the desired parameter ranges.

The configuration of the GA is shown in table 4.

Table 4. GA configuration.

Population size	50 chromosomes
Chromosome length	Dependent on control algorithm
Randomisation	Normal distribution $\mu = 0.5$, $\sigma^2 = 0.5$
Selection	Tournament selection, $p = 0.9$
Crossover	Genes crossed $p = 0.01$
Mutation	Creep mutation from $N(0.5, 0.5)$, $p = 0.04$
Elitism	Best 4 chromosome copied unaltered

3.2. Particle swarm optimisation (PSO)

PSO (Eberhart and Kennedy, 1995) uses a swarm of particles to move through the search space, with each particle's position changing according to its own experience, and the experience of other particles in the swarm. By recording the best position found for each particle, and for the whole swarm or for the sets of neighbouring (by index) particles, candidates for optimal solutions can be exploited. By using momentum, each particle can explore the search space and potentially find better solutions which can then be exploited.

The candidate solutions are represented by the location of each particle in an n -dimensional space with n equal to the number of parameters to optimise. The position \mathbf{x}_i of each particle i is updated according to:

$$\mathbf{x}_i(t+1) = \mathbf{x}_i(t) + \mathbf{v}_i(t+1), \quad (4)$$

where \mathbf{v}_i is the velocity of each particle i . The velocity for each particle i and dimension j is calculated according to:

$$v_{ij}(t+1) = v_{ij}(t) + c_1 r_{1j}(t)[y_{ij}(t) - x_{ij}(t)] + c_2 r_{2j}(t)[\hat{y}_{ij}(t) - x_{ij}(t)] \quad (5)$$

$$r_{1j}(t), r_{2j}(t) \sim U(0,1)$$

where y_{ij} is the best solution found so far by particle i , \hat{y}_{ij} is the best solution found in the set of local (by particle index) neighbouring particles, c_1, c_2 are constants balance exploration (ability to find useful areas of the search space) versus exploitation (ability to refine solutions), and r_{1j}, r_{2j} are random numbers calculated for each iteration t . The use of a local neighbourhood characterises this as the Local Best PSO variant. The size of the neighbourhood around each particle controls how information about good solutions is communicated throughout the swarm.

The PSO was specified with the parameter set given in table 5.

Table 5. PSO configuration.

c_1	c_2	$v_{ij}(t=0)$
1.49445	1.49445	0.0

3.3. Fitness functions

The fitness function for the vertical jump was kept as simple as possible as the purpose of this paper is to prove a concept. It is outlined in table 6.

Table 6. Fitness functions and termination criteria.

Skill	Fitness score	Termination criteria
Vertical jump	Maximum height attained by chest	3 seconds of simulated time

	Modified to favour counter-movement jumps by measuring the maximum height achieved by the chest after the waist had lowered below 0.65m	
--	---	--

Termination criteria can be used to frame the skill in time, to quickly stop failing attempts such as falling over, or to avoid undesirable movements such as jumping while attempting running. In this case, only the time frame was considered as important.

It was found that a fitness function based on chest height was preferable to one based on waist height. This was because the waist based function ignored toppling of the upper body. The modified version emphasising counter-movement is discussed in the next section.

4. Results

All control and optimisation combinations tested were able to produce satisfactory vertical jumps. An optimisation run was deemed to be successful if the final output consisted of a counter movement jump. A counter-movement element in a vertical jump has been identified as important in human jumping (Finni et al., 2000). The physical simulation consisted of a rigid body system and so the stretch-shortening cycle in a real human jump could not be modelled. However, the increased acceleration phase made possible by bending the knees before jumping still favours a counter-movement jump for maximum height.

The frequency of successes and best jump heights varied across the different combinations of control and optimisation algorithms. The results from 25 optimisation runs are summarised in table 7.

Table 7. Results for vertical jump – number of successful runs, maximum height after 100 and 200 generations for CPG and RNN algorithms, optimised by GA and PSO.

Optimiser	GA			PSO		
	In 25 runs	Successes	Best height after 100 generations	Best height after 200 generations	Successes	Best height after 100 iterations
CPG	25	2.26m	Not simulated	25	2.55m	Not simulated
RNN	4	2.23m	2.31m	7	1.54m	1.63m

CPG runs were always successful but RNN controlled runs often failed to produce a success. When failure occurred, optimisation tended to hit a local maximum – typically using ankle plantarflexion to jump up a little. In an attempt to combat this, the fitness function was modified to include a threshold criterion. The waist had to lower below 0.65m before maximum height was measured. This favoured counter-movement jumps over ankle jumps. Even so, RNN control had a very poor success rate. Using a normal distribution approximation there was little evidence supporting a difference in the success rates of RNN optimisation between GA and PSO ($p>0.3$). However, this approximation is unreliable as the GA optimised RNN had fewer than 5 successes.

The results are suggestive of PSO to optimisation produces more successes for RNN controlled jumps, with GA better at refining solutions to give greater heights. More simulations are needed to establish strong evidence of the PSO having better exploration (finding successful solutions) and the GA better exploitation (refining those solutions to get a near optimal height).

It should be noted that both the GA and PSO algorithms have several variables and sub-functions that can be changed. This will all have an impact on exploitation versus exploration. For example, some versions of PSO systematically vary the constants over the course of the run, so that they favour exploration in the beginning but exploitation towards the end. Therefore, conclusions on the overall properties of GA versus PSO in this domain cannot be made just on the experiment presented in this paper.

The use of this modified fitness function includes a priori expert knowledge – that a counter-movement jump is preferential to maximum height. This is problematic as expert knowledge for other skills may be missing or wrong,

and highlights the need to further develop the control and optimisation algorithms so that they can consistently find the global maximum. In this experiment, only the CPG was able to find suitable solutions without expert knowledge.

The higher success rate of the CPG can be attributed to its explicit oscillatory form. Counter movements (cycling knee flexion to knee extension) are generally present in the first, random, population. Optimisation is then simply a process of identifying and refining these patterns. For the RNN, the optimisation task is more complicated as many configurations of the network do not produce any periods of oscillation at all.

Although very successful, the ease of optimising the CPG created instances of solutions that would not be appropriate for real life – even though they scored very high heights. These solutions involved one or more preparatory mini-jumps before the final big jump. For a real world application, further modification to the fitness function would be needed to measure only the first jump. Both CPG and RNN algorithms were capable of producing unnecessary movements after the launch phase. This was often pronounced for the CPG as oscillations tended to continue after the launch phase. To improve the form, there would need to be some cut-off or transition control added to the CPG to control for these movements, and these form elements to be accounted for in the fitness function.

Lastly, it was observed that the CPG was capable of producing very explosive movements. This was possible because large amplitudes of oscillation were allowed (but clipped to control joint ranges). The movement pattern was probably not realistic but attempts to constrain the range more appropriately were less successful. Further work is needed to address this problem.

The simulation software can be used to produce a video of the optimised movement (see figure 3). This video can then be shown to coaches and athletes to help them visually understand the optimised pattern, or to be used as a reference in video analysis.

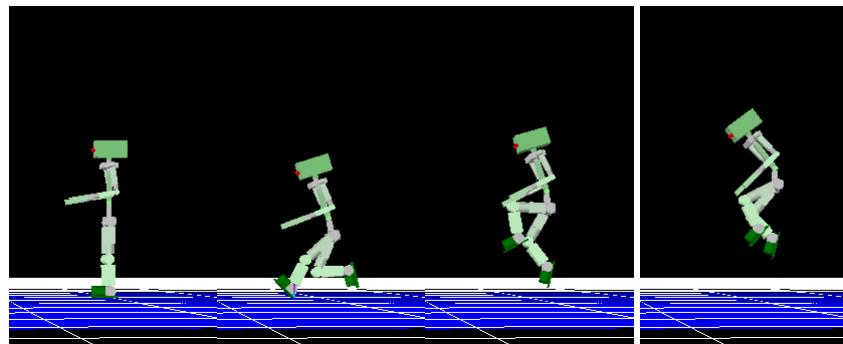


Figure 3. Stills from a video of an optimised vertical jump.

Alternatively, the raw output of the control algorithms (joint angles over time – figure 4) or the calculated torques over time can be used in training. To do this, the traces can be compared to data collected using motion tracking techniques.

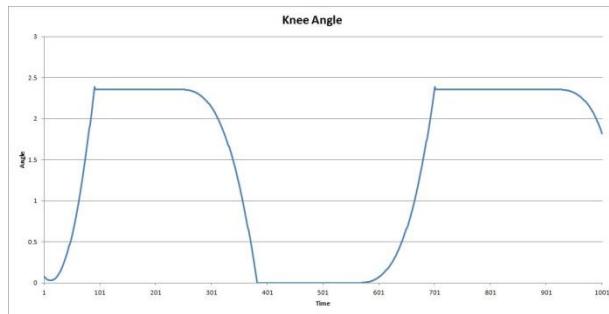


Figure 4. CPG output trace for the right knee angle during a vertical jump.

Finally, the researcher could extract key features from either the video or control algorithm output data. Key features may include extrema positions of joints, timing information, and movement sequence description. Identification of important features can be done by comparing to pre-existing technique and highlighting major differences.

Fitness functions have to be carefully chosen because they can affect the ability of the system to optimise and fundamentally define the form of the final solution. For example, in the vertical jump, choosing to measure height achieved by the waist is successful in producing vertical jumps but, in general, the upper body tends to rotate towards the horizontal. Measuring height achieved by the chest corrects this problem. For a more realistic jump skill, it would be appropriate to add horizontal plane factors to the fitness function. This is to ensure that the direction of the jump is task appropriate.

In this experiment, the CPG consistently outperformed the RNN. It was always successful whereas the RNN sometimes failed to produce satisfactory results. However, the CPG does have issues with unrealistic movements that are less of a problem with RNN controlled movement. For GA optimised RNN jumping, the success rate was 1 in 10. This compares to a similar success rate for GA optimised RNN controlling a walk gait found in (Reil and Husbands, 2002). Although the movement skills are different, we speculate that this represents a similar difficulty in finding oscillatory patterns in the RNN using GA. The PSO was more successful in finding optimal solutions for the RNN.

5. Conclusion and future work

The procedures investigated in this paper can form the basis of an in-depth optimisation process for sport skills. It has been shown that sport skills can be produced and optimised in a bipedal system using appropriate fitness functions. To bring this work to a practical level, the simulated system should be replaced with a more realistic human model. Then the produced patterns can be evaluated in a real life setting to see if the optimisation process produces valid results. Refinement of the techniques involves improving the physical modelling, developing control systems that are capable of a full range of skills, and discovering the best optimisation algorithms for this domain.

Finally, future work should also focus on producing appropriate fitness functions for a range of sport skills.

6. References

- ARNOLD, E., WARD, S., LIEBER, R. & DELP, S. 2010. A Model of the Lower Limb for Analysis of Human Movement. *Annals of Biomedical Engineering*, 38, 269-279.
- DAMSGAARD, M., RASMUSSEN, J., CHRISTENSEN, S. T., SURMA, E. & DE ZEE, M. 2006. Analysis of musculoskeletal systems in the AnyBody Modeling System. *Simulation Modelling Practice and Theory*, 14, 1100-1111.
- EBERHART, R. & KENNEDY, J. A new optimizer using particle swarm theory. *Micro Machine and Human Science*, 1995. MHS '95., Proceedings of the Sixth International Symposium on, 4-6 Oct 1995 1995. 39-43.
- FINNI, T., KOMI, P. V. & LEPOLA, V. 2000. In vivo human triceps surae and quadriceps femoris muscle function in a squat jump and counter movement jump. *European Journal of Applied Physiology*, 83, 416-426.
- GLAZIER, P. S. & DAVIDS, K. 2009. Constraints on the Complete Optimization of Human Motion. *Sports Medicine*, 39, 15-28.
- GOLDBERG, D. & HOLLAND, J. 1988. Genetic Algorithms and Machine Learning. *Machine Learning*, 3, 95-99.
- GUAN, Y., NEO, E. S., YOKOI, K. & TANIE, K. 2006. Stepping over obstacles with humanoid robots. *Robotics, IEEE Transactions on*, 22, 958 -973.
- HORE, J. & WATTS, S. 2005. Timing Finger Opening in Overarm Throwing Based on a Spatial Representation of Hand Path. *Journal of Neurophysiology*, 93, 3189-3199.
- KANEHIRO, F., HIRUKAWA, H. & KAJITA, S. 2004. Openhrp: Open architecture humanoid robotics platform. *The International Journal of Robotics Research*, 23, 155.
- KIAM HEONG, A., CHONG, G. & YUN, L. 2005. PID control system analysis, design, and technology. *Control Systems Technology, IEEE Transactions on*, 13, 559-576.
- KIM, J.-J. & LEE, J.-J. Gait adaptation method of biped robot for various terrains using central pattern generator (CPG) and learning mechanism. *Control, Automation and Systems*, 2007. ICCAS '07. International Conference on, Oct. 2007. 10-14.
- REIL, T. & HUSBANDS, P. 2002. Evolution of central pattern generators for bipedal walking in a real-time physics environment. *Evolutionary Computation, IEEE Transactions on*, 6, 159-168.
- VUNDAVILLI, P. R. & PRATIHAR, D. K. 2010. Dynamically balanced optimal gaits of a ditch-crossing biped robot. *Robotics and Autonomous Systems*, 58, 349 - 361.
- WATANABE, K., TAJIMA, A. & IZUMI, K. Locomotion pattern generation of semi-looper type robots using central pattern generators based on van der Pol oscillators. *Industrial Informatics, 2008. INDIN 2008. 6th IEEE International Conference on*, July 2008. 377-382.