Proceedings of the
Sixth Australian Conference on

MATHEMATICS AND COMPUTERS IN SPORT

Bond University
Queensland

Edited by
Graeme Cohen and Tim Langtry
Department of Mathematical Sciences
Faculty of Science
University of Technology, Sydney

6M&CS

1 – 3 July 2002

Proceedings of the Sixth Australian Conference on MATHEMATICS AND COMPUTERS IN SPORT

Bond University, Queensland

Edited by Graeme Cohen and Tim Langtry
Department of Mathematical Sciences
Faculty of Science
University of Technology, Sydney

6M&CS

1 – 3 July 2002

Typeset by LaTeX 2ε

Proceedings of the Sixth Australian Conference on Mathematics and Computers in Sport, 1–3 July 2002, at Bond University, Gold Coast, Queensland 4229, Australia.

ISBN 1 86365 532 8

Printed by UTS Printing Services

Contents

Abstract

Conference Director's Report

Welcome to the Sixth Australian Conference on Mathematics and Computers in Sport. This year we return to Bond University after the successful Fifth Conference in 2000 held in Sydney because of the Olympic Year in that city. It is a pleasure to renew acquaintances with one of our principal speakers, Ray Stefani who was at the First Conference in 1992, and has been collaborating with Steve Clarke from Swinburne for many years now. Our second principal speaker, Steve Gray from Queensland, will add a new dimension to the Sixth Conference with his interest in the economical aspects of sport. Graeme Cohen (now retired from UTS) is our third principal speaker, but he and Tim Langtry have once again taken over the responsibilities of producing the printed *Proceedings*. I thank Graeme and Tim for relieving me of this major task.

The conference has once again attracted academics from New Zealand, the United Kingdom, the United States and Canada. I welcome them all, including many familiar faces. I hope that you all find the conference rewarding in many aspects, including the content and presentation of the talks, the many discussions that are generated, and the close social contact with like-minded academics. We now have a website www.anziam.org.au/mathsport due to Elliot Tonkes, who is the webmaster. This site contains information about this and all previous conferences.

All the papers in these *Proceedings* have undergone a detailed refereeing process. I am indebted to the referees for their time and comments to improve the quality of all papers. The *Proceedings* begin with the papers of our principal speakers, followed by the contributed papers in alphabetical order of author, or first author. The *Proceedings* conclude with an abstract—the paper was accepted for presentation but was received too late to be submitted to the full refereeing process.

Neville de Mestre
June 2002

Participants

Mr Paul ALLSOPP	Swinburne University
Mr Michael BAILEY	Swinburne University
Mr Tristan BARNETT	Swinburne University
Mr David BEAUDOIN	Simon Fraser University, Canada
Emeritus Professor Maurice BREARLEY	Clifton Springs, Victoria
Assistant Professor Kyung-Jiu BYUN	Bond University
Dr George CHRISTOS	Curtin University of Technology
Associate Professor Stephen CLARKE	Swinburne University
Dr Graeme COHEN	University of Technology, Sydney
Professor John CROUCHER	Macquarie University
Dr Michel DE LARA	École Nationale des Ponts et Chaussées, France
Professor Neville DE MESTRE	Bond University
Mr Donald FORBES	Swinburne University
Professor Stephen GRAY	University of Queensland
Associate Professor Chris HARMAN	University of Southern Queensland
Ms Kate HOGARTH	University of Queensland
Associate Professor Kuldeep KUMAR	Bond University
Dr Tony LEWIS	Oxford Brookes University, UK
Dr Graham McMAHON	Bond University
Associate Professor R. Hugh MORTON	Massey University
Mr Emil MURESAN	Katholieke Universiteit Leuven, Belgium
Dr Robert NEAL	Golf Biodynamics Pty Ltd, Queensland
Dr Pam NORTON	Monash University
Dr David PANTON	University of South Australia
Professor Graham POLLARD	University of Canberra
Professor Ray STEFANI	University of California, USA
Associate Professor Steve SUGDEN	Bond University
Professor Tim SWARTZ	Simon Fraser University, Canada
Dr Elliot TONKES	University of Queensland

Program

Monday 1 July

8.30	**Registration and opening remarks**	
9.00	Ray Stefani	The fundamental nature of differential male/female world and Olympic winning performances, sports and rating systems
10.00	Graham Pollard* and Ken Noble	The characteristics of some new scoring systems in tennis
10.30	**Morning tea and coffee**	
11.00	Frank Duckworth and Tony Lewis*	Review of the application of the Duckworth/Lewis method of target resetting in one-day cricket
11.30	Paul Allsopp* and Stephen Clarke	Factors affecting outcomes in test-match cricket
12.00	David Beaudoin, Robin Insley and Tim Swartz*	Studying the bankroll in sports gambling
12.30	Maurice Brearley* and Neville de Mestre*	How do lawn bowls and golf balls slow down on grass?
1.00	**Lunch**	
2.30	Stephen Clarke* and Pam Norton	Collecting statistics at the Australian Open Tennis Championship
3.00	Hugh Morton	Team ratings and home advantage in SANZAR rugby union 2000–2001
3.30	Pam Norton* and Stephen Clarke	Serving up some grand slam tennis statistics
4.00	Kate Hogarth*, M. Yeardon and Dave Stump	Analysis of a twisting dive

Tuesday 2 July

8.30	Michel de Lara	Diving into mathematics or some mathematics of SCUBA diving
9.00	Stephen Gray* and Ahn Le	How to fix a one-day international cricket match
10.00	Graham McMahon* and Neville de Mestre*	Tennis serving strategies
10.30	**Morning tea and coffee and PHOTO**	
11.00	John Croucher	Analysing scores in English premier league soccer
11.30	David Panton*, Kylie Bryant and Jan Schreuder	Optimisation tools for round-robin and partial round-robin sporting fixtures
12.00	Michael Bailey* and Stephen Clarke	Predicting the Brownlow medal winner
12.30	Emil Muresan*, Jelle Geerits and Bart De Moor	Soccer: from match taping to analysis
1.00	**Lunch**	
2.30	Emil Muresan*, Jelle Geerits and Bart De Moor	Video analysis in sports: Video-coach©
3.00	Shawn Clowes, Graeme Cohen* and Ljiljana Tomljanovic	Dynamic evaluation of conditional probabilities of winning a tennis match
3.30	Graham Pollard	The effect of a variation to the assumption that the probability of winning a point in tennis is constant
4.00	Frank Duckworth and Tony Lewis*	Can the probability of winning a one-day cricket match be maintained across a stoppage?
4.30	**Mathsport meeting**	

Wednesday 3 July

8.30	**Forum on communicating sports science to the public**	
9.00	Graeme Cohen	Cricketing chances
10.00	Chris Harman	A biomechanical power model for world-class 400 metre running
10.30	**Morning tea and coffee**	
11.00	Graham Pollard* and Ken Noble	A solution to the unfairness of the tiebreak game when used in tennis doubles
11.30	George Christos* and Jamie Simpson	Fixing the fixtures with genetic algorithms
12.00	Tristan Barnett* and Stephen Clarke	Using Microsoft® Excel to model a tennis match
12.30	Elliot Tonkes	Dartboard arrangements with a concave penalty function

CRICKETING CHANCES

G. L. Cohen
Department of Mathematical Sciences
Faculty of Science
University of Technology, Sydney
PO Box 123, Broadway
NSW 2007, Australia
graeme.cohen@uts.edu.au

Abstract

Two distinct aspects of the application of probabilistic reasoning to cricket are considered here.

First, the career bowling figures of the members of one team in a limited-overs competition are used to determine the team bowling strike rate and hence the probability of dismissing the other team. This takes account of the chances of running out an opposing batsman and demonstrates that the probability of dismissing the other team is approximately doubled when there is a good likelihood of a run-out.

Second, we show that under suitable assumptions the probability distribution of the number of scoring strokes made by a given batsman in any innings is geometric. With the further assumption (which we show to be tenable) that the ratio of runs made to number of scoring strokes is a constant, we are able to derive the expression $(A/(A+2))^{c/2}$ as the approximate probability of the batsman scoring at least c runs ($c \geq 1$), where A is the batsman's average score over all past innings.

In both cases, the results are compared favourably with results from the history of cricket.

1 Introduction

In an excellent survey of papers written on statistics (the more mathematical kind) applied to cricket, Clarke [2] writes that cricket "has the distinction of being the first sport used for the illustration of statistics", but: "In contrast to baseball, few papers in the professional literature analyse cricket, and two rarely analyse the same topic."

This paper analyses two aspects of cricket. The first is an apparently novel investigation of bowling strike rates to determine the probability of bowling out the other team in one-day cricket. The likelihood of running out one or more of the opposing batsmen is then incorporated for greater accuracy, and leads to the useful conclusion that the probability of dismissing the other team is approximately doubled when there is the likelihood of at least one run-out. These ideas were developed in the papers Cohen [4, 5], and are presented here using an improved model and additional comments. (The opportunity is also taken to make some minor corrections to the earlier papers.)

The second aspect, quite distinct from the first and not previously written up, is a further discussion of a topic described in Clarke [2]. It concerns the distribution of scores in the traditional game. Rather than seek directly a probability distribution for the number of runs scored by a particular batsman, we derive instead the distribution of the number of scoring strokes. Scoring strokes are related to the number of runs scored by assuming the ratio of these quantities to be (approximately) constant. Having a probability distribution for the number of scoring strokes then allows the probability to be determined of a batsman scoring a century, say, even if he has not previously made such a score.

2 An application of bowling strike rates

We begin by showing that the strike rates of the bowlers on one team allow an estimate to be made of the probability of getting an opposing batsman out in some manner that is credited to the bowler (so we exclude run-outs for the moment). For one-day cricket, where there is a limit to the number of balls to be bowled in an innings, this can be used to obtain the probability of getting the whole team out. When we include the possibility of run-outs, we get a much better estimate for the probability of dismissing the other side, as confirmed by comparisons with actual results from cricket's World Cup.

Let b and w stand, respectively, for the number of balls bowled by a certain bowler (excluding wides and no-balls) and the number of wickets taken from his bowling in a season, or in his career, or against a particular team, say. Then that bowler, for our purposes, has a strike rate given by b/w. If $w = 0$ (which is hardly likely, for our purposes) then the bowler is deemed not to have a strike rate. A bowler's strike rate, along with his average and his economy rate (neither of which is used here by us), are in common use when analysing the effectiveness of various bowlers; the better bowler has the smaller strike rate. The reciprocal of the strike rate can be interpreted as the probability that the bowler subsequently takes a wicket with each ball bowled.

For a complete team, suppose we have n bowlers so that, in one-day competition, $5 \le n \le 11$. Let their strike rates based on previous experience be s_k, for $k = 1, \ldots, n$. If the kth bowler is to bowl b_k balls in a coming match (excluding wides and no-balls), then experience suggests he will take w_k wickets, where

$$w_k = \frac{b_k}{s_k}, \qquad \text{for } k = 1, \ldots, n.$$

Let B be the total number of balls to be bowled by that team in the match (excluding wides and no-balls), and let W be the total number of wickets taken. Then

$$B = \sum_{k=1}^{n} b_k \qquad \text{and} \qquad W = \sum_{k=1}^{n} w_k = \sum_{k=1}^{n} \frac{b_k}{s_k},$$

and the team's strike rate S for that match may be predicted to be

$$S = \frac{B}{W} = \sum_{k=1}^{n} b_k \Bigg/ \sum_{k=1}^{n} \frac{b_k}{s_k}\,. \tag{1}$$

This is a weighted harmonic mean of $s_1, \ldots, s_n$, the weights being $b_1, \ldots, b_n$.

For example, if four bowlers are to bowl ten overs each, and two others five overs each, then $b_1 = \cdots = b_4 = 60$, $b_5 = b_6 = 30$, and

$$S = \frac{10}{\dfrac{2}{s_1} + \dfrac{2}{s_2} + \dfrac{2}{s_3} + \dfrac{2}{s_4} + \dfrac{1}{s_5} + \dfrac{1}{s_6}}.$$

A bowling combination that allows $S \le 30$, since $B \le 300$ and $W = 10$ in a completed innings, would be most desirable, though rarely achievable in practice.

Typically good individual strike rates satisfy $25 \le s_k \le 50$. If selectors were to choose only that combination of bowlers that allows S to be least, then they would accomplish this by taking the five bowlers with smallest strike rates. However, economy rates or bowling averages and batting and "all round" skills would also all be taken into account, and the captain has his tactical considerations, so it is usually necessary to have a much more varied bowling attack. It would be useful to know then what chances the various bowling combinations have of bowling the other side out.

The quantity $p = 1/S$ represents wickets per ball during the opposing team's innings of at most 50 overs and is the empirical probability with each ball of a bowler taking a wicket, by any of the means

that allow a wicket to be credited to the bowler. The probability of the bowlers taking w wickets in 50 overs is

$$\binom{300}{w} p^w q^{300-w},$$

where $q = 1 - p$, on the assumption that each ball bowled is an independent event. It was argued in [4] that the other team has not been bowled out if $w \leq 9$, so the probability of bowling them out is

$$P_1 = 1 - \sum_{w=0}^{9} \binom{300}{w} p^w q^{300-w}. \tag{2}$$

A more detailed analysis is given in [4] in terms of the bowlers' individual strike rates, but a numerical argument there shows that, for practical purposes, it is sufficient to make use of (2).

However, because $P_1 = \sum_{w=10}^{300} \binom{300}{w} p^w q^{300-w}$, the use of (2) would seem to suggest that the rules of the game in fact allow for ten or more wickets to be taken, but that the game is to be abandoned after ten wickets, the others being defaulted. This whimsy is avoided with the following alternative approach.

To bowl the other team out, ten wickets must be taken and this may be done in anything from ten to 300 balls. If k balls are required, $10 \leq k \leq 300$, then the tenth wicket must be taken with the kth ball, and the first nine wickets with any of the first $k-1$ balls. No wicket is taken with the remaining $k-10$ balls. Hence the probability of bowling the other team out is

$$P_2 = \sum_{k=10}^{300} \binom{k-1}{9} p^{10} q^{k-10} = \left(\frac{p}{q}\right)^{10} \sum_{k=10}^{300} \binom{k-1}{9} q^k.$$

It is reassuring to calculate that values of P_1 and P_2 are, for practical purposes, very close. They agree to four decimal places for S up to 42. In fact, for integer values of S, the greatest difference $P_1 - P_2$ is 0.00335, at $S = 85$. (Always, $P_1 > P_2$ since P_1 is, whimsically, the probability of taking ten or more wickets in 300 balls.)

Coincidentally, the article [4] appeared just as the 1999 World Cup of one-day cricket was about to get under way in England, and it was noticed by the science writer in *The Times*. He wrote a column [12] describing the ideas above and giving his own calculations regarding the English team. Two days later, on the morning of the first match in the World Cup and having seen the English article, *The Australian* [16], in more journalistic style, prevailed upon the author to rank the twelve competing teams in order of the probabilities of bowling their opponents out, even though bowling out the other team does not ensure a win. These probabilities were compared with odds then being offered for each team, and so the ideas in [4] were promoted to a level somewhat above the original conception. (The author's top six ranked teams included five that made the Super Six, who then played off to determine the finalists. This is praiseworthy but not relevant.)

The calculation of S in (1) is described above as being for predictive purposes, based on bowlers' strike rates prior to a match. It may subsequently be compared with the strike rate actually attained in an innings, calculated as the number of balls bowled (not including wides and no-balls) divided by the number of wickets credited to the bowlers. It seems to be standard, if perhaps wrong, that wides and no-balls are not included when determining bowlers' strike rates, so we follow that practice. Moreover, a team's actual overall bowling performance may be based on calculations made following a series of games, such as in the World Cup. It was apparent to the author that the probabilities based on the model in (2) and actual games played in the 1999 World Cup, underestimated the proportion of times that each team in fact bowled out its opponents.

One presumed reason for this was clear: "bowling out" the opponents (as we will use the term) is not the same as "dismissing" them, since the latter includes wickets lost by batsmen who are run out. These are not credited to the bowler. In [5], the methods of [4] were made more realistic by allowing for run-outs, including the possibility that run-outs may occur off wides and no-balls.

World Cup	Matches (α)	Good balls (β)	Bad balls (γ)	Bowled out (δ)	Run out (ϵ)	Teams dismissed (ζ)
1987	27	15413	363	321	64	12
1992	37	20206	661	439	67	18
1996	35	19461	508	411	63	14
1999	42	22721	1218	549	49	27
Totals	141	77801	2750	1720	243	71

Table 1: Data from previous four World Cups.

All match results from the preceding four World Cups (in the years 1987, 1992, 1996 and 1999) were scanned to arrive at estimates for the probability of a run-out with each ball bowled and the average number of wides and no-balls in a 50-over innings. (The World Cups prior to 1987 were 60 overs a side, and not considered for that reason, although the model could be easily adjusted to take this into account.)

The resulting data are given in Table 1. We use the term "good ball" for any delivery not resulting in a wide or no-ball, and "bad ball" for a wide or no-ball. Wickets resulting from good balls, but not run-outs, are credited to the bowler. A batsman can be run out from any ball, good or bad. There are other means of getting out off bad balls (such as being stumped off a wide, in which case the wicket is credited to the bowler), or off good balls with the result not credited to the bowler, but these are very rare and ignored for our purposes. The columns in Table 1 are labelled α, β, ..., ζ for later use.

The 9th and 13th matches in the 1992 World Cup were abandoned due to rain, and the 5th and 14th matches in the 1996 World Cup were forfeited. These have not been included in Table 1. The 16th match in 1996 was replayed after the first attempt was washed out, and only the replayed match has been included. It is possible that some of the figures for balls bowled, both good and bad, in Table 1 may be off by a few from the true numbers, since, for example, umpires' errors (such as allowing a few seven-ball overs) and rule changes for the 1999 World Cup that allowed penalty runs have not always been easy to take into account. We have used the scorecards from CricInfo at www.cricket.org. "Bowled out" refers to wickets credited to bowlers, and in this table includes batsmen who retired hurt or were absent ill, so that they may be taken into account in determining overall bowling strike rates.

Suppose, in a completed innings of 50 overs, there are y bad balls bowled. Re-define B by $B = 300+y$, the total number of balls bowled, so that the probability of any particular ball being good is $300/B = g$, say. Since the strike rate S is based only on good balls bowled, we can give the actual probability of a bowler taking a wicket as gp, where $p = 1/S$, as before. Let r be the probability of a run-out with each ball bowled. Then the probability that the bowlers take w_b wickets, and that a further w_r batsmen are run out, follows a multinomial distribution. It is

$$\binom{B}{w_b, w_r, B - w_b - w_r} (gp)^{w_b} r^{w_r} (1 - gp - r)^{B - w_b - w_r}$$
$$= \frac{B!}{w_b!\, w_r!\, (B - w_b - w_r)!} (gp)^{w_b} r^{w_r} (1 - gp - r)^{B - w_b - w_r},$$

where, in practice, $0 \le w_b + w_r \le 10$, assuming that each ball bowled is an independent event. If $w_b + w_r \le 9$, then the team has not been dismissed, so the probability of dismissing the other side is

$$P_3 = 1 - \sum_{0 \le w_b + w_r \le 9} \frac{B!}{w_b!\, w_r!\, (B - w_b - w_r)!} (gp)^{w_b} r^{w_r} (1 - gp - r)^{B - w_b - w_r}.$$

When $y = 0$ and $r = 0$, so that $g = 1$ and $w_r = 0$, this reduces to the result in (2) (with the understanding that then $r^{w_r} = 1$).

Although numerically accurate, this formula has the same conceptual drawback as for P_1. Instead, we may argue as follows.

Let w be the number of wickets taken by the bowlers, $0 \le w \le 10$, so that $10 - w$ is the number of run-outs. The taking of wickets and the bowling of balls are considered to be independent events, except that a wicket must be taken with the last of k balls bowled, $10 \le k \le B$. Then the probability of dismissing the other side is

$$P_4 = \sum_{w=0}^{10} \binom{10}{w} (gp)^w r^{10-w} \cdot \sum_{k=10}^{B} \binom{k-1}{9} (1 - gp - r)^{k-10}. \tag{3}$$

When $r = 0$, we must have only the summand with $w = 10$ and then, as above, must interpret 0^0 as 1. With $y = 0$, then P_4 reduces to the expression for P_2.

We can now demonstrate that this model approximates well the actual results from the four World Cups.

World Cup	S	r	y'	Π	P
1987	48.016	0.00406	7.07	0.222	0.220
1992	46.027	0.00321	9.81	0.243	0.222
1996	47.350	0.00315	7.83	0.200	0.199
1999	41.386	0.00205	16.08	0.321	0.268
Combined	45.233	0.00302	10.60	0.252	0.229

Table 2: Actual proportion (Π) of teams dismissed and predicted probability (P) of dismissing a team, given Cup bowling strike rate (S), run-outs per ball (r), and average number of wides and no-balls (y').

From the data given in Table 1, we may calculate combined bowling strike rates $S = \beta/\delta$ for each World Cup, the proportion $r = \epsilon/(\beta + \gamma)$ of run-outs, and the average number $y' = 300\gamma/\beta$ of bad balls in a 50-over innings. We also have from Table 1 the actual proportion $\Pi = \zeta/(2\alpha)$ of teams dismissed. We put $p = 1/S$, r and y (equal to y', rounded to the nearest integer) into (3) to obtain the values $P = P_4$ in Table 2. Compare the values of P and Π.

Notice that the values for r in Table 2 show that there are on average about three run-outs per 1000 balls in world class one-day cricket, which equates to about one per innings of 50 overs. The values for y' show that there are, say, seven to ten wides or no-balls altogether in a 50-over innings. (The 1999 World Cup seems to be exceptional in the latter regard—this was the time when accusations of corrupt practice in cricket were rife and in many cases subsequently shown to be justified, and perhaps here we see some evidence for the accusations.)

Finally in this section, we give Table 3. For team bowling strike rates S from 20 to 62, incremented by 2, and three values (0.002, 0.003 and 0.004) for the probability r of a run-out with each ball (pick the probability that matches the team's fielding skills or the opponents' lapses in running), we give the probability of dismissing the other team. We have taken $y = 10$, although it turns out that, whether $y = 0$ or $y = 20$, the computed values are rarely affected even in the second decimal place. (Because P_4 is used rather than P_3, Table 3 differs in a few entries, but not at all substantially, from the corresponding table in [5].)

The first row of Table 3, with $r = 0$, corresponds to the probability of dismissing the opponents if run-outs are not to be considered. This should still be seen as important to assist a captain or selector to estimate the ability of their chosen team to bowl out the opponents (the theme of the paper [4]), as other considerations would not then be taken into account. Perhaps of more interest is a comparison between the entries corresponding to $r = 0$ (no run-outs) and $r = 0.003$ (close enough to one run-out in 300 balls) for $46 \le S \le 60$ (a range for the team bowling strike rate that would be common in practice). They may be interpreted as showing that the probability of dismissing the other team is approximately doubled if there is a good likelihood of a run-out.

r	S										
	20	22	24	26	28	30	32	34	36	38	40
0	0.93	0.88	0.80	0.72	0.63	0.54	0.46	0.39	0.32	0.27	0.22
0.002	0.95	0.91	0.85	0.78	0.70	0.62	0.54	0.47	0.41	0.35	0.30
0.003	0.96	0.92	0.87	0.80	0.73	0.66	0.58	0.51	0.45	0.39	0.34
0.004	0.97	0.93	0.88	0.82	0.76	0.69	0.62	0.55	0.49	0.43	0.38
r	S										
	42	44	46	48	50	52	54	56	58	60	62
0	0.18	0.15	0.12	0.10	0.08	0.07	0.06	0.04	0.04	0.03	0.02
0.002	0.25	0.21	0.18	0.15	0.13	0.11	0.09	0.08	0.07	0.06	0.05
0.003	0.29	0.25	0.22	0.18	0.16	0.14	0.12	0.10	0.09	0.07	0.06
0.004	0.33	0.29	0.25	0.22	0.19	0.17	0.15	0.13	0.11	0.10	0.08

Table 3: Probability of dismissing the other team, given the probability of a run-out (r) and the team bowling strike rate (S), and assuming 10 wides or no-balls are bowled per 50 overs.

There was further newspaper interest in these ideas in January 2001, culminating in an article [6] in Sydney's *Daily Telegraph*. That article included also a description of the main results in de Mestre and Cohen [10], and it was reprinted with a little more mathematical detail in Cohen and de Mestre [7]. The newspaper article included probabilities of dismissing the other team for the triangular one-day series about to commence between Australia, Zimbabwe and the West Indies. The predictions were acceptably accurate, as detailed in [7].

3 An application of batting averages

As we have indicated above, the work of this section is quite distinct from the preceding work. It will be convenient to use a similar notation to before, but now from a batsman's point of view. For example, we will use b for the number of balls faced by a particular batsman, rather than the number bowled by a particular bowler.

3.1 The distribution of scoring strokes

In cricket, a batsman's average is the number of runs he has scored divided by the number of times he was out. If a tail-end batsman scores five in each of ten innings in a season and is not out nine times, then he finishes the season with an average of 50. For good reason, this is not seen to be a properly representative score.

It seems that there are two separate questions that people expect the one batting average to answer.

- First, how good is the batsman? What score would we expect of him if he were allowed to bat on, leaving the field for the final time only when he is given out in a standard manner?
- Second, what score do we expect of him given the possibilities also of his retiring hurt, or running out of batting partners, or having the team's innings declared closed or the match interrupted for some reason such as rain, in all of which cases he would remain not out?

We will try to answer both questions.

At least two papers have attempted to determine more significant single measures of batting performance, generally being more intent on answering the first of the above questions. Danaher [8] used analogies with survival analysis to find an estimate of a cricketer's "true but unknown batting average" based on the product limit estimator. Kimber and Hansford [14] also adopted an approach "akin to that used in reliability and survival analysis", and also based on product limit estimation, to arrive at

a different nonparametric estimator. The latter gives values generally much closer to the traditional average than Danaher's estimator (with both always giving smaller values), and both have the property that the fewer the number of not-outs, the closer their estimator is to the traditional average. On the other hand, Davis [9, pages 96–98], argues from an empirical viewpoint for the worth of the traditional average.

It seems reasonable that the score you might expect a batsman to attain would be his "true average", based on all relevant previous innings. To define this term, we take data pertaining to a particular batsman over a particular period, such as his career or the previous season, or in a particular position, or against a particular team. Let i, n, w and r be, respectively, the number of innings, the number of not-out innings, the number of dismissals, and the number of runs scored. Then $w = i - n$. The batsman's traditional and true averages are

$$B = \frac{r}{w}, \qquad A = \frac{r}{i},$$

respectively. Notice that

$$A = \frac{w}{i}B = \frac{i-n}{i}B, \tag{4}$$

so that, for overall career results, say, the true average may be determined from the usual published batting statistics. Of course, $A = B$ when $n = 0$.

We will justify our use of the term "true average" by obtaining in a theoretical fashion the probability distribution of the number of scoring strokes and, based on this, showing that the expected value of the batsman's score (in the statistical sense) equals this true average.

The same approach will allow us to find the probability of the batsman making 100 runs, or any other score. Thus we answer the intriguing question: how do you estimate a batsman's probability of making a century if he has not yet made one? We will see that our probability compares well with the actual frequency of century scores by batsmen who have made a few centuries.

Our method relies on the new concept of the *strike constant*. This is the ratio of runs made to number of scoring strokes and its introduction may be viewed as a device to serve our end: it is a first approximation to a comparison of runs made and scoring strokes which indeed (as we will see) leads to plausible and testable results. An investigation of this ratio for a large number of Sydney grade cricketers by Cochran [3] came up with the value 2.16, with standard deviation 0.25, for traditional cricket, and 1.82, standard deviation 0.43, for limited-overs cricket. (He also investigated indoor cricket: ignoring runs subtracted for loss of wicket, the mean strike constant was 2.08 with standard deviation 0.41.)

We consider the main application of this work to be to the traditional form of cricket. Strike constants for individual cricketers over a small number of matches might range between 1.9 and 2.4, say, but this will be seen in any case to have little effect on the final calculations.

We will show that, subject to certain assumptions, the number of scoring strokes follows a geometric distribution. The distribution of runs scored is related to this through the strike constant. The possibility that cricket scores are geometrically distributed goes back at least to the writings of Elderton [11]. Wood [15] gives further numerical evidence to support this. Both these papers are dismissed by Kimber and Hansford [14] as "flawed because the authors treated not-out scores as if they were completed innings", despite the evidence of the data. The details are summarised by Clarke [2]. *Inter alia*, Clarke states: "If a batsman scores only singles and his probability of dismissal is constant, his scores should follow a geometric distribution, the discrete equivalent of the negative exponential." This observation appears to be based on a viewing of the empirical data, but will be a direct consequence of our work below.

3.2 Expected values

In addition to the quantities i, n, $w = i - n$ and r introduced above, let b and s be, respectively, the number of balls faced and the number of scoring strokes made. We must have $r \geq s \geq 0$ and we will assume that $b > i > n \geq 0$. Then $w > 0$. Recall that $A = r/i$.

We define

$$\begin{aligned} p_w &= \Pr\,(\text{the batsman's innings ends, out or not out, with each ball faced}) = \frac{i}{b},\\ q_w &= 1 - p_w,\\ p_s &= \Pr(\text{the batsman makes a scoring stroke with each ball faced} \mid \text{the batsman's}\\ &\qquad \text{innings does not end with that ball}) = \frac{s}{b-i},\\ q_s &= 1 - p_s. \end{aligned}$$

These probabilities are considered to be constant throughout a subsequent innings.

Notice that we have made an assumption that no scoring stroke is made from the ball on which the batsman's innings ends (so that $s \le b - i$). Therefore, we do not take into account the rare instance in which the batsman makes at least one run and is then run out on the same ball while attempting a further run, or the admittedly more common instance in which a captain declares an innings closed following the batsman's final scoring stroke.

In *any* period, the ratio of the number of runs obtained to the number of scoring strokes made is considered to be constant. This is the simplification described above. The ratio is the strike constant, denoted by κ. Then

$$\kappa = \frac{r}{s}.$$

Let the random variable X be the number of scoring strokes made by the batsman in a subsequent innings, and let R be the score (number of runs) obtained. In order that $X = k$ for integer $k \ge 0$, the batsman must face $j + 1$ balls, for some $j \ge k$, scoring on k of these and having his innings end on the $(j + 1)$th ball. (If the team's innings ends or the batsman is run out while not facing, some number of balls after last facing a ball himself, this is still effectively the case.) Whether or not he scores off any of the first j balls bowled are considered to be independent events, and so the distribution of the k scoring strokes among the first j balls bowled to him will be binomial(j, p_s). Write $\Pr\,(X = k)$ for the probability that the batsman makes k scoring strokes ($k \ge 0$), before being dismissed. (Later notations will have a corresponding meaning.) Then

$$\begin{aligned} \Pr\,(X = k) &= \sum_{j=k}^{\infty} q_w^j p_w \cdot \binom{j}{k} p_s^k q_s^{j-k}\\ &= \frac{p_s^k p_w q_w^k}{k!} \sum_{j=k}^{\infty} \frac{j!}{(j-k)!} (q_s q_w)^{j-k}\\ &= \frac{p_s^k p_w q_w^k}{(1 - q_s q_w)^{k+1}} = \frac{p_w}{1 - q_s q_w} \left(\frac{p_s q_w}{1 - q_s q_w}\right)^k = P Q^k, \end{aligned}$$

where

$$P = \frac{p_w}{1 - q_s q_w}, \quad Q = \frac{p_s q_w}{1 - q_s q_w} = 1 - P.$$

Thus the number of scoring strokes made follows a geometric distribution. (Notice, for example, that if $p_s = 1$ then this reduces to $\Pr\,(X = k) = q_w^k p_w$, for $k \ge 0$.) Using the definitions of p_w and p_s, we find that

$$P = \frac{i}{i+s} = \frac{\kappa}{A + \kappa}, \quad Q = \frac{A}{A + \kappa}.$$

The expected number of scoring strokes is then easily determined, or it may be obtained as a particular case from results in Johnson *et al.* [13]. We have

$$\mathrm{E}\,(X) = \sum_{k=0}^{\infty} k \cdot \Pr\,(X = k) = \frac{Q}{P} = \frac{A}{\kappa}.$$

Then, since $X = 0$ if and only if $R = 0$,

$$\mathrm{E}(R) = \mathrm{E}(R \mid R > 0) = \mathrm{E}\left(X \cdot \frac{R}{X} \,\middle|\, X > 0\right) = \mathrm{E}(\kappa X \mid X > 0) = \kappa \mathrm{E}(X) = A.$$

This is our theoretical justification of the use of the true average. For a batsman with no not-out innings ($n = 0$), his expected number of runs will be the traditional average B.

It is well known, however, that the average is a very imprecise measure of the score you might expect a batsman to achieve. A technical reason for this may be demonstrated as follows.

The geometric distribution, $\Pr(X = k) = PQ^k$ for $k \geq 0$, has mean and variance

$$\mu_X = \mathrm{E}(X) = \frac{Q}{P}, \quad \sigma_X^2 = \mathrm{var}(X) = \frac{Q}{P^2},$$

respectively (see [13]), from which $\sigma_X = \mu_X/\sqrt{Q}$. We have $\mu_X = A/\kappa$, so

$$\sigma_X = \frac{A}{\kappa}\sqrt{1 + \frac{\kappa}{A}}.$$

Then, for the variance of the batsman's score, we have $\mathrm{var}(R) = \mathrm{var}(\kappa X) = \kappa^2 \mathrm{var}(X)$, so that

$$\sigma_R = \kappa \sigma_X = A\sqrt{1 + \frac{\kappa}{A}}.$$

The standard deviation of the number of runs scored will thus be a little greater than the expected value of that number. Because of that, we cannot really expect the batsman to make anything very close to his true average score at any appearance. We shall return later to this situation.

3.3 The probability of scoring a century (or a duck)

By "scoring c runs", we shall mean the usual notion of actually making c or more runs, except below when we give the probability of a duck (zero runs).

Writing $d = \lfloor (c-1)/\kappa \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor (or greatest integer) function, we have, for $c \geq 1$,

$$\begin{aligned}\Pr(R \geq c) &= 1 - \Pr(R \leq c-1) = 1 - \Pr(\kappa X \leq c-1) = 1 - \Pr(X \leq d) \\ &= \sum_{k=d+1}^{\infty} \Pr(X = k) = Q^{d+1} = \left(\frac{A}{A+\kappa}\right)^{d+1}.\end{aligned}$$

If $\kappa = 2$, say, then these probabilities are equal for $c = 2m$ and for $c = 2m - 1$, for any positive integer m. In practice that should be considered as acceptable, especially for large c, and in theory (*this* theory) it is unavoidable.

An application of the formula is given in Table 4. This gives in particular the number of century scores and over-50 scores by the all-time top twenty Test batsmen (as at 7 February 2002; qualification: at least 20 innings), and the estimates of these values resulting from our formula. (The referee lamented that Adam Gilchrist was not on the list; by early May 2002, he was in fact fifth on the list with a traditional average of 60.00.)

The first four columns from left to right in Table 4, following the batsman's name, are a selection of standard published data: the number of innings; the number of not-out innings; the number of runs scored; and the traditional average (previously denoted by B). Then come values for our true average A, which may be calculated from the data in the preceding columns using equation (4), and values of A_w, which is a batsman's "wicket-average" to be discussed below. Finally, there are, for each batsman, the number of century scores (headed 100+); the expected values of these using the above probability calculation multiplied by the number of innings; the number of scores of 50 or more (headed 50+ and

Name	Inn.	NO	Runs	Ave.	A	A_w	100+	E (100+)	50+	E (50+)
DG Bradman	80	10	6996	99.94	87.45	85.04	29	25.8	42	45.5
RG Pollock	41	4	2256	60.97	55.02	54.43	7	6.9	18	16.8
GA Headley	40	4	2190	60.83	54.75	45.61	10	6.7	15	16.3
H Sutcliffe	84	9	4555	60.73	54.23	54.64	16	13.7	39	34.0
E Paynter	31	5	1540	59.23	49.68	48.31	4	4.3	11	11.6
KF Barrington	131	15	6806	58.67	51.95	50.37	20	19.8	55	50.9
EdeC Weekes	81	5	4455	58.61	55.00	54.88	15	13.6	34	33.2
WR Hammond	140	16	7249	58.46	51.78	46.19	22	21.0	46	54.3
SR Tendulkar	143	15	7419	57.96	51.88	48.85	27	21.6	57	55.5
GStA Sobers	160	21	8032	57.78	50.20	44.06	26	22.7	56	60.2
JB Hobbs	102	7	5410	56.95	53.04	53.34	15	16.0	43	40.4
CL Walcott	74	7	3798	56.69	51.32	51.03	15	10.9	29	28.5
L Hutton	138	15	6971	56.67	50.51	47.89	19	19.8	52	52.3
GE Tyldesley	20	2	990	55.00	49.50	47.22	3	2.8	9	7.4
MH Richardson	21	1	1088	54.40	51.81	50.75	2	3.2	10	8.1
DR Martyn	35	9	1413	54.35	40.37	35.88	4	3.1	9	10.4
CA Davis	29	5	1301	54.21	44.86	40.83	4	3.3	8	9.7
VG Kambli	21	1	1084	54.20	51.62	53.30	4	3.1	7	8.1
GS Chappell	151	19	7110	53.86	47.09	44.57	24	21.8	55	57.3
AD Nourse	62	7	2960	53.82	47.74	47.49	9	8.0	23	22.2

Table 4: The all-time top twenty Test batting averages, at 7 February 2002, with approximate expected values of number of scores of 100 or more, or 50 or more (E (100+) and E (50+), respectively).

differing from the usual lists which give the number of scores from 50 to 99, inclusive); and their expected values similarly calculated.

Unless it is possible to have access to the original score sheets, it is most unlikely that actual values of κ, the ratio over the past of runs made to number of scoring strokes, could be obtained. The easy approach is to set $\kappa = 2$, and this was done in Table 4. (The expected values, whether we took $\kappa = 1.9$, 2 or 2.1 were not appreciably different.)

A large proportion of the expected values in Table 4 are observed to match their actual values very well, so that, in this case at least, the model fits the data acceptably. The table suggests that our model, with $\kappa = 2$, will allow reasonable predictions to be made.

Taking $\kappa = 2$ allows a further simplification. By assuming that c is even, as in the common cases $c = 100$ or $c = 50$, we obtain

$$\Pr(R \geq c) \approx \left(\frac{A}{A+2}\right)^{c/2}, \tag{5}$$

and, if desired, this may be adopted as a useful approximation for all $c \geq 1$.

The wicket-average A_w in Table 4 is the average of only those innings in which the batsman was out. (These values were obtained by going back to the lists of all Test scores, for each batsman.) This information has been included to show that the true average and the wicket-average are in most cases very close, as one would expect if a batsman averaged much the same in his completed innings as in his not-out innings. However, the true average is greater in all but two cases, indicating that not-out scores tend on average to be greater than completed innings. Sometimes this is emphatically so, as in Headley's case: his not-out Test scores were 102, 169, 270 and 7.

The point of tabulating A_w is to give an example of other averages that might be determined for more accurate predictions. Using equation (5) with $A = A_w$ and $c = 100$ will give an estimate of the chance of scoring a century, with the batsman getting out. (The earlier theory needs to be adjusted in a minor way to allow for the different sample space: b, r and s now relate only to completed innings, and i in the definitions of p_w and p_s must be replaced by w. Then, in particular, p_w is the probability of the batsman losing his wicket, out, with each ball faced. The subsequent analysis would then refer only to completed innings.)

We return now to the question of determining a more useful means of estimating a batsman's future score than simply giving the expected value. We will instead find "50% probability intervals" for the

score, for differing values of A. That is, for each A, we will determine an approximate interval with the property that the batsman would obtain a score in it with probability 0.5, with equal probabilities of smaller or greater scores outside the interval.

We make use of (5). For a given probability p, the score c required to ensure that $\Pr(R \geq c) = p$ is obtained approximately by solving

$$\left(\frac{A}{A+2}\right)^{c/2} = p.$$

We obtain

$$c = \frac{2\log(1/p)}{\log(1+2/A)}, \tag{6}$$

where the logarithms may be to any suitable base. Taking the ceiling value when $p = 0.75$ and the floor value when $p = 0.25$, we obtain our approximate interval.

Examples of these intervals appear in Table 5. We have taken values of A from 10 to 65, incremented by 5, and, in case the ghost of Sir Donald is watching, also $A = 90$. Notice that any sensible average can be used. Thus, for Damien Martyn with $A_w \approx 35$ and $A \approx 40$ (see Table 4), we could say he has a 25% chance of scoring 50 or more, getting out, but the same chance of scoring more than 56, out or not out.

batting average	10	15	20	25	30	35	40
50% probability interval	[4, 15]	[5, 22]	[7, 29]	[8, 36]	[9, 42]	[11, 49]	[12, 56]
batting average	45	50	55	60	65	...	90
50% probability interval	[14, 63]	[15, 70]	[17, 77]	[18, 84]	[19,91]	...	[27, 126]

Table 5: A batsman with true batting average shown (*not* the traditional average) has probability 0.5 of making a score in the given interval, with equal probabilities of smaller or greater scores outside the interval.

Using equation (6) with $p = 0.5$ allows us to use a batsman's traditional average number of runs scored to estimate his median number of runs scored. In fact, for $A \geq 20$, say, we have $\ln(1+2/A) \approx 2/A$, so that the median score is about $A \ln 2$. Thus $0.7A$ would be an easy approximate formula for the median.

Data on median scores is almost nonexistent, but Wood [15, Table B] gives this information for 22 "leading batsmen" to September 1939. Their ratio of median score to traditional average score (B) ranges from 0.61 to 0.71. Wood's list does not allow direct calculations of the true average A, since he does not give the numbers of not-out innings. The CricInfo web site allowed this to be done (although it differed from Wood on *every* occasion in the number of first class innings for the batsmen on his list). This exercise suggested that taking $A \approx 0.9B$ would be acceptable in general (and is the rule of thumb following the observation by Clarke [2] that "more than 10% of scores are not outs"), so that $0.63B$ would be a useful theoretical estimate of a batsman's median score over a long career.

We also note that the formula (5) retains the non-aging (or Markovian, or lack of memory) property of the geometric distribution (see Johnson *et al.* [13, page 201]). Thus, for example, $\Pr(R \geq 100) = (\Pr(R \geq 50))^2$.

Finally, we consider the probability of a batsman getting a duck. From our early work, the probability of a batsman making no scoring stroke is

$$\Pr(X = 0) = P = \frac{\kappa}{A+\kappa}.$$

But this includes the probability of scoring 0, not out. Our earlier discussion suggests the following as the way to go:

$$\Pr(\text{duck}) \approx \frac{2}{A_w + 2}.$$

Name	Inn.	NO	A_w	ducks	E(ducks)
DG Bradman	80	10	85.04	7	1.8
RG Pollock	41	4	54.43	1	1.4
GA Headley	40	4	45.61	2	1.7
H Sutcliffe	84	9	54.64	2	3.0
E Paynter	31	5	48.31	3	1.2
KF Barrington	131	15	50.37	5	5.0
EdeC Weekes	81	5	54.88	6	2.8
WR Hammond	140	16	46.19	4	5.8
SR Tendulkar	143	15	48.85	7	5.6
GStA Sobers	160	21	44.06	12	6.9
JB Hobbs	102	7	53.34	4	3.7
CL Walcott	74	7	51.03	1	2.8
L Hutton	138	15	47.89	5	5.5
GE Tyldesley	20	2	47.22	2	0.8
MH Richardson	21	1	50.75	0	0.8
DR Martyn	35	9	35.88	2	1.8
CA Davis	29	5	40.83	1	1.3
VG Kambli	21	1	53.30	3	0.8
GS Chappell	151	19	44.57	12	6.4
AD Nourse	62	7	47.49	3	2.5

Table 6: The all-time top twenty Test batsmen, by traditional batting average, at 7 February 2002, with number of ducks scored and the approximate expected value of this number.

The world's top twenty batsmen have been known to score a duck or two. Table 6 repeats some information from Table 4, and gives the number of ducks scored by those batsmen in Tests and our suggested expected value of this number using the probability estimate $2/(A_w + 2)$ multiplied by the number $i - n$ of completed innings. The table indicates some level of agreement between the actual and expected values; any attempt to model small numbers like these would be generally acknowledged as difficult.

4 Conclusion

Many papers concerned with tennis have exploited the fact that the proportion of points won by a player in some situation allows estimates of the probability of winning a game, set or match in a similar future situation. Considering separately points won on service and points won when receiving leads to refined estimates. In Bennett [1], there are references to probabilistic analysis in tennis, baseball, basketball and American football, and numerous other relevant references. Yet, as we have already quoted Stephen Clarke as saying, hardly any such analysis has previously taken place in cricket.

A "winning" ball in a game of cricket is one that takes a wicket from the bowler's point of view, or allows a scoring stroke from the batsman's point of view. The proportion of winning balls has been used in this paper to give the probability of bowling out a team, in the former case, or scoring a century, in the latter case. Along the way, refinements and other applications have been given.

Bowling strike rates, along with estimates of the probability of running out an opposing batsman, have been used in Section 2 not only to find the probability of dismissing the other side in one-day cricket, but to demonstrate that this chance is approximately doubled when there is a good likelihood of obtaining at least one run-out.

At the beginning of Section 3, two questions were posed regarding conclusions to be drawn from the traditional batting average. But, to make a pun of it, this average is a very demeaned statistic. Even in cricketing circles, it is not seen as being properly representative of a batsman's past scores because of the "ad hoc" treatment of not-outs.

We prefer instead the true batting average A: simply the average of all scores, out or not out. The wicket average A_w, which refers specifically to completed innings, is approximately the same as A and should be used for questions concerning completed innings (such as the first of those at the beginning of

Section 3). Use A otherwise. Among other things, we have justified the simple formula $(A/(A+2))^{c/2}$ as the probability of scoring at least c runs, and the formula $2/(A_w+2)$ as the probability of a duck. Both of these have been compared favourably with results from the history of cricket.

The work of Section 3 depends crucially on the concept of the strike constant, although less crucially on the value chosen for it. As a theoretical if hypothetical construct, its worth seems clear, and further investigation of the notion would be extremely welcome.

Acknowledgment

I am grateful for discussions with Dr Peter Wright, formerly of the University of Technology, Sydney, with regard to Section 3. In particular, he suggested the use of the 50% probability interval to quantify a batsman's chances.

References

[1] J. Bennett (editor), *Statistics in Sport*, London, Arnold (1998).

[2] S. R. Clarke, "Test statistics", in *Statistics in Sport*, J. Bennett (editor), Arnold, London (1998), 83–103.

[3] K. Cochran, *Comparison of the Score to Scoring Strokes Ratio in Various Forms of Cricket*, undergraduate essay, Department of Mathematical Sciences, University of Technology, Sydney (2000).

[4] G. L. Cohen, "One-day cricket: inferences from bowlers' strike rates", *Math. Today*, **35** (1999), 45–47.

[5] G. L. Cohen, "One-day cricket: the effect of running out an opposing batsman", *Math. Today*, **36** (2000), 75–77.

[6] G. Cohen, "For openers, let's try a little maths", *The Daily Telegraph* (11 January 2001), 18.

[7] G. Cohen and N. de Mestre, "Mathematics and cricket", *Reflections*, **26** (May 2001), 11–13.

[8] P. J. Danaher, "Estimating a cricketer's batting average using the product limit estimator", *N. Z. Statist.*, **24** (1989), 2–5.

[9] C. Davis, *The Best of the Best*, Sydney, ABC Books (2000).

[10] N. de Mestre and G. Cohen, "The flight of a cricket ball", in *Fifth Conference on Mathematics and Computers in Sport*, G. Cohen and T. Langtry (editors), University of Technology, Sydney, Australia (2000), 107–112.

[11] W. Elderton, "Cricket scores and some skew correlation distributions", *J. Roy. Statist. Soc.*, **108** (1945), 1–11.

[12] N. Hawkes, "Howzat for statistics", *The Times* (12 May 1999), downloaded from the world wide web.

[13] N. L. Johnson, S. Kotz and A. W. Kemp, *Univariate Discrete Distributions* (second edition), New York, Wiley (1993).

[14] A. C. Kimber and A. R. Hansford, "A statistical analysis of batting in cricket", *J. Roy. Statist. Soc. Ser. A*, **156**, (1993) 443–455.

[15] G. H. Wood, "Cricket scores and geometrical progression", *J. Roy. Statist. Soc.*, **108** (1945), 12–22.

[16] J. Zubrzycki, "Professor's figures add up for punters", *The Australian* (14 May 1999), 18.

HOW TO FIX A ONE-DAY INTERNATIONAL CRICKET MATCH

Stephen Gray and Tuan Anh Le*
Centre for Valuation and Venture Capital
University of Queensland Business School
Brisbane, Queensland 4072, Australia.
gray@commerce.uq.edu.au, l.le@commerce.uq.edu.au

Abstract

In this paper, we develop a Gaussian kernel non-parametric regression model for one-day international cricket matches. The model can be used to forecast the total score of the team that bats first as a function of overs and wickets remaining and runs scored so far. It can also be used to determine the probability of the second-batting team winning the match, as a function of overs and wickets remaining and runs required. We also develop a filter that allows us to impose monotonicity constraints in an optimal way. For example, the predicted score must be higher if there are more overs and/or wickets remaining, but an unconstrained kernel regression model admits violations of such logical constraints. Moreover, the framework we develop can be used to assess the impact of individual events on the likely outcome of the match. For example, it is possible to compare the relative impact of a lost wicket versus a number of slow-scoring overs. Thus, the model has applications to coaches, commentators, and match-fixers.

1 Introduction

In a standard one-day international cricket match, each team has 50 overs and ten wickets with which to score runs. Wickets and overs remaining can be thought of as the resources available to the team, with the objective of converting these resources into runs. In this paper, we seek to model and predict how efficiently these resources will be converted into runs. In particular, we seek to estimate two quantities. For the team that bats first, we seek a prediction of their total score, conditional on the runs scored so far, wickets remaining, and overs remaining. For the team that bats second, we seek to measure the probability that they will win the match, conditional on the runs still required, wickets remaining, and overs remaining.

The framework we develop involves a two-stage approach. First, we use a data-intensive non-parametric regression technique. This essentially involves basing predictions on outcomes from similar observations in the data set. Suppose, for example, that we seek to predict the final score of a team that has lost three wickets for 100 runs in the first 20 overs of the match. We could examine our data set to find any instance of this score in past games, and this could form the basis of our prediction. Of course, the problem is that there are likely to be very few (if any) instances of this *exact* score in our data set. But there may be observations that are close (according to a metric to be defined). Information about the final score of a team that was three for 102 off 21 overs, would presumably be relevant. Our technique would also give this observation some weight, and observations that are less close will receive less weight.

*We are grateful to Wally Boudry and Aaron Macksey for expert research assistance and help in collecting data. The first-named author is employed also at the Fuqua School of Business, Duke University.

The strength of this approach is that it allows the data, and the data alone, to guide us in making predictions. But this is also a weakness as it admits logical contradictions. For example, the predicted score should be higher if there are more overs and/or wickets remaining, but an unconstrained non-parametric regression model admits violations of such constraints. Consider predicting the final score of a team with 170 runs and ten overs remaining. It may be that in the actual data set, teams in this circumstance with four wickets remaining do better than teams with ive wickets remaining. A non-parametric regression model may then predict that, in this region, losing another wicket would be beneficial. This type of counter-intuitive result is particularly likely if the data are sparse in the relevant region.

To avoid such logical contradictions, we develop a filtering algorithm that allows us to impose monotonicity constraints in an optimal way. We impose two constraints: (i) the predicted score must be higher if there are more overs and/or wickets remaining (other things equal), and (ii) the predicted probability of the second team winning the match must be greater if they have more overs and/or wickets remaining and if they require fewer runs. Our filtering algorithm adjusts the predictions from the non-parametric regression model such that these constraints are satisfied. The algorithm minimises a weighted sum of the squared movements and provides a unique transformation. The result is a set of conditional predictions that is driven by a large data set, but conforms to this set of logical monotonicity constraints.

The remainder of the paper is structured as follows. Section 2 provides an overview of our modelling approach. It describes the non-parametric regression procedure and the filtering algorithm that we develop. Section 3 describes our data set. Section 4 presents the empirical results and Section 5 concludes.

2 Model development

The first stage of our approach involves estimating a Gaussian kernel non-parametric regression model. In this stage we seek an estimate of two quantities. For the team that bats first, we seek a prediction of their total score, conditional on the runs scored so far, wickets remaining, and overs remaining. For the team that bats second, we seek to measure the probability that they will win the match, conditional on the runs still required, wickets remaining, and overs remaining. We describe how the non-parametric regression technique can be used to estimate each of these quantities in turn below.

2.1 Predicting the score of the team that bats first

The objective of this component of the model is to forecast the total score of the first-batting team. A flexible model is required so that such a forecast can be made at any time during the first team's innings. At any point throughout this innings, the three most obvious influences on the first team's total score are (i) the number of runs scored so far, (ii) the number of overs remaining, and (iii) the number of wickets remaining. Therefore, the total score should be modelled as a function of (at least) these three predictor variables. What makes this exercise difficult is the *interaction* among the variables. An extra five overs of batting time, for example, is worth more if there are eight wickets remaining than if there is only one. Duckworth and Lewis (1998) recognise this interaction in considering both overs and wickets to be "remaining resources" in their exponential function that is used for target-adjustment in rain-affected matches.

Rather than attempting to fit the data to a concise mathematical formula, a non-parametric data-driven approach is adopted here. The goal is to form an estimate of the Team 1[1] total as a function of runs scored so far and overs and wickets remaining. This is done by finding a subset of the data for which the predictor variables are similar to those of the current match. For example, suppose the current score is three wickets for 100 runs in the 25th over. The non-parametric kernel regression technique

[1]Henceforth "Team 1" will be used to denote the team that bats first and "Team 2" will be used to denote the team that bats second.

essentially searches through the data set and selects all observations for which the number of wickets is close to three, the number of runs is close to 100 and the number of overs is close to 25. The total score of teams in this subset of the data then forms the basis of the estimate of the conditional expectation of this team's total score. When using this technique, "close" is defined in terms of a Gaussian weighting function.

More formally, T is defined to be the total score of Team 1 and the goal is to compute the conditional expectation of T, given values of the predictor variables, $E[T \mid r, w, o]$, where r represents the run-rate per over achieved so far, w represents the number of wickets remaining, and o represents the number of overs remaining. Throughout the paper we estimate run-rates per over rather than runs in absolute terms. In this way, quantities are comparable at different stages throughout the match. Thus, we estimate the expected run-rate per over for the rest of the innings, rather than the number of runs yet to be scored. In this case, for a particular observation i,

$$E[T_i \mid r_i, w_i, o_i] = r_i + o_i \times E[R_i \mid r_i, w_i, o_i],$$

where R_i is the run-rate per over for the remainder of the innings.

This conditional expectation is estimated using the Nadaraya–Watson Gaussian kernel technique as described in Nadaraya (1965):

$$E[R_i \mid r_i, w_i, o_i] = \frac{\sum_{j=1}^{N} R_j N((r_j - r_i)/h_r) N((w_j - w_i)/h_w) N((o_j - o_i)/h_o)}{\sum_{j=1}^{N} N((r_j - r_i)/h_r) N((w_j - w_i)/h_w) N((o_j - o_i)/h_o)}.$$

Here we choose a triplet (r_i, w_i, o_i) and we want an estimate of the total score, conditional on these values. To compute this, we sum the total score for each of the $j = 1, \ldots, N$ observations in our data set, each weighted by how "close" it is to the triplet being evaluated. Closeness is defined in terms of a Gaussian distance function in each of the three conditioning variables. Thus, $N(\cdot)$ represents the standard normal probability density function. In each case, we use a smoothing parameter, h. Large values of this parameter result in smoother estimates as a more diffuse weighting function means that more observations from the data set are given more weight in the estimation. Smaller values of h result in only data observations very close to the triplet entering into the estimation with any significant weight. In all of our calculations, we use the optimal bin width of Scott (1979). The implementation of this technique is described further in Silverman (1986).

2.2 Predicting the probability that the second-batting team wins the match

The objective of this component of the model is to measure the conditional probability of the second-batting team winning the match. The model must be flexible enough so that this forecast can be made at any time during the second team's innings. At any point throughout this innings, the four most obvious influences on the outcome of the match are (i) the first team's total score, (ii) the number of runs scored by the second team so far, (iii) the number of overs remaining for the second team, and (iv) the number of wickets remaining for the second team. The first two of these can be combined as the run-rate per over required to reach the target set by the first team.

The result of the match is defined as:

$$P = \begin{cases} 0, & \text{if Team 1 wins,} \\ 1, & \text{if Team 2 wins.} \end{cases}$$

The goal is to compute the conditional expectation of P, given particular values of the predictor variables. This is estimated using the Nadaraya–Watson Gaussian kernel technique:

$$E[P_i \mid R_i, w_i, o_i] = \frac{\sum_{j=1}^{N} P_j N((R_j - R_i)/h_R) N((w_j - w_i)/h_w) N((o_j - o_i)/h_o)}{\sum_{j=1}^{N} N((R_j - R_i)/h_R) N((w_j - w_i)/h_w) N((o_j - o_i)/h_o)}.$$

In this setting, P_i is the binary game outcome and all other terms are as defined in the previous section.

2.3 Using a filter to impose monotonicity constraints

In this section, we develop a filter that allows us to impose monotonicity constraints in an optimal way. For example, the predicted score must be higher if there are more overs and/or wickets remaining, but an unconstrained kernel regression model admits violations of such logical constraints. We begin by developing the filter for a one-dimensional problem to develop intuition, followed by a more general representation.

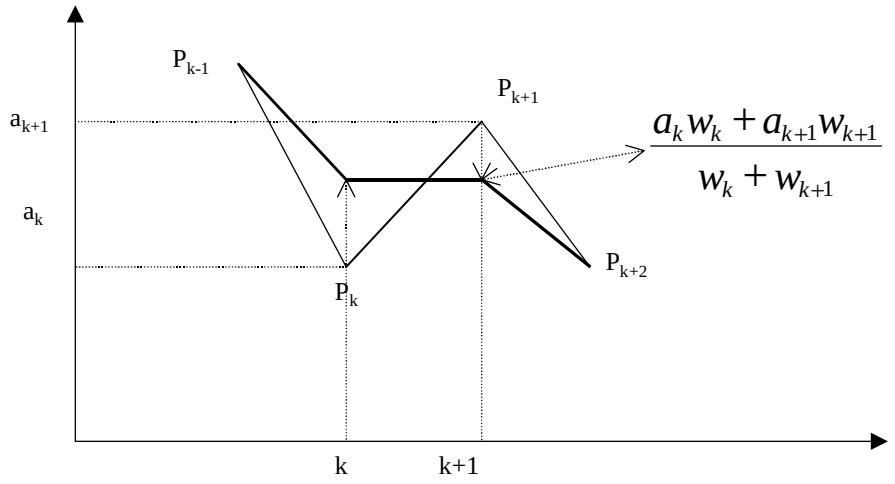


Figure 1: The filtering algorithm.

The one-dimensional problem

For given vectors **a** and **w** having length n, we want to perform the following operation:

$$\begin{aligned} \text{minimise:} \quad & \sum (b_i - a_i)^2 w_i \\ \text{subject to:} \quad & b_1 \geq b_2 \geq \cdots \geq b_n. \end{aligned}$$

In this setting, **a** represents the output from the kernel regression model (e.g. the predicted score, given the overs and wickets remaining) and **w** represents the weight given to that point (e.g. there are many observations of ten wickets and 49 overs remaining so that point on the grid receives high weight, there are no observations in our sample of two wickets and 49 overs remaining, so that point on the grid receives little weight). Geometrically, each element of vector **a** can be represented by a point in a 2-dimensional space. Each point is weighted by the corresponding element of vector **w**. Finally, **b** represents the output of the filter—the points in **a** are moved in a way that imposes the monotonicity but requires the minimum possible weighted-average movement in the elements of **a**. This geometric representation is illustrated in Figure 1.

Lemma 1 *If $a_k < a_{k+1}$ for some k, the optimal solution is such that $b_k = b_{k+1}$.*

Proof: Assume the optimal solution **b** gives $b_k > b_{k+1}$. We will prove that there exists a $\mathbf{b}'$ that satisfies the constraint condition, yet gives a smaller value for the objective function.

Consider $\mathbf{b}'$, where

$$\begin{aligned} b'_i &= b_i, \text{ for } i \neq k,\ k+1, \\ b'_k = b'_{k+1} &= \frac{b_k w_k + b_{k+1} w_{k+1}}{w_k + w_{k+1}}. \end{aligned}$$

As $\mathbf{w} > \mathbf{0}$ and $b_k > b_{k+1}$, therefore $b_k > b'_k = b'_{k+1} > b_{k+1}$. Since $b'_i = b_i$ $(i \neq k,\, k+1)$, it is clear that $\mathbf{b}'$ satisfies the constraint condition.

In terms of the objective function,

$$\sum (b'_i - a_i)^2 w_i - \sum (b_i - a_i)^2 w_i = A + B + C,$$

where

$$\begin{aligned}
A &= \sum_{i=1}^{k-1} (b'_i - a_i)^2 w_i - \sum_{i=1}^{k-1} (b_i - a_i)^2 w_i, \\
B &= w_k (b'_k - a_k)^2 + w_{k+1}(b'_{k+1} - a_{k+1})^2 - w_k (b_k - a_k)^2 - w_{k+1}(b_{k+1} - a_{k+1})^2 \\
&= w_k ((b'_k - a_k)^2 - (b_k - a_k)^2) + w_{k+1}((b'_{k+1} - a_{k+1})^2 - (b_{k+1} - a_{k+1})^2) \\
&= w_k (b'_k - b_k)(b'_k + b_k - 2a_k) + w_{k+1}(b'_{k+1} - b_{k+1})(b'_{k+1} + b_{k+1} - 2a_{k+1}), \\
C &= \sum_{i=k+1}^{n} (b'_i - a_i)^2 w_i - \sum_{i=k+1}^{n} (b_i - a_i)^2 w_i,
\end{aligned}$$

By construction, $b'_i = b_i$ for $i \neq k,\, k+1$. Therefore, it is easy to see that $A = C = 0$. From $b'_k = b'_{k+1} = (b_k w_k + b_{k+1} w_{k+1})/(w_k + w_{k+1})$, we have

$$w_k (b'_k - b_k) = -w_{k+1}(b'_{k+1} - b_{k+1}) = \frac{w_k w_{k+1}}{w_k + w_{k+1}} (b_{k+1} - b_k).$$

Substitute this in B. We have

$$B = \frac{w_k w_{k+1}}{w_k + w_{k+1}} (b_{k+1} - b_k)((b'_k + b_k - 2a_k) + w_{k+1}(b'_{k+1} - b_{k+1})(b'_{k+1} + b_{k+1} - 2a_{k+1})).$$

Because $b'_k = b'_{k+1}$, we have

$$B = \frac{w_k w_{k+1}}{w_k + w_{k+1}} (b_{k+1} - b_k)((b_k - b_{k+1}) + 2(a_{k+1} - a_k)).$$

Since $\mathbf{w} > \mathbf{0}$, $b_k > b_{k+1}$ and $a_k < a_{k+1}$, it follows that $B < 0$. This leads to

$$\sum (b'_i - a_i)^2 w_i - \sum (b_i - a_i)^2 w_i < 0,$$

or $\sum (b'_i - a_i)^2 w_i < \sum (b_i - a_i)^2 w_i$. That is, $\mathbf{b}'$ results in a smaller value for the objective function. This completes the proof.

Using Lemma 1, if $a_k < a_{k+1}$ we have $b_k = b_{k+1}$. Our objective function becomes

$$\begin{aligned}
&\left(\sum_{i \neq k,\, k+1} (b_i - a_i)^2 w_i \right) + w_k (b_k - a_k)^2 + w_{k+1} (b_k - a_{k+1})^2 \\
&= \left(\sum_{i \neq k,\, k+1} (b_i - a_i)^2 w_i \right) + (w_k + w_{k+1}) \left(b_k - \frac{w_k a_k + w_{k+1} a_{k+1}}{w_k + w_{k+1}} \right)^2 + D.
\end{aligned}$$

Here, D is a function of $\mathbf{a}$ and $\mathbf{w}$, and therefore can be ignored for optimisation purposes.

To this point, the algorithm for our problem should be obvious:

- If $a_1 \geq a_2 \geq \cdots \geq a_n$, the optimal solution is clearly $\mathbf{b} = \mathbf{a}$; the objective function is minimised at zero.

- If $a_k < a_{k+1}$ for any k, the optimisation problem becomes one where

$$\begin{aligned} \mathbf{a} &= \left(a_1, \ldots, a_{k-1}, \frac{w_k a_k + w_{k+1} a_{k+1}}{w_k + w_{k+1}}, a_{k+2}, \ldots, a_n\right), \\ \mathbf{w} &= (w_1, \ldots, w_{k-1}, w_k + w_{k+1}, w_{k+2}, \ldots, w_n), \\ b_{k+1} &= b_k. \end{aligned}$$

This is a simpler optimisation problem because the length of each of the vectors $\mathbf{a}$ and $\mathbf{w}$ has reduced from n to $n-1$. The same procedure can be repeated until a solution is found. Obviously, the algorithm converges because the procedure would be repeated no more than $n-1$ times.

Geometrically, the above algorithm can be understood as follows:

- *Step 1*: If the starting set of points is monotonically decreasing, no transformation is required.
- *Step 2*: If monotonicity is not observed at, say, P_k and P_{k+1}, we will move P_k up and P_{k+1} down to the weighted average of P_k and P_{k+1}; and go back to *Step 1*. However, in subsequent steps, the two points P_k and P_{k+1} will be treated as *one group of points*. That is, they will have the same value, move together, and be weighted by the total weight of the constituent points.

Each vector $\mathbf{a}$ can be seen as consisting of some *groups of points*. Points in each group have the same value, move together and are weighted by the total weight of the constituent points. Initially, vector $\mathbf{a}$ consists of n groups, each has one point—weighted by the corresponding element of $\mathbf{w}$. Our algorithm can be restated as follows:

- *Step 1*: If the starting groups decrease monotonically from left to right, they form the needed solution.
- *Step 2*: If monotonicity is not observed at two *adjacent* groups, these two groups will be merged to form one group, whose value will be the weighted average of the two constituent groups; and we go back to *Step 1*. Two groups are *adjacent* if there is no other point standing between them. For example, groups of points $\{P_1, P_2\}$ and $\{P_3, P_4, P_5\}$ are adjacent and they can be *merged* to form a group of $\{P_1, P_2, P_3, P_4, P_5\}$. This procedure can be repeated until a solution is found.

The two-dimensional problem

For given $n \times m$ matrices $\mathbf{a}$ and $\mathbf{w}$,

$$\begin{aligned} \text{minimise:} \quad & \sum (b_{ij} - a_{ij})^2 w_{ij} \\ \text{subject to:} \quad & b_{i_1 i_2} \geq b_{j_1 j_2} \text{ for every } i_1 \geq j_1,\ i_2 \geq j_2. \end{aligned} \tag{1}$$

Geometrically, each element of matrix $\mathbf{a}$ can be represented by a point in a 3-dimensional space. Each point is weighted by the corresponding element of matrix $\mathbf{w}$. Again, our problem is to find a new set of points, closest to the given set, that forms a monotonic decreasing trend along the two dimensions of matrix $\mathbf{a}$.

Analogous to the one-dimensional problem above, each matrix $\mathbf{a}$ can be seen as consisting of some groups of points. Points in each group have the same value, move together and are weighted by the total weight of the constituent points. It should be noted however that while in the one-dimensional case, each group of points will correspond to a vector, each two-dimensional group of points may not necessarily represent a rectangular or square matrix. For example, points $\{P_{11}, P_{12}, P_{21}\}$ can form a group. Given the form each group of points can take, the concept of *adjacency* in the two-dimensional setting also needs further clarification.

Two groups of points G_1 and G_2 are said to be *adjacent* if there exists no other group G_3 such that: $G_1 \geq G_3 \geq G_2$ or $G_2 \geq G_3 \geq G_1$ by the operation of inequality (1). For example, if G_1 consists of $\{P_{11}, P_{21}\}$ and G_2 consists of $\{P_{22}\}$, they are not adjacent because there exists G_3 consisting of $\{P_{12}\}$ and by the operation of (1):

- value of $P_{22}(b_{22}) \geq$ value of $P_{12}(b_{12})$, that is, $G_2 \geq G_3$; and
- value of $P_{12}(b_{12}) \geq$ value of $P_{11}(b_{11})$, that is, $G_3 \geq G_1$.

On the other hand, if $G_1 = \{P_{11}, P_{21}\}$ and $G_2 = \{P_{12}\}$, they are *adjacent*.

Two adjacent groups can be merged in a similar way as in the one-dimensional case. For example, $G_1 = \{P_{11}, P_{21}\}$ and $G_2 = \{P_{12}\}$ can be merged to form $G_3 = \{P_{11}, P_{21}, P_{12}\}$.

So in the two-dimensional case, the algorithm we use is:

Initially, matrix **a** consists of $n \times m$ groups; each group has one point, weighted by the corresponding element of matrix **w**.

- *Step 1*: If the starting groups decrease monotonically along the two dimensions of matrix **a**, they form the needed solution.
- *Step 2*: If the constraint condition (1) is not satisfied by any two *adjacent* groups, these two groups will be merged to form one group, whose value will be the weighted average of the two constituent groups; and we go back to *Step 1*. This procedure can be repeated until a solution is found.

It is important to note that Lemma 1 does not hold if **a** and **w** are matrices rather than vectors. Therefore, while our algorithm always gives the exact solution in the one-dimensional case, it may not always give the optimal solution in multi-dimensional cases.

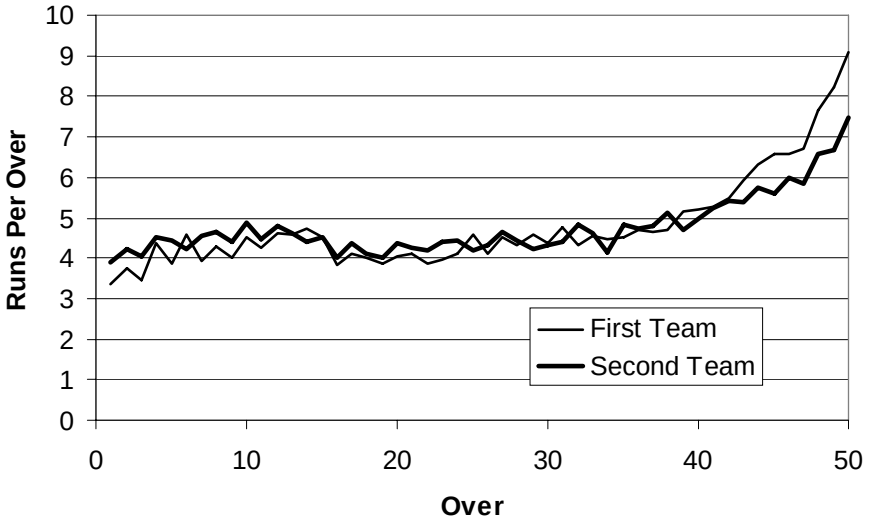


Figure 2: Average runs scored per over.

The advantage of our algorithm is, however, its relative simplicity and its generalisability to cases of multi-dimensional problems. Though there exist several methods (such as Lagrange multiplier) that can deliver exact solutions, the computational burden often renders such approaches intractable. The Lagrange multiplier method, for example, requires a number of calculations that increases exponentially with the number of points in our data matrix **a**. That is, for a 50×50 matrix, the Lagrange method requires at least 2^{2500} calculations. Our algorithm, on the other hand, delivers a reasonable solution within less than five seconds. In addition, if there are not many violations of the constraint conditions (1) in the data matrix **a**, our algorithm tends to return solutions that are intuitively and visually appealing.

The three-dimensional problem

For given $n \times m \times k$ matrices $\mathbf{a}$ and $\mathbf{w}$,

$$\begin{aligned} \text{minimise:} \quad & \sum (b_{ijh} - a_{ijh})^2 w_{ijh} \\ \text{subject to:} \quad & b_{i_1 i_2 i_3} \geq b_{j_1 j_2 j_3} \text{ for every } i_1 \geq j_1,\ i_2 \geq j_2,\ i_3 \geq j_3. \end{aligned} \tag{2}$$

Geometrically, each element of matrix $\mathbf{a}$ can be represented by a point in a 4-dimensional space. Each point is weighted by the corresponding element of matrix $\mathbf{w}$. Again, our problem is to find a new set of points, closest to the given set, that forms a monotonic decreasing trend along the three dimensions of matrix $\mathbf{a}$.

Again, each matrix $\mathbf{a}$ can be seen as consisting of some groups of points. Points in each group have the same value, move together and are weighted by the total weight of the constituent points. Initially, matrix $\mathbf{a}$ consists of $n \times m \times k$ groups; each group has one point, weighted by the corresponding element of matrix $\mathbf{w}$.

The algorithm in this setting follows in an analogous way to the two-dimensional case above. The extension to higher dimensions then follows in the same way.

3 Data

Data on all one-day internationals from November 1998 to August 2001 were collected from the *Statistical Archive* section of the cricinfo.com web site. We collected data in the form of a *Run-Rate Comparison*. These data include the score (wickets and runs) of the team that is batting. The data are reported as at the end of every over of the match. In the vast majority of games, both teams have 50 overs available to them when they bat. We exclude games that are interrupted by rain to the extent that one or both teams have less than 50 overs available. Only games involving national teams are included in our data set. Games involving "A" teams or provincial teams are excluded, as are games involving non-Test-playing nations such as Kenya. The result is a sample of 317 games. Table 1 contains some summary statistics.

Mean	233
Standard deviation	53
Median	236
Minimum	68
Maximum	376

Table 1: Summary statistics: score of first-batting team.

Table 1 indicates that there is a wide range of outcomes for the score of the first-batting team. Within our sample, the second-batting team won 51.42% of the matches. Another interesting feature of the data is the pattern in which runs are scored. Figure 2 shows the average runs per over scored by each team. To compute these averages, we use all available observations. For the first over, we have the full sample of 317 games. In our sample, there are ten instances of the first-batting team being bowled out within the first forty overs, so we have only 307 observations for over number forty, and so on. The figure shows that the second-batting team begins the innings at a faster rate, on average than the first-batting team. The average score of the second-batting team is lower in the later overs for two reasons. First, there are a number of games in which the second team is well-placed by the 40th over and scores at a low rate as they cruise to victory. Second, in the cases where the second team requires a very high rate to win the game in the later overs, there are many instances of the team being bowled out and hence dropping out of the sample for the remaining overs.

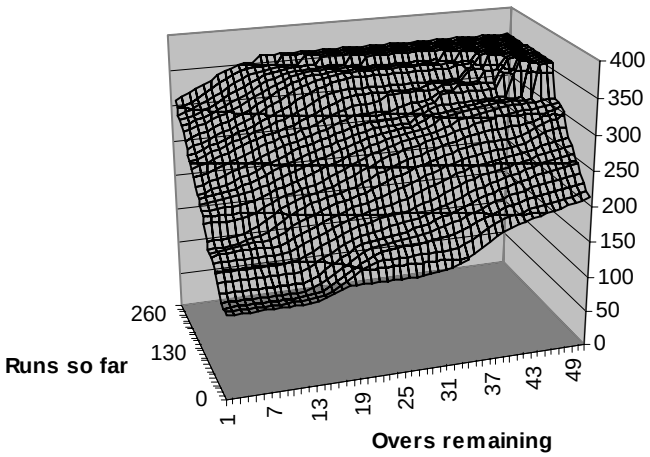


Figure 3: Predicted total scores—8 wickets remaining.

4 Results

4.1 Predicting the score of the team that bats first

The output of our filtered kernel regression procedure is a series of surfaces of predicted total scores. Each surface relates to a particular number of wickets remaining. Within each surface, the predicted total score depends upon the runs scored so far and the number of overs remaining. In each case, only part of the surface is relevant. For example, the surface for eight wickets remaining appears in Figure 3.

Clearly, there are very few observations in our data set of eight wickets remaining with 49 or more overs remaining. Almost as rare is eight wickets remaining in the final few overs of the innings. Thus, the extreme left and right areas of the surface above are based on scant data and are unlikely to be required in practice. Moreover, the back right corner of this figure can also be ignored as it relates to observations where there are large amounts of runs scored from very few overs. Symmetrically, the front left corner is equally irrelevant as it relates to extremely small scores acquired over many overs. That is, it is the middle portion of the surface that relates to observations that are likely to be observed in practice.

Figure 4 demonstrates that, point-by-point, predicted scores are lower when there are fewer wickets remaining. The difference becomes less when there are fewer overs remaining (near the left-hand edge of the surface) as wickets in hand are less valuable in that instance.

More intuition can be obtained by contrasting the predictions of total scores across regions of these surfaces that are more heavily populated with data. Figure 5, for example, shows the predictions of total scores when there are ten overs remaining. Obviously, more runs scored so far and more wickets remaining suggest a higher total score.

A similar picture emerges when examining total scores predicted when there are 20 overs remaining, as illustrated in Figure 6. In this case, we compare the model predictions (which are based exclusively on actual data) with a common rule of thumb that predicts that the total score will be twice the score after 30 overs. This rule of thumb is also marked in Figure 6. Our procedure suggests that total scores are less sensitive to the number of runs scored so far than the rule of thumb suggests.

4.2 Predicting the probability that the second-batting team wins the match

The output of our filtered kernel regression procedure for the probability of the second team winning the match is a series of surfaces of probabilities. Each surface relates to a particular number of wickets remaining. Within each surface, the probability of winning depends upon the run-rate required and the

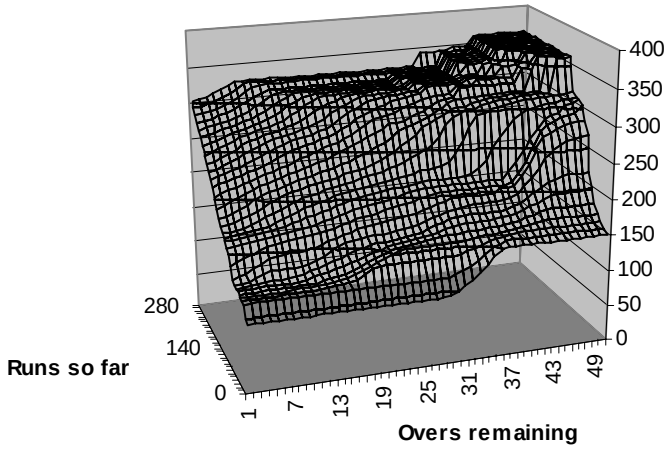


Figure 4: Predicted total scores—5 wickets remaining.

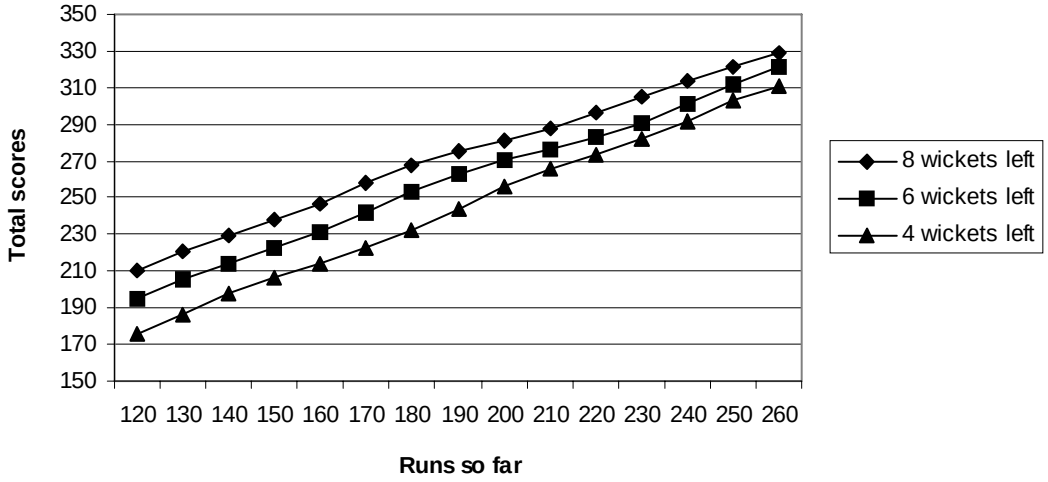


Figure 5: Predicted total scores—10 overs remaining.

number of overs remaining in the match. For example, the surface for eight wickets remaining appears in Figure 7.

More intuition can be obtained by contrasting the probabilities across various regions of these surfaces. Figure 8, for example, shows the probabilities of the second team winning when there are ten overs remaining. Obviously, a lower required run-rate and more wickets remaining suggests a higher probability of victory. But this relationship is not linear or uniform in either dimension. Wickets in hand are less important when the required run-rate is very low, but very important when the required run-rate is between 6 and 8.5 runs per over. Within this range, there is almost 80% chance of winning if eight wickets are in hand, and less than 20% chance if only four wickets remain.

A similar picture emerges when examining probabilities when there are 30 overs remaining, as illustrated in Figure 9.

Figure 10 illustrates the success rate of the model at different stages during the second innings of

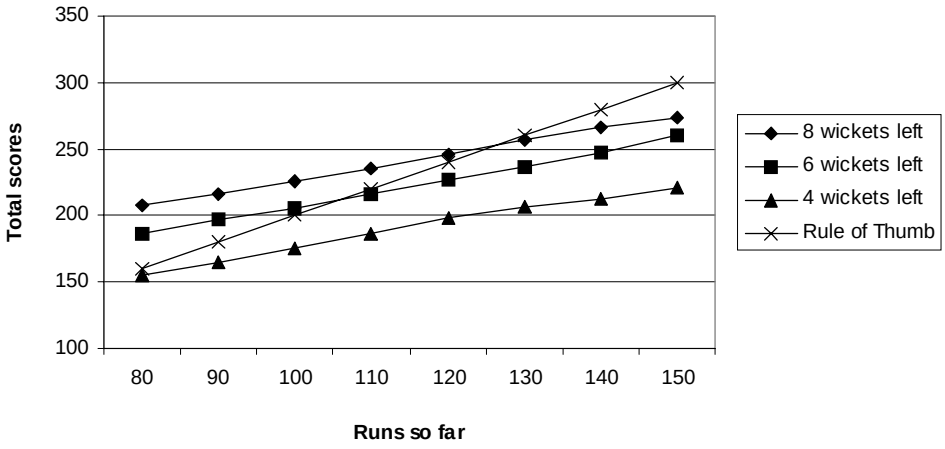


Figure 6: Predicted total scores—20 overs remaining.

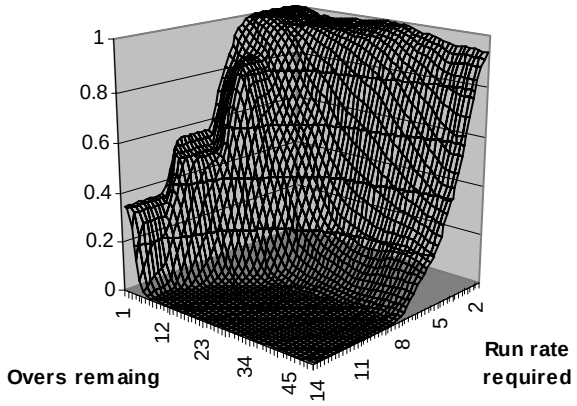


Figure 7: Probability of second team winning—8 wickets remaining.

the match. The model is able to predict outcomes with better than 70% success at the beginning of the second innings. The success rate is over 85% midway through the innings. In all cases, the success rates are based on matches that are incomplete. Many matches are complete before the 50th over, either due to the second team being bowled out or having already passed the first team's score. Thus, the success rate for the model in the 50th over is based on a relatively smaller number of observations.

4.3 Application to actual matches

To further illustrate the operation of the model, we apply it to two actual matches from the 2001–2002 Australian season.

Our first example is the match between New Zealand and Australia at the Melbourne Cricket Ground (MCG) on 11 January 2002. In this match, New Zealand batted first, slumped to 7/96 in the 28th over and recovered to post a total of 199. In Figure 11, we plot the model's prediction of the total score

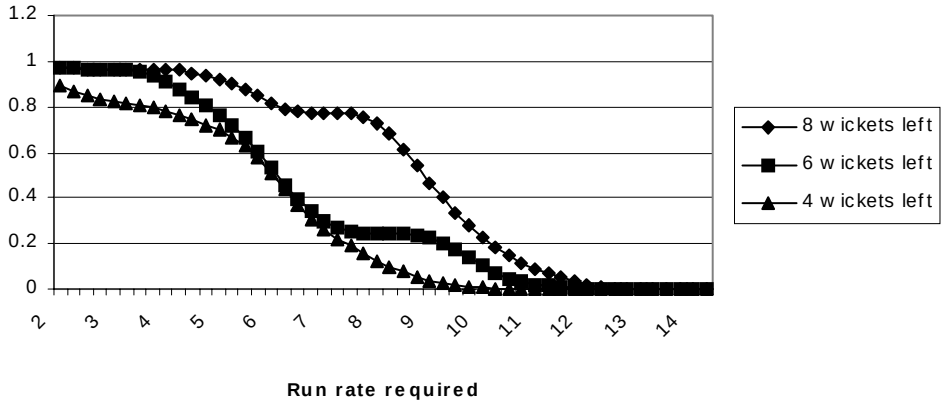


Figure 8: Probability of second team winning—10 overs remaining.

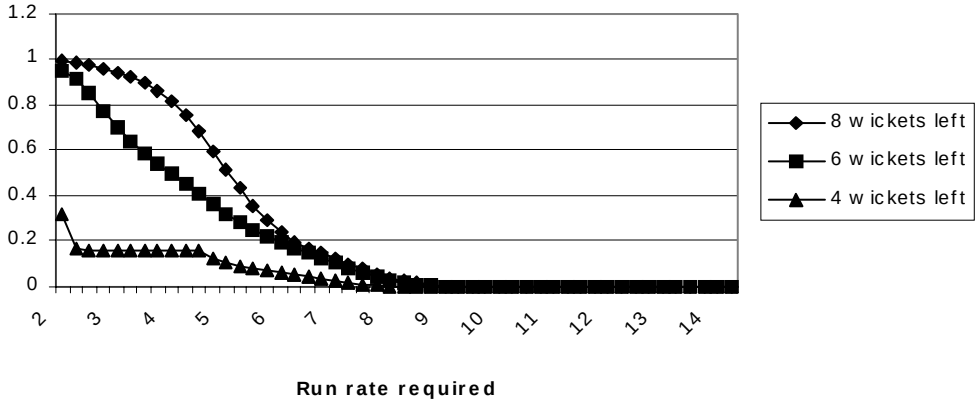


Figure 9: Probability of second team winning—30 overs remaining.

throughout the innings. At the outset, the predicted score is around 230, which is the unconditional mean total score of the matches in our data set. The predicted score falls as wickets are lost. At the end of the 30th over, the score is 7/103, primarily due to the fact that seven wickets have already been lost. This suggests that in matches in which teams are in this sort of position, the average total posted is in the order of 160. In this particular match, New Zealand recovered remarkably, performing much better than the average team in this situation. The predicted total therefore rises over the last 20 overs of this innings.

Figure 12 illustrates the model's assessment of the likelihood of the second team (Australia) winning the match. Given the low target and the excellent start made by the Australian team, the probability of success is over 90% for the first 20 overs. After this point, the large number of wickets lost causes a sharp reduction in this win probability. By the 30th over, the score is, 5/137, and the probability of a win is less than 10%.

The second match we analyse is also between New Zealand and Australia. This game was played at the MCG on 29 January 2002. Again, New Zealand batted first and posted a total of 245. Figure 13 plots the models prediction of the total score throughout the innings. This innings is more "standard"

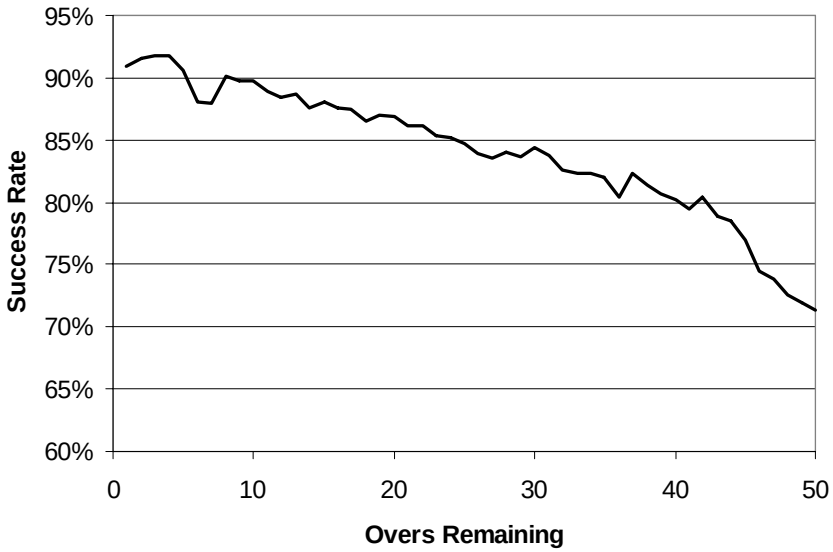
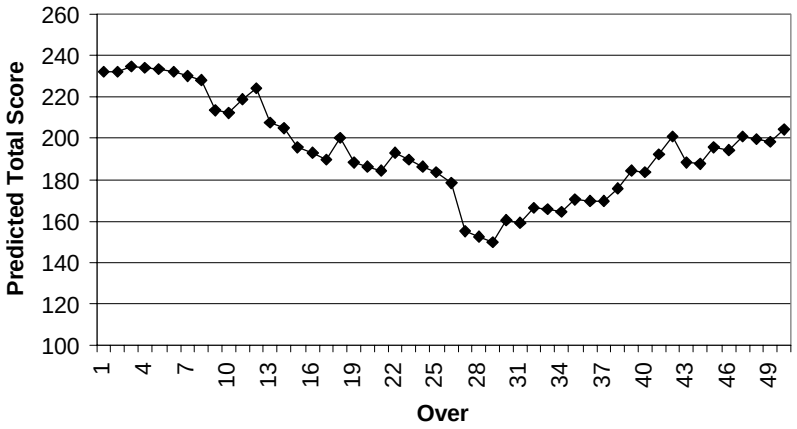


Figure 10: Success rate of model predictions.

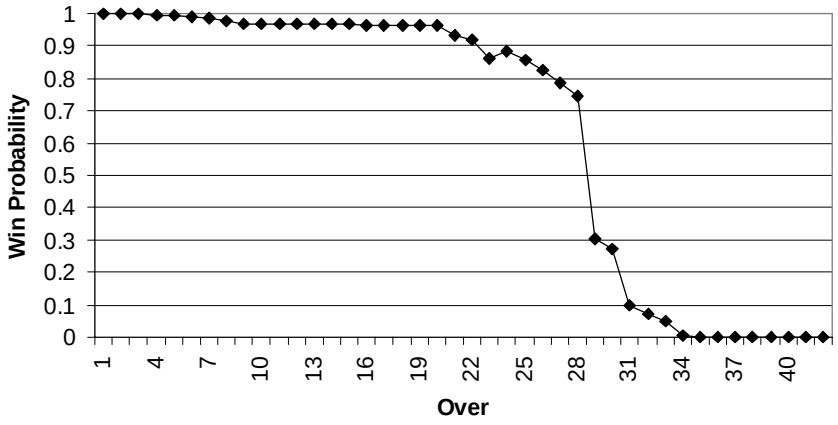


Over	5	10	15	20	25	30	35	40	45	50
Score	1/20	2/23	4/54	5/71	5/89	7/103	7/121	7/148	8/171	8/199

Figure 11: Australia v New Zealand, 11 January 2002, MCG, first innings (New Zealand)—model predictions of total score.

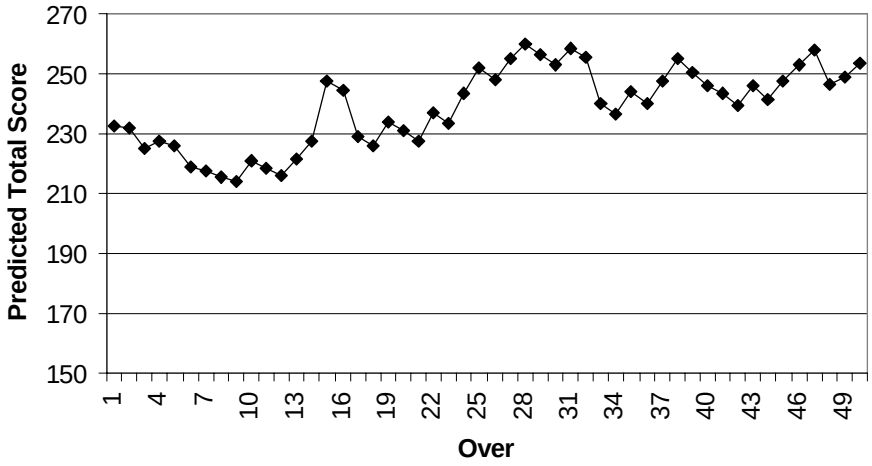
in that there is no dramatic collapse or any spectacular recovery. Thus, the predicted score is relatively flat throughout the innings.

Figure 14 plots the probability of Australia winning the match. This win probability is initially a little below 50%, consistent with the target being somewhat above average. As wickets fall, however, the likelihood of an Australian victory falls. By the 25th over, six wickets had fallen and the win probability is less than 10%. By the 43rd over, however, Australia had reached 7/195 and required 7.28 per over for seven overs with three wickets in hand. At this point, the model assesses a better than 20% chance



Over	5	10	15	20	25	30	35	40	45	50
Score	1/44	2/71	2/95	3/104	4/122	5/137	8/151	9/169		

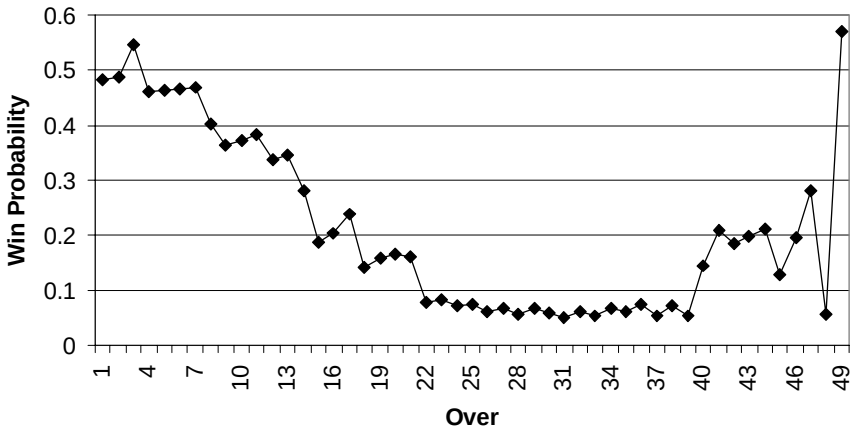
Figure 12: Australia v New Zealand, 11 January 2002, MCG, second innings (Australia)—probability of second team victory.



Over	5	10	15	20	25	30	35	40	45	50
Score	1/13	2/35	2/70	3/87	3/110	3/135	4/152	4/178	5/202	8/245

Figure 13: Australia v New Zealand, 29 January 2002, MCG, first innings (New Zealand)—model predictions of total score.

of an Australian victory. By the last over, Australia required only six runs, with two wickets remaining, and the win probability is in the order of 60%.



Over	5	10	15	20	25	30	35	40	45	50
Score	1/28	2/42	4/53	5/76	6/90	6/108	6/133	7/172	7/206	

Figure 14: Australia v New Zealand, 29 January 2002, MCG, second innings (Australia)—probability of second team victory.

4.4 Comparison with the Duckworth–Lewis method

Duckworth and Lewis (1998) proposed a method for adjusting targets in rain-interrupted one-day matches.[2] Their technique is based on a mathematical description of the runs that are expected to be scored, conditional on the overs and wickets remaining. In this sense, their framework provides a prediction of the total score and can therefore be compared with our method. The Duckworth–Lewis approach models the runs to be scored as an exponential function:

$$E[R_i - r_i \mid w_i, o_i] = Z_0(w_i)(1 - \exp(-b(w_i)o_i)).$$

In this setting, $R_i - r_i$ represents the runs yet to be scored, $Z_0(w_i)$ represents the asymptotic total runs to be scored (as the number of overs becomes large) conditional on w_i wickets remaining, and $b(w_i)$ is an exponential decay factor. Figure 15 illustrates the Duckworth–Lewis forecasts of runs to be scored.

Our kernel regression method provides conditional forecasts of the total score that will be scored and is therefore comparable to this aspect of the Duckworth–Lewis approach. There are, however, some key differences between the two approaches. First, our approach is data-based. We make no assumptions about the functional form of any relationship between any of the variables. The Duckworth–Lewis approach is based on a specific mathematical relationship, albeit one that is calibrated to the data. Second, in our kernel regression forecasts, we condition on runs scored so far. Consider two cases: one in which a team has reached 5/120 in the 40th over, and another in which a team has reached 5/200 in the 40th over. The Duckworth–Lewis framework would forecast that 63 more runs will be scored in both cases, since the amount of resources is identical—ten overs and five wickets. Our model predicts that an additional 51 runs will be scored in the first case and 72 in the second. This seems to make intuitive sense in that a higher number of runs scored so far is likely to be an indication of good batting conditions available. In our data set, there are six matches in which teams had lost five wickets at the 40th over and had scored between 115 and 125 runs. On average these teams scored an additional 48

[2] The Duckworth–Lewis method has been published in the form of a set of tables and as a simple computer program. It is intended to adopt the same practice with the results of this paper, after some further refinement of the model and an extension of the data set.

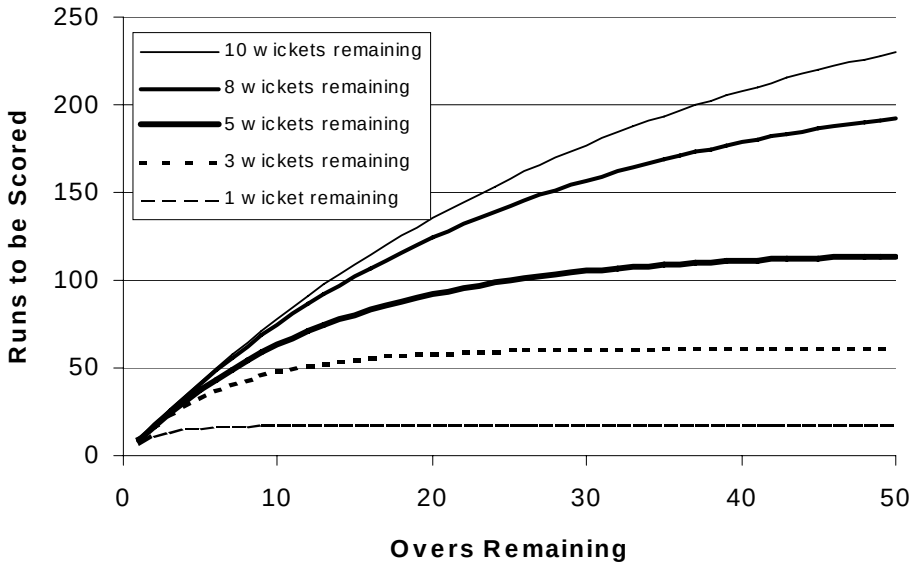


Figure 15: Runs to be scored according to Duckworth–Lewis method.

runs over the remaining ten overs. Similarly, there are seven matches in which teams had scored between 195 and 205 and had lost five wickets at the 40th over. These teams scored an additional 70 runs over the final ten overs, on average. Of course, it is not surprising that our model calibrates more closely to the data, given that our approach is entirely data-driven. Moreover, in designing a system to be used for target-adjustment in rain-affected matches, Duckworth and Lewis are constrained to produce a simple and transparent method.

Therefore, to make the two approaches more comparable, we apply our kernel method to the task of estimating the runs yet to be scored as a function only of wickets and overs remaining. The results are shown in Figure 16.

In this figure, we do not display estimates over regions for which we have no close observations in our data set. For example, we have no observations of one wicket and 49 overs remaining, nor any observation close to this, so we remove this from the figure.

To gauge the differences between the Duckworth–Lewis approach and our data-driven kernel approach, we plot the difference between the two procedures in Figure 17.

Figure 17 displays the forecast total from our kernel method minus the forecast from the Duckworth-Lewis method. When there are ten wickets remaining, the Duckworth–Lewis method underpredicts relative to our method when there are 50 overs remaining. This is caused by differences in the data sets. Our data consists of one-day internationals, in which the average score of the first-batting team is around 233. The Duckworth–Lewis data set includes lower-scoring first-class matches and the average total score in their data set is 225. The Duckworth–Lewis method overpredicts relative to our method between the 25th and 42nd overs. This is caused by the greater curvature of the Duckworth–Lewis function relative to ours in the figures above. Since it is extremely unlikely that a team would still have ten wickets intact near the end of the innings, comparisons when there are few overs remaining are difficult to interpret. A similar shaped pattern exists when there are eight wickets remaining, although in this case the Duckworth–Lewis method almost uniformly underpredicts relative to our kernel model. Again, this is probably attributable to differences in data sets. First-class teams usually have less batting depth than international teams, so the loss of two wickets has a more detrimental effect in the Duckworth–Lewis sample than in ours.

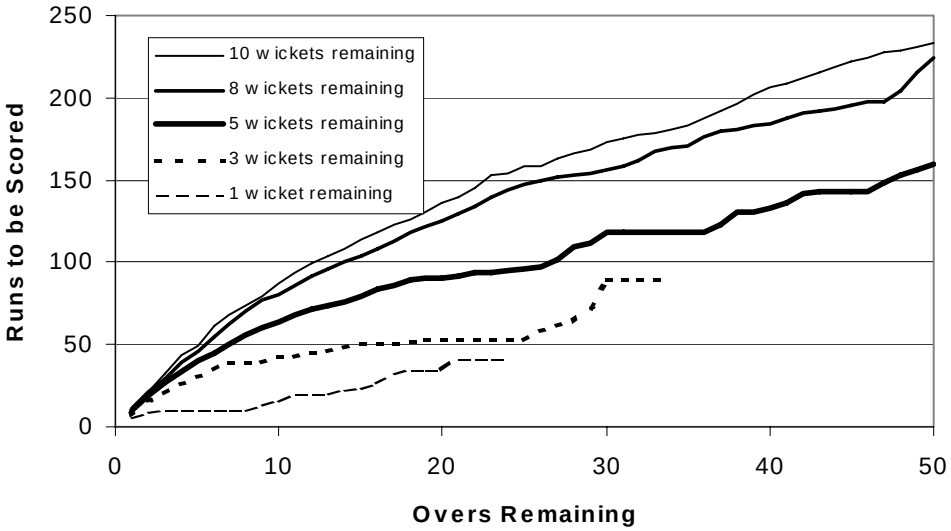


Figure 16: Runs to be scored according to kernel method.

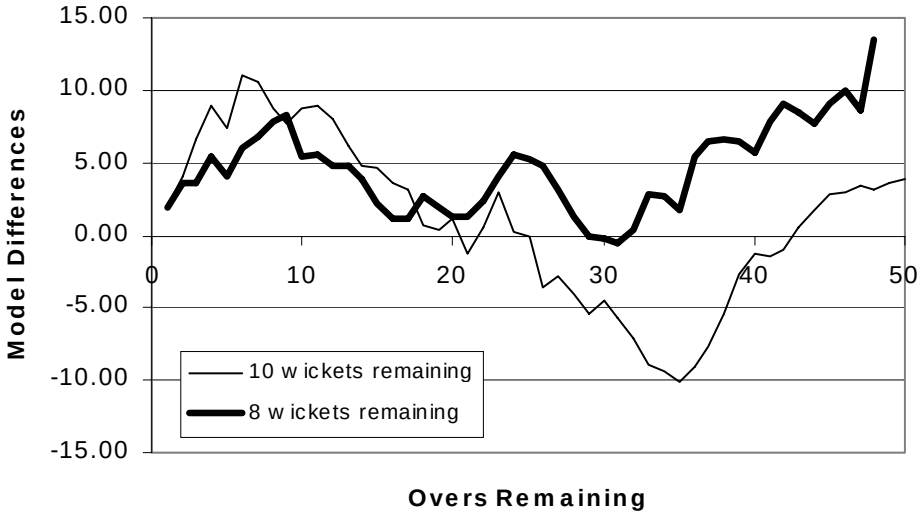


Figure 17: Runs to be scored according to kernel method.

5 Conclusions

Our Gaussian kernel non-parametric regression model provides a framework for forecasting the total score of the first-batting team and the probability of the second team winning the match. These are conditional estimates, based on information about the current score.

Future extensions of the model would include conditioning on a larger information set. For example, average scores in India and Pakistan are higher than those in the UK, so location could also be included in the conditioning information. Even within geographical regions, some grounds consistently produce higher scores than others.

We could also condition on the recent results of the two teams to incorporate any persistence in results.

Finally, we could condition on the score *and* further game details. Consider two cases, that both involve scores of 8/210 after 45 overs. In one case, the two not-out batsman are low-order players who have just begun batting. In the other case, one of the batsmen is a top-order player who has already scored 75 runs. Presumably, we would expect a higher score in the second case. This kind of additional conditioning information is likely to further improve the performance of the model.

References

F. C. Duckworth and A. J. Lewis (1998), "A fair method for resetting the target in interrupted one-day cricket matches," *J. Oper. Res. Soc.*, **49**, 220–227.

E. A. Nadaraya (1965), "On nonparametric estimates of density functions and regression curves", *Theory Probab. Appl.*, **10**, 186–190.

D. W. Scott (1979), "On optimal and data-based histograms", *Biometrika*, **66**, 605–610.

B. W. Silverman (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

THE FUNDAMENTAL NATURE OF DIFFERENTIAL MALE/FEMALE WORLD AND OLYMPIC WINNING PERFORMANCES, SPORTS AND RATING SYSTEMS

Ray Stefani
Department of Electrical Engineering
California State University
Long Beach, USA
raystefani@aol.com

Abstract

Some fundamental, and hopefully interesting, results are presented in three areas of sport science: male/female differential performance, sport categorisation and rating systems. The laws of physics are used to convert differential male/female running, jumping and swimming winning performances into differential power output as reflected in those performances. In turn, those power differences disappear when corrected for the fact that men and women burn the same number of calories per unit of lean body mass. That is, those performance-based power differences are consistent with differences in lean body mass. It is shown that there are only three types of sports, depending on the interaction of the competitors: contact sports with direct contact where each competitor tries to control the other (as for boxing, judo), object sports with indirect contact where each competitor tries to control a single object (as for basketball, soccer) and independent sports where each competitor performs independently and needs only to control the athlete's own performance (as for running, swimming, shooting and golf). It is suggested that there are three types of rating systems each of which align with one specific audience. The general non-technical audience trusts subjective systems (polls), those seeding tournaments trust accumulative systems where points can only increase due to entering many tournaments while performing well and the technical audience trusts adjustive rating systems in which statistics and theory create ratings which can increase or decrease in predictor–corrector fashion.

1 Introduction

In the 30 years in which I have observed and analysed sports statistics, a number of general conclusions have presented themselves. These conclusions and the ensuing insights are shared in this paper. Perhaps the most significant conclusions relate to the very nature of male versus female athletic achievement. Much has been written on that subject. In one infamous paper [21], female runners were projected to defeat male runners midway through the 21st century, based on the then rate of improvement, which was not sustained, as we shall soon see. In an earlier conference paper [8], I reported on the differential performance, based on time, of male and female Olympic champions in running, jumping and swimming events, leaving open the question of the origin of those differences. An incentive for one part of this paper comes from a study of the power output of Olympic rowing champions [12], using laws of hydrodynamics and data on boats and athletes.

In Section 2, the laws of physics are applied to running (viscous friction), jumping (ballistic trajectory), pole vault (transfer of kinetic energy to potential energy) and swimming (hydrodynamics), so as

to estimate power output as reflected in winning times. Those performance-based differences are compared with physiological differences in lean body mass, given that male and female burn equal calories per unit of lean body mass. Section 3 covers the very nature of sport. The incentive for this section came from categorising [10] 80 major sports, recognised by the International Olympic Committee or cited in widely used sports references. Section 4 covers one aspect of rating systems that is generally overlooked: the audience that trusts that type of rating system. While a person creating a rating system may demonstrate very good properties of that system, an important question is "What audience will accept that type of system?".

2 Fundamental nature of differential male/female world and Olympic winning performances

2.1 Measures of differential performance

The goal is to establish an understanding of the very nature of differences in male and female athletic performance. To delimit the study, elite athletes are considered, in particular Olympic and world champions, about whom much is said and to whom athletes look for inspiration. An advantage of this delimiting is that performance data is unambiguous and readily available via any almanac or record book. A method of comparison in needed. Percent differences are used here. If t_W is the time for the winning woman while t_M is the time for the winning man, the percent difference ($\%D$) in time (T) is

$$\%DT = 100\left(1 - \frac{t_M}{t_W}\right). \tag{1}$$

Instead, suppose velocities are to be used, where v_W is the average velocity of the winning woman while v_M is the average velocity of the winning man. The percent difference in velocity (V) is

$$\%DV = 100\left(1 - \frac{v_W}{v_M}\right). \tag{2}$$

In each case, the larger quantity appears as the base. These equations seem to leave a choice (and invoke a discussion) as to preference. In fact, no choice is needed. Over a distance d with elapsed time t, average velocity is d/t. Thus, (2) can be equated to

$$\%DV = 100\left(1 - \frac{d/t_W}{d/t_M}\right) = 100\left(1 - \frac{t_M}{t_W}\right) = \%DT. \tag{3}$$

It is useful to note that $\%DT$ and $\%DV$ are the same. Another useful measure of differential performance is the fraction of the race remaining to be run by the winning woman, when the winning man crosses the finish line (understanding, of course, that the races are not run at the same time). Let d_R be the distance remaining (R) to be run. Using average velocity, and noting that the woman must run for $t_W - t_M$ seconds after the man finishes,

$$\%DR = 100\frac{d_R}{d} = 100\frac{(t_W - t_M)(d/t_W)}{d} = 100\left(1 - \frac{t_M}{t_W}\right) = \%DT = \%DV. \tag{4}$$

Thus, all three measures are exactly the same. If, for a given race, we obtain a 10% difference, that simultaneously means that there is a 10% difference in time, a 10% difference in velocity and 10% of the race remains for the woman to run. If the same measure occurs for a longer race, then in proportion to the length of the race, the woman would remain equally as far behind.

Clearly (1) to (4) are appropriate for running and swimming. What about events determined by distance or height (as for the long jump and pole vault)? For computational convenience (which was far more important than mathematical consistency when this analysis was first employed 30 years ago

in the time of punched cards and mainframe computers), I modify (1) with a sign change, realising that division is by the lesser value, in contrast with (1) to (4) where division is by the larger value. For the $\%D$ in distance (Di),

$$\%DDi = 100\left(\frac{d_M}{d_W} - 1\right). \tag{5}$$

Throwing events such as the shot put, discus and javelin are not considered. There are a number of reasons for this omission. First, the weight thrown by female competitors is about half of the weight thrown by males. Second, it is difficult to accurately gauge the aerodynamic effect on the differential weights, which becomes important if power differential is to be estimated. Third, the weight of both men's and women's javelins have been redistributed rendering older results meaningless.

A significant contribution of this paper is the estimation of the percent difference in power (the product of force and velocity) based on each winning performance, and then its interpretation using physiological differences between men and women. Power is measured in watts, representing the rate at which energy is consumed. Power is analogous to the rate that an automobile travels in kilometres per hour, while energy, often measured in kilocalories, is analogous to the distance traveled in kilometres. Most people are familiar with light bulb wattage. An average person can produce about 100 W of output power applied to a step machine or bicycle. The human body is a fairly inefficient machine, in that about four calories of internal energy must be burned for every calorie of useful output. That 100 W of output power requires the burning of about 350 k calories per hour by the athlete's body. An elite athlete can produce more than 400 W of output power and burn over 1400 k calories per hour internally.

The $\%D$ in power (P) is calculated from the power ratio

$$\%DP = 100\left(1 - \frac{p_W}{p_M}\right). \tag{6}$$

In the rest of this paper, mass and weight will be used interchangeably, although strictly speaking, mass is measured in kilograms while weight (the product of mass times the acceleration of gravity) is measured in newtons. When you stand on a scale, the acceleration of gravity pulls down on your mass, deflecting the scale, which is normally calibrated to read kilograms instead of newtons. Certainly, weight is proportional to mass, so in that context the terms will be used interchangeably.

2.2 Observed percent differences in running, jumping and swimming

Data were collected for the last eight Olympic games (1972–2000), the last eight IAAF track and field (athletics) world championships (1983–2001) and the last eight FINA swimming world championships (1973–1998). Data for $\%DT$ are shown in Table 1 for eight running events and data for $\%DDi$ are shown for four jumping events, averaged over the 16 world and Olympic competitions. The third column will be discussed shortly.

The average $\%D$ in the second column of Table 1 is just over 10% for the running events and about twice that for the jumping and pole vault events. The pole vault has only been contested three times (once in Olympic and twice in IAAF competition), hence the $\%D$ may drop as happened in the triple jump when that event was introduced and then mastered by female athletes. For timed events, $\%DT$ increases from 100 m to 10000 m then drops for the marathon.

Table 2 shows $\%DT$ data for 12 swimming events averaged over the 16 world and Olympic competitions.

For swimming, average $\%D$ in the second column of Table 2 is just under 10%. There a downward trend as distance increases in all events except breaststroke. In the next section, the laws of physics are employed to convert $\%D$ in time and distance to $\%D$ in power to create column three in both tables.

Running events	Percent difference in time ($\%DT$)	Percent difference in power based on $\%DT$
100 m	8.53	25.45
200 m	9.30	26.70
400 m	10.44	28.53
800 m	10.57	28.74
1500 m	10.56	28.72
5000 m	11.25	29.82
10000 m	11.47	30.17
Marathon	10.14	28.05
Average	10.28	28.27
Jumping events	**Percent difference in distance ($\%DDi$)**	**Percent difference in power based on $\%DDi$**
High jump	17.22	26.66
Triple jump	18.81	27.16
Long jump	20.82	27.76
Pole vault	28.81	31.08
Average	21.41	28.16

Table 1: Running and jumping events—eight Olympic Games (1972–2000), eight IAAF World Championships (1983–2001).

Swimming	Percent difference in time ($\%DT$)	Percent difference in power based on $\%DT$
50 m freestyle	11.29	32.14
100 m freestyle	10.35	29.96
200 m freestyle	8.84	26.37
400 m freestyle	7.98	24.26
100 m backstroke	10.23	29.68
200 m backstroke	8.69	26.00
100 m breaststroke	10.34	29.94
200 m breaststroke	10.39	30.06
100 m butterfly	10.14	29.47
200 m butterfly	8.97	26.68
200 m individual medley	9.08	26.95
400 m individual medley	8.02	24.36
Average	9.52	27.99

Table 2: Swimming events—eight Olympic Games (1972–2000), eight FINA World Championships (1973–1998).

2.3 Using physics to estimate power differences

Some approximations

In the use of physical laws to follow, approximations are needed for the relative surface area of male and female swimmers that are wetted (in contact with the flow of water) and for the cross-sectional area of an athlete, needed to estimate the foot area of a runner that is in contact with the ground during each running cycle. The ratio of the wetted surface area for swimmers can be estimated similarly to

finding the relative wetted surface area of boats; that is, by taking a ratio of volume raised to the exponent of 2/3. Volume may have units of m^3 while area may have units of m^2. Certainly, dimensional considerations suggest applying the exponent 2/3 to volume when approximating relative surface area, but the accuracy of that approximation depends on the actual shape. This approximation works well for a slender cylinder, which approximates the shape of the human body and of many boats. For example, suppose one cylinder has radius of 5 in and length of 72 in. The volume V_1 is $5655\,\text{in}^3$ and the area A_1 is $2419\,\text{in}^2$. Suppose a second cylinder has radius 4.45 in and length 60 in. The volume V_2 is $3733\,\text{in}^3$ and the area A_2 is $1802\,\text{in}^2$. Now, the actual ratio of areas is 2419/1802 or 1.34. If we apply the exponent 2/3 to the volume ratio, we get $(5655/3733)^{2/3}$ or 1.32. Clearly, that is a good approximation.

Regarding the physics of running, an approximation is needed for the cross-sectional area of an athlete to estimate the foot area in contact with the ground. Using the slender cylinder approximation, cross-sectional area may be approximated by taking the square root of volume, that is, the half-power of volume. For the first cylinder, the cross-sectional area is $78.5\,\text{in}^2$ whereas $V_1^{1/2}$ is $75.2\,\text{in}^2$, an accurate approximation. For the second cylinder the cross-sectional area is $62.2\,\text{in}^2$ compared to $V_2^{1/2}$ or $61.1\,\text{in}^2$, again an accurate approximation.

Running

A runner moves forward with a force generated by the friction between foot and ground. The laws of viscous friction [20] may be employed here. In Figure 1, as a runner pushes off, the forward velocity of the foot is assumed to increase linearly from 0 to v. The backward friction force against the ground is balanced by the opposing force f, propelling the runner forward. Viscous friction is given by $(\mu A/h)v$, where μ is viscosity (how slippery the surface is), A is the cross-sectional foot area in contact with the ground, h is the amount of vertical compression of the shoe and running surface, while v is velocity. The equation for f agrees with intuition in that a soft spongy surface will become compressed significantly (large h) implying reduced friction and slower running. Similarly, a slippery surface will have a low viscosity μ, also slowing the runner. Clearly, the stiffness (or softness) of a track shoe must be tuned to the running surface and the runner's preferences to avoid shin splints or wasted energy. The interest here is in the ratio of male and female output power in terms of $\%DT$. Since power p equals fv, and since like constants can be cancelled,

$$\frac{p_W}{p_M} = \frac{A_W}{A_M}\left(\frac{v_W}{v_M}\right)^2. \tag{7}$$

The ratio of foot area is needed. Data in [15] by Toussaint regarding swimmers show that the ratio of cross sectional body area may be approximated by the square root of the mass ratio, thus, a logical approximation to A_W/A_M is $(m_W/m_M)^{1/2}$, making the power ratio for runners proportional to $m^{1/2}v^2$. (See the numerical example above for a slender cylinder). The resulting approximation for the power ratio is

$$\frac{p_W}{p_M} = \left(\frac{m_W}{m_M}\right)^{1/2}\left(\frac{v_W}{v_M}\right)^2 = \left(\frac{m_W}{m_M}\right)^{1/2}\left(1-\frac{\%DT}{100}\right)^2. \tag{8}$$

Next, data are needed to calculate the mass (weight) ratio for men and women. Data were obtained for the weights of the male and female USA Olympic rowing team members for 2000 [23], the Australian Olympic swimming team members from 1988–1994 [6] and for elite runners, jumpers and swimmers competing in the Tokyo and Mexico City Olympics [4]. The data are remarkably consistent. The male athletes are 26% heavier than the female athletes for each category. Thus, the ratio 1/1.26 is a reliable estimate of m_W/m_M for Olympic champions, resulting in the $\%DP$ values in column three of Table 1 for running events. The average $\%DP$ is about 28%. There is about a 5% rise of $\%DP$ up to the 10000 m run and a 2% drop from the 10000 m run to the marathon.

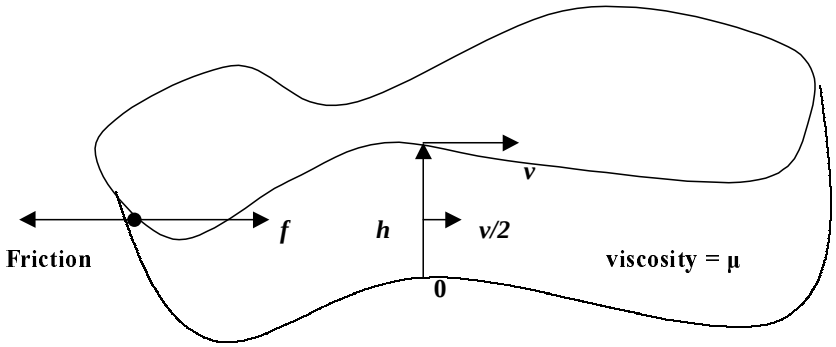


Figure 1: Viscous friction. Top surface area = A. Material beneath compresses by a distance h. Viscosity is μ. The forward velocity drops from v to 0 linearly as material compresses over distance h. Forward force is equal to the force of friction.

Jumping and pole vault

It is doubtful that any long jumper thinks of the laws of ballistic trajectory while speeding down the runway and then taking off, but clearly the jumper's physical motion is intended to maximise the length of the jump. If we apply ballistics [3] here, we consider the flight of a jumper who takes off with velocity v at some angle θ with respect to the ground as in Figure 2. The jumper starts with a horizontal velocity of $v\cos\theta$ and a vertical velocity of $v\sin\theta$. The acceleration of gravity (g) pulls down on the jumper, so the vertical velocity becomes $v\sin\theta - gt$. The jumper is at the top of the jump when the vertical velocity is zero, which happens when t is $(v\sin\theta)/g$. The maximum vertical height is given by $vt\sin\theta - \frac{1}{2}gt^2$ for that t, or $(\frac{1}{2}v^2\sin^2\theta)/g$. This time (to reach the maximum height) is half the total flight time, so the horizontal distance is found by multiplying the horizontal velocity by twice that time. The result is a distance $d = (2v^2\sin\theta\cos\theta)/g$. The resulting horizontal distance equation may be used for the long jump (LJ) and the maximum vertical height may be used for the high jump (HJ), recognising that the takeoff angle will be higher for the HJ than for the LJ. That is,

$$d(\mathrm{LJ}) = \frac{2v^2\sin\theta\cos\theta}{g}, \tag{9a}$$

$$d(\mathrm{HJ}) = \frac{v^2\sin^2\theta}{2g}. \tag{9b}$$

The applicability of these equations can be tested using HJ and LJ data. The take off angle for an elite long jumper, measured about the centre of gravity (CG), is about 20° [5]. To use (9a) and (9b) for a ballistic trajectory, the angle with the ground is needed. Projecting backwards from the take off point about 1.5 m, the take off angle with the ground is about 30°. A high jumper leaps at a higher angle: about 30° with the CG and 40° with the ground. Using an angle of 30° for (9b) and 40° for (9b), and assuming that the velocities are equal, the ratio $d(\mathrm{HJ})/d(\mathrm{LJ})$ is 0.238. The ratio of average HJ to (average LJ plus 1.5 m) for the last five men's Olympic competitions is 2.36 m /(8.60 m + 1.5 m) or 0.234, in good agreement.

For the pole vault (PV), kinetic energy (due to the run up and jump onto the pole) is converted to potential energy, which propels the vaulter upward [7]. From physics, kinetic energy is given by $\frac{1}{2}mv^2$, while potential energy is mgd, where m is the vaulter's mass, g is the acceleration of gravity and d is the vertical rise; that is, is the difference in height from the jumper's CG (when the potential energy

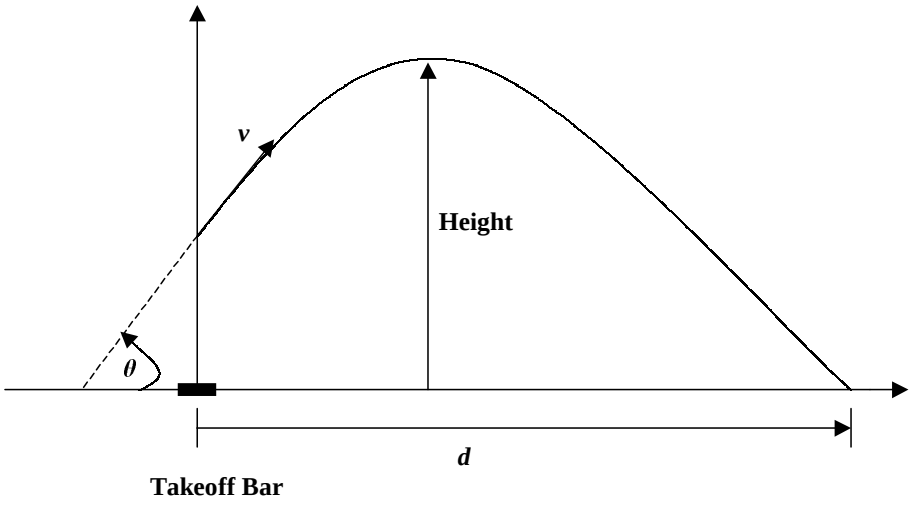


Figure 2: Ballistic trajectory of the centre of gravity. Distance d is measured from takeoff bar. Takeoff velocity is v. Takeoff angle is θ.

from the pole begins to lift the pole-vaulter) up to the height cleared. That is

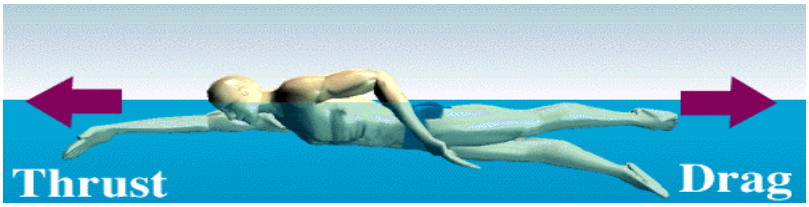
$$d(\mathrm{PV}) = \frac{v^2}{2g}. \tag{10}$$

Some approximate values may be used to test the reasonableness of (10). This equation can be tested using the men's average winning performance in the PV for the last eight Olympic and eight world championships. Given that g is $9.8\,\mathrm{ms}^{-2}$, if a male pole-vaulter runs about $9.5\,\mathrm{ms}^{-1}$ with a CG that is 1.2 m above the ground when the pole begins to rebound, the vault would be about 5.80 m, closely agreeing with the actual average of 5.99 m. If the female pole-vaulter runs about 10% slower, at $8.5\,\mathrm{ms}^{-1}$, and her CG is about 1 m off the ground when the pole begins to rebound, then the expected vault is 4.76 m, as compared to the average of 4.65 m for two world championships and one Olympic championship. These speeds and CG locations are reasonable approximations; hence, the analytical process seems to be reasonable.

The goal is to estimate the power ratio for these events. As before, power is the product of force and velocity. Here, force is the weight of the jumper/vaulter, which is proportional to mass. The relationship between distance and velocity is the same for the long jump (triple jump can be included with the long jump), high jump, and pole vault; that is, velocity is proportional to the square root of distance. Thus, the power ratio of force times velocity becomes proportional to $md^{1/2}$, where like constants are cancelled, assuming that male and female athletes take off at the same angle. So

$$\frac{p_W}{p_M} = \frac{m_W}{m_M}\left(\frac{d_W}{d_M}\right)^{1/2} = \frac{m_W/m_M}{(1+PDDi/100)^{1/2}}. \tag{11}$$

Using the mass ratio of 1/1.26, (11) is used to estimate the power output for the jumping and pole vault events. For the pole vault, 1.2 m is subtracted from the male performance and 1 m is subtracted from the female performance, such that distance becomes height above the CG. See Table 1, column three. The $\%DP$ again average about 28%. The higher value for the PV may be due to the newness of the event (women may become more efficient in power use in future competitions) or that difference may persist and be attributed to some loss of useful power due to lower arm strength for a typical female pole-vaulter. More data is needed. Next comes an analysis of the power ratio for swimming.



power output (p) = drag force × velocity (p_D) + power losses (p_L)

Figure 3: The physical forces acting on a swimmer. Power losses increase kinetic energy of water around arms and legs.

Swimming

The power output of an Olympic swim champion is of particular interest today, given the recent introduction of bodysuits, and the emergence of many claims of significantly improved swimming performance. Figure 3 helps to illustrate the physics of swimming. Output power from the swimmer is applied to the arms and legs, which produce some power loss (increasing the kinetic energy of the surrounding water) and useful power (overcoming drag, causing the swimmer to move forward at velocity v) [17, 13]. Here, p is the total output power applied by the swimmer to the water which includes power loss p_L and power applied to overcome drag, p_D; that is, $p = p_L + p_D$. Efficiency (e) may be defined as p_D/p, making $p_D = pe$. Thus any power ratio based only on p_D does not necessarily reflect the body's output power ratio, because e may be different for men and women. How them is drag to be analysed? In general [18, 19], p_D is the product of drag force f_D times velocity v. Drag force f_D equals dynamic pressure ($\frac{1}{2}\rho v^2$) times wetted surface area (A) times the drag coefficient (C_D), which corrects for differing shapes and velocities. Here, ρ is the density of water.

Drag has three main causes: form drag caused by shape, wave drag formed as the water surface is broken by the motion of the swimmer and skin drag formed as the stream of water moves past the skin surface. For convenience, all three forms of drag may be combined into one total drag force. To measure that drag force, a swimmer may be connected to a force-measuring device while floating in a moving stream of water called a flume. The force exerted on the tether is the drag force, which can be measured for various conditions of interest. For example, passive and active drag conditions may be measured. In passive drag, the swimmer (or a dummy) is motionless. Active drag occurs when the swimmer is employing some form of a swimming stroke.

Early data given by manufacturers of body suits, measured using passive drag, suggested significant reduction in drag. In [19], five experts pool their knowledge to evaluate the effectiveness of the bodysuit, based on active drag force measured under various conditions. They conclude that backstroke and breaststroke swimmers are not enhanced. In freestyle and butterfly swimming, an unshaved swimmer with a conventional swimsuit might have improvement, when switching to the bodysuit. Conversely, there is no improvement in any stroke using a bodysuit compared to a shaved swimmer wearing a conventional tight swimsuit. Some tight bodysuits could hamper performance by restricting range of motion. A wet bodysuit, during a race of longer than 200 m, could slow the swimmer. Thus, a body suit is not at all a guarantee of improved performance, illustrating the importance of measuring and interpreting drag force.

Let us return to the task of estimating the power ratio of male and female swimmers, where we

estimate the total power output as applied to the water, some of which does not propel the swimmer forward. The ratio of power for $ep = p_D$ results when common constants are canceled:

$$\frac{e_W}{e_M}\frac{p_W}{p_M} = \frac{C_{DW}}{C_{DM}}\frac{A_W}{A_M}\left(\frac{v_W}{v_M}\right)^3. \tag{12}$$

The ratio of wetted surface area for male and female swimmers in (12) can be estimated by taking the volume ratio raised to the exponent of 2/3, as discussed earlier. How may the displaced volume (also called displacement) be calculated? A swimmer that is less dense than water floats because the weight of the volume of water displaced equals the weight of the swimmer and the volume displaced is smaller than the volume of the swimmer. The displacement volume of a swimmer is given by the mass of the swimmer, m, divided by the density of water, ρ, giving m/ρ. Density cancels out when taking the volume ratio, therefore we can estimate A_W/A_M by $(m_W/m_M)^{2/3}$.

Next, how about efficiency of male and female swimmers? Toussaint [16] observes that men are more efficient because they apply greater output power, losing less to the surrounding water. Greater output power may be assumed to be proportional to mass. Then, we can approximate e_W/e_M by m_W/m_M, so (12) becomes

$$\frac{m_W}{m_M}\frac{p_W}{p_M} = \frac{C_{DW}}{C_{DM}}\left(\frac{m_W}{m_M}\right)^{2/3}\left(\frac{v_W}{v_M}\right)^3. \tag{13}$$

The output power ratio can be found from (13):

$$\begin{aligned}\frac{p_W}{p_M} &= \left(\frac{m_W}{m_M}\right)^{-1/3}\frac{C_{DW}}{C_{DM}}\left(\frac{v_W}{v_M}\right)^3 \\ &= \left(\frac{m_W}{m_M}\right)^{-1/3}\frac{C_{DW}}{C_{DM}}\left(1-\frac{\%DT}{100}\right)^3.\end{aligned} \tag{14}$$

The output power ratio is proportional to $m^{-1/3}C_D v^3$. The power ratio proportionalities are summarised in Table 3. Suppose that we use (13) or (14) to solve for the velocity ratio for a given power ratio. The mass ratio then has a plus 1/3 exponent when moved to the left side of the equation, in direct proportion to the male swimmers mass/weight. The implication is that for a given power ratio, a heavier male swimmer has greater velocity advantage and not a disadvantage due to weight. This apparently counter-intuitive result, consistent with a similar analysis of Olympic rowers [12], is a result of buoyancy. A boat of heavier rowers will defeat a boat with the same number of lighter rowers having the same skill level, as evidenced by Olympic winning times, because the heavier rowers have greater drag by the 2/3 exponent of their weight (a disadvantage) but greater forward thrust by the 3/3 exponent of their weight (an advantage). The result is a relative gain equal to the 1/3 exponent of their weight, as evidenced also in (13) and (14). Velocity increases by the 1/3 exponent of the power ratio, which is the 1/9 exponent of the mass ratio giving an advantage in velocity/time to the heavier athlete. Olympic winning times in rowing fit that relationship closely.

The mass/weight ratio may be approximated by 1/1.26 as discussed earlier. It remains to estimate the ratio of drag coefficients. In [14], Toussaint published a ratio C_{DW}/C_{DM} of drag coefficients equal to 0.83 for velocities in the $1\,\mathrm{ms}^{-1}$ to $1.8\,\mathrm{ms}^{-1}$ range. He noted that the drag difference becomes less at higher velocities. Since Olympic winning velocities are in the $2\,\mathrm{ms}^{-1}$ to $2.5\,\mathrm{ms}^{-1}$ range, setting C_{DW}/C_{DM} to about 0.9 seems reasonable, especially when (13) results in $\%DP$ values in column 3 of Table 2 averaging about 28%, consistent with all the other events studied so far.

In the next section, a physiological explanation is proposed for that 28% average $\%DP$.

2.4 Physiological explanation of power differences

Male and female athletes have a number of measurable physiological differences affecting athletic performance. For example, male athletes are heavier, taller and clear a greater volume of oxygen. In

Event	Power ratio is proportional to
Running	$m^{1/2}v^2$
Jumping and pole vault	$md^{1/2}$
Swimming	$m^{-1/3}C_D v^3$

Table 3: Power ratio proportionalities.

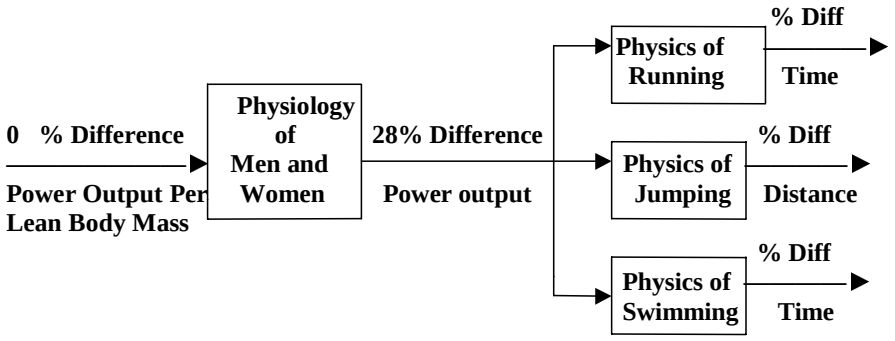


Figure 4: Summary of power output differences between male and female athletes.

what way can physiology be used to explain the 28% difference in power that exists for the various running, jumping and swimming events just considered? Oxygen consumption is often used to measure the amount of internal energy the body consumes. It has been observed [1] that the typical value of 40% greater maximum oxygen consumption ($VO_{2\,\max}$) for elite male athletes, drops to less than 10% when expressed per unit of lean body mass. The remaining 10% difference is explainable by known physical and hematological factors. It is also well known that men and women burn the same number of calories per unit of lean body mass [1]; thus a calculation of the ratio of lean body mass for female and male athletes is a logical next step in estimating the relative quantity of calories available for generating power, the rate that the calories are burned per unit of time. Lean body mass consists of everything except body fat. A person having 15% body fat has 85% lean body mass.

Table 4 summarises physiological data for elite athletes. Total body weight data in [23, 6, 4], as discussed earlier, pertains to the 2000 USA Olympic rowing team [23], the Australian Olympic swimming team from 1988–1994 [6] and for a cross section of track and field athletes competing in the Tokyo and Mexico City Olympics [4]. The male athletes are consistently 26% heavier than the female counterparts; hence that result is a good estimate for Olympic champions. Similarly, body fat percentages are quite consistent for elite swimmers as cited in [2, 22], where male swimmers have an average 11.5% body fat compared to 20.5% for female swimmers. For runners, the studies in [1, 2] are similar, where male runners have an average of 7.25% body fat compared to 15.5% for females.

The physiological data in Table 4 can be used to estimate the ratio of power output for female and male Olympic champions, by using the ratio of lean body mass for women divided by lean body mass for men. The estimated power difference is 28.7% for swimmers and 27.7% for runners. Both figures agree quite closely with the average power differences of Tables 1–3.

The cause-effect flows of Figure 4 produce the power output differences in male/female performance as evaluated herein. A zero $\%D$ in power output per unit of lean body mass, together with physiological data for typical elite athletes (lean body mass and percent body fat), are consistent with a 28% difference in power output, which (using physical laws for each sport) in turn, are consistent with all the observed

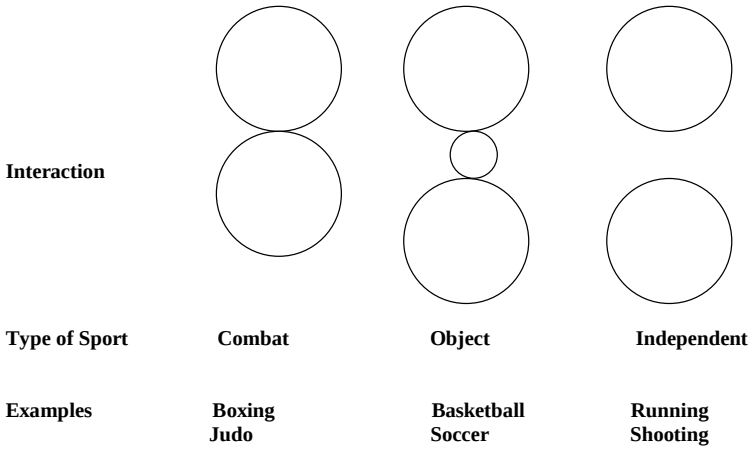


Figure 5: Three ways for competitors to interact.

$\%D$ values in time (for running and swimming events) and in distance (for jumping events).

It is interesting to examine the recent history of $\%DP$ for running events. Table 5 shows the average $\%D$ in time and average $\%D$ in power for Olympic running events from 1964 (four common events for men and women) through 2000 (eight common events). The $\%D$ in power clearly dropped rapidly from 1964 (32%) through 1980 (24%), a period of rapid growth in women's sports. The pool of female athletes increased rapidly as it became much more socially acceptable for a woman to be a successful athlete. That pool of female athletes became increasingly more fit and had access to better coaching and training facilities. Using data similar to that in Table 5, for the period when women were improving more rapidly than men, it was projected by at least one paper [21] that women would be running as fast as men by the mid-21st century. It was misleading in 1980 that there was only a 7.86% difference in time (24% difference in power). That difference is unusually low due to the boycott of Western-bloc male runners who tended to be the best at the time, while Eastern-bloc women, who also tended to be the best at that time, did compete. Further, as suggested in this current paper, the $\%DP$ was actually moving to a sort of natural equilibrium at around 28%, and not to a linear projection of 0%. Since 1980, the $\%DP$ has moved somewhat randomly around 28%.

It is important to note that while women improved faster than men from around 1964 to 1980, it is unfair to criticise women for lower rates of improvement today. Both men and women continue to improve, but they both improve at the about the same rate, as it should be for equality of opportunity.

3 Fundamental nature of sports

In most basic terms, a sport consists of the interaction of two teams or competitors attempting to accomplish the goal of that sport as defined by a set of rules for deciding the winner. Similarly, each activity aimed at that goal is further governed by additional rules. Assuming that team members cannot change affiliation during competition, there are only three types of sports possible and only three types of activities during each sport. That is, two competitors can interact in only three ways as in Figure 5. First, competitors can be in direct contact with each other, where the goal is to control the opponent as in boxing, fencing and judo. Second, competitors can be in indirect contact via some intervening object, where the goal is to control that object and score points, as in basketball, soccer and volleyball. Third, competitors can operate independently, where the goal is to control the athlete's own performance for optimal result, as in running, swimming, shooting and golf.

Reference	Men	Women	Ratio women/men
Body weight in kilograms			
[23]	92.4	73.3	1/1.26
[6]	81.8	64.8	1/1.26
[4]	72.2	57.8	1/1.25
Average			1/1.26
Physiology of swimmers			
Percent body fat [2]	12.0	20.0	
Percent body fat [22]	11.0	21.0	
Percent body fat average	11.5	20.5	
Percent lean body mass	88.5	79.5	
$\frac{\text{Female lean body mass}}{\text{Male lean body mass}}$ = Estimated power ratio			$\frac{79.5/88.5}{1.26} = 0.713$
Estimated percent difference in power			28.7
Physiology of runners and jumpers			
Percent body fat [1]	7.5	16.0	
Percent body fat [2]	7.0	15.0	
Percent body fat average	7.25	15.5	
Percent lean body mass	92.75	84.5	
$\frac{\text{Female lean body mass}}{\text{Male lean body mass}}$ = Estimated power ratio			$\frac{84.5/92.75}{1.26} = 0.723$
Estimated percent difference in power			27.7

Table 4: Physiological differences between elite male and female athletes and estimated percent difference in power.

Year	Common events	Percent difference in time	Percent difference in power
1964	4	12.62	31.97
1968	4	13.13	32.76
1972	5	10.56	28.72
1976	5	9.94	27.73
1980	5	7.86	24.36
1984	7	10.91	29.28
1988	7	9.16	26.48
1992	7	9.59	27.17
1996	8	11.26	29.84
2000	8	9.54	27.09

Table 5: Percent difference in power for winning male and female Olympic champions (1964–2000).

Activities performed during competition can also be classified into the same three categories. In soccer (an object sport), for example, while an attacking player is setting up a corner kick, the other players may jostle for position around the goal (combat activity). Next, when the ball is played toward

the goal, key attacking players attempt to shoot on goal (object activity). If the goalkeeper makes a save in front of the goal, that goalkeeper can distribute the ball into play without interference being allowed (independent activity).

Table 6 contains 93 widely played sports, organised into combat, independent and object categories. Sports are drawn from three sources: sports on the Olympic program, sports not on the Olympic program but recognised by the International Olympic Committee for extensive international competition and widely-played regional sports appearing in widely-published sports references. There are ten combat sports, 42 object sports and 42 independent sports. Note that the five-sport modern pentathlon appears twice because that sport has both combat and independent components, resulting in 93 different sports but 94 entries.

The three types of sport possess fundamentally different properties that must be considered when attempting to rate teams and individuals. For an independent sport, the individual's average independent performance may be used as a reliable predictor of future competition. Conversely, a combat sport is usually decided by subjective judging, so that further subjective rankings are often used. A competitor in an object sport has an objective score as an outcome, but that score depends on interaction with the other competitors; hence, the most sophisticated rating systems tend to be employed for object sports to establish a separate ratings for each competitor. Sports rating systems are discussed in the next section.

10 Combat Sports	42 Independent Sports		42 Object Sports	
Aikido	Alpine Skiing	Nordic Combined	American Football	Lacrosse
Boxing	Archery	Orienteering	Australian Rules	Netball
Fencing	Athletics (Track	Roller Skating	Football	Paddleball
Judo	and Field)	Rowing	Badminton	Pelota Vasca
Karate	Automobile Racing	Rhythmic	Bandy	(Jai Alai)
Kendo	Ballroom Dancing	Gymnastics	Baseball	Pool/billiards
Modern Pentathlon	Biathlon	Sailing	Basketball	Raquetball
(Fencing)	Bobsled	Shooting	Bowls	Roller Hockey
Taekwondo	Bowling	Short Track	Bridge	Rugby Union and
Wrestling	Canoe/Kayak	Skating	Canadian Football	Rugby League
Wushsu	Cross-country Ski	Skeleton Sled	Checkers	Shinty
	Cycling	Ski Jumping	Chess	Soccer
	Diving	Skydiving	Court Handball	Softball
	Equestrian	Snowboarding	Cricket	Squash
	Figure Skating	Speed Skating	Croquet	Table Tennis
	Freestyle Skiing	Surfing	Curling	Team Handball
	Golf	Swimming	Darts	Tennis
	Gymnastics	Synchronised	Field Hockey	Tug of War
	Lifesaving	Swimming	Gaelic Football	Underwater
	Luge	Trampoline	Horseshoes	Hockey
	Modern Pentathlon	Triathlon	Hurling	Underwater Rugby
	Motorcycle Racing	Water Skiing	Ice Hockey	Volleyball
	Mountaineering	Weightlifting	Korfball	Water Polo

Table 6: Widely played sports (Modern Pentathlon appears twice).

4 Fundamental nature of rating systems

This paper seems to be organised in sets of three. We examined three types of events (running, jumping and swimming) and three types of sports (combat, object and independent). In this section, a triumvirate again appears in that there are three types of audiences and three types of rating systems. It is important to recognise that the success of a rating system is affected by the likes and dislikes of each audience.

Rating systems have been extensively analysed as in [10, 9, 11]. There are at least three types: subjective, accumulative and adjustive. There are also three types of audiences: the general non-technical public, organisers of tournaments and the technical audience seeking sophisticated ratings. Each audience tends to be most comfortable with one of the rating systems.

A subjective rating system utilises some poll of authorities as in boxing, American college football and soccer. In the USA, two polls are widely accepted indicators of the top 25 American college football teams: one poll is of sports writers (by Associated Press) and one poll is of coaches (by *USA Today*/ESPN). Since 1956, France Football has used a poll of sports writers to select the best male soccer player in Europe. Since 1991 the international soccer federation, FIFA, has used a poll of national team coaches to select the best male soccer player of the world. Since 2001, FIFA also honours the best female soccer player in the world by a similar poll. It is common for sports writers to rate the performance of soccer players on a 10-point scale. The wide acceptance of these ratings speaks to the acceptance of subjective polls by the general non-technical audience. It is also true that this audience has an interest in the simplest of adjustive systems, a simple average. It is likely that this audience trusts those systems with which each person is most familiar: calculating simple averages for salary and expenses, for example, and the typical sports enthusiast subjectively selects favourite players and teams.

An accumulative system causes points to increase monotonically due to success in competition. Points may be assigned due to the final position in the competition and the importance of the competition, for example. The current men's and women's professional tennis tours (ATP and WTA respectively) seed players into tournaments with accumulative systems. The world ski federation, FIS, uses accumulative systems for alpine skiing and cross-county skiing. The systems are used for start listing and for recognising yearly champions. There is only one way an accumulative rating can become smaller: when a moving window is used (perhaps for one calendar year), in which a rating can drop when a good result is aged away and replaced by a lesser result. Tournament organisers favor these systems. Such systems encourage good athletes to enter many competitions to maximise available points. Although these systems can put opponents in a rank order, that order does not predict the expected margin of victory, but that is not a stated purpose.

Adjustive ratings are the most sophisticated. These ratings rise or fall as future performance is compared to expected performance and the ratings are adjusted (up or down) appropriately. The technical audience understands that the goal is accurate prediction, perhaps for gambling, fantasy game play or just better enjoyment. These systems utilise such methods as least squares and exponential smoothing.

Two case studies serve to illustrate the part played by an audience. The ATP and the WTA professional tennis tournament directors both changed from adjustive systems to the current accumulative systems in 1990 and 1996, respectively. ATP and WTA directors cited the same reason. Under an adjustive system, they felt that top rated players were entering fewer tournaments later in the tour year, especially if there had been any injury or minor illness, fearing that their averages would drop, affecting ranking and appearance fees. The accumulative system does not penalise a poor result, since the rating does not change in that case. A player can safely enter a tournament to regain form. Directors of both tours released follow up evaluations indicating increased entries by top professionals.

The competition among (supposedly) amateur American college football teams generates prodigious television revenues. At the end of the regular season, over 20 "bowl games" are played. These games do not constitute a tournament, since each team plays only one game. The four most prestigious of those games (called the Rose, Orange, Sugar and Cotton Bowls) distribute to each team about US$12 million, indeed big business. The credibility of those four games lies in the public acceptance that the competing teams are the eight best of more than 100 major college teams, and television advertising revenue is similarly sensitive to viewer acceptance. One bowl game, on a rotating basis, is designated the national championship game. Since 1998, Bowl Championship Series (BCS) ratings have been used to select the top eight teams, including the top two teams who play for the national championship. That system includes the two major subjective polls, a panel of computer ratings and a third component intended to include strength of schedule. In spite of the supposed importance of using a broad-based system, three times that composite system differed from the polls and three times efforts were made (or suggested)

to bring the composite system into alignment with the subjective polls.

In 1998, one computer rating system preferred Kansas State (third overall but fourth in the polls) and the next year, additional computer ratings were added so that no one system (supposedly) would be important enough to overrule the polls. In 2000, a late season victory over a top team by the University of Miami went unrewarded by the computers, considering each game equal, while the polls, sensitive to late season victories, moved Miami to second place and into public acceptance. Overall, Miami was placed third and not included in the national championship game. For 2001, a special reward was added for victory over top rated teams, justified by the Miami case of the previous year. In 2001, there were four evenly matched teams. The final overall rating placed the first poll-rated team (Miami) and the fourth poll-rated team (Nebraska) into the national championship game, won by Miami, partially justifying the system, but leaving concern that the fourth poll-rated team might have won. As of April 2002, the BCS Committee is considering schemes to further reduce the influence of computer rankings as compared to the poll ratings. Newspaper commentary is highly negative toward computer-based systems. Evidently, the general public considers the polls to be the accurate measure of ability.

As we develop ratings that make sense to us, the technical audience, it is well to consider the likes and dislikes of other audiences.

5 Conclusions

Men and women produce the same energy and power output per unit of lean body mass. Based on data for elite athletes, including total body mass and percentage of body fat, women produce 28% less power output than men, exactly the value arrived at by applying the laws of physics to running, jumping and swimming events. Thus, observed performance differences disappear when corrected for differences in lean body mass. Put another way, if events could be handicapped based on lean body mass, men and women should finish in a dead heat.

Only three types of sports and three types of sports activities are possible. These are combat sports with direct contact between opponents, object sports with indirect contact via an object (such as a basketball) and independent sports with separate performance. It might be said the athlete must control the opponent, an object or the athlete's own self, respectively.

There are three types of rating systems and three audiences. The general public trusts subjective rating systems (and simple averages). Tournament organisers favour accumulative systems. The technical audience favours sophisticated adjustive systems, capable of predicting future performance. The creator of a rating system should understand the likes and dislikes of these audiences.

References

[1] A. Foster and F. Korkia, "Women and cycling—what makes women different than men", University of Luton, freespace.virgin.net/martin.shakeshaft/women.html.

[2] K. T. Jang *et al.*, "Energy balance in competitive swimmers and runners", *J. Swimming Research*, **3** (1987), 19–23.

[3] L. Lerner, *Physics for Scientists and Engineers*, Jones and Bartlwett (1996).

[4] W. McArdle *et al.*, *Exercise Physiology*, Lea & Febiger (1981).

[5] P. McGinnis, "Eight elements of an effective takeoff", *National Pole Vault Summit* (January 21–22, 2000); www.vaultworld.com/prsport/Articles/peterhandout.html.

[6] D. Pyne D *et al.*, "Laboratory environment/subject preparation, equipment checklist, blood testing and anthropometry", Chapter 27, *Physiological Test for Elite Athletes (Australian Sports Commission)*, C. Gore (editor), Human Kinetics, Champaign, Illinois (2000), 372–382; www.education.ed.ac.uk/swim/papers3/wg3a.html.

[7] *Physics of Sports, Lecture 5: Center of Mass*, courses.Washington.edu/phys208/notes/lect5.html.

[8] R. Stefani, "Applying least squares to team sports and Olympic winning performances", in *First Conference on Mathematics and Computers in Sport*, N. de Mestre (editor), Bond University, Queensland, Australia (1992), 43–49.

[9] R. Stefani, "Survey of the major world sports rating systems", *J. Appl. Statist.*, **24** (1997), 635–646.

[10] R. Stefani, "Predicting outcomes", in *Statistics in Sport*, J. Bennett (editor), Arnold, London (1998), 249–275.

[11] R. Stefani, "A taxonomy of sports rating systems", *IEEE Transaction on Systems, Man, and Cybernetics*, Part A, **29** (January 1999), 116–120.

[12] R. Stefani and D. Stefani, "Power output of Olympic rowing champions", *Olympic Review*, **XXVI-30** (December 1999 – January 2000), 59–63.

[13] H. Toussaint, "Propelling efficiency in front crawl swimming", *J. Appl. Physiol.*, **65** (1988), 2506–2512; www.ifkb.nl/B4/propeff.html.

[14] H. Toussaint *et al.*, "Active drag related to velocity in male and female swimmers", *J. Biomech.*, **21** (1988), 435–438.

[15] H. Toussaint *et al.*, "The effect of growth on drag in young swimmers", *J. Biomech.*, **6** (1990), 18–28; www.ifbk.nl/B4/growthondrag.html.

[16] H. Toussaint, "The mechanical efficiency of front crawl swimming", *Medicine and Science in Sports and Exercise*, **22** (1990), 402–408; www.ifkb.nl/B4/mechanicaleff.html.

[17] H. Toussaint, "Research line B4: cyclic movements, mechanics and energetics of swimming, propelling efficiency", Faculty of Movement Sciences, Vrije Universiteit, Amsterdam; www.ifkb.nl.B4/swimmingpropellingeff.html. See also "Effect of propelling surface size on the mechanics and energetics of front crawl swimming", *J. Biomech.*, **24** (1991), 205–211; www.ikfb.nl/B4/paddles.html.

[18] H. Toussaint, "Hydrodynamics makes a splash", *Physics World*, **13** (9 September 2000), 39–43; takagi.edu.mie-u.ac.jp/takagi/abstruct/PhysicsWorld/hydrodyanamic.htm.

[19] H. Toussaint, "Bodysuit yourself: but first think about it", *J. Turbulence*, downloaded from the web site www.iop.org/Journals/Jo/extra/20.

[20] E. Umez-Eronini, *System Dynamics and Control*, Brooks/Cole (1999), 37–38.

[21] B. Whipp and S. Ward, "Will women soon outrun men?", *Nature*, **355** (January 1992), 25.

[22] "The effect of varying body composition on swimming performance", *J. Strength and Conditioning Research*, **8** (1994), 149–154; www.pponline.co.uk/encyc/0346.htm.

[23] "2000 Olympic rowing roster", www.usrowing.org/olympic/olympicroster.html.

FACTORS AFFECTING OUTCOMES IN TEST MATCH CRICKET

Paul Allsopp and Stephen R. Clarke
School of Mathematical Sciences
Swinburne University
PO Box 218, Hawthorn
Victoria 3122, Australia
paulalls@bigpond.com, sclarke@groupwise.swin.edu.au

Abstract

Least squares regression is used to model the first innings performances of teams in test cricket in order to establish batting and bowling ratings, a common home advantage and a country effect. Logistic regression techniques are then used to model match outcomes based on a team's first innings lead, innings duration, home advantage, batting and bowling ratings and the country effect. It is shown that the factors that impact most significantly on the outcome of a match are a team's first innings lead home team performance and innings duration. A team's first innings lead is found to more likely shape a win rather than a draw or a loss whereas the longer the duration of the first innings the more likely a match will end in a draw. It is shown that the home team, on average, needs to establish a lead in excess of 93 runs to have a better than even chance of winning, whereas the away team needs to establish a lead in excess of 115 runs to have the same chance. There is a better than an even chance of a draw for a first innings duration in excess of 1165 minutes (or approximately 277 overs). It is also shown that the home team is more likely to win a match rather than lose or draw, which suggests that the home team has a distinct winning advantage over the away team. There is some evidence suggesting that teams gained an advantage by batting last.

1 Introduction

Test match cricket is currently played between the ten International Cricket Committee (ICC) test-playing nations, with Bangladesh being a very recent inclusion. A test match is scheduled to finish within a five-day period and comprises a maximum of two innings per team. There are four possible outcomes: a win, loss, draw or tie. A tied result is an extremely rare event and has only occurred a handful of times throughout the history of test cricket.

Outcomes in a test cricket match are difficult to predict because they are dependent on a wide range of interrelated factors. By applying standard modelling techniques we will initially focus on the factors that affect the performance of teams in their first innings and then determine which of these factors, if any, have an impact on the outcome of a match. The factors to be analysed are a team's first innings lead, home advantage, team batting order, innings duration, attack and defence ratings and the country effect We have considered all 371 completed matches from seasons 1990 through to 2001. By "completed matches", we mean those matches that produced a result independent of weather conditions. Note that as Bangladesh had only played in three matches throughout the study period their results have not been included in the analysis.

2 Exploratory data analysis

Throughout the analysis Team 1 and Team 2 refer to the teams batting first and second respectively in the first innings. Subsequently, unless Team 2 has been forced to follow on it will also be the team batting last in the second innings. Table 1 provides a descriptive summary of the first innings results for each team. Overall, Team 2 won 140 matches and lost 121. There were 110 draws. The results show that Team 2 had a winning advantage over Team 1. However, using a chi-square goodness of fit test to compare the expected and actual number of wins, losses and draws for Teams 1 and 2 suggests that the observed differences are not significant (p-value = 0.251). The observed differences are due solely to random variation. However, the data refute the accepted wisdom that being first in the batting order provides a team with a winning advantage. Conversely, the data show that the team batting first has a tendency to lose rather than win or draw.

The first innings batting performances by India (as Team 1 and 2), on average, are substantially higher than the majority of nations but are much more variable. This underscores India's lack of batting consistency.

Team	Mean first innings score		Standard deviation	
	Team 1	Team 2	Team 1	Team 2
Australia	358	368	131	134
South Africa	354	333	106	106
India	352	342	165	108
England	298	309	121	119
Pakistan	296	326	110	129
Sri Lanka	289	320	107	165
West Indies	284	312	131	133
New Zealand	283	286	111	104
Zimbabwe	262	263	131	70
Overall	312	321	128	125

Table 1: Descriptive summary of the first innings in test cricket.

3 Modelling the first innings

In a typical test match the first innings batting side aims to score as many runs as possible before losing the ten wickets at their disposal, whereas the bowling side aims to dismiss the batting side by taking all ten wickets for a total that is as small as possible. Assuming that both teams are endeavouring to maximise their first innings lead the score that is achieved provides a measure of the relative batting and bowling strength of the two teams. Beyond the first innings, however, playing strategies are harder to predict because teams become more reactionary and tend to customise their style of play. Using techniques similar to those adopted by Harville and Smith (1994), Clarke and Norman (1995), de Silva, Pond and Swartz (2000) and Clarke and Allsopp (2001), a team's first innings score in a test match played between the batting team i and the bowling team j at a location k with home ground l and batting order m is modelled as

$$S_{ijklm} = A + a_i - d_j + c_k + h_{il} + b_m + \epsilon_{ijklm}, \tag{1}$$

where the indices i, j, k, $l = 1, \ldots, 9$ represent the nine ICC test-playing nations and $m = 1, 2$ indicates whether a team batted first or second. The response variable s_{ijklm} signifies a team's first innings score; A represents the expected score between average teams on a neutral ground; a_i and d_j signify the first innings batting (attack) and bowling (defence) ratings of teams i and j, respectively;

c_k is the country effect term, which represents the advantage gained by teams playing in a particular country. The common home advantage enjoyed by the batting side is represented by h_{il}, such that

$$h_{il} = \begin{cases} h, & \text{if } l = i, \\ 0, & \text{otherwise.} \end{cases}$$

The team batting first is indicated by b_m, such that

$$b_m = \begin{cases} b, & \text{if } m = i, \\ 0, & \text{otherwise.} \end{cases}$$

Finally, ϵ_{ijklm} is a zero-mean random error. The error term is included because two competing teams will not necessarily repeat their first innings performances the next time they meet. Subsequently, a least squares regression model is fitted to the scores to quantify the parameter estimates for each of the explanatory variables. For convenience, $\sum_{i=1}^{9} a_i = 900$, $\sum_{j=1}^{9} d_j = 900$ and $\sum_{k=1}^{9} c_k = 0$, which assumes that a team's first innings average batting and bowling ratings are each 100 and the country effect rating is 0. Accordingly, a rating greater than 100 signifies that a team has performed above average whereas a rating less than 100 signifies that a team has performed below average.

The batting, bowling and country effect estimates are outlined in Table 2. The parameter estimates associated with the expected score by an average team on a neutral ground, the common home advantage and any advantage gained by batting first are estimated to be 306, 28 and -11 runs, respectively. The p-value for the common home advantage parameter is 0.002, which suggests that the home team, on average, gained a significant first innings runs advantage, whereas the p-value for the batting first parameter is 0.208, which suggests that there is no significant batting order effect.

The long-term dominance of Australia and South Africa in test match cricket is clearly evident, with both teams enjoying batting and bowling ratings substantially above average. All other teams have under-performed in one or both of these areas. Notably, India has performed exceptionally well with the bat but has been let down by relatively poor bowling performances. The negative country effect ratings for India, South Africa and the West Indies suggest that the batting teams playing in these countries were disadvantaged to some degree by the conditions. This latter point highlights India's excellent batting form, particularly when playing at home.

Team	Batting rating	Bowling rating	Country effect rating
India	148	73	−14
Australia	142	145	7
South Africa	132	143	−16
England	96	75	10
Pakistan	94	105	1
West Indies	93	118	−2
Sri Lanka	84	96	8
New Zealand	70	81	3
Zimbabwe	42	63	3

Table 2: First innings ratings.

To show how the model can be applied, assume Australia is playing South Africa at home, with Australia batting first. The model predicts Australia's first innings score to be $306 + 142 - 143 + 28 - 11 + 7 = 329$ runs, whereas South Africa's first innings score is estimated to be $306 + 132 - 145 + 7 = 300$ runs, an advantage to Australia of 29 runs. However, if the match were to be played in South Africa and Australia remains the team batting first, the model predicts Australia's first innings score to be $306 + 142 - 143 - 11 - 16 = 278$ runs, whereas South Africa's first innings score is estimated to be

$306 + 132 - 145 + 28 - 16 = 305$ runs. This time there is an advantage to South Africa of 27 runs. This indicates that with Australia and South Africa being both highly rated teams factors that are indirectly related to a team's batting and bowling performance such as winning the toss or home advantage could have a potentially significant impact on match outcomes.

4 Modelling the second innings

Since the match result is a categorical variable, a logistic regression model is used to model the outcome of a match. The response variable is the match outcome for Team 1 and the explanatory variables are Team 1's lead, the cumulative duration of Team 1's and Team 2's innings, a home team indicator where 1 indicates that Team 1 is the home team and 0 otherwise, the difference in the batting and bowling ratings for competing teams and the country effect. The model is expressed as

$$\ln\left(\frac{\gamma}{1-\gamma}\right) = \beta_0 + \beta_1 l + \beta_2 t + \beta_3 h + \beta_4 d_A + \beta_5 d_B + \beta_6 c + \epsilon, \tag{2}$$

where the response variable γ represents the probability of a win for Team 1. The parameter l signifies the lead enjoyed by Team 1; t represents the innings duration parameter; h indicates whether Team 1 was the home team; d_A and d_B represent the rating differential parameters for the batting and bowling teams; c is the country effect parameter; and ϵ is a zero-mean random error. We will use nominal logistic regression to investigate the three comparisons: win/loss, draw/loss and win/draw. The results are outlined in Tables 3 to 5.

The analysis suggests that the first innings lead contributes significantly to the shaping of a win rather than a loss or a draw. There is also strong evidence suggesting that the home team is also significantly more likely to generate a win rather than a loss or a draw. There is also evidence to suggest that the cumulative time taken to complete each of the first innings is more likely to produce a drawn result rather than a win or a loss. There is some marginal evidence suggesting that Team 2, which generally bats last, is more likely to manufacture a loss rather than a drawn result.

	Parameter	Coefficient	p-value
β_0	Intercept term	-0.1598	0.818
β_1	Lead	0.013797	0.000
β_2	Time	-0.0005405	0.499
β_3	Home	1.0362	0.003
β_4	Rating differential (bat 1 − bowl 2)	0.004729	0.401
β_5	Rating differential (bat 2 − bowl 1)	-0.008445	0.123
β_6	Country effect	0.02493	0.241

Table 3: Results for comparison of win/loss.

	Parameter	Coefficient	p-value
β_0	Intercept term	-4.5721	0.000
β_1	Lead	0.007060	0.000
β_2	Time	0.0045660	0.000
β_3	Home	0.9953	0.002
β_4	Rating differential (bat 1 − bowl 2)	0.001076	0.835
β_5	Rating differential (bat 2 − bowl 1)	-0.009724	0.063
β_6	Country effect	0.02504	0.209

Table 4: Results for comparison of draw/loss.

	Parameter	Coefficient	p-value
β_0	Intercept term	4.4123	0.000
β_1	Lead	0.006736	0.000
β_2	Time	−0.0051065	0.000
β_3	Home	0.0409	0.899
β_4	Rating differential (bat 1 − bowl 2)	0.003654	0.475
β_5	Rating differential (bat 2 − bowl 1)	0.001279	0.800
β_6	Country effect	−0.00011	0.995

Table 5: Results for comparison of win/draw.

To get a sense of the effect of batting first we need to restate the model without the time parameter. This is necessary since when the time parameter is set to zero it has little meaning in the context of investigating any perceived advantage for Team 1. The model is re-expressed as

$$\ln\left(\frac{\gamma}{1-\gamma}\right) = \beta_0 + \beta_1 l + \beta_2 h + \beta_3 d_A + \beta_4 d_B + \beta_5 c + \epsilon. \tag{3}$$

Setting all parameters to zero in effect represents Team 1's advantage at the completion of the first innings with all things being equal, i.e. no lead, playing on a neutral ground in a neutral country and equal ratings in the batting and bowling departments. Using nominal logistic regression, the parameter estimates for the comparison of a win/loss, draw/loss and win/draw are, respectively, −0.5446 (p-value = 0.026), −0.2837 (p-value = 0.194) and −0.2610 (p-value = 0.283). With all things being equal, there is a significant batting order effect, with Team 1 more likely to lose a match rather than win. This suggests that generally the team batting last in the match shows a tendency to win rather than lose. Note that this slightly contradicts the notion outlined in Section 2, which suggested that the effect was not significant. This possibly highlights the cumulative effects such as home advantage, a first innings lead and batting and bowling strength may have on a team's overall performance.

5 Analysis of the first innings lead and innings duration

The influence of the first innings lead established by Team 1 can be modelled as

$$\ln\left(\frac{\gamma_m}{1-\gamma_m}\right) = \beta_0 + \beta_1 l + \beta_2 h + \epsilon_m, \tag{4}$$

where γ_m is the probability of a particular match outcome for Team 1 with $m = 1, 2, 3$ for a win, draw and loss respectively. If a loss for Team 1 signifies the reference event, then, using nominal logistic regression, the probability of a win for Team 1 when Team 1 is the home team is expressed as

$$\gamma_1 = \frac{\exp(\beta_0 + \beta_1 l_1 + \beta_2 h_1 + \epsilon_1)}{1 + \sum_{m=1}^{2} \exp(\beta_0 + \beta_1 l_m + \beta_2 h_m + \epsilon_m)}. \tag{5}$$

The probability of a draw is

$$\gamma_2 = \frac{\exp(\beta_0 + \beta_1 l_2 + \beta_2 h_2 + \epsilon_2)}{1 + \sum_{m=1}^{2} \exp(\beta_0 + \beta_1 l_m + \beta_2 h_m + \epsilon_m)}. \tag{6}$$

The probability of a loss (the reference event) is

$$\gamma_3 = \frac{1}{1 + \sum_{m=1}^{2} \exp(\beta_0 + \beta_1 l_m + \beta_2 h_m + \epsilon_m)} \tag{7}$$

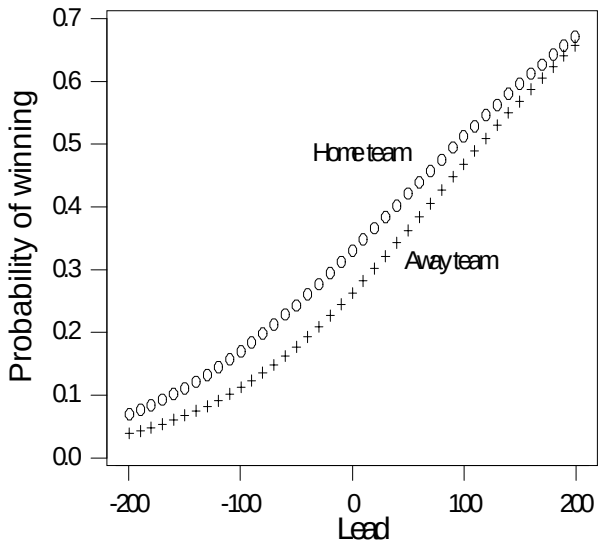


Figure 1: Plot of the probability of winning for the home and away teams.

The results of the analysis are provided in Tables 6 and 7. Notably, the parameter estimate for the lead term in both instances is significant, which confirms that after a team has established a first innings lead they are more likely to win or draw rather than lose a match. Similarly, the home team term is also significant, which suggests that the home team is also more likely to win or draw rather than lose a match. This suggests that the home team has a distinct winning advantage over the away team.

	Parameter	Coefficient	p-value
β_0	Intercept term	-0.4458	0.056
β_1	Lead	0.013040	0.000
β_2	Home	0.8160	0.011

Table 6: Comparison of win/loss.

	Parameter	Coefficient	p-value
β_0	Intercept term	-0.2261	0.285
β_1	Lead	0.007436	0.000
β_2	Home	0.8916	0.002

Table 7: Comparison of draw/loss.

If we let the lead term be zero, so that both teams have the same first innings score, and we apply formulas (5), (6) and (7), then the respective probabilities of a win, draw and loss for Team 1, when Team 1 is the home team, are 0.330, 0.443 and 0.228. These results suggest that after the completion of the first innings, with all things being equal, the home team displays a tendency to win or draw rather than lose a match. A drawn result is the more likely outcome. However, if the lead is increased to 100 runs, say, then the respective probabilities of a win, draw or loss are 0.512, 0.392 and 0.096. As the first innings lead increases, the probability of a win for the home team markedly increases, whereas the probabilities of a draw or loss decrease. To determine the lead the home team needs to establish in order to have a better than even chance of winning we let $\gamma_1 = 0.50$ and the lead in runs be x, such

	Parameter	Coefficient	p-value
β_0	Intercept term	4.2245	0.000
β_1	Time	−0.004312	0.000

Table 8: Comparison of win/draw.

	Parameter	Coefficient	p-value
β_0	Intercept term	4.2242	0.000
β_1	Time	−0.0041396	0.000

Table 9: Comparison of loss/draw.

that

$$0.50 = \frac{\exp(-0.4458 + 0.013040x + 0.8160)}{1 + \exp(-0.4458 + 0.013040x + 0.8160) + \exp(-0.2261 + 0.007436x + 0.8916)}$$

This gives $x = 93$ runs. This suggests that the home team, on average, needs to establish a lead in excess of 93 runs to have a better than even chance of winning. If we repeat this for Team 1 when Team 1 is the away team, this gives $x = 115$ runs. Figure 1 provides a plot of the probabilities of winning for leads up to 200 runs. Clearly, the probability of winning increases for both the home and away teams as the first innings lead increases, with the home team having the upper hand.

The first innings duration can be modelled as

$$\ln\left(\frac{\gamma_m}{1-\gamma_m}\right) = \beta_0 + \beta_1 t_m + \epsilon_m, \tag{8}$$

where γ_m is the probability of a particular match outcome at home with $m = 1, 2, 3$ for a win, loss or draw. If a draw signifies the reference event then using nominal logistic regression the probability of a draw is

$$\gamma_3 = \frac{1}{1 + \sum_{m=1}^{2} \exp(\beta_0 + \beta_1 t_m + \epsilon_m)} \tag{9}$$

Tables 8 and 9 provide a summary of the parameter estimates. Both the parameter estimates for duration are highly significant, with the negative coefficients confirming that the combined duration of the first innings is more likely to shape a draw than a win or a loss. Figure 2 provides a plot of the probability of a draw for Team 1 for durations up to 2000 minutes. Clearly, the likelihood of a draw increases as the duration increases. To determine when there is a better than even chance of a draw let the duration be t, such that

$$0.50 = \frac{1}{1 + \exp(4.2245 - 0.004312t) + \exp(4.2242 - 0.0041396t)}$$

giving $t = 1165$ minutes. We calculated the average duration of an over to be 4.2 minutes, with 1165 minutes equating to approximately 277 overs. This suggests that there is a better than even chance of a draw for a duration in excess of approximately 1165 minutes (or about 277 overs). Conversely, there is a better than even chance of a result for durations less than 1165 minutes.

6 Conclusions

Of the many factors shaping test cricket, there is strong evidence to suggest that the factors which impact most significantly on the outcome of a match are a team's first innings lead, home team performance and the duration of the first innings. Clearly, a team is more likely to win a match after they have established a first innings lead, with the probability of winning increasing as the lead increases. To have

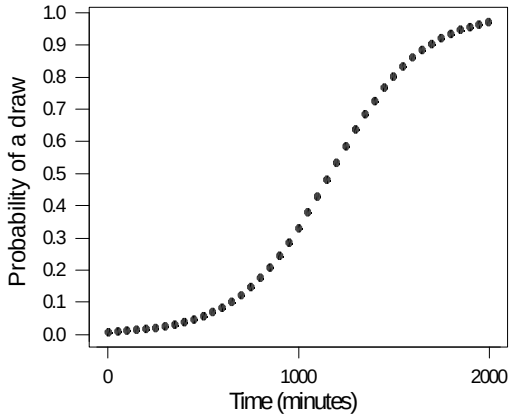


Figure 2: Plot of the probability of a draw against the total time for the first innings.

a better than even chance of winning the home team, on average, needs to establish a lead in excess of 93 runs, whereas the away team needs a lead in excess of 115 runs to have the same chance.

The home team is more likely to win a match rather than lose or draw, which suggests that the home team has a distinct winning advantage over the away team. There is some evidence to suggest that teams gain an advantage by batting second and so are possibly advantaged by batting last in the second innings. This contradicts the accepted wisdom that batting last is a disadvantage. This is an area that encourages more detailed research.

The time taken to complete each of the first innings contributes more to the shaping of a drawn result rather than a win or a loss, with the likelihood of a draw increasing as the duration of the first innings increases. For durations in excess of 1165 minutes (or approximately 277 overs) there is shown to be a better than an even chance of a draw.

References

S. R. Clarke and P. Allsopp (2001), "Fair measures of performance: The World Cup of cricket", *J. Oper. Res. Soc.*, **52**, 471–479.

S. R. Clarke and J. M. Norman (1995), "Home ground advantage of individual clubs in English soccer", *The Statist.*, **44**, 509–521.

B. M. de Silva, G. R. Pond and T. B. Swartz (2000), "Applications of the Duckworth–Lewis method", in *Proceedings of the Fifth Australian Conference on Mathematics and Computers in Sport*, G. Cohen and T. Langtry (editors), University of Technology, Sydney, 113–117.

D. A. Harville and M. H. Smith (1994), "The home-court advantage: How large is it, and does it vary from team to team?", *Amer. Statist.*, **48**, 22–28.

PREDICTING THE BROWNLOW MEDAL WINNER

Michael Bailey and Stephen R. Clarke
School of Mathematical Sciences
Swinburne University of Technology
PO Box 218, Hawthorn
Victoria 3122, Australia
mbailey@groupwise.swin.edu.au, sclarke@groupwise.swin.edu.au

Abstract

The Brownlow medal is the highest individual honour that can be achieved by Australian Football League (AFL) players. It is based on the umpires votes to the three best players (3 for first, 2 for second, 1 for third) in each of the 176 home and away matches for a season. An ordinal logistic regression model retrospectively applied to past data has been used to identify specific player performance statistics from each match that can aid in the prediction of votes polled. By applying this model to present data it is possible to objectively assign leading players a probability of winning the Brownlow medal. Over the past two AFL seasons (2000 and 2001), the authors have successfully used this approach to identify the leading contenders for the Brownlow medal.

1 Introduction

During the 2000 Australian Football League (AFL) season, discussion arose between the authors as to the best possible way to predict the winner of the Brownlow medal, both before and during the actual count. The aim was to complement existing predictions for football, tennis and other sports on our web site www.swin.edu.au/sport. The original idea was to collect votes from the media, and then perform a resampling simulation to estimate the chances of winning. The concept was to update predictions on the night, by gradually replacing simulated results with real ones. However, we decided to include a large amount of match performance statistics readily available, as we felt that a mathematical modelling process might well assist in the objective assignment of a player's probability of polling votes.

Based on data collected from the 1997, 1998, and 1999 seasons an ordinal logistic regression model was constructed and applied to the 2000 season. Predicted votes for each match were then tallied over the season to provide players predicted totals for the year. During the course of the 2000 Brownlow count, Swinburne Sports Statistics provided updated online predictions that combined predicted and actual totals throughout the course of the evening. Following considerable success and media attention [1, 2] this modelling process was further enhanced for the 2001 season. With the addition of an extra year's data and several additional variables, this modelling process was able to clearly identify the three leading candidates for the 2001 Brownlow medal, and was widely publicised prior to the count [3, 4]. This paper describes the modelling process. All analysis was performed using SAS version 8.0.[1]

[1] SAS Institute Inc, Cary, NC, USA.

2 Database

A database was constructed that comprised data collected from each regular season AFL match played between 1997 and 2001 (880 games). For each game, in addition to team scores, an array of individual match statistics are readily available, both in the newspapers and via the internet. An example is given in Table 1.

Name	TK	TH	DI	RE	IN50	MA	HO	CL	TO	FF	FA	TK	G
J. Akermanis	9	10	19	1	0	5	1	2	1	1	0	0	1
M. Ashcroft	10	8	18	1	2	2	0	9	0	1	1	0	0
C. Bolton	7	2	9	3	2	1	0	1	1	0	0	0	0
D. Bradshaw	4	1	5	1	0	0	0	0	1	1	3	0	3
R. Champion	5	2	7	3	0	3	0	0	3	1	2	3	1
D. Cupido	2	2	4	2	0	0	0	0	2	0	2	2	1
S. Hart	6	5	11	3	1	0	0	1	0	0	0	1	0

Key: TK = total kicks, TH = total handballs, DI = disposals, RE = rebounds, IN50 = times inside 50 metre zone, MA = marks, HO = hit outs, CL = clearances, TO = turnovers, FF = frees for, FA = frees against, TK = tackles, G = goals.

Table 1: Example of individual match statistics.

In addition, the AFL web site gives a list of the six best players from each side. The data were combined with team statistics, such as match result (win or loss), team score, margin and number of scoring shots.

3 Univariate analysis

Initially, a descriptive statistical analysis was undertaken, with each possible statistic investigated in isolation to determine its predictive capability.

Disposals: The number of disposals (kicks + handballs) that a player accumulates during the course of a match is the strongest predictor for polling votes. This is reflected in the fact that the leading possession winner for each match has a 51% chance of polling votes, with the leading possession winner of the winning side having a 64% chance of polling votes.

Goals: Not surprisingly, the number of goals that a player kicks during the course of a match is a significant predictor for votes. What is surprising is that the number of goals does not carry as much weight as some may think, with only 40% of players kicking five goals for the game managing to poll votes! If a player kicks six goals, his chance of polling increases to 62%, and with seven goals, chances of polling increase to 79%. Only one player in the past five years has kicked eight or more goals and not been awarded a vote. Figure 1 shows the chance of polling votes depending on the number of goals kicked.

Won the game / Margin of victory: Umpires have a clear tendency to award votes to the winning side, with 92% of 3-votes being awarded to a player from the winning side. Similarly, 83% of 2-votes and 76% of 1-votes are awarded to players from the winning side. To a lesser extent, the margin of victory is also important in determining the voting probabilities.

Hit outs: Ruckmen have traditionally polled well in the Brownlow medal, as a good ruckman can have a big influence on the match outcome without gathering a lot of possessions. Not surprisingly, the number of hit outs that a ruckman has during a match is a strong predictor for polling votes.

Quality of disposals: Although the number of disposals that a player has is the strongest predictor for votes, this variable fails to take into account the quality of the disposals. As the quality of a disposal

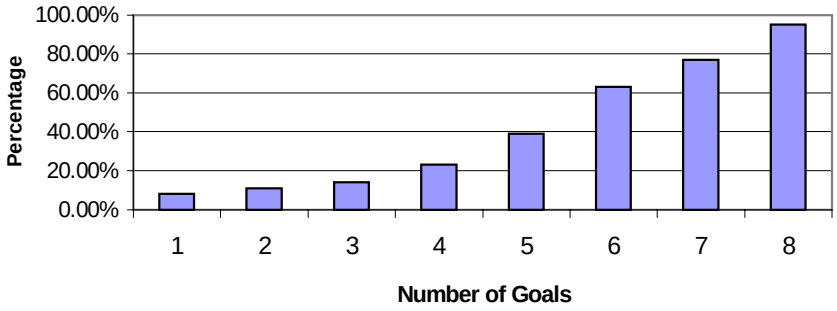


Figure 1: Relationship between polling votes and kicking goals.

is subjective, it will always be difficult to measure accurately. Rather than measure the quality of the disposal, the authors have endeavoured to measure the quality of the player having the disposal. By constructing a prediction model based solely on the number of disposals, it is possible to identify players that poll more votes than quantity alone predicts. By then measuring the proportion of times that this occurs for each player, it is possible to gain an indirect measurement of the quality of the player in question. This measure was included as a predictor.

Best players: Any prediction model based solely on player statistics would fail to take into account how well a player has performed in comparison to his direct opponent, which would in turn clearly bias against defenders. To compensate against this, the best players in each team as given by the AFL web site are included in the model, and are a significant predictor for votes, with only 12% of players awarded 3-votes not being named in their team's best six players. Similarly, the order in which players are named is also of importance. This can be seen from Figure 2, with the best named player from the winning side having a 60% chance of polling votes, whereas the best named player from the losing side only has a 21% chance of polling votes.

Distinct appearance: It can be statistically shown that any player with a distinctive appearance has a slightly higher chance of polling votes in comparison to their non-distinctive teammates. Distinctive appearance has been quantified as any player with red or blond hair or a significantly darker or lighter skin colour. An indicator variable was used to flag these players.

Team scoring shots: If a team has a lot of scoring shots on goal, this acts as an indirect measure of the number of players in the side that are playing well, as both midfielders and forwards are required

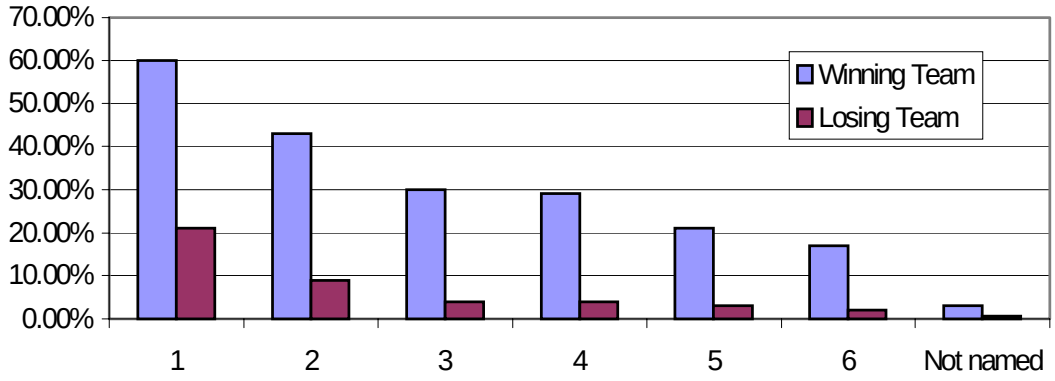


Figure 2: Chance of AFL website best players polling votes.

to be successful for shots on goal to eventuate. Consequently, the greater the number of scoring shots, the more players that are playing well and thus the more difficult it is for any given player from that side to poll votes.

Rebounds / Frees for / Marks / Inside 50s: Rebounds — the number of times a player rebounds the ball from defence into attack. Frees for — the number of free kicks awarded to a player. Marks — the number of marks that a player takes. Inside 50 — the number of times that a player propels the ball inside the forward 50 metre arc. Although only having small effects on a player's probability of polling votes, all of the above four variables are statistically significant predictors.

Variable	Univariate OR	Multivariate OR*	Lower95†	Upper95†
Disposals	1.21	1.13	1.12	1.15
Goals	1.67	1.45[1]	1.40	1.50
Best players	1.76	1.29	1.26	1.32
Win	5.67	4.72	3.98	5.56
Hit-outs	1.03	1.05	1.04	1.06
Quality of disposal	1.01	1.01	1.01	1.01
Marks	1.39	1.08	1.06	1.10
Margin of victory	1.02	1.03	1.02	1.03
Team scoring shots	1.08	0.97	0.96	0.98
Rebounds	1.17	1.08	1.05	1.11
Distinct	2.49	1.41	1.24	1.61
Frees for	1.50	1.11	1.06	1.16
Inside 50	1.43	1.05	1.02	1.07

OR represents the odds ratio associated with a one unit increase in the variable in question.
*Multivariate odds ratio adjusting for all other variables.
†Upper and lower 95% confidence intervals for the multivariate odds ratio.

Table 2: Odds ratios for polling votes.

4 Ordinal logistic regression

Using the number of votes polled as the outcome, an ordinal logistic regression was applied to the past data to ascertain the relative importance of all predictors in predicting votes polled. Each variable in the model was found to be statistically significant at a level of $p = 0.001$. The percentage of the explained variation attributable to each predictor is shown in Figure 3. Disposals, goals, best players and being on the winning side are the most important variables, contributing between them over 75% of the explained variation. Table 2 gives the odds ratios for each predictor. Thus playing in the winning team increases a player's odds of polling nearly five fold, while every additional goal that a player kicks during a match increases his odds of polling votes by 45%. Having a distinctive appearance also increases the odds of polling by over 40%.

5 Predicting the winner

By applying this model to current season data it was possible to assign to each player for each match a probability that they would poll one, two or three votes. Probabilities were standardised so that the match probabilities summed to one. A predicted value for the number of votes was created by the

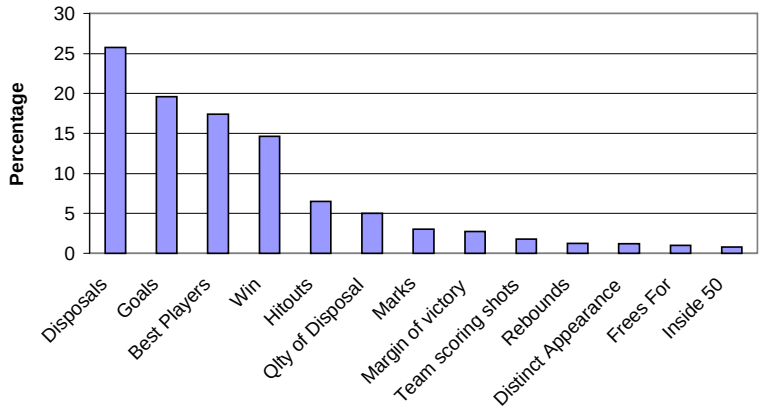


Figure 3: The relative importance of variables in the multivariate model as determined by the WALD chi-square statistic.

following formula:

$$\text{Predicted votes} = 3 \times \text{Prob}(3 \text{ votes}) + 2 \times \text{Prob}(2 \text{ votes}) + \text{Prob}(1 \text{ vote}).$$

By then tallying each player's predicted number of votes, it was possible to derive a predicted total for the season. Investigation of the predicted totals for the leading 25 players indicated that the error in the number of votes was found to be approximated by a normal distribution with a standard deviation of approximately five votes. By simulating 10,000 seasons and counting the number of occasions that each player would have won based on his predicted total, it was possible to give each of the leading players a probability that they would win.

6 Model accuracy

By ranking players in each game according to their predicted probability, it is possible to measure the accuracy of the modelling process. For the 2001 season, Figure 4 shows the chance of a player ranked

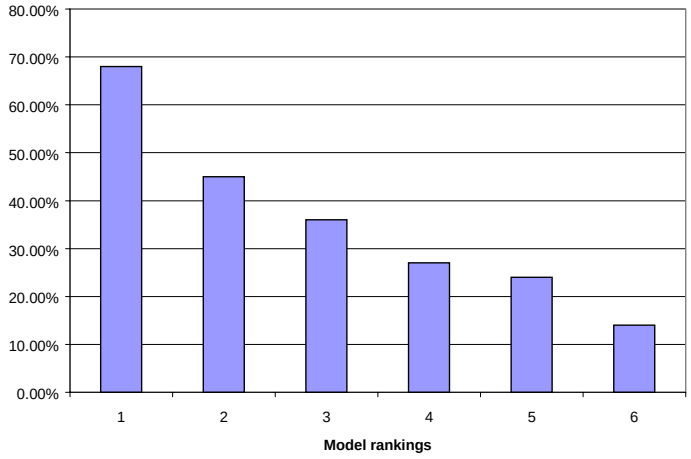


Figure 4: Probability of polling votes according to model rankings.

Name	Chance of winning (%)	Predicted total	Actual total	Ranked 1st by model	three votes received	Highly rated	2 & 3 votes received
A. McLeod	17	18.6	21	6	6	5	7
M. Voss	15	18.1	19	4	3	3	7
J. Akermanis	14	17.9	23	6	5	7	9
N. Buckley	8	16.6	14	6	3	7	5
S. West	6	15.6	8	4	0	3	4
B. Cousins	6	15.5	18	5	4	5	6
B. Ratten	5	15.2	8	5	2	3	3
J. Bowden	5	15.0	9	6	2	5	3
W. Campbell	4	14.3	4	3	0	1	1
J. Hird	4	14.2	5	3	0	3	2
N. Stevens	4	14.0	11	3	3	3	3
N. Lappin	3	13.6	4	3	1	3	1
S. Crawford*	0*	12.6	14	5	3	4	5
S. Black	2	12.6	12	1	2	2	4
M. Lloyd*	0*	12.0	15	5	3	4	5
B. Ottens	1	12.0	6	3	2	3	2
J. Johnson	1	11.9	15	2	4	2	5
B. Harvey	1	11.4	16	1	2	1	6
A. Koutoufidis	1	11.3	8	4	2	4	2
P. Bell	0.5	10.5	11	3	3	0	4
S. Grant	0.5	10.5	14	3	4	2	5
W. Tredrea	0.5	10.4	8	3	2	1	3
S. Goodwin	0.5	10.3	10	1	1	2	4
L. Darcy	0.4	9.3	11	1	3	1	4
D. Cresswell	0.3	9.2	8	4	2	3	2

*Players suspended during the 2001 season are ineligible to win the Brownlow.

- *Predicted total*: By aggregating the predicted probability for each player for each match, it is possible to derive a predicted total for the year. Although the final vote total must be an integer, the predicted total has been left to one decimal place to indicated slight differences between players.
- *Chance of winning*: By simulating results 10,000 times it is possible to derive a player's probability of winning based on his predicted vote total.
- *Ranked 1st*: Based on the fitted model, it is possible to rank the 44 players in each match according to their probability of polling votes. Ranked 1st represents the number of matches in which each player is ranked as the leading player on the ground.
- *Highly rated*: If a player has a greater than 50% chance of polling two or three votes for a match then he is considered to have played a game that is highly rated.

Table 3: Predicted leading vote getters for 2001 Brownlow medal.

by the model polling votes. Thus the leading player on the field, as ranked by the modelling process, had a 68% chance of polling votes, with the second ranked player having a 45% chance of polling votes.

The predicted total also offered a clear indication towards performance, with any player with a total of two or more having an 87% chance of polling votes, and a 74% chance of polling at least two votes.

7 Results and discussion

Table 3 shows a comparison of the model prediction and actual results for the leading vote getters in the 2001 Brownlow. It can be seen that the ordinal logistic regression approach offers an excellent insight into the leading players for the Brownlow medal with 15 of the leading 25 predicted players actually finishing in the top 25. Based on these predictions it was the authors' belief that the ultimate winner would come from one of the top three ranked players. As it turned out, the top three predicted players were the top three vote pollers. Although Akermanis was given a slightly lesser chance than both Voss and McLeod to win, he was quickly elevated to favouritism as the count progressed, as he polled more votes than predicted early in the count. For the second year in a row, the interactive process of updating a player's predicted total by combining his predicted with actual votes proved to be a tremendous success, with the accuracy of the prediction process constantly improving as the count unfolded.

Although interactive betting throughout the course of the Brownlow was not available for the 2001 Brownlow medal count, the ordinal logistic regression model provided an objective price setting approach prior the commencement of the count. Of the three leading contenders, Andrew McLeod was a 6 to 4 favourite (40% chance) with the bookmakers, Michael Voss was second favourite at 3 to 1 (25% chance) while Akermanis was the model's recommended value bet of the threesome at the odds of 11 to 1 (8% chance). A further significant application of the ordinal model was the determination of team totals for the Brownlow. Although relatively large differences can occur between individuals' predicted and actual totals, when aggregated over the team the level of accuracy was found to improve substantially.

It was interesting to note that several of the leading players that did not poll as many votes as predicted by the model were in actual fact high vote pollers from the previous season (Buckley, West, Ratten and Hird). This suggests the possibility that following a successful year the expectations of quality players may subconsciously be raised by umpires.

Further improvement can be expected in future years, with preliminary results indicating that model improvement may be achieved with the use of an Inverse Normal link function for the logistic regression as opposed to the default Logit link function. When considering all players, the modelling approach is quite accurate with the predicted seasonal total explaining over 50% of the variation associated with the actual total. Unfortunately, amongst the leading players, this process is not quite as accurate, as the winning players will inevitable poll more votes that the model can predict. This modelling process can be seen to have clear applications for bookmakers, punters and fans who are keen to ascertain the final result prior to the completion of the count.

References

[1] Anon., "Being blond gets votes", *Herald Sun*, Melbourne (29 August 2000).

[2] Anon., "Hair's to all those Brownlow hopefuls", *The Age* (Sport section), Melbourne (23 August 2000), 3.

[3] Anon., "Following the countdown to the Brownlow", *Campus Review*, Brisbane (5–11 September 2001), 22.

[4] M. Bailey, "Who'll win the Brownlow: Permutations, combinations, probabilities and guesswork", *AFL Record* (2001), 80–81.

USING MICROSOFT® EXCEL TO MODEL A TENNIS MATCH

Tristan J. Barnett and Stephen R. Clarke
School of Mathematical Sciences
Swinburne University
PO Box 218, Hawthorn
Victoria 3122, Australia
tbarnett@groupwise.swin.edu.au, sclarke@groupwise.swin.edu.au

Abstract

This paper demonstrates the use of Microsoft® Excel to generate the probability of winning a tennis match and mean length of the remainder of the match conditional on the state of the match. Previous models treat games, sets and matches independently. We show how a series of interconnected sheets can be used to repeat these results. We also set up a sheet which can give the required statistics using the full match, set and game score as the state. The exercise could form an interesting and useful teaching example, and allow students to investigate the properties of tennis scoring systems.

1 Introduction

Many papers have investigated the characteristics of various scoring systems, particularly in racquet sports. The game of Lawn Tennis has a history of changing formats. Although it uses a basic nested system of points, games, sets and matches, there is much variation within that basic structure. Matches can be one set, or best of three or five sets, sets can be first to six, eight or ten games, advantage or non-advantage, and games can be the traditional four point advantage, or seven or ten point tiebreakers. We even have the recent innovation at the 2001 Australian Open mixed, where the third set becomes a single tiebreaker game. This innovation is extending into the Victorian Tennis Association pennant this year.

The basic principles involved in modelling a tennis match are well known. A Markov chain model with a constant probability of winning a point was set up by Schutz [4]. While such a model is acceptable within a game, a model which allows a player a different probability of winning depending on whether they are serving or receiving is essential for tennis. Statistics of interest are usually the chance of each player winning, and the expected length of the match. Croucher [1] looks at the conditional probabilities for either player winning a single game from any position. Pollard [3] uses a more analytic approach to calculate the probability for either player winning a game or set along with the expected number of points or games to be played and their corresponding variance.

Most of the previous work uses analytical methods, and treats each scoring unit independently. This results in limited tables of statistics. Thus the chance of winning a game and the expected number of points remaining in the game is calculated at the various scores within a game. The chance of winning a set and the expected number of games remaining in the set is calculated only after a completed game and would not show for example the probability of a player's chance of winning from three games to two, 15–30.

This paper discusses the use of Microsoft® Excel to repeat these applications using a set of interrelated spreadsheets. This allows any probabilities to be entered and the resultant statistics automatically

calculated or tabulated. In addition, more complicated workbooks can be set up which allow the calculation of the chance of winning a match and the expected number of points in the remainder of the match at any stage of the match given by the match, set and game score. These allow the dynamic updating of player's chances as a match progresses.

2 Simple Model

We explain the method by first looking at the simple model where we have two players, A and B, and player A has a constant probability p of winning a point. We set up a Markov chain model of a game where the state of the game is the current game score in points (thus 40–30 is 3–2). With probability p the state changes from (a, b) to $(a+1, b)$ and with probability $1-p$ it changes from (a, b) to $(a, b+1)$. Thus if $P(a, b)$ is the probability that player A wins when the score is (a, b), we have

$$P(a, b) = pP(a+1, b) + (1-p)P(a, b+1).$$

The boundary values are $P(a, b) = 1$ if $a = 4$, $b \leq 2$, $P(a, b) = 0$ if $b = 4$, $a \leq 2$. The boundary values and formula can be entered on a simple spreadsheet. The problem of deuce can be handled in two ways. Since deuce is logically equivalent to 30–30, a formula for this can be entered in the deuce cell. This creates a circular cell reference, but the iterative function of Excel can be turned on, and Excel will iterate to a solution. Alternatively, it is easily shown that the chance of winning from deuce is

$$\frac{p^2}{p^2 + (1-p)^2}.$$

Entering this formula removes the need for iteration. Table 1 shows the results obtained, using a value of $p = 0.54$. It indicates that a player with a 54% chance of winning a point has a 60% chance of winning the game. Note that since advantage server is logically equivalent to 40–30, and advantage receiver is logically equivalent to 30–40, the required statistics can be found from these cells.

		A score				
		0	15	30	40	game
	0	0.60	0.74	0.87	0.96	1
	15	0.44	0.59	0.76	0.91	1
B score	30	0.25	0.39	0.58	0.81	1
	40	0.09	0.17	0.31	0.58	
	game	0	0	0		

Table 1: The conditional probabilities of A winning the game from various scorelines.

The mean number of points in the remainder of the game, $M(a, b)$, is handled similarly as shown in Table 2. All boundary points become 0 and the recurrence relation becomes

$$M(a, b) = 1 + pM(a+1, b) + (1-p)M(a, b+1).$$

The formula for the number of points from deuce is

$$\frac{2}{p^2 + (1-p)^2}.$$

A similar spreadsheet can be set up for a set. The constant probability of winning a game can be obtained from the (0,0) cell of the previous sheet. Similarly for a match. Thus we obtain six connected sheets—entering the chance of winning a point results in the chance of winning a game and expected

		A score				
		0	15	30	40	game
	0	6.7	5.6	4.0	2.1	0
	15	5.9	5.2	4.1	2.3	0
B score	30	4.5	4.4	4.0	2.8	0
	40	2.5	2.7	3.1	4.0	
	game	0	0	0		

Table 2: The expected number of points remaining in a game from various scorelines.

number of points in the remainder of the game conditional on the game score, the chance of winning a set and expected number of games in the remainder of the set conditional on the set score, and the chance of winning a match and expected number of sets in the remainder of the match conditional on the match score.

In this case, the sheet suggests a player with a 54% chance of winning a point, has a 59.9% chance of winning a game, a 76.3% chance of winning a tiebreak set, a 85.9% chance of winning a three-set match and a 91% chance of winning a five-set match.

Excel has excellent facilities for building up one or two way tables. Table 3 shows some of the match statistics for chances and expected lengths of winning games, sets and matches with the probability of winning a point ranging from 0.5 to 1. We use the following notation:

$$\begin{aligned} p(\text{Game}) &= \text{probability of winning a game,} \\ M(\text{Game}) &= \text{expected number of points in a game,} \\ p(\text{Set}) &= \text{probability of winning a tiebreak set,} \\ M(\text{Set}) &= \text{expected number of games in a tiebreak set,} \\ p(\text{Match}) &= \text{probability of winning a 5 set match,} \\ M(\text{Match}) &= \text{expected number of sets in a 5 set match.} \end{aligned}$$

Magnus and Klassen [2] tested some often-heard hypotheses relating to the service in tennis. They collected data on 481 matches played in the men's singles and women's singles championships at Wimbledon from 1992 to 1995. We will compare some of their statistics with our model. In men's singles with two seeded players, on service they won 67% of the points and 86% of the games. In women's singles with two seeded players, on service they won 57% of the points and 67% of the games. Both these results agree with our model as highlighted in Table 3.

3 Two-parameter model

For more realistic models the process is essentially the same, just a little more complicated. We now have two parameters, p_A and p_B:

$$\begin{aligned} p_A &= \text{probability of A winning a point if A is serving,} \\ p_B &= \text{probability of B winning a point if B is serving.} \end{aligned}$$

Since the chance of a player winning now depends on who is serving, we need to introduce this into the state description. The state of a game becomes the score and whether the player is serving or receiving. Thus we need two sheets for a game, one for player A serving and one for player B serving. Similarly we need two sheets for the set score, and one for the match score. The formula and boundary conditions are similar. The main difference is that adjacent cells in terms of the Markov chain are not

p	p(Game)	M(Game)	p(Set)	M(Set)	p(Match)	M(Match)
0.50	0.50	6.8	0.50	9.7	0.50	4.1
0.52	0.55	6.7	0.56	9.6	0.75	4.0
0.54	0.60	6.7	0.62	9.3	0.91	3.7
0.56	0.65	6.7	0.68	9.0	0.98	3.4
0.57	0.67	6.6	0.73	8.8	1	3.3
0.58	0.69	6.6	0.74	8.6	1	3.2
0.60	0.74	6.5	0.79	8.1	1	3.1
0.62	0.78	6.4	0.83	7.7	1	3
0.64	0.81	6.3	0.87	7.4	1	3
0.66	0.85	6.1	0.90	7.1	1	3
0.67	0.86	6.1	0.92	7	1	3
0.68	0.88	6.0	0.93	6.9	1	3
0.70	0.90	5.8	0.95	6.7	1	3
0.80	0.98	5.1	1	6.1	1	3
0.90	1	4.5	1	6	1	3
1	1	4	1	6	1	3

Table 3: Match statistics depending on the probability of winning a point.

necessarily adjacent on the spreadsheet, since players alternate service games. We will also use the following notation:

$$p'_A = \text{probability of A winning a game if A is serving,}$$
$$p'_B = \text{probability of B winning a game if B is serving.}$$

The following probabilities are applied: $p_A = 0.6$, $p_B = 0.58$.

The conditional probabilities for an advantage set are given in Tables 4 and 5 for each player serving. The probability of A winning an advantage set from 6 games all, A or B serving, is

$$\frac{p'_A p'_B}{1 - p'_A p'_B + (1 - p'_A)(1 - p'_B)}$$

By substituting $p_A = 0.6$ in our simple model, $p'_A = 0.74$. Similarly by letting $p_B = 0.58$, $p'_B = 0.69$. Substituting these figures into the equation above gives 0.55.

		A score						
		0	1	2	3	4	5	6
	0	0.57	0.75	0.82	0.93	0.97	1	1
	1	0.49	0.57	0.77	0.84	0.96	0.99	1
	2	0.29	0.48	0.56	0.79	0.87	0.98	1
B score	3	0.20	0.25	0.46	0.58	0.82	0.92	1
	4	0.06	0.15	0.20	0.44	0.55	0.88	1
	5	0.02	0.03	0.09	0.13	0.41	0.55	0.88
	6	0	0	0	0	0	0.41	0.55

Table 4: The conditional probabilities of A winning an advantage set from various games scores if A is serving.

		A score						
		0	1	2	3	4	5	6
B score	0	0.57	0.64	0.82	0.88	0.97	0.99	1
	1	0.37	0.57	0.65	0.84	0.91	0.99	1
	2	0.29	0.35	0.56	0.65	0.87	0.94	1
	3	0.12	0.25	0.31	0.56	0.67	0.92	1
	4	0.06	0.08	0.20	0.26	0.55	0.69	1
	5	0.01	0.03	0.04	0.13	0.17	0.55	0.69
	6	0	0	0	0	0	0.17	0.55

Table 5: The conditional probabilities of A winning an advantage set from various game scores if B is serving.

As expected, the probabilities of winning from 6–6, 5–5 or 4–4 are the same regardless of who is serving. Also worth noting is that $P(\text{winning set} \mid \text{a score with an even number of games with A serving}) = P(\text{winning set} \mid \text{a score with an even number of games with B serving})$.

4 Six-parameter model

Instead of building up independent sheets linked through the probability of winning a point, game or set, we can build a model in which the state of the game is match score, set score, game score and whether the player is serving or receiving. Also included are first and second serves. With this model, a typical question could be: What is the probability of player A winning a five-set match with the current score at 2 sets to 1, 3 games to 4, 30–15 and fault one?

We now have six parameters: first serve % for each player, winning % on first serve for each player and winning % on second serve for each player.

The same techniques of deriving a formula for the last point of a game or set and allowing recurrence relations to work out the remaining points are applied.

We will take the statistics from the 2002 Australian Open men's singles final, where Thomas Johannson defeated Marat Safin 3–6, 6–4, 6–4, 7–6 (7–4). Let Johannson = player A, Safin = player B. The results were:

$$\begin{aligned}
\text{1st serve \% for player A} &= 56\%,\\
\text{1st serve \% for player B} &= 63\%,\\
\text{Winning \% on 1st Serve for player A} &= 86\%,\\
\text{Winning \% on 1st Serve for player B} &= 67\%,\\
\text{Winning \% on 2nd Serve for player A} &= 53\%,\\
\text{Winning \% on 2nd Serve for player B} &= 54\%.
\end{aligned}$$

Table 6 represents the conditional probabilities of player A winning a five-set match at a set all, 4–4 in the third set with player A serving. It highlights the importance of a first serve on certain points. At 30–40 a first serve in would give a 77% chance of winning the match, whereas 30–40 fault drops to 74%. Particularly in tiebreaks, this margin has greater emphasis on the outcome of a match.

5 Conclusions

Through the use of Excel we have shown how to produce the conditional probabilities of either player winning the match and the expected number of points, games and sets remaining from any position within the match.

		A score								
		0	0/2nd	15	15/2nd	30	30/2nd	40	40/2nd	game
	0	0.84	0.83	0.84	0.84	0.85	0.85	0.85	0.85	1
	15	0.81	0.80	0.83	0.82	0.84	0.84	0.85	0.85	1
B score	30	0.77	0.76	0.80	0.78	0.82	0.81	0.84	0.84	1
	40	0.71	0.70	0.74	0.71	0.77	0.74	0.82	0.81	
	game	0	0	0	0	0	0	0		

Table 6: Conditional probabilities of winning a game from various scorelines.

The use of "if statements" often used in Excel would allow us to enter what type of match it is (tiebreaker or advantage, three or five sets) and the current position in the match, to give us the corresponding probabilities and the expected length remaining. This idea would make accessible the relevant statistics for a match in progress. Microsoft® Excel now incorporates Visual Basic for Applications (VBA), enabling the modelling of a tennis match to be powerful and still relatively easy to implement.

References

[1] J. S. Croucher, "The conditional probability of winning games of tennis", *Res. Quart. for Exercise and Sport*, **57** (1986), 23–26.

[2] J. R. Magnus and F. J. G. M Klaassen, "On the advantage of serving first in a tennis set: four years at Wimbledon", *The Statist.*, **48** (1999), 247–256.

[3] G. H. Pollard, "An analysis of classical and tie-breaker tennis", *Austral. J. Statist.*, **25** (1983), 496–505.

[4] R. W. Schutz, "A mathematical model for evaluating scoring systems with specific reference to tennis", *Res. Quart. for Exercise and Sport*, **41** (1970), 552–561.

STUDYING THE BANKROLL IN SPORTS GAMBLING

D. Beaudoin, R. Insley and T. B. Swartz*
Department of Statistics and Actuarial Science
Simon Fraser University
Burnaby, British Columbia
Canada V5A 1S6
dbeaudoi@stat.sfu.ca, insley@stat.sfu.ca, tim@stat.sfu.ca

Abstract

This paper looks at various issues that are of interest to the sports gambler. We begin with an overview of sports betting and how two simple ideas can be used in the context of sports wagering. We then turn to questions concerning the management of one's bankroll. In particular, an expression is obtained for the distribution of the final bankroll using fixed wagers with both infinite and finite resources. Also, the Kelly method is discussed in the case of simultaneous bets placed under various gambling systems; a computational algorithm is presented to obtain the Kelly fractions.

1 Introduction

Despite its illegality in many jurisdictions, gambling on the outcome of sporting events is a common activity. For example, Crist (1998) states that Americans illegally wager over $100 billion annually on professional and college sports. On the 1999 Superbowl alone (the championship game of the National Football League), approximately $87 million was wagered legally in Las Vegas Sportsbooks (Ordine, 2000).

Sports gambling has been increasing over the years and it is likely that the trend will continue as more states and provinces are eager to share in the huge profits that have and can be made. One particular avenue for growth is sports gambling via the internet (Haywood, 2000). Whereas such internet sites are illegal in the United States and in Canada, they are legal and operational in various Caribbean and Latin American countries and in Australia. Internet gambling yields various murky legal issues such as the determination of where the wager is actually placed (e.g. in one's home in North America where the activity is illegal or in the offshore country).

In this paper, we put ethics and legal issues aside and take a look at various practical problems associated with sports betting. We are primarily concerned with issues involving wagering systems that have a positive expected gain. To a lesser extent, the results are also applicable to certain financial investments corresponding to an investor who engages in numerous transactions. Many fundamental probabilistic results concerning optimal systems for favourable games are reviewed in Thorp (1969).

In Section 2, we provide an introduction to pointspread wagering where we emphasise that it is possible to develop winning systems. Two simple ideas are provided that may assist in the development of winning systems. In Section 3, we investigate the results that one might achieve with fixed wagers under various conditions. Although the results only involve binomial calculations and a simple extension of the classical drunkard's walk problem, they are practical and we have not seen them recorded

*Swartz's work was partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada. This work was initiated during Beaudoin's tenure as a 2000 NSERC summer student at Simon Fraser University.

in the sports gambling literature. Moreover, much of the work concerning the gambler's ruin and related problems was developed long ago with an emphasis on challenging mathematics and approximate solutions. Our approach is based on the realisation that today nearly everyone has a computer on their desk. In Section 4, we turn to fixed percentage wagering and how one ought to wager in various practical scenarios. The results in this section are extensions of the Kelly system (Kelly, 1956) and require some computation. In Section 5, we provide a concluding discussion with advice applicable to typical Sportsbook scenarios where upper limits on wagering are imposed.

2 A primer on betting the pointspread

There are many types of wagers that can be placed on sporting events (McCune, 1989). Perhaps the most common is a wager placed against the pointspread. For example, consider a contest between a strong team (Team A) and a weak team (Team B). Whereas popular sentiment may overwhelmingly favour Team A to win, there is typically no concensus on the magnitude of the victory. To facilitate interest in wagering on such a match, a posted *line* may appear as

$$\begin{array}{lll} \text{Team A} & -l & -110 \\ \text{Team B} & +l & -110. \end{array} \tag{1}$$

The line in (1) is based on *American odds* and stipulates that a wager of \$110 placed on Team A returns the original \$110 plus an additional \$100 if Team A wins by more than l points. Alternatively, a wager of \$110 placed on Team B returns the original \$110 plus an additional \$100 if Team B wins or if Team B loses by less than l points. In the case where Team A wins by exactly l points, the original bets are returned. The quantity l is referred to as the *pointspread*.

We point out a few variations in the above example. First, different odds may be given. For example, odds of -120 stipulate that a winning wager of \$120 returns the original \$120 plus an additional \$100. When the odds are positive, this means that an event is less likely. For example, odds of $+140$ stipulate that a winning wager of \$100 returns the original \$100 plus an additional \$140. Second, wagers can be made in multiples or fractions of the amounts discussed. Third, a nearly equivalent expression of (1) is based on *European odds* and appears as

$$\begin{array}{lll} \text{Team A} & -l & 1.91 \\ \text{Team B} & +l & 1.91. \end{array} \tag{2}$$

Here, a winning wager of x dollars returns $x(1.91)$ dollars. In the case of a \$110 wager, the return is $\$110(1.91) = \$210.10$ which is nearly equivalent to the $\$110 + \$100 = \$210$ situation described above.

Having discussed the betting procedure, it is important to understand the objective of the *bookmaker*, i.e. the individual or organisation that posts the line and collects the bets. In (2), suppose that a total of y dollars is wagered on Team A and a total of y dollars is wagered on Team B. In this case, the bookmaker collects $2y$ dollars, and given that a winner is decided, the bookmaker pays out $y(1.91)$ dollars. The bookmaker has made a profit of $y(0.09)$ dollars regardless of the winning team and the percentage profit (*vigorish*) is calculated as $y(0.09)/(2y) \to 4.5\%$. Thus, in the case of the line in (2), the bookmaker's objective is to balance the total bets placed on Team A and on Team B for this guarantees a profit. It is this fact that suggests opportunities for the gambler since the bookmaker is not trying to achieve an optimal line from the point of view of prediction. Rather, the bookmaker is trying to assess public opinion (i.e. determine the pointspread l) so as to balance the bets. A gambler then "simply" needs to (a) have more insight on reality than the rest of the gambling public and (b) win often enough to overcome the vigorish. Unlike most mechanical games (e.g. roulette), it is evident that the possibility exists to develop winning strategies when gambling on the outcome of sporting events.

A simple heuristic for wagering is to develop some procedure for constructing pointspreads. When one's personal pointspread differs sufficiently from the pointspread in the posted line, this signals a condition to wager. We note in passing that the stated heuristic of comparing one's personal pointspread

to a bookmaker's pointspread is an application of subjective probability. In fact, it is difficult to imagine how such an application can be reconciled in terms of a frequentist notion of probability. The discussion highlights one of the distinctions between sports gambling and typical casino gambling. In typical casino gambling, mechanical devices (e.g. roulette wheels, card shuffling, etc.) are the underlying mechanisms that generate probabilities. In the absence of mechanical errors, the probabilities are known, and the games are constructed such that a gambler's expected returns are negative. There are few exceptions to this (e.g. card counting in poker) and the exceptions seem difficult to capitalise upon.

Finally, we mention two other types of wagers commonly available in the same match involving Team A and Team B. First, odds are usually posted for Team A winning outright (i.e. without a pointspread). Odds are also posted for Team B winning outright. Second, there is usually a *total* t associated with the match together with *under* and *over* odds. An under wager wins if fewer than t total points are scored in the match and an over wager wins if more than t total points are scored in the match. If exactly t total points are scored, the wagers are refunded.

2.1 Making use of rounding

In basketball and football it is common to have lines such as (2) with only the pointspread l varying from game to game. With wagers placed on these lines it is natural to ask with what probability p does a bettor need to pick winners in order to realise a long term profit? Suppose that a bettor places wagers of x dollars. The answer is obtained by solving the expected value inequality

$$(1.91x - x)p + (-x)(1-p) > 0,$$

which yields $p > 1/1.91 \approx 0.52356$. Now, given that tossing a coin to determine whether you bet on Team A or Team B (i.e. no knowledge whatsoever) gives $p = 0.5$, it appears plausible to certain sports enthusiasts that it should not be too difficult to pick winners slightly more than 52% of the time. More often than not, this intuition is misleading.

However, something can be done to lower the probability p required to attain long term profits. Certain internet sites for sports gambling round their payouts to dollar amounts but require a minimal wager (suppose \$6). In this case, it is not difficult to show that the optimal wager is \$16. Note that this is not restrictive to the bettor who places large wagers as the bettor can simply make consecutive bets of \$16. Under a successful \$16 wager, the return is $\$16(1.91) = \$30.56 \to \$31$. In this scenario, solving the expected value inequality

$$(31 - 16)p + (-16)(1-p) > 0$$

yields $p \approx 0.51613$. Although the improvement $52.4\% - 51.6\% = 0.8\%$ may appear small, it can lead to meaningful differences over hundreds of wagers.

We note that this observation concerning rounding may be exploited further in sports such as baseball where different odds are often given for Team A and Team B. For example, consider the actual line

$$\begin{array}{lll} \text{Team A} & -l & 1.5882 \\ \text{Team B} & +l & 2.5. \end{array}$$

A winning wager on Team A yields $\$6(1.5882) = \$9.5292 \to \$10$. From the perspective of the general betting public, a wager on Team A should be just as profitable as a wager on Team B. Therefore Team A should win with probability q where $(1.5882)q = (2.5)(1-q)$ giving $q = 0.6115161$ and the expected gain from betting \$6 on Team A is $(\$10 - \$6)q + (-\$6)(1-q) = \$0.1151607$. This results in a positive expected return of 1.92.

2.2 Constructing pointspreads in basketball

In certain sports, it is often sudden and unpredictable events that determine the outcome of a game. For example, consider the impact of fumbles and interceptions in football, and home runs and errors in

baseball. For a sports bettor who wants to get a probabilistic handle on a game, it seems that basketball is an ideal sport. A game of basketball can be modelled for the most part as a series of independent and identical trials (i.e. possessions) where the trials alternate between two teams and on each trial, the team with the ball tries to score. In a National Basketball Association (NBA) game, each team has roughly 100 possessions and scores with a success percentage ranging from 35% to 65%. Another probabilistic attraction of the NBA is that many games are played (82 for each of the 29 teams in the regular season) resulting in many opportunities to test a betting strategy.

Suppose then that Team A is playing Team B on Team B's home court. A bookmaker constructs an initial pointspread by calculating Team A's average score $\overline{A}$ over a specified number of games, by calculating Team B's average score $\overline{B}$ over a specified number of games and then letting the initial pointspread for Team A be $\overline{B} - \overline{A}$. The initial pointspread for Team B is $\overline{A} - \overline{B}$. Now recalling that the bookmakers objective is to balance the total bets on Team A and Team B, he uses his personal experience to modify the initial pointspread. Factors that modify the pointspread include the home court advantage, key injuries to players, results of previous matches between the two teams, recent performance of the teams, team popularity, etc.

One of the things that we have noticed is that the initial pointspread calculation does not adequately account for the strength of the opposition. This seems particularly important for the 1999–2002 NBA seasons where it is widely accepted that the Western Conference has been stronger and we note that a given team plays teams from within its conference more often than teams from outside its conference. The statistical methodology which can address this issue is regression analysis. We therefore propose the model

$$y_{ijk} = \tau_j - \tau_i + \Delta + \epsilon_{ijk},$$

where y_{ijk} is the points scored by home team j minus the points scored by the visiting team i in its kth match, τ_j is a strength variable for the jth team, $\sum \tau_i = 0$, Δ is the home court advantage and ϵ_{ijk} is an error term. Therefore an improved initial estimate of the pointspread by which home team j is favoured over team i is $\widehat{\tau}_j - \widehat{\tau}_i + \widehat{\Delta}$ where the estimates are obtained using multiple linear regression. We note that we have experienced some success in modifications of this general approach.

3 Fixed wagers

In fixed wagering, without loss of generality, a gambler bets a fixed amount \$1 on the outcome of each match. Given a gambling system with probability $0 < p \leq 1$ of choosing winners and European odds $\theta > 1$ in each of the matches, we are interested in the profit P_m that is realised after placing m bets.

Consider first the simplest situation where the gambler has an infinite bankroll. In practice, this may correspond to a gambler who only bets within his means (i.e. is always prepared to lose). For example, a gambler may be willing to lose m dollars over a fixed period consisting of m bets. In this case, given m independent bets, it is straightforward to write the binomial probability

$$\text{Prob}\left[P_m = j(\theta - 1) - (m - j)\right] = \binom{m}{j} p^j (1-p)^{m-j}, \qquad j = 0, \ \ldots, \ m.$$

From the binomial distribution, we have $\text{E}(P_m) = m(\theta p - 1)$ and $\text{Var}(P_m) = \theta^2 mp(1-p)$. The normal approximation to the binomial then suggests

$$P_m \sim \text{Normal}\left(m(\theta p - 1), \theta^2 mp(1-p)\right)$$

from which probabilities may be calculated. For example, the probability of realising a profit after m bets is

$$\text{Prob}[P_m > 0] \approx \Phi\left(\frac{m(\theta p - 1)}{\sqrt{\theta^2 mp(1-p)}}\right).$$

It is also possible to introduce a continuity correction of $-m + \lfloor (m/\theta) \rfloor \theta + \theta/2$ in the calculation of $\text{Prob}[P_m > 0]$.

We now turn to the more complicated situation where the gambler has an initial bankroll of B_0 dollars. We make calculations a little simpler by assuming that if, after bet $i = 1, \ldots, m$, the gambler's bankroll is $0 < B_i < 1$, then the gambler's next bet remains \$1 and he is able to cope with a loss on this bet by borrowing $1 - B_i$ dollars. We say that a gambler is *ruined* and ceases betting if $B_i \leq 0$ for any $i = 1, \ldots, m$.

We concentrate on the probability of ruin and observe that this is a simple extension of the classical drunkard's walk problem. In this case, the drunkard, instead of taking steps of size ± 1 each with probability $\frac{1}{2}$, takes an upward step of size $\theta - 1$ with probability p and a downward step of size -1 with probability $1 - p$. For a discussion of a more general version of the problem, see Section 14.8 of Feller (1968). The probability of ruin within m bets corresponds to the probability that the drunkard hits the absorbing barrier $-B_0$ within m steps. We therefore define $P_{m,n,j}$ as the probability that the drunkard hits the barrier within the next n steps given that j of the first $m - n$ steps are upward steps. By the law of total probability, it is easy to obtain the recursive relationship

$$P_{m,n,j} = pP_{m,n-1,j+1} + (1-p)P_{m,n-1,j}, \tag{3}$$

provided that $P_{m,n,j} \neq 1$. Therefore, the probability of ruin within m bets is given by $P_{m,m,0}$ and the probability is obtained via the recursion (3) and using the conditions $P_{m,n,j} = 1$ if $j \leq (m - n - B_0)/\theta$ (i.e. the drunkard has currently hit the barrier) and $P_{m,n,j} = 0$ if $j > (m - B_0)/\theta$ (i.e. it is impossible for the drunkard to hit the barrier in the remaining n steps). Naturally, recursions can be computationally intensive (both memory and time), and that is the case here. For example, using $\theta = 1.91$, $B_0 = 5$ and $p = 0.56$, we obtain $P_{35,35,0} = 0.216$ and this requires 20 minutes of computation on a SUN workstation based on a Fortan implementation of the recursion.

A source of the tremendous inefficiency in the recursion programming is the repetition of calculations. For example, the calculation of $P_{m,n,j}$ and $P_{m,n,j-1}$ may both invoke $P_{m,n-1,j}$. We avoid this problem by considering the entries $\{P_{m,n,j}\}$ as a tree structure where the quantity of interest $P_{m,m,0}$ is the parent at the top of the tree and the first row of children are $P_{m,m-1,0}$ and $P_{m,m-1,1}$. We construct the tree by first obtaining the bottom row $P_{m,0,j}$, $j = 0, \ldots, m$ and working upward along rows using the recursion (3) and the accompanying conditions. Note that memory issues are minimised by overwriting the current row of the tree on the row of its children. Repeating the calculation from above using a Fortran implementation of the tree structure gives an immediate solution on a SUN workstation. In fact, increasing the number of bets to $m = 500$ still gives a nearly immediate solution $P_{500,500,0} = 0.428$.

The distribution of the final bankroll B_m is also expressible via a recursion. We assume $B_0 > 0$ and, for $i = 1, \ldots, m$ and $j = 0, \ldots, i$, we define

$$\begin{aligned} Q_{j,i-j} &= \text{Prob}\,[B_i = B_0 + j(\theta - 1) - (i - j)] \\ &= \text{Prob}\,[j \text{ upward steps, } i - j \text{ downward steps, no ruin}] \\ &= \begin{cases} pQ_{j-1,i-j} + (1-p)Q_{j,i-j-1}, & j > (i - B_0)/\theta, \\ 0, & j \leq (i - B_0)/\theta, \end{cases} \end{aligned}$$

where $Q_{0,0} = 1$ and $Q_{-1,k} = Q_{k,-1}$ for all k. This also forms a tree structure where $Q_{j,i-j}$ is the jth entry along the ith row from the top of the tree. In this case, we construct the tree beginning with the top element $Q_{0,0}$ and work downward along rows, where the last row $\{Q_{j,m-j}, j = 0, \ldots, m\}$ gives the probabilities of interest. We note that the probability of ruin $P_{m,m,0}$ is also available in this approach since $P_{m,m,0} + \sum_{j=0}^{m} Q_{j,m-j} = 1$.

4 Fixed percentage wagers

In simple fixed percentage wagering, a gambler bets a fraction $0 \leq f \leq 1$ of the bankroll on the outcome of a single match. When the outcome of the match is determined, the gambler resumes fixed percentage wagering on the new balance. Assuming the infinite divisibility of money, one of the

immediate attractions of fixed percentage wagering is that the bettor never goes bankrupt, provided that $f < 1$.

In the context of information theory, Kelly (1956) provided a neat result that is often quoted but misused in gambling circles. Given a gambling system with probability $0 < p \leq 1$ of choosing winners and European odds $\theta > 1$, Kelly showed that the "optimal" betting fraction is

$$f^* = \frac{p\theta - 1}{\theta - 1}, \tag{4}$$

provided that $p > 1/\theta$. For example, consider a system that historically picks winners 54% of the time with the standard European odds payout of $\theta = 1.91$. In this case, the optimal betting fraction is 3.45% of the bankroll. The Kelly criterion is optimal from several points of view; for example, it maximises the exponential rate of growth and it provides the minimal expected time to reach a preassigned balance (Breiman, 1961).

Breiman (1961) investigated the properties of betting systems and considered a more general wagering scenario than the simple situation discussed above. The generality of Breiman's problem did not yield the derivation of solutions nor sharp statements concerning the uniqueness of solutions.

We are concerned with restricted problems that are of real interest to the sports bettor. For example, it is likely that a sports bettor would like to bet on several matches in a single day. If the bettor uses the Kelly fraction (4) on n such matches where $nf^* > 1$, then it would be possible to go bankrupt on that day.

We therefore consider the situation where on day $j = 1, \ldots, m$, the gambler wishes to place n_{ji} wagers on matches with European odds θ_i and where the probability of picking winners is p_i, $i = 1, \ldots, k$. For example, it is possible that a bettor has a system for betting the pointspread and another system for betting totals (i.e. $k = 2$). The question arises as to what are the optimal betting fractions $f_{j1}^*, \ldots, f_{jk}^*$ on day j, where n_{ji} wagers are placed with a fraction f_{ji} of the bankroll, $i = 1, \ldots, k$? Given an initial bankroll B_0, the bankroll at the completion of day j is

$$B_j = \prod_{x_{j1}=0}^{n_{j1}} \cdots \prod_{x_{jk}=0}^{n_{jk}} \left(\left(1 - \sum_{i=1}^{k} n_{ji} f_{ji}\right) B_{j-1} + \sum_{i=1}^{k} x_{ji} \theta_i f_{ji} B_{j-1} \right)^{Y(x)}, \tag{5}$$

where

$$Y(x) = \begin{cases} 1, & \text{if } X_{ji} = x_{ji}, \quad i = 1, \ldots, k, \\ 0, & \text{otherwise,} \end{cases}$$

and the random variable X_{ji} denotes the number of winning wagers of type i on day j. Therefore, we see that except for one term in the product (5), all remaining terms are equal to unity. Note also that the first term in the outer parentheses is the balance of the bankroll not bet in a given day and we require $\sum_{i=1}^{k} n_{ji} f_{ji} < 1$ to prevent the possibility of bankruptcy.

Assuming (often reasonably) that $X_{j1}, \ldots, X_{jk}$ are independent with $X_{ji} \sim \text{Binomial}(n_{ji}, p_i)$, $i = 1, \ldots, k$, we are concerned with the maximisation of $G = G(f_{j1}, \ldots, f_{jk}) = \text{E}[\log(B_j/B_{j-1})]$, where

$$\begin{aligned} G &= \text{E}\left[\sum_{x_{j1}=0}^{n_{j1}} \cdots \sum_{x_{jk}=0}^{n_{jk}} Y(x) \log\left(1 - \sum_{i=1}^{k} n_{ji} f_{ji} + \sum_{i=1}^{k} x_{ji} \theta_i f_{ji}\right) \right] \\ &= \sum_{x_{j1}=0}^{n_{j1}} \cdots \sum_{x_{jk}=0}^{n_{jk}} \left(\prod_{i=1}^{k} \binom{n_{ji}}{x_{ji}} p_i^{x_{ji}} (1-p_i)^{n_{ji}-x_{ji}} \right) \log\left(1 + \sum_{i=1}^{k} f_{ji}(x_{ji}\theta_i - n_{ji})\right). \end{aligned} \tag{6}$$

From (6), the first and second derivatives of G are given by

$$G_i = \frac{\partial G}{\partial f_{ji}} = \sum_{x_{j1}=0}^{n_{j1}} \cdots \sum_{x_{jk}=0}^{n_{jk}} \left(\prod_{i=1}^{k} \binom{n_{ji}}{x_{ji}} p_i^{x_{ji}} (1-p_i)^{n_{ji}-x_{ji}} \right) \frac{x_{ji}\theta_i - n_{ji}}{1 + \sum_{i=1}^{k} f_{ji}(x_{ji}\theta_i - n_{ji})}$$

and

$$G_{i_1 i_2} = \frac{\partial^2 G}{\partial f_{ji_1} \partial f_{ji_2}}$$
$$= \sum_{x_{j1}=0}^{n_{j1}} \cdots \sum_{x_{jk}=0}^{n_{jk}} \left(\prod_{i=1}^{k} \binom{n_{ji}}{x_{ji}} p_i^{x_{ji}} (1-p_i)^{n_{ji}-x_{ji}} \right) \frac{-(x_{ji_1}\theta_{i_1} - n_{ji_1})(x_{ji_2}\theta_{i_2} - n_{ji_2})}{(1+\sum_{i=1}^{k} f_{ji}(x_{ji}\theta_i - n_{ji}))^2},$$

for i, i_1, $i_2 = 1, \ldots, k$. Now, G is a continuous function defined on the intersection of $[0,1)^k$ and the halfspace $\sum_{i=1}^{k} n_{ji} f_{ji} < 1$. When $p_i > 1/\theta_i$, $i = 1, \ldots, k$, it is not difficult to establish that

(a) $G_i > 0$ when $f_{ji} = 0$, $i = 1, \ldots, k$;

(b) $G_{i_1 i_2} < 0$ everywhere for i_1, $i_2 = 1, \ldots, k$; and

(c) $G = -\infty$ on the bounding plane $\sum_{i=1}^{k} n_{ji} f_{ji} = 1$.

A little analysis then shows that there is a single critical point lying in the interior of the region and this point is a global maximum. This establishes the uniqueness of $f_{j1}^*, \ldots, f_{jk}^*$. Since $G(0) = 0$, we also have that $G(f^*) > 0$.

The shape of G yields a simple algorithm which is guaranteed to find $f_{j1}^*, \ldots, f_{jk}^*$. We begin by initialising an interior point $f_{ji} = 1/\sum_{i=1}^{k} n_{ji}$, $i = 1, \ldots, k$. To obtain the root of G_1 along the first coordinate direction, bisection is carried out using the lower starting value $f_{j1}^{(l)} = 0$ and the upper starting value $f_{j1}^{(u)}$ which intersects the plane. After the first coordinate is updated, the procedure is repeated along the coordinate directions $2, \ldots, k$. The loop in this sequential procedure is repeated until the movement in the point is sufficiently small. The maximum has then been obtained. For example, consider a betting system where $k = 2$, $n_{j1} = 2$, $n_{j2} = 3$, $p_1 = 0.545$, $p_2 = 0.585$ and $\theta_1 = \theta_2 = 1.91$. The algorithm gives $f_{j1}^* = 0.0428$ and $f_{j2}^* = 0.1245$. In passing, we comment that convergence difficulties were experienced with more sophisticated algorithms taken from certain numerical libraries. The lesson, as always, is that it is good to use available information (e.g. the shape of G) in optimisation problems.

As previously discussed, ruin (i.e. bankruptcy) is not really an issue with fixed percentage wagering. In fact, what is more important to a gambler is the distribution of the final bankroll B_m. Referring to (5), we can express the final bankroll as

$$B_m = \left(1 - \sum_{i=1}^{k} n_{mi} f_{mi}\right) B_{m-1} + \sum_{i=1}^{k} X_{mi}\theta_i f_{mi} B_{m-1}$$
$$= B_0 \prod_{j=1}^{m} \left(1 + \sum_{i=1}^{k} f_{ji}(X_{ji}\theta_i - n_{ji})\right). \tag{7}$$

In assessing a system before the season begins, expression (7) is useful for simulation purposes. Of course, the bettor does not know in advance the number n_{ji} of wagers of type i on a given day j, and therefore some distribution on n_{ji} is assigned. The simulation then proceeds by generating n_{ji} and then generating $X_{ji} \sim \text{Binomial}(n_{ji}, p_i)$, $j = 1, \ldots, m$, $i = 1, \ldots, k$. Using (7), this determines a single variate B_m; whence the procedure is repeated to build up the distribution of B_m.

Now one might casually assume that, under repeated wagering, a large sample result may hold and that the distribution of the final bankroll B_m in (7) is approximately normal. To see that this may be far from the truth, consider the following simple situation where $m = 162$, $k = 2$, $p_i = 0.56$, $\theta_i = 1.91$ and $n_{ji} = 1$ for all i, j. This corresponds to a successful betting system for the 2000/2001 NBA season where the bettor places a single wager on each of two types of bets every day of the season. The optimal Kelly fraction (4) is $f_{ji}^* = 0.0761$. Using the simulation procedure based on 1000 simulations, Figure 1 provides a histogram of the final bankroll B_{162} using an initial bankroll $B_0 = \$100$. We observe a distinctly non-normal distribution with a very long right tail. However, it is clear that $\log B_m$ is a sum

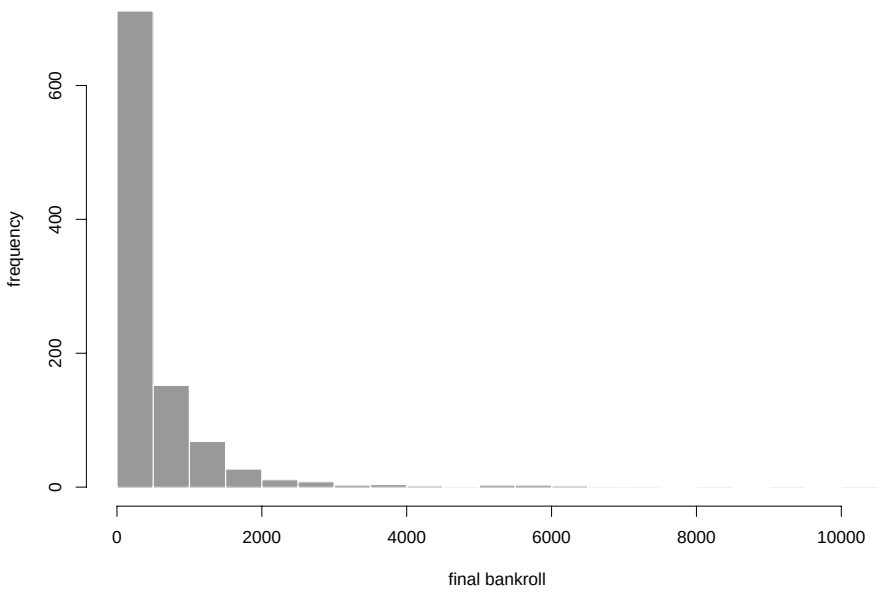


Figure 1: Histogram of the final bankroll based on 1000 simulations.

of random variables. Therefore the Central Limit Theorem suggests that $\log B_m$ may be approximately normal. Using the same example as above, Figure 2 provides a histogram of $\log B_{162}$. The histogram appears more normal and the log data passes the Anderson–Darling goodness-of-fit test for normality yielding a p-value exceeding 0.5.

To obtain some quick insight on the final bankroll B_m without simulation, we calculate the mean

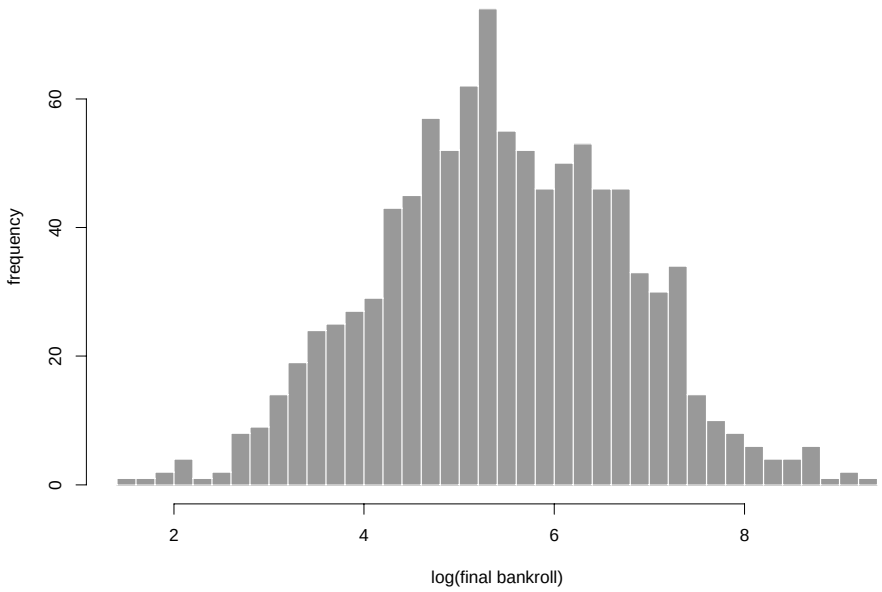


Figure 2: Histogram of the logarithm of the final bankroll based on 1000 simulations.

$\mathrm{E}(B_m)$. Using the formula involving conditional expectations,

$$\mathrm{E}(B_m) = B_0\left(1 + \sum_{i=1}^{k} \mathrm{E}(n_{ji}f_{ji})(p_i\theta_i - 1)\right)^m, \tag{8}$$

where it is assumed that the expectation in (8) does not depend on the day j.

5 Discussion

Now suppose that you have a "winning" system. How should you bet? Typically, Sportsbooks assign an upper limit on wagering. Since the Kelly approach has an optimal rate of growth, it seems logical to begin with the Kelly system using fixed percentage wagering until the upper limit is attained. As long as the Kelly system prescribes a bet exceeding the upper limit, use fixed wagering with the upper limit.

Sportsbooks also typically assign a lower limit on wagering. For some internet sites, the lower limit is so low (e.g. \$1 at www.intertops.com) that it can be practically ignored. We recommend an alternative betting strategy that also allows us to ignore the lower limit on wagering. Suppose that you have an initial bankroll $B_0 = \$500$ and that you are wagering at a Sportsbook with a lower limit of \$10 and an upper limit of \$3000. Extensive personal simulations have shown that it is better to begin with the Kelly system using a bankroll of $x_1 = \$400$, and if the bankroll drops below $x_2 = \$200$, add the final \$100 to the bankroll. The idea is to "kickstart" the system since a very small bankroll grows slowly with fixed percentage wagering. It would be interesting to see if optimal values for x_1 and x_2 could be obtained. Again, as long as the prescribed value of Kelly wagering exceeds \$3000, you would maintain \$3000 betting.

The results that we have presented in this paper are readily applicable to sports betting. The real difficulty is coming up with a winning system (i.e. a system where p is sufficiently large to overcome the vigorish). Naturally p is unknown, and therefore one might estimate p for a proposed system using past data. Of course, there is no guarantee that results will replicate from year to year, and one need also be wary of multiple comparisons issues when considering various systems. Good luck!

References

L. Breiman (1961), "Optimal gambling systems for favorable games", in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, Jerzy Neyman (editor), 65–78.

S. Crist (1998), "All bets are off", *Sports Illustrated*, **88**, 82–92.

W. Feller (1968), *An Introduction to Probability Theory and its Applications*, Volume 1 (third edition), John Wiley and Sons, Inc.

H. Haywood (2000), *BeatWebCasinos.Com: The Shrewd Player's Guide to Internet Gambling*, RGE Publishing, Oakland, California.

J. L. Kelly (1956), "A new interpretation of information rate", *Bell System Tech. J.*, **35**, 917–926.

B. McCune (1989), *Education of a Sports Bettor*, McCune Sports Investments, Las Vegas, Nevada.

B. Ordine (2000), "Super bowl Sunday means lots of action in Las Vegas", *Seattle Times*, January 23.

E. O. Thorp (1969), "Optimal gambling systems for favorable games", *Review of the International Statistical Institute*, **37**, 273–293.

HOW DO LAWN BOWLS AND GOLF BALLS SLOW DOWN ON GRASS?

Maurice N. Brearley
85 Dandarriga Drive
Clifton Springs
Victoria 3222, Australia

Neville J. de Mestre
School of Information Technology
Bond University
Gold Coast
Queensland 4229, Australia
ndemestr@staff.bond.edu.au

Abstract

The rolling of a ball on a horizontal deformable surface was investigated on the assumptions that the ball was a rigid sphere and the surface was elastic. Finite strain theory was used to develop theoretical results which were found to match observations well in cases where the ball and surface involved were such as to ensure no slipping at the region of contact, including a lawn bowl rolling on a grass rink. The theory did not match well the behaviour of a golf ball on a grass green because the ball was too light to enforce the no-slipping condition.

1 Introduction

Experiments were performed with a billiard ball rolling on carpet, a lawn bowl rolling on a grass rink, and a golf ball rolling on a grass putting green. In each case the ball was launched from an inclined ramp, and accurate measurements were made of a number of distances travelled at measured times throughout each run.

In every case the rolling surface was plane and horizontal. Under the very rapidly changing stresses induced by a rolling ball, the behaviour of the surface approximates to that of an elastic solid. It is reasonable to assume that the rigidity μ of the surface is constant and that inertial forces induced in it are negligible. Calculations showed that, in the situations considered, air resistance to the motion of the ball was small enough to be neglected.

Finite strain theory of an elastic solid was found to be adequate to account for the observed nonlinear nature of the deceleration of the ball as a function of time. The resulting theory was found to match closely the results observed for both a billiard ball on carpet and a lawn bowl on grass. This was not so, however, for a golf ball on grass; this was because the light golf ball did not enforce a no-slip condition over the region of contact in the way that the heavier balls did.

The approach used here differs from that used by most investigators of this subject. The relevant elasticity theory is shown to lead to a nonlinear differential equation governing the deceleration of the ball. The problem is thus reduced to solving this equation and evaluating the constants which occur in it and in its solution. (For other approaches to the subject, see Bueche and Flom [1], which also contains a list of sixteen other references.)

2 The experimental arrangements

It was necessary to measure the distances x travelled by the ball, and the corresponding times t, at a number of locations spaced along the path, including the end position. For the billiard ball on carpet

this was achieved by recording the progress of the ball by a video camera which was mounted on rails to enable it to move with the ball. A zoom lens enabled accurate distance readings to be made every 0.04 seconds from a metric tape laid beside the path of the ball. It was found that the distances recorded every 0.08 seconds were adequate for analytic purposes.

For the golf ball and lawn bowl the outdoor situation made the use of a moving video camera difficult. Instead, a multi-recording electronic stopwatch was used to record the times for the ball to reach several measured locations, including the end point, the distances being given by a metric tape laid beside the path.

Because only the distances travelled by the bowl along its path were needed in this investigation, the bias causing a curved path was offset by two 20 cent coins taped to the side of the bowl. This was found to be ideal for producing a straight path. (The effect of bias on the path of a bowl has been described elsewhere, e.g. Brearley and Bolt [2], Brearley [3].)

All of the balls were launched down inclined ramps. For the billiard ball on carpet, a short aluminium ramp was used, with a curve at the bottom to ensure smooth transition to the carpet. The lawn bowl and golf ball were launched from a 2 metres long wooden ramp of 30° angle of inclination. A curved steel plate at the foot of the ramp gave smooth access to the grass surface. A number of different starting positions were marked along the ramp to enable different velocities to be attained by the ball at the foot of the ramp.

The vertical heights of all three balls were measured at each starting position, and from these the velocities of the balls at the bottom of the ramps could be determined. Calculations showed that air resistance had a negligible effect on the motion, and that the lawn bowl was close enough to spherical for it to be treated as a sphere when applying the principle of conservation of energy to its motion down the ramp. For the billiard ball and lawn bowl, let

$$h = \text{the vertical starting height of the ball on the ramp,}$$
$$V = \text{the velocity of the ball on reaching the foot of the ramp.}$$

Then (in obvious notation) the mechanical energy conservation principle gives

$$\frac{1}{2}mV^2 + \frac{1}{2}\left(\frac{2}{5}mr^2\right)\left(\frac{V}{r}\right)^2 = mgh. \tag{1}$$

Using $g = 9.8\,\mathrm{ms}^{-2}$, with h expressed in metres, this gives

$$V = 3.74166\sqrt{h}\,\mathrm{ms}^{-1}. \tag{2}$$

This formula was used to calculate the initial velocity of the bowl and billiard ball in each experiment.

Golf balls are not homogeneous as they have dense inner cores surrounded by less dense cores and light urethane covers. In Section 7, equations (1) and (2) will be modified to deal with the golf ball experiments described there.

For the experiments on grass, every run from each height h was performed several times, the results compared, and the means of the observed times calculated, with the object of reducing experimental errors. Between each run the ramp was moved sideways by several centimetres so as to prevent the ball from running over a track which had been compressed by a previous run.

3 The solution of the deceleration equation

It is shown in Appendix A that the equation of motion of the ball in terms of the distance x travelled in time t is

$$\ddot{x} = -a - b\dot{x} + c\dot{x}^2. \tag{3}$$

The solution of (3) is needed under the initial conditions

$$t = 0, \quad x = 0, \quad \dot{x} = V,$$

and is shown in Appendix B to be

$$x = Kt - c^{-1}\ln(1 + Fe^{-Dt}) + G, \tag{4}$$

where

$$cK^2 - bK - a = 0, \tag{5}$$

$$D = b - 2Kc, \tag{6}$$

$$F = \frac{c(V-K)}{D - c(V-K)}, \tag{7}$$

$$G = c^{-1}\ln(F+1). \tag{8}$$

These equations show that the values of K and D are the same for all initial velocities V, but that the values of F and G vary with V.

A method of calculating the values of the constants in the equation of motion (3) and in the solution (4) from the results of experimental runs of the ball is described in detail in Appendix B. It requires knowing the values of the initial velocity V and the total time T and distance X for three different runs. Experience shows that more accurate results are obtained by using results from a single run to calculate the values of the constants, rather than from three entirely different runs. To obtain the three data triples V, T, X necessary for evaluation of the constants, the following device is used.

The longest run is chosen because any errors in measuring its parameters will be proportionally smaller than for shorter runs. Let V_1, T_1, X_1 denoted its measured initial velocity and observed total time and distance. A graph can be drawn carefully through all corresponding time and distance pairs (t, x) observed during the run, including $(0, 0)$ and (T_1, X_1).

An intermediate point on this graph is selected, at which the coordinates, (t_2, x_2) say, can be read off. The slope of the graph at this point is determined carefully; its magnitude represents the initial velocity V_2 of an intermediate "run", for which the total time and distance are

$$T_2 = T_1 - t_2, \qquad X_2 = X_1 - x_2.$$

The process is then repeated for another selected intermediate point on the graph, having measured coordinates (t_3, x_3). The measured slope there represents the initial velocity V_3 of another intermediate "run", for which the total time and distance are

$$T_3 = T_1 - t_3, \qquad X_3 = X_1 - x_3.$$

In this way the magnitudes V, T, X for three runs are obtained, enabling the values of the constants in the solution to be found. The process will be illustrated in the following sections.

4 Lawn bowl rolling on a grass rink

Experiments were performed with a lawn bowl of diameter 5 in (= 0.127 m) on a grass rink. The mass of the bowl (including the two 20 cent coins taped to its side to counteract the bias and make it run straight) was 1.573 kg. This is large enough to ensure that no slipping could occur between the bowl and the grass, so the theory expounded in the previous section should be applicable. The rink was dry, with newly mown grass, and was classed as one of medium speed.

Runs were made from four different starting heights on the ramp. Equation (2) enabled the initial velocity V on the grass to be calculated in each case. Table 1 shows some of the data, including the intermediate times t_n and distances x_n, total times T and total distances X recorded by stopwatch and measuring tape. The times and total distances are the averages of those recorded over several trials. The intermediate distances x_n were indicated by pieces of wooden dowel laid on the grass at right angles to the measuring tape. The run numbers are those marked on the ramp at the different starting heights h. The times are in seconds and the distances in metres.

Run no.	1	2	3	4
h (m)	0.867	0.681	0.491	0.281
V (ms^{-1})	3.484	3.088	2.622	2.018
t_1	1.935	2.26	1.19	1.014
x_1	6.00	6.00	3.00	2.00
t_2	3.075	3.27	2.256	1.76
x_2	9.00	8.00	5.00	3.00
t_3	4.4175	4.436	3.548	2.637
x_3	12.00	10.00	7.00	4.00
t_4	6.215	6.032	5.534	3.805
x_4	15.00	12.00	9.00	5.00
t_5	9.5625			
x_5	18.00			
T	11.02	9.81	8.31	6.51
X	18.33	13.82	9.91	5.85

Table 1: Lawn bowl experimental results.

In all of the numerical work, the units are metric throughout, and will usually not be stated. The number of figures carried in many parameters is for purposes of calculation only, and is not indicative of the accuracy with which their values are known.

Three sets of data were obtained from Run 1 by the method described in Section 3, using (t_2, x_2) and (t_4, x_4) as the intermediate points. The results obtained were

$$K = 1.08372, \quad c = 0.03748, \quad D = 0.23351.$$

The values of F and G for each run were then calculated from equations (7) and (8). The results are as shown in Table 2.

Run no.	1	2	3	4
F	2.74752	2.02655	1.46788	0.99372
G	35.247	29.547	24.102	18.410

Table 2: Values of F and G for the lawn bowl.

The value of b was then found from equation (6), and that of a from equation (5); they are

$$a = 0.209\,\mathrm{ms}^{-2}, \quad b = 0.152\,\mathrm{s}^{-1}.$$

The numerical forms of the solution (4) are now known for all of the four runs investigated. They enable the graphs of x versus t shown in Figure 1 to be plotted. This figure also shows the observed points (t, x) found during the experiments. The agreement between calculated and observed results is excellent, lending support to the theory on which the calculations were based.

5 The velocity and acceleration of the lawn bowl

So far it has been tacitly assumed that the acceleration formula (3) applies throughout the whole run of the bowl. The acceleration actually undergoes very sharp changes at the start and finish of a run, their durations being so brief that their influences on the path of the bowl are imperceptible. These changes will now be considered, together with the velocity of the bowl, using Run 1 as an illustrative example.

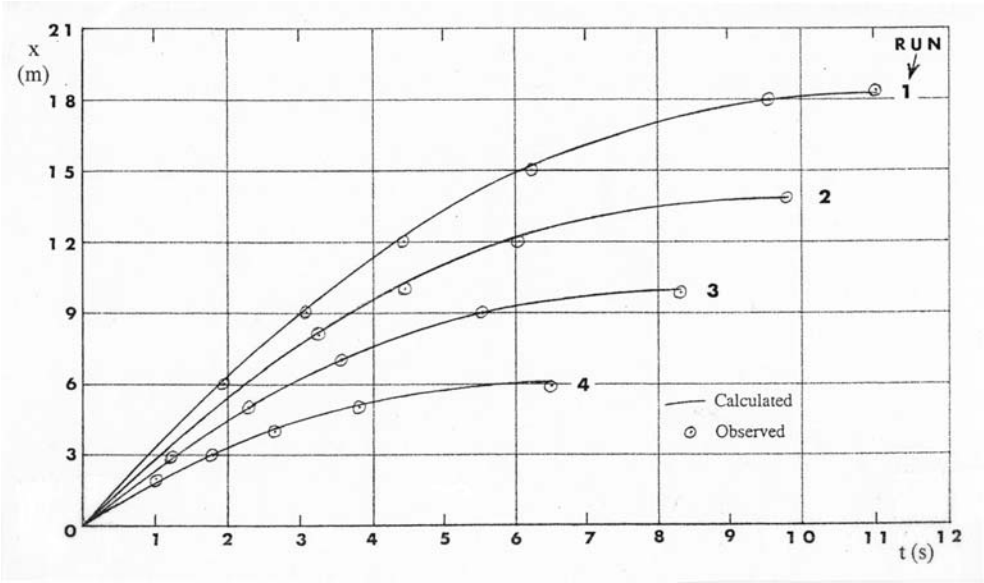


Figure 1: Distance x versus time t for lawn bowl on a grass rink.

The velocity at any instant is easily found by differentiating the solution (4). This gives the velocity as

$$\dot{x} = K + c^{-1}D(1 + F^{-1}e^{Dt})^{-1}. \tag{9}$$

After insertion of the values of the constants found in Section 4, this equation enables the graph of $\dot{x}$ versus t for Run 1 to be drawn in the relevant interval $0 \leq t \leq 11.02$. It is shown in Figure 2.

During its travel down the launching ramp the acceleration of the bowl is constant. The curved metal piece at the foot of the ramp is horizontal at its lower end, which rests on the grass, so the acceleration of the bowl reduces to zero at this point.

On leaving the metal piece the bowl sinks into the grass to a very small depth. It is not hard to show that, if this small depth is estimated to be 1 mm, the duration of the transition of the bowl from metal piece to grass is about 3 milliseconds in the case of Run 1, for which the initial bowl velocity V is $3.484\,\mathrm{ms}^{-1}$. During this time the acceleration of the bowl decreases sharply from zero to a negative value.

On the grass the deceleration is given by the derivative of equation (9), namely

$$\ddot{x} = -\frac{e^{Dt}}{cF}\left(\frac{D}{1 + F^{-1}e^{Dt}}\right)^2, \tag{10}$$

and this governs the motion of the bowl for most of its run. Very near the end of the run, the horizontal displacements of the surface, which have been induced by the motion of the bowl, are decreasing to zero. The force F_2 described in Appendix A is therefore decreasing to zero, and the same is true of the force F_1 which vanishes when the forward movement of the bowl ceases. The acceleration $\ddot{x}$ is therefore tending to zero rapidly, rather than undergoing a finite discontinuity where $\dot{x} = 0$ as suggested by equation (3). Calculations suggest that the duration of this end phase is of the order of 30 milliseconds.

When the values of the constants appropriate for Run 1 of the bowl are inserted in (10) this equation enables the graph of $\ddot{x}$ versus t shown in Figure 3 to be drawn. The slopes of the graph at the start and end have been reduced to make the terminal features visible.

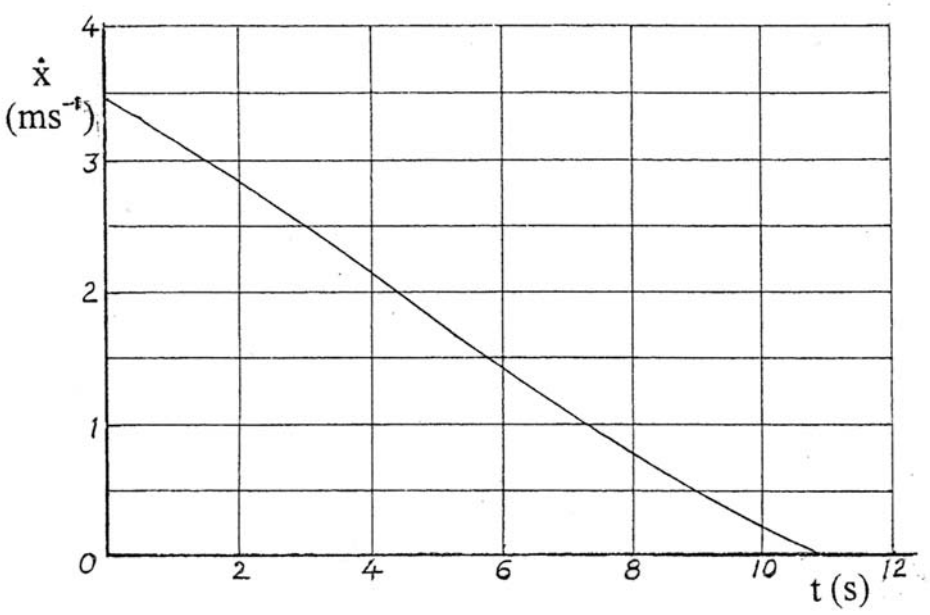


Figure 2: Velocity $\dot{x}$ of lawn bowl versus t, Run 1.

6 Billiard ball rolling on carpet

The experimental arrangements described in Section 2 were used for the billiard ball rolling on loop pile carpet. A run was made over only one distance of 241 cm. Because of the shortness of this run, the observations and calculations were all made in centimetre units rather than metric.

Analysis of the results showed that the billiard ball on carpet behaved in the same way as the lawn bowl on grass, in the sense that its deceleration is given by equation (3), so that the solution (4) is relevant. With an initial velocity of $V = 120.0\,\mathrm{cm\,s^{-1}}$, the form of the solution was found to be

$$x = -135.49t - 701.70\ln(1 + 4.1799e^{-0.44717t}) + 1154.2.$$

The associated values of the constants in the deceleration equation were found to be

$$a = 34.4\,\mathrm{cm\,s^{-2}}, \qquad b = 0.0610\,\mathrm{s^{-1}}, \qquad c = 0.00143\,\mathrm{cm^{-1}},$$

or, in metric units,

$$a = 0.344\,\mathrm{ms^{-2}}, \qquad b = 0.0610\,\mathrm{s^{-1}}, \qquad c = 0.143\,\mathrm{m^{-1}},$$

Figure 4 shows the calculated form of the path of the billiard ball predicted by the above solution, and also some of the observed values (t, x). The agreement between calculated and observed results is exceedingly close, lending support to the theory on which the solution was based.

7 Golf ball rolling on grass

It is well know that golf balls are not homogeneous but consist of a dense inner core, a less dense outer core, and a thin plastic cover. They cannot be treated as homogeneous spheres. The moment of inertia I can be calculated from the conservation equation

$$\frac{1}{2}mV^2 + \frac{1}{2}I\left(\frac{V}{r}\right)^2 = mgh. \tag{11}$$

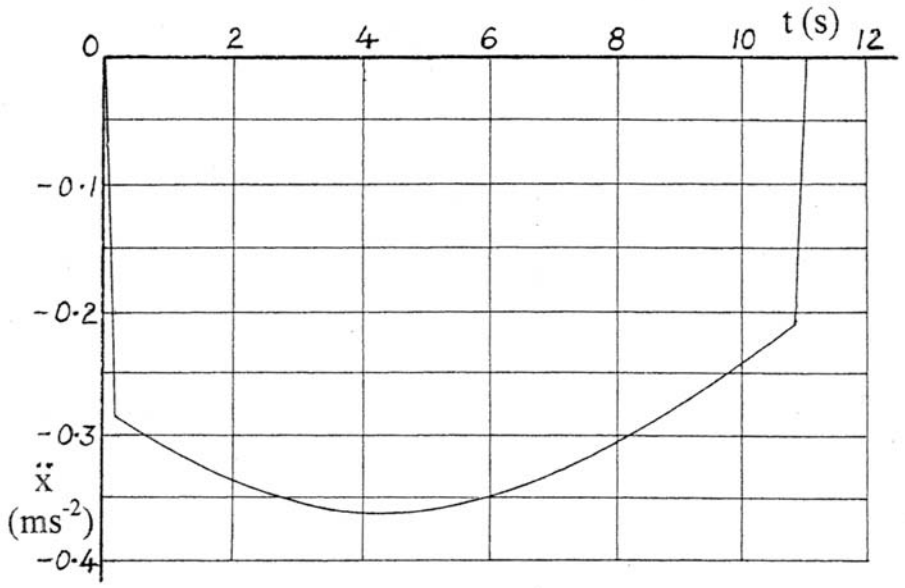


Figure 3: Acceleration of lawn bowl versus t, Run 1.

By measuring the time for the golf ball to roll down an inclined plane, and assuming the acceleration to be constant, the value of V can be found. Since $m = 0.0450\,\text{kg}$ and $r = 0.02135\,\text{m}$, (11) enables I to be calculated for the golf ball. The result obtained was $71.7\,\text{gm}\,\text{cm}^2$, which is 12.6% less than if the ball were homogeneous. Equation (11) then gives

$$V = 3.8110\sqrt{h}\,\text{ms}^{-1} \tag{12}$$

for the velocity of the ball at the bottom of any ramp, the starting height h being expressed in metres.

Equation (12) enables Table 3 to be constructed, giving the initial velocity of the golf ball on the grass for each of the four starting positions investigated. This table also lists the total time T and distance X observed for each run, these values being averages over several trials on each run. The putting green was one of slow speed.

Run no.	1	2	3	4
h (m)	0.883	0.698	0.508	0.297
V (ms^{-1})	3.581	3.184	2.716	2.077
T (s)	4.663	3.72	3.23	3.07
X (m)	6.777	5.59	4.25	2.89

Table 3: Golf ball experimental results.

The results in Table 3 were analysed in the same way as was done for the lawn bowl in Section 4. The values of the constants in the deceleration equation were found to be

$$a = 0.195\,\text{ms}^{-2}, \qquad b = 0.779\,\text{s}^{-1}, \qquad c = 0.152\,\text{m}^{-1},$$

For Run 1, the form of the solution (4) is

$$x = -0.23926t - 6.5604\ln(1 + 2.1581e^{-0.85215t}) + 7.544.$$

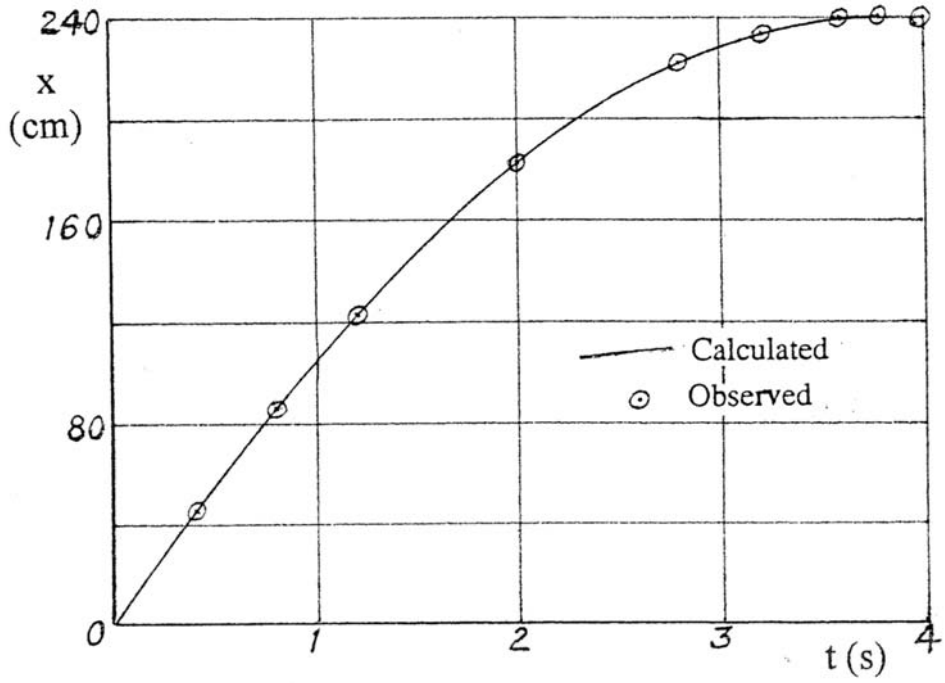


Figure 4: Billiard ball on loop pile carpet.

The solutions for the other three runs differ from this in the values of the constants F and G.

The calculated forms of the solutions for all four runs are shown in Figure 5, together with the observed values of the (t, x) pairs.

The agreement between the calculated and observed results shown in Figure 5 is poor, the calculated value of X being 9% less than the observed value in Run 1, and this deficit increases to 16% in Run 4. The reason for this is that the golf ball is too light to ensure the no-slip condition between it and the surface, on which the theory is based. The light ball skips along on the blades of grass in a way that the massive bowl did not.

It is, of course, possible to obtain a good match between calculated and observed results by modifying the values of the constants a, b and c from those which were calculated from the data, but there is no point in such an artificial move. The lack of agreement shown in Figure 5 lends support to the theory on which the calculated results are based, for it shows that the no-slip condition is an essential part of it.

The experiments with a golf ball were repeated on the medium speed bowling rink which was used earlier for the lawn bowl (see Section 4). The results were similar to those displayed in Figure 5, though the distances travelled by the ball were greater by about 30% than those on the putting green.

8 Summary and conclusions

The rolling of a ball on a plane horizontal surface was investigated by treating the ball as rigid and the surface as deformable. Experiments were performed by launching from ramps a lawn bowl on a grass bowling rink, a golf ball on a grass putting green, and a billiard ball on a loop pile carpet.

It was adequate to treat the bowl and billiard ball as homogeneous spheres, but investigations showed that the density of the golf ball varied internally to a degree that required its moment of inertia being evaluated accurately.

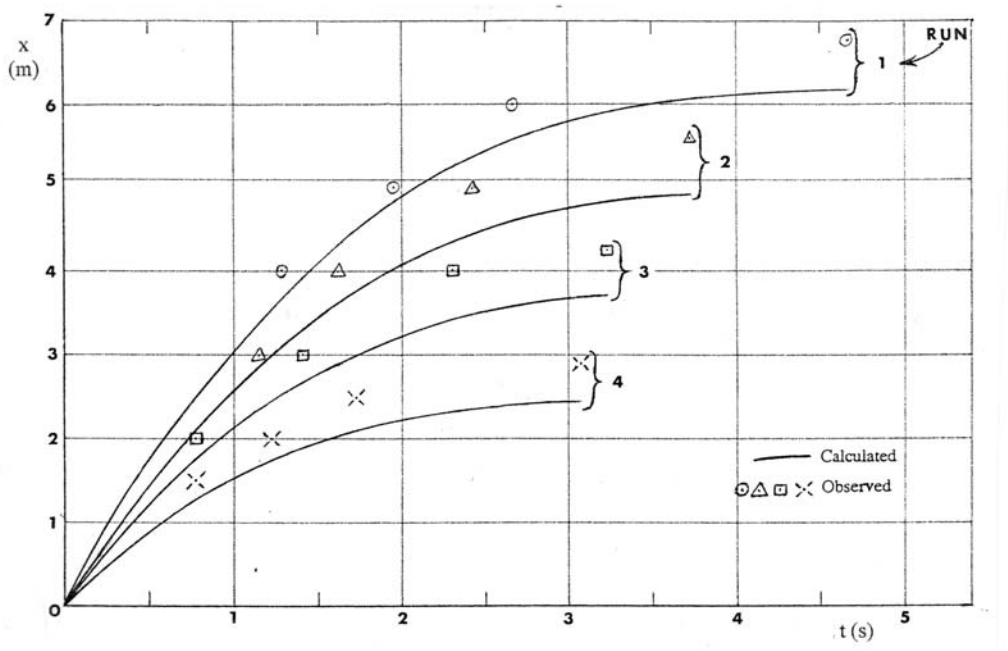


Figure 5: Golf ball on grass putting green.

For the lawn bowl and billiard ball the agreement between calculated and observed results was excellent, indicating the appropriateness of the theory. Because the golf ball is too light to ensure the no-slip condition on grass, accurate predictions could not be made about its travel distances on this surface.

Appendix A The mechanics of ball retardation

A1 Introduction

The ball is assumed to be rigid, its retardation being caused by its deformation of the elastic rolling surface. It is assumed that there is no slipping between the ball and the rolling surface.

The retarding force provided by the surface consists of two parts. One is the result of the ball being in a shallow depression caused by its weight; it is virtually constant, and will be denoted by F_1. The other part, denoted by F_2, is induced by the elastic strain caused in the surface by the motion of the ball. Since air resistance is neglected, the equation of motion of the ball in the forward direction is

$$m\dot{v} = -F_1 - F_2, \tag{13}$$

where m is the mass of the ball and v is the velocity of its centre of mass G.

The form of F_2 will be determined on the assumption that the rolling surface obeys the laws of elastic finite strain theory.

A2 The response of the deformable surface

Because there is no slipping between the ball and the surface at any point of the region S of contact between them, and the displacements of the surface caused by the ball are very small, the path of every point P on the ball is very nearly cycloidal, as shown in Figure 6 (a).

The position of the ball will be referred to rectangular axes Oxy, with Ox in the direction of motion and lying in the surface on which the ball is rolling, as in Figure 6.

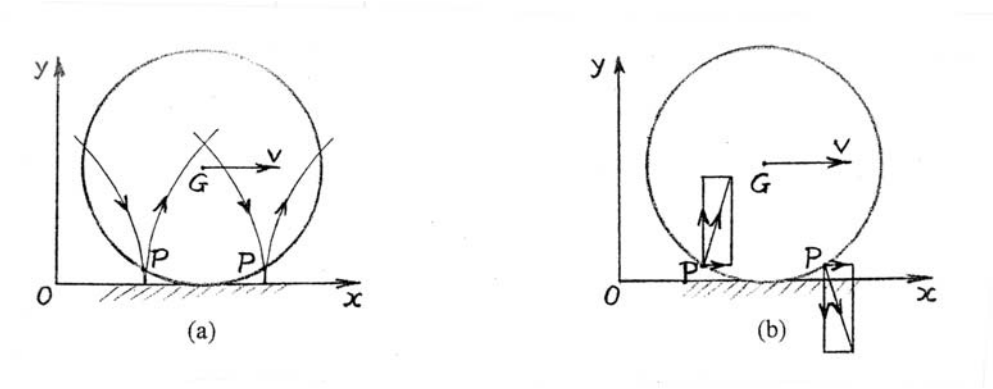


Figure 6: Cycloidal paths and velocity components of points of the rolling ball.

The cycloidal path of P has horizontal and vertical components of velocity as shown in Figure 6 (b). When P is in contact with the surface on which the ball is rolling, it induces in it a shear strain which is very nearly horizontal because the depression is small, and a compression strain which is very nearly vertical. In the following section the horizontal strain will be calculated. From this the associated stress can be found, and when summed over all points of S it will be the reverse of the force F_2 acting on the ball.

At the end of the travel of the ball, the forces F_1 and F_2 drop to zero, as they are caused purely by the motion of the ball. This will not occur as finite discontinuities of the forces, but as very sharp declines accompanying relaxation of the strains of the deformed surface.

A3 The stress–strain relations for the surface

Every vertical "slice" of the ball parallel to the plane Oxy in Figure 6, which has a circumference making contact with the region S, is moving only in a direction parallel to Oxy. Each such slice can therefore reasonably be expected to cause stresses and strains which have no components perpendicular to Oxy. The elasticity problem involved is thus one known as "plane strain", with no dependence on a third space dimension perpendicular to Oxy.

It is convenient to rename the axes Oxy as Ox_1x_2. The general form of the stress–strain relations is (Sokolnikoff [4])

$$p_{ij} = \lambda\theta\delta_{ij} + 2\mu e_{ij}, \tag{14}$$

where λ, μ are Lamé parameters and θ is the dilatation. There are grounds for believing that finite strain theory is appropriate, for which the form of the strain tensor is

$$e_{ij} = \frac{1}{2}\left(\frac{\partial u_j}{\partial x_i} + \frac{\partial u_i}{\partial x_j}\right) - \frac{1}{2}\frac{\partial u_k}{\partial x_i}\frac{\partial u_k}{\partial x_j}, \tag{15}$$

where the u_i are displacements, and summation over k is intended in the last term.

The shear stress of interest in equation (14) is

$$p_{21} = 2\mu e_{21},$$

since this is responsible for the shear force

$$F_2 = \iint_S p_{21}\, dS = 2\mu \iint_S e_{21}\, dS, \tag{16}$$

where the integration is over the whole region S of contact.

A4 Consideration of the surface displacements

It is necessary to consider separately the cases when the points under strain are ahead of and behind the lowest point of the ball, as shown in Figure 7.

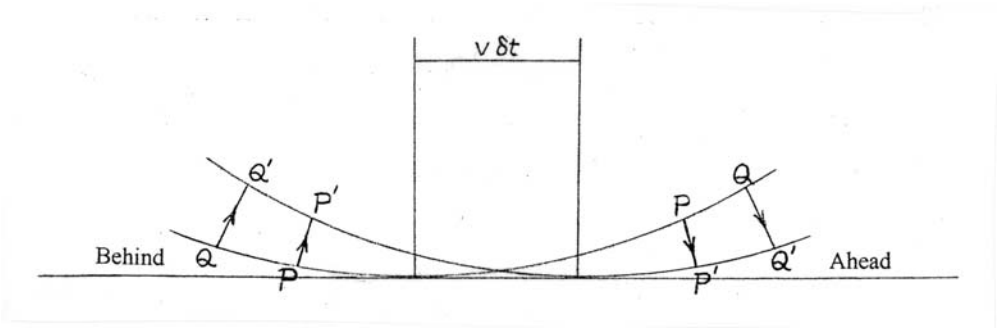


Figure 7: Movements of points P, Q to P', Q' during strain.

Figure 7 depicts the movement of the bottom of the ball during a small time interval δt, in the course of which the centre moves a distance $v\delta t$. The displacements PP' and QQ' experienced by points P and Q during δt are shown for the cases in which the points are ahead of and behind the lowest point of the ball.

Before considering these displacements in detail it can be shown that they are proportional to the ball velocity v. Writing $v\delta t = d$ for brevity, the situation involved may be represented by Figure 8.

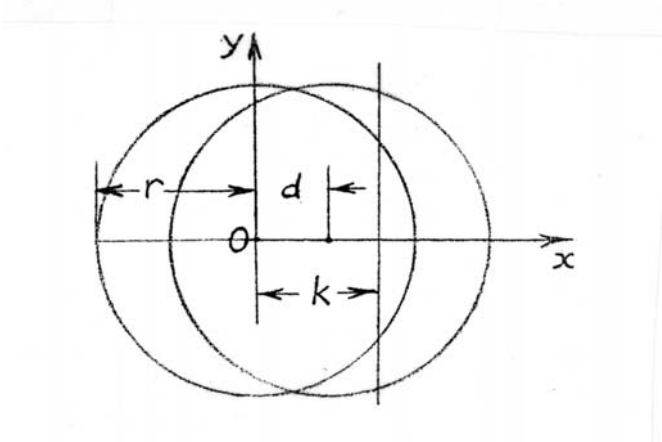


Figure 8: Two successive ball positions a distance d apart.

Let the equations with respect to the axes Oxy of the circular cross-sections of the ball shown in Figure 8 be

$$x^2 + y^2 = r^2, \qquad (x-d)^2 + y^2 = r^2.$$

The two circles intersect the vertical line with equation $x = k$ at values of y given by

$$y = (r^2 - k^2)^{1/2}, \qquad y = (r^2 - (k-d)^2)^{1/2},$$

respectively. Although not shown in Figure 8, it is intended that $d \ll k \ll r$. It is easily seen that the difference between the values of y given by the last two equations is approximately $kd/r = kv\delta t/r$. When interpreted in the context of Figure 7 this means that the small displacements PP' and QQ' are proportional to v.

To consider these displacements it is desirable to enlarge the relevant parts of Figure 7, as shown in Figure 9.

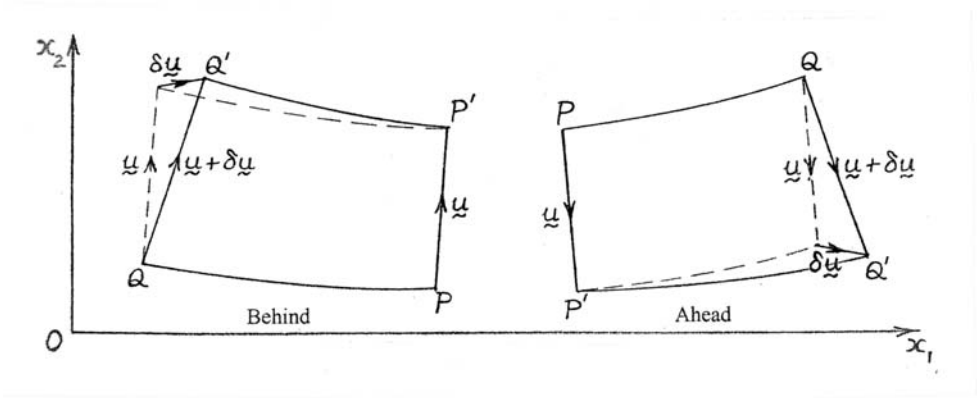


Figure 9: Displacements at points P, Q ahead of and behind the lowest point of the ball.

The displacements of P and Q are shown in vector notation in Figure 9. The components of $\delta\mathbf{u}$ with respect to the axes Ox_1x_2 are denoted as $\delta\mathbf{u} = (\delta u_1, \delta u_2)$.

If the coordinates of P are (x_1, x_2), and those of Q are $(x_1 + \delta x_1, x_2 + \delta x_2)$, then $\mathbf{PQ} = (\delta x_1, \delta x_2)$. From Figure 9, it can be seen that $|\delta x_1| \gg |\delta x_2|$ for points within the small region S of contact. It is also clear that:

$$\begin{aligned} &\text{behind the ball:} \quad \delta u_1 > 0,\ \delta u_2 > 0,\ \delta x_1 < 0,\ \delta x_2 < 0; \\ &\text{ahead of the ball:} \quad \delta u_1 > 0,\ \delta u_2 < 0,\ \delta x_1 > 0,\ \delta x_2 > 0. \end{aligned}$$

A5 The surface strain and stress

From equation (15), we have

$$e_{21} = \frac{1}{2}\left(\frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1}\right) - \frac{1}{2}\left(\frac{\partial u_1}{\partial x_2}\frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2}\frac{\partial u_2}{\partial x_1}\right). \tag{17}$$

From the work in Appendix A4 it follows that the magnitudes of all the partial derivatives in (17) are proportional to v. Also, in the first term on the right hand side of (17), the derivative $\partial u_1/\partial x_2$ is the dominant one, and this term is positive both ahead of and behind the ball. Hence

$$\frac{1}{2}\left(\frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1}\right) = f_1 v, \tag{18}$$

where f_1 is a function of position in the contact area S which is positive throughout S.

From Appendix A4 it also follows that the magnitude of the second term on the right-hand side of (17) is proportional to v^2, and that its sign changes on moving from behind the ball to ahead of it. So

$$-\frac{1}{2}\left(\frac{\partial u_1}{\partial x_2}\frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2}\frac{\partial u_2}{\partial x_1}\right) = \pm f_2 v^2, \tag{19}$$

where f_2 is a positive function of position in S, and the sign is $+$ behind the ball and $-$ ahead of it.

The form of (19) suggests that its contributions to the strain e_{21} would annul one another when summed over the whole of S. The deformations in the rolling surface are, however, not symmetrical about the lowest point of the ball. The ball compresses the region ahead of it, increasing the magnitudes of all four positive derivatives in the last term in equation (17); and behind the ball the strain is reduced by the ball's progress, thus reducing in this region the magnitudes of both negative products in the last

term in (17). The net result is a negative value when the last term (including its coefficient $-\frac{1}{2}$) is integrated over the whole of the region S.

On using (17), (18) and (19), equation (16) yields

$$F_2 = 2\mu \iint_S (f_1 v \pm f_2 v^2)\, dS = A_1 v - A_2 v^2, \tag{20}$$

where A_1 and A_2 are positive constants. Their values will, of course, depend on the size and weight of the ball and on the character of the rolling surface. Substitution from (20) in (13) yields

$$m\dot{v} = -F_1 - (A_1 v - A_2 v^2),$$

whence

$$\dot{v} = -a - bv + cv^2, \tag{21}$$

where a, b, c are positive constants.

Appendix B

B1 The solution of the equation of motion

The equation of motion is

$$\ddot{x} = -a - b\dot{x} + c\dot{x}^2.$$

On substituting $y = x - Kt$, where

$$cK^2 - bK - a = 0, \tag{22}$$

and then putting $w = \dot{y}$, it is found that

$$\dot{w} = cw^2 - Dw,$$

where

$$D = b - 2Kc. \tag{23}$$

Solving by separation of variables yields

$$w = \frac{Dc^{-1}}{1 + F^{-1}e^{Dt}},$$

where F is constant.

Since $w(0) = V - K$, where $V = \dot{x}(0)$, we find that

$$F = \frac{c(V - K)}{D - c(V - K)}. \tag{24}$$

Integrating again to find x leads to

$$x = Kt + c^{-1}Dt - c^{-1}\ln(e^{Dt} + F) + G = Kt - c^{-1}\ln(1 + Fe^{-Dt}) + G,$$

where G is a constant. Since $x(0) = 0$, we have

$$G = c^{-1}\ln(F + 1). \tag{25}$$

B2 Evaluation of the constants

Differentiating the above solution for x gives

$$\dot{x} = K + \frac{c^{-1}FDe^{-Dt}}{1 - Fe^{-Dt}}.$$

This result, and the above form for F, enable $\ln(1 + Fe^{-Dt})$ to be eliminated from x, giving

$$(Kc + D)t = cx - \ln(\dot{x} - K) + \ln(V - K).$$

At the end of a run, we have $t = T$, $x = X$, $\dot{x} = 0$, and so

$$(Kc + D)T = cX + \ln(1 - VK^{-1}).$$

Using data (T_1, V_1), (T_2, V_2), (T_3, V_3) from three runs gives three equations for the three unknowns K, c, D. By eliminating $Kc+D$ between two pairs of these equations, and then c between the two resulting equations, an equation for K is obtained. This can be solved numerically by Newton's method.

The constants c and D can then be found, after which equation (23) yields the value of b, and then (22) gives a. These constants depend only on the properties of the ball and the surface.

For a given run, the constants F and G can be found from equations (24) and (25), their values depending on the initial velocity V. The function $x(t)$ then gives values which can be compared with experimental results.

Acknowledgments

The writers are grateful to the staff of the Audio Visual Services at Bond University for assistance with the billiard ball experiments, and to the Presidents of the Boomerang Park Golf Club and Mudgeeraba Bowls Club on the Gold Coast, Queensland, for making their facilities available for experiments. They are also grateful to a referee for extensive suggestions which greatly improved the presentation of the paper.

References

[1] A. M. Bueche and D. G. Flom, "Surface friction and dynamic mechanical properties of polymers", *Wear*, **2** (1958/59), 168–182.

[2] M. N. Brearley and B. A. Bolt, "The dynamics of a bowl", *Quart. J. Mech. Appl. Math.*, **11** (1958), 351–363.

[3] M. N. Brearley, "The motion of a biased bowl with perturbing projection conditions", *Proc. Camb. Phil. Soc.*, **57** (1961), 131–151.

[4] I. S. Sokolnikoff, *Mathematical Theory of Elasticity* (second edition), McGraw–Hill, New York (1956).

FIXING THE FIXTURES WITH GENETIC ALGORITHMS

George Christos and Jamie Simpson
Department of Mathematics and Statistics
Curtin University of Technology
GPO Box U1987, Perth
Western Australia 6845
christos@maths.curtin.edu.au, simpson@maths.curtin.edu.au

Abstract

In 1998 the West Australian Football Commission (WAFC) asked us to help them set the fixtures for the premier Australian Rules football competition in Western Australia, The West Australian Football League (WAFL). In this competition there are nine teams, and our task was to try to come up with a set of fixtures in which each team had (as much as possible) a sequence of alternating home and away games, while trying to satisfy an enormous number of constraints. The problem was separated into two parts, the listing problem (or who plays whom and when?), and the scheduling problem (where they play, that is, at whose home-ground?). The listing problem was tackled by using a standard "big wheel" generator, while the scheduling problem was dealt with by using a genetic algorithm. Our work serves as a useful example of how to apply genetic algorithms to solve practical optimisation problems.

1 Introduction

When the WAFL went from eight teams to nine teams in 1997, they encountered a major problem with setting their fixtures. When there were only eight teams, every team could play every other team three times over 21 rounds. One of the main problems with having nine teams in the competition is that one team must stand out each week. For 1999, we were asked to come up with a set of fixtures over 23 rounds, with one bye in each of 21 rounds, and three byes in the two remaining rounds. This means that there are $(21 \times 4) + (2 \times 3) = 90$ matches, so each team plays 20 games in all and has three byes. These matches had to be arranged so that each team had ten home (H) games, ten away (A) games, and three byes, with teams playing as near as possible to an alternating sequence of home and away games. Note that since each team plays a total of 20 matches, they must play four of the other teams twice and the other four teams three times during the season, $20 = (4 \times 2) + (4 \times 3)$.

The nine clubs in the WAFL are: Claremont (CL), East Fremantle (EF), East Perth (EP), Peel Thunder (PT), Perth (PE), South Fremantle (SF), Subiaco (SU), Swan Districts (SD), and West Perth (WP).

In 1999, the fixtures were actually set by the Football Operations Manager at the WAFL, using a commercial fixturing program available on the market, combined with some switches by hand. Our solution was not used, because there was insufficient time to implement our algorithm. There were some major problems with the fixtures used in 1999. The two rounds with three byes were too close to each other, and some teams did not enjoy an alternating sequence of home and away games. For example, one team had four consecutive away games and another team had only one home game in the space of seven weeks. One pair of teams also played each other twice in three weeks (rounds 14 and 16). Our expertise in this field won us a contract to set the fixtures for seasons 2000, 2001 and 2002. However, we

did not have to use genetic algorithms to arrive at a reasonable solution. One of the main reasons for this was that the WAFL, on our advice, decided to relax some of its constraints. As the Olympic Games were held in Australia in 2000, that year's season was shortened to 21 rounds, with each team playing 18 games (nine home and nine away) with three byes. The WAFC decided to stick with 21 rounds in 2001 and 2002. Although our methods developed for 1999 were never used, it is instructive to review our proposed method of solution.

2 The wish list for 1999

In 1998, each of the clubs presented the WAFC with a wish list for the 1999 season. Below we have listed some of these wishes. Some (not all) of them were incorporated into our model. Some of the wishes were designated to be hard constraints (that had to be satisfied) while others were nominated to be soft constraints that were incorporated into the fitness function of our genetic algorithm. As will become clear below, the complexity of the fixturing problem increases dramatically with the number of constraints.

Hard constraints

General/WAFC

1. Each team is to have ten home games, ten away games, and three byes played over the 23 rounds.

2. If two teams play each other exactly twice in the season, each must have one home game, that is, each team must have either an HA or AH arrangement for these matches.

3. If two teams play each other three times in the season, they must have at least one home game each. This means they can have one of the following arrangements: HAH, HHA, AHH, AHA, AAH, or AHH.

4. SF must play EF three times, as this "local derby" is always well attended.

5. SF must play EF at SF's home ground (SF v EF) on the Foundation Day holiday in round 10.

6. WP must play EP three times, another "local derby".

7. EP v WP on Foundation Day in round 10.

East Perth (see also points 6 and 7)

8. For 1999, EP required its first 4 games to be away from its home-ground, which was being used by Perth Glory, Perth's interstate soccer team. At a later date this requirement was extended to the first six weeks.

Peel Thunder

9. PT requested home games on all public holidays (in rounds 1, 4 and 10), given that Peel is based in Mandurah, which is a holiday destination south of Perth. This would assist to maximise crowds.

South Fremantle

10. SF requested byes in rounds 9, 12 and 19. As this was SF's centenary season this was considered to be a hard constraint.

Soft constraints

General/WAFC

11. To avoid, where possible, the scheduling of two consecutive home or away games.

12. To achieve an even spread of byes for each team.

13. Scheduling of a general bye for the interstate matches (formally in round 12). The only effect of this constraint is that one must try to avoid having a sequence like BGB, BGA, AGB, or AGA (where B = bye, G = general bye, A = Away), so that no team has a prolonged absence from their home-ground.

14. With respect to point 3 above, AHA and HAH are the preferred arrangements.

15. Attempt to rotate clubs that play each other three times, with the exception of the local derbies, points 4 and 6 above.

Claremont

16. Bye in middle of August (round 19 or 20).

17. CL v SF in round one.

East Fremantle (see also points 4 and 5 above)

18. EF requested that home games should not occur in consecutive weeks so that it could more effectively promote its home games.

Perth

19. PE to play WP three times in 1999 with two of the matches played at Perth's home ground.

20. Bye in late April or early May (rounds 4, 5 or 6).

21. PE v EP to be scheduled approximately four weeks after the Perth's first bye.

South Fremantle (see also points 4, 5 and 10 above.)

22. SF to play SD 3 times during 1999.

Subiaco

23. Avoid scheduling a home-game for Subiaco at Subiaco Oval on Saturday during the weeks when one of the interstate Australian Football League (AFL) teams West Coast or Fremantle is playing at Subiaco Oval on a Friday night. This was not really a problem as the match could always be moved to a Sunday in this event. One also needs to ensure that Subiaco Oval is avoided on Saturdays or Sundays when it is required by the interstate teams.

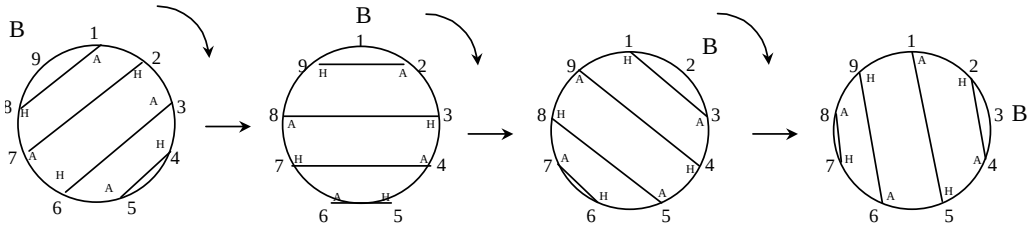
24. SU requested a bye in round 1.

Swan Districts

25. SD v WP in round 1. This conflicts with request 26 below.

West Perth

26. WP v EP in round 1.



Round 1	Round 2	Round 3	Round 4
3 v 5	9 v 2	1 v 3	2 v 4
6 v 2	3 v 8	4 v 9	5 v 1
1 v 7	7 v 4	8 v 5	9 v 6
8 v 9	5 v 6	6 v 7	7 v 8
4 bye	1 bye	2 bye	3 bye

Figure 1: The "big wheel" algorithm.

3 Solving the listing problem

We approached the fixturing problem by separating it into two parts, the listing part of the problem, that is, who plays whom and when, and the scheduling part of the problem, that is, where the game is played. In this section we show how the listing part of the problem can be solved by using the so-called "big wheel" algorithm. The big wheel can sometimes be used to solve the scheduling part of the problem as well, but this is generally not practical when there are a number of extra constraints. Imperfect scheduling solutions generated in this way can be used as starting configurations for our genetic algorithm to solve the scheduling problem. In other work, we tried to solve both the listing and the scheduling problem by using genetic algorithms "in one hit", but this proved to be much too difficult. With so many constraints it was not possible to perform small manipulations of our chromosomes, which is an essential ingredient in a good genetic algorithm.

The diagrams in Figure 1 illustrate how the listed pairs are generated by the big wheel algorithm. Each number on the circle represents one of the nine possible teams involved in the tournament. The numbers will be later associated with particular teams when we try to incorporate the constraints. The parallel bars show how teams are paired in each round. The team that is not adjacent to the end of a bar (in the diagram on the left, this is team number 9) has a bye. To generate the next round, one rotates the parallel bars through one ninth of a revolution, as shown in the second diagram. If this process is repeated for nine rounds then each team plays each other team exactly once, and has one bye. One can also assign a home (H) and away (A) to the ends of the bars, as shown in Figure 1, to generate a scheduling solution. Notice that each team has an alternating sequence of H and A for each group of nine rounds. One generates $27 = 3 \times 9$ rounds in this way, with the order of H and A reversed for rounds 10–18 with respect to rounds 1–9. This is obtained from the big wheel algorithm by swapping the H and A labels on the bars. Rounds 19–27 are identical to rounds 1–9.

To get 23 rounds of fixtures using this scheme one deletes the last four rounds. See Table 1. At this stage, four teams (5, 6, 7 and 8) do not have their allotted three byes. One must then find places in the remaining list where these teams play each other in pairs. See highlighted cells in Table 1. The appropriate set of byes is then achieved by eliminating two appropriate pairs of matches involving these teams. One way that this can be achieved is to eliminate the round 3 match between teams 8 and 5 and the round 21 match between teams 6 and 7, as shown in Table 1. This leaves 21 rounds with four matches each and two rounds with three matches (and three byes). One needs to ensure that each team has ten home and ten away games. If we had instead eliminated the 5 v 8 match in round 12, we would have found that team 5 now had nine home games and 11 away games, and we would need to correct the balance by making an appropriate adjustment somewhere else. Note that we could have also eliminated both matches in the same round (3, 12 or 21). This would generate a listing where there are 22 rounds

with four matches and one round with two matches and five byes.

r1	r2	r3	r4	r5	r6	r7	r8	r9
8 v 1	9 v 2	1 v 3	2 v 4	3 v 5	4 v 6	5 v 7	6 v 8	7 v 9
2 v 7	3 v 8	4 v 9	5 v 1	6 v 2	7 v 3	8 v 4	9 v 5	1 v 6
6 v 3	7 v 4	~~8 v 5~~	9 v 6	1 v 7	2 v 8	3 v 9	4 v 1	5 v 2
4 v 5	5 v 6	6 v 7	7 v 8	8 v 9	9 v 1	1 v 2	2 v 3	3 v 4
9b	1b	2b	3b	4b	5b	6b	7b	8b
r10	r11	r12	r13	r14	r15	r16	r17	r18
1 v 8	2 v 9	3 v 1	4 v 2	5 v 3	6 v 4	7 v 5	8 v 6	9 v 7
7 v 2	8 v 3	9 v 4	1 v 5	2 v 6	3 v 7	4 v 8	5 v 9	6 v 1
3 v 6	4 v 7	5 v 8	6 v 9	7 v 1	8 v 2	9 v 3	1 v 4	2 v 5
5 v 4	6 v 5	7 v 6	8 v 7	9 v 8	1 v 9	2 v 1	3 v 2	4 v 3
9b	1b	2b	3b	4b	5b	6b	7b	8b
r19	r20	r21	r22	r23	~~r24~~	~~r25~~	~~r26~~	~~r27~~
8 v 1	9 v 2	1 v 3	2 v 4	3 v 5	~~4 v 6~~	~~5 v 7~~	~~6 v 8~~	~~7 v 9~~
2 v 7	3 v 8	4 v 9	5 v 1	6 v 2	~~7 v 3~~	~~8 v 4~~	~~9 v 5~~	~~1 v 6~~
6 v 3	7 v 4	8 v 5	9 v 6	1 v 7	~~2 v 8~~	~~3 v 9~~	~~4 v 1~~	~~5 v 2~~
4 v 5	5 v 6	~~6 v 7~~	7 v 8	8 v 9	~~9 v 1~~	~~1 v 2~~	~~2 v 3~~	~~3 v 4~~
9b	1b	2b	3b	4b	~~5b~~	~~6b~~	~~7b~~	~~8b~~

Table 1: A possible listing of matches devised by the big wheel algorithm.

There are a number of ways in which one can manipulate the listing in Table 1, while preserving the hard listing constraints. These have been outlined elsewhere [1]. For example, one can swap two rounds so long as this does not generate a solution in which a team has byes too close together or the same two teams play each other too close together. One can also swap the labels of teams in certain sections of the listing, or move the byes around.

The big wheel algorithm leads to a reasonably satisfactory solution to the listing problem, as well as the scheduling problem as most teams have an alternating home-away sequence like HAHAHA or AHAHAH, except for a possible hitch with the switch from round 9 to 10 and from round 18 to 19, where some of the teams have consecutive home or away games. Much of this natural flow in the scheduling is however disrupted when we try to impose the constraints.

From this listing one can see that the following pairs of teams play each other twice:

$$(1,2),(1,4),(1,6),(1,9),(2,3),(2,5),(2,8),(3,4),(3,7),$$
$$(3,9),(4,6),(4,8),(5,7),(5,8),(5,9),(6,7),(6,8),(7,9),$$

and the following pairs of teams play each other three times:

$$(1,3),(1,5),(1,7),(1,8),(2,4),(2,6),(2,7),(2,9),(3,5),$$
$$(3,6),(3,8),(4,5),(4,7),(4,9),(5,6),(6,9),(7,8),(8,9).$$

In developing the fixtures one needs to determine which pairs of teams should play together three times during the season. Ideally one should rotate them from year to year but financial considerations require that the popular local derbies (SF v EF, and WP v EP) should occur three times during the season.

4 Assigning teams to numbers

In the first instance one should try to assign the labels to teams in such a way as to satisfy most of the constraints listed in Section 2. For example, referring to the fixtures in Table 1, one could assign SF = 1 because SF wants a bye in round two. As SF wants to play EF in round 10 we could set EF = 8. As Perth wants a bye in rounds 4, 5 or 6, this implies that PE = 3, 4 or 5. Suppose we take PE = 3. As PE is to play EP four rounds later, we could take EP = 2. Requiring that WP v EP in the 10th round, implies WP = 7. Note also in our choice of label assignments above we were conscious of the fact that EF and SF (and EP and WP) need to play each other three times.

Continuing in this way, we were able to satisfy most of the constraints [1]. A possible listing is given in Table 2. Our solution is, however, not without its own problems because in the course of satisfying the various constraints we have upset the natural flow of alternating home and away games generated by the big wheel algorithm. Generally it only takes a few such manipulations of the listings to dramatically alter this alternating sequence. In practice it may not be possible to satisfy all of the constraints and one needs to prioritise them. In the end however, we hope that our genetic algorithm can restore the fixtures to an almost alternating sequence of H and A for each team.

5 Encoding the scheduling part of the problem on a chromosome

The scheduling part of the fixtures is solved by using a genetic algorithm [2, 3, 4]. Here we take a given listing, such as that given in Table 1 or Table 2 and we encode the information pertaining to the scheduling onto a "chromosome". Special consideration must be given to the pairs of teams that play each other twice and the pairs of teams that play each other three times. See Section 3. Recall that for this particular set of fixtures there are 18 pairs of teams that play each other twice and 18 pairs of teams that play each other three times. We will call these "A pairs" and "B pairs", respectively. All A pairs must play each other in a HA or AH arrangement. We can label these two possibilities by using a binary digit $x = \{0, 1\}$. We can let 0 denote the case where the team with the lowest label plays the first home game, and 1 denote the other possibility. For the pair $(1, 2)$ in Table 1 the home/away sequence will then be encoded as a 0, as 1 plays the first home game. One can encode the corresponding sequences for the rest of the A pairs in this way. These binary digits will form the first 18 genes of our chromosome. For the set of fixtures given in Table 1 these genes would be assigned the values 010101001001001000. These genes tell us precisely which team from each pair plays the first game at home. The genes are ordered in the same way as the previously given list of A pairs. For example, the tenth gene in this sequence refers to the pair $(3, 9)$ and the value 0 tells us that the games are played in the order 3 v 9 followed by 9 v 3.

For the B pairs, there are six allowed possibilities (HAA, AHA, AAH, HHA, AHH, HAH). We will encode these six possibilities onto our chromosome by using a binary and a ternary digit. The binary digit ($z = \pm 1$) tells us which of the teams in the pair has the most H games. We will use the convention that the lower z value -1 corresponds to the lowest numbered team having only one home game and $z = 1$ corresponds to the lowest numbered team having two home games. For the pair $(1, 3)$ the matches are played in the order 1 v 3, 3 v 1, and 1 v 3 which we would label as HAH, which would attract the value $z = 1$. For each of these z values there are three possibilities, shown in Table 3. We will encode these possibilities by using a ternary digit $y = 1, 2, 3$. Note that for a fixed y value, going from one z value to another corresponds to changing exactly one of the allowed H's into an A, or vice versa.

The B pairs in Table 1 would be given the z/y pairs: $1/3$, $-1/2$, $1/3$, $-1/2$, $1/3$, $-1/2$, $1/3$, $-1/2$, $1/3$, $-1/2$, $1/3$, $1/3$, $-1/2$, $1/3$, $1/3$, $-1/2$, $1/3$, $1/3$. The fourth pair of values, for example, which is associated with the $(1, 8)$ pair corresponds to the arrangement AHA, or 8 v 1, 1 v 8, 8 v 1.

A typical chromosome will then consist of the 18 binary genes representing the pairs of teams playing twice and the 18 binary and 18 ternary gene pairs representing the teams that play each other three times. There are however certain restrictions which apply to how we can manipulate these chromosomes.

	3/4	10/4	17/4	24/4	1/5	8/5	15/5	22/5	29/5	5/6	12/6	26/6	3/7	10/7	17/7	24/7	31/7	7/8	14/8	21/8	28/8	4/9	11/9
Club	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
WP	A SD	H PE	A SF	B	A EP	H PT	A SU	B	H CL	H EP	A PE	H SD	A EF	H SF	A PT	H SU	B	A CL	A EP	H PE	H SU	H EF	A SF
PT	H EF	A WP	B	H EP	A SD	H SU	A CL	H SF	A PE	H SD	H WP	A EF	A EP	B	A SU	H CL	A SF	H PE	H SD	A WP	H CL	A EP	B
PE	H CL	H EF	H SD	B	A SU	A WP	A SF	H EP	H PT	H SU	A EF	A CL	B	A SD	H WP	H SF	A EP	A PT	A SU	H EF	A SF	B	H SD
EF	A PE	A PT	H CL	B	H SF	A EP	H SD	A SU	B	A SF	H PT	H PE	H WP	A CL	H EP	A SD	H SU	B	H SF	A PT	A PE	A WP	H CL
EP	A SF	A CL	A SU	A PE	H WP	H EF	B	A PT	A SD	A WP	H CL	H SF	H PE	H SU	A EF	B	H PT	H SD	H WP	A CL	B	H PE	A SU
SU	B	A SD	H EP	A CL	H PT	A PE	H WP	H EF	A SF	A PT	H SD	B	H CL	A EP	H PE	A WP	A EF	B	H PT	H SF	A WP	A CL	H EP
SD	H WP	H SU	A PT	H SF	H PE	B	A EF	A CL	H EP	A PE	A SU	A WP	A SF	H PT	B	H EF	H CL	A EP	A PE	B	H EF	H SF	A PT
SF	H EP	B	H WP	A SD	A EF	A CL	H PT	A PE	H SU	H EF	B	A EP	H SD	A WP	H CL	A PT	H PE	B	A EF	A SU	H SF	A SD	H WP
CL	A PT	H EP	A EF	H SU	B	H SF	H PE	H SD	A WP	B	A EP	H PT	A SU	H EF	A SF	A PE	A SD	H WP	B	H EP	A PE	H SU	A EF

Table 2: A specific set of fixtures generated by the big wheel algorithm with minor manipulations. The first row gives the dates of fixtures as day/month. No matches were played on 19/6 (general bye).

	$z=-1$	$z=1$
$y=1$	HAA	HHA
$y=2$	AHA	AHH
$y=3$	AAH	HAH

Table 3: The assignment of y and z genes for B pairs.

With the B pairs, one can freely change the value of the ternary genes. The same applies to the binary gene for A pairs, however we cannot freely change the value of the z gene without interfering with the constraint that each team plays ten H and ten A games. If, for example, we were to change from the configuration $y=2$, $z=-1$ (or AHA) to $y=2$, $z=1$ (AHH), then we have changed the number of home games that are played by each of these pairs. Clearly one needs to make another adjustment to the other B genes to compensate for this, but these additional changes may induce further required changes. To solve this problem one needs to identify cycles which maintain the allocation of ten H games to each team. This is discussed in the next section.

6 The matrix gene for B pairs

Each team plays a total of ten home games: four of these will be part of an A pair, two will be a part of a B pair in which the team has one home game ($z=-1$) and two will be a part of a B pair where it has two home games ($z=+1$). We can describe the B pair situation with a 9×9 matrix, whose rows and columns are labelled by the team numbers. The matrix will have a 0 entry for each row/column pair that corresponds to an A pair, and along the diagonal. The matrix will have $+1$ in each position, where the row team plays a column B pair team home twice and a -1 where the column team plays the row B pair team twice at home.

For our starting configuration of Table 1 this matrix would correspond to

$$M=\begin{pmatrix} 0 & 0 & 1 & 0 & -1 & 0 & 1 & -1 & 0\\ 0 & 0 & 0 & 1 & 0 & -1 & 1 & 0 & -1\\ -1 & 0 & 0 & 0 & 1 & -1 & 0 & 1 & 0\\ 0 & -1 & 0 & 0 & 1 & 0 & -1 & 0 & 1\\ 1 & 0 & -1 & -1 & 0 & 1 & 0 & 0 & 0\\ 0 & 1 & 1 & 0 & -1 & 0 & 0 & 0 & -1\\ -1 & -1 & 0 & 1 & 0 & 0 & 0 & 1 & 0\\ 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1\\ 0 & 1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 \end{pmatrix}.$$

Consider for example row 2. The B pairs involving team 2 include $(2,4)$, $(2,6)$, $(2,7)$ and $(2,9)$. With the pairs $(2,4)$ and $(2,7)$ team 2 has two H games, whereas with the pairs $(2,6)$ and $(2,9)$ it has one H game.

The matrix M is antisymmetric, that is,

$$M_{ij}=-M_{ji},$$

and for any row or column there are two $+1$'s and two -1's, with the other entries zero.

We can now describe a legal change in the z values, that is, one that ensures that each team has ten home games and ten away games. Choose a row, say $i=3$, choose a $+1$ on that row, say entry M_{38}, and change it to -1. Then change M_{83} to $+1$. The row $i=3$ then has three -1's and row $j=8$ has three $+1$'s. Choose a $+1$ in row $j=8$, say M_{89}, and change it to -1, and then change M_{98} to $+1$. Now take a $+1$ in row 9 and change it to a -1. Say we change M_{96} to -1; this forces M_{69} to change to $+1$. Then change a $+1$ in row 6, say M_{63}, to a -1, which forces a change of M_{36} to $+1$. As we have come back

to row 3, we now have an alternate matrix M that has two $+1$'s in each row and two -1's in each row, so we can stop. This is another legal matrix that has all teams playing ten home and ten away games. This is deemed a legal manipulation of the z values corresponding to the B pairs. Technically we could get rid of listing the z values on the chromosome and instead give the x values, the y values and the associated matrix. The entries in the matrix containing the z values also tell us how we are allowed to manipulate them.

Another way to look at the allowed manipulations of the z values arising from the matrix M is to look at it as a directed graph, which essentially summarises the allowed cycles [1].

7 Hard constraints in the scheduling part of the problem

There are also some hard constraints that we may need to satisfy in the scheduling part of the problem. For example EP = 2 want to have its first four games in rounds r_1 to r_4 away from home. In these weeks EP plays SF = 1, CL = 9, SU = 6 and PE = 4. The $(1, 2)$ pair is an A pair while the other pairs $(2, 9)$, $(2, 6)$, and $(2, 4)$ are B pairs. The arrangement of matches for these pairs are controlled by the x_1, z_8/y_8, z_6/y_6 and z_5/y_5 genes respectively. To ensure that EP = 2 has an away game for the first match for each of these pairs restricts the allowed values of these genes to be equal to $x_1 = 0$ ($z_8 = -1$, $y_8 = 2$ or 3) or ($z_8 = +1$, $y_8 = 2$) ($z_6 = -1$, $y_6 = 2$ or 3) or ($z_6 = +1$, $y_6 = 2$) ($z_5 = -1$, $y_5 = 2$ or 3) or ($z_5 = +1$, $y_5 = 2$).

These genes could be frozen at these values to ensure that EP has its first four games away from its home ground. Also, PT = 4 has requested that it should have home games in rounds r_1, r_4 and r_{10}. In these rounds, PT = 4 plays EF = 8, EP = 2 and SD = 5. This restricts the value of the genes as follows: $x_{12} = 0$ (corresponding to the A pair $(4, 8)$) ($z_5 = -1$, $y_5 = 2$ or 3) or ($z_5 = +1$, $y_5 = 2$) (same as for the EP request above) ($z_{12} = -1$, $y_{12} = 2$) or ($z_{12} = 1$, $y_{12} = 1$ or 2).

The other hard constraints imposed on the scheduling part of the problem include:

- SF(1) v EF(8) in r_{10} implies ($z_4 = -1$, $y_4 = 2$) or ($z_4 = 1$, $y_4 = 1$ or 2),
- WP(7) v EP(2) in r_{10} implies ($z_7 = -1$, $y_7 = 1$ or 3) or ($z_7 = 1$, $y_7 = 3$),
- SD(5) v WP(7) in r_1 implies $x_{14} = 0$,
- P(3) v EP(2) in r_8 implies $x_5 = 1$.

In the genetic algorithm below we would ensure that these genes are not altered in our chromosomes, and that all individuals in our starting configuration had the required values for these genes.

8 The genetic algorithm

From the previous sections we have managed to encode the scheduling part of the problem onto a chromosome which has 18 binary x-genes (for the A pairs), 18 ternary y-genes (for the B pairs) and a 9×9 matrix gene that tells us how we can manipulate the binary z-genes for B pairs to ensure that each team still has ten H and ten A games. Such a chromosome is shown in Figure 2 Our genetic algorithm uses a population of around 100 such chromosomes, each with the same ascribed listing. A new population is formed by using the three genetic operators: mutation, crossover and selection.

- Mutation consists of changing, with low probability, randomly selected digits from the chromosome. This can be done freely for any of the x or y values but a change to a z value must be incorporated using a legal manipulation on the matrix as outlined previously. Changing an x-gene forces two changes to the schedule, either HAAH or AHHA. Changing the value of a y-gene also results in changing the schedule for two matches for the B pairs. Any change to the matrix gene requires a change to at least three of the z-genes as the lowest order cycle within the directed graph is of order three. Hence an allowed mutation of the matrix z-genes results in changing at least three matches. One would consequently wish to make mutations to the z-genes with a lower probability than mutations to the x and y genes.

- Crossover is a technique whereby pairs of chromosomes mate and produce offspring whose genetic material is a combination of the chromosomes of both parents. Crossover can only occur anywhere along the xy part of the chromosome as indicated in Figure 3. One can also have multiple point crossovers (Figure 4). Mutation and crossover produce an altered population, which is hopefully not too different from the original population.

- In selection a new population is generated by choosing copies of the better members (but not necessarily the best) of this altered population using a probability distribution weighted by the fitness of the chromosomes. The fitness of a chromosome is a number determined by how well the associated home/away sequence for each team approaches an alternating sequence and how well the individual satisfies the other requirements requested by the WAFC. This will be discussed in more detail below.

9 The fitness function

The main task of the fitness function is to try to give each team a sequence of home and away games which is approximately alternating. To this end we need to penalise the fitness function if there are consecutive home or away games. Since a sequence of A games in some part of a team's H/A sequence would imply a sequence of H games elsewhere, it is enough to consider away games only. Typically one might want to penalise occurrences of longer strings of A's more strongly than shorter strings of A's. One should also factor into these considerations the likelihood of sequences like ABG, where G = general bye (in the week before round 12) and B = bye, which are not desirable because this team is away from home for three consecutive weeks. The fitness of an individual could then be determined to be, say 500, minus the penalties for each of the sequences listed in Table 4. Note that this penalisation is simultaneously applied to all nine teams for each individual in the population for each generation. The initial figure from which we subtract is arbitrary and may be varied during simulation depending on how close we are to an optimal solution. Notice in Table 4 that one would want to typically penalise an occurrence of AAA much more than two occurrences of AA, hence it is given four demerit points in this particular assignment.

Penalty	Occurrence
1	AA, GBA, GAB, AGB, ABG, BGA, BAG
2	AAG, AGA, GAA
3	AAB, ABA, BAA
4	AAA, AAGB, AABG, ...
5	AAAG, AAGA, AGAA, GAAA
6	AAAB, AABA, ABAA, BAAA
8	AAAA
10	AAAAG, AAAAB, ...
15	AAAAA

Table 4: A possible penalisation scheme for occurrences of consecutive sequences of away games for the nine clubs; A = away, B = bye and G = general bye.

The H/A sequences for each team are given in Table 2. SF for example has the H/A sequence

HBH<u>AAA</u>HAHH<u>BGA</u>HAHAH<u>BAA</u>HAH,

which would attract penalties for each of the underlined sections. In this particular case, according to Table 4, SF would attract the penalty −8.

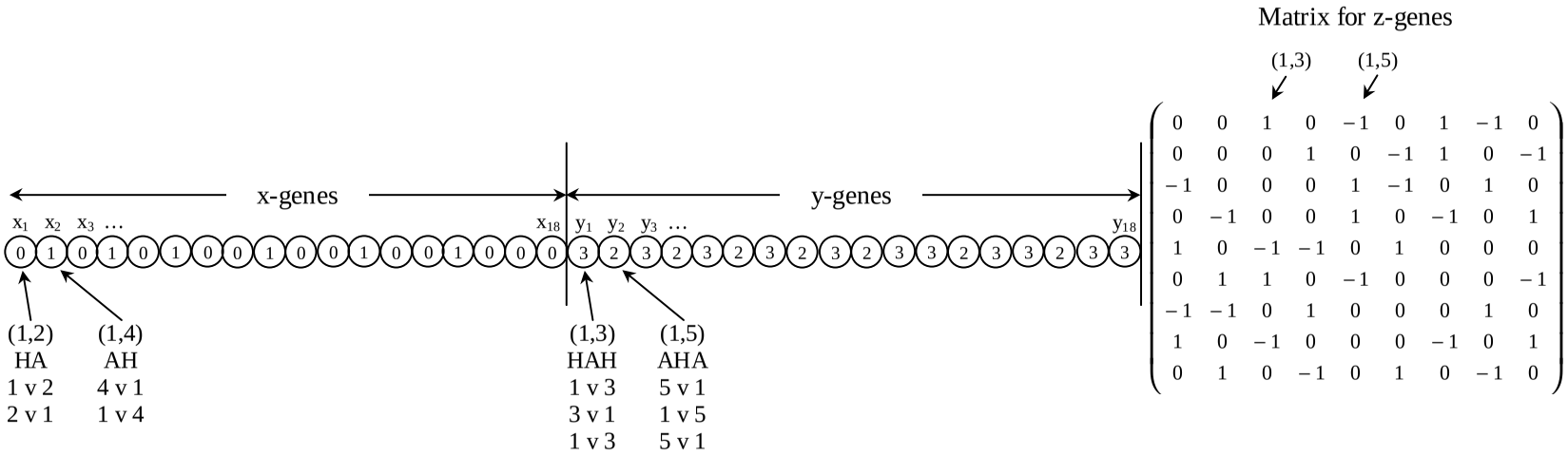


Figure 2: A typical chromosome for the scheduling problem with 18 binary x-genes, 18 ternary y-genes and a matrix gene containing the 18 z-genes. The actual values given to the genes here are specific to the listing and schedule given in Table 1.

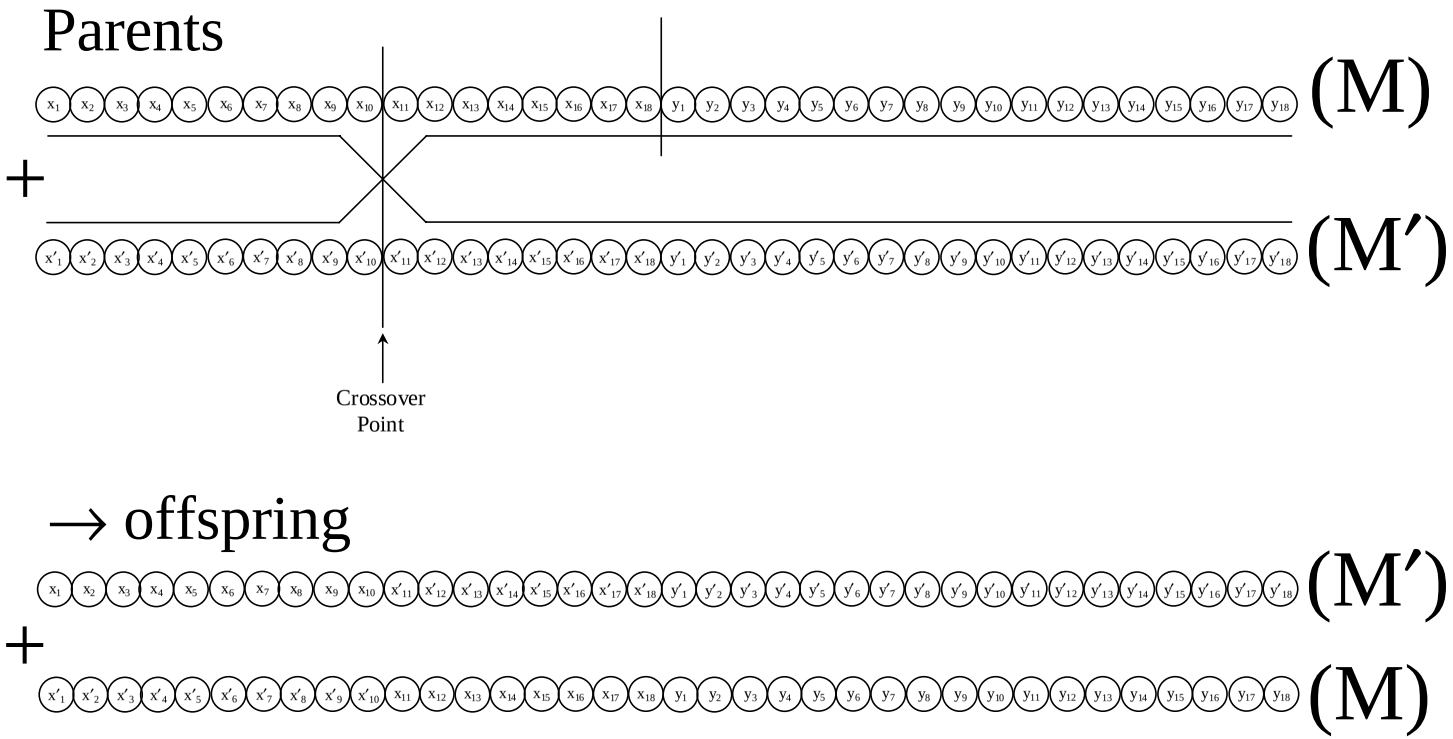


Figure 3: An example of a crossover between two individuals to produce offspring that have genetic sequences derived from both parents.

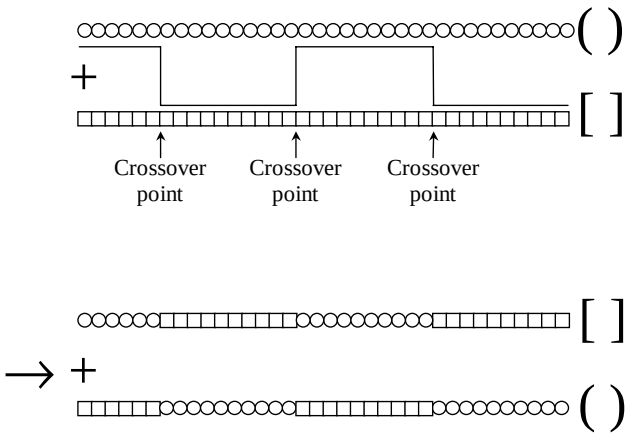


Figure 4: Multiple point crossover at various places on the xy portion of the parent chromosomes.

10 The population dynamics

In this genetic algorithm simulation we may typically proceed with an initial population of 100 individuals, which has allowed values of x, y and z genes. These chromosomes may have been generated by the big wheel algorithm, or by manipulations of such solutions. These individuals all have the same listing component and hence the same A pairs and B pairs of teams that play each other twice or thee times respectively. Each individual also has a matrix gene.

These 100 individuals would be grouped into 50 pairs and would then be allowed to crossover, typically with a crossover probability of approximately 0.05. This will produce 100 "new" individuals, some of which may be unchanged from the parent population. Mutation can be added either before or after crossover, and typically operates with a probability of the order of 0.003. This would mean that one in every three individuals in our population of 100 would be mutated. This mutation would take the form of a change in the values in the x or y genes, or a contorted change in z-genes through the matrix gene.

The effect of mutation and crossover is to ensure that we get a new population that is not very different from the previous one. The fitness of some individuals will have increased slightly while that of others will have decreased. The next step in the algorithm is to apply selection. Because individuals are selected with a probability weighted according to their fitness, the new population will consist predominantly of fitter individuals from the previous population, and is likely to have a greater average fitness.

The whole procedure of mutation, crossover and selection is now repeated. With each iteration we hope for a small increase in fitness, and when no further increase seems likely we halt the process and use the best individual as a schedule for the competition.

References

[1] G. A. Christos and R. J. Simpson, "Using Genetic Algorithms to set the fixtures for the 'Westar Rules' football competition", Curtin University of Technology, Mathematics and Statistics, Technical Report 3/00, not submitted for publication (2000).

[2] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison–Wesley, Reading (1989).

[3] J. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor (1975).

[4] J. Holland, "Genetic algorithms", *Scientific American* (July 1992), 66–72.

COLLECTING STATISTICS AT THE AUSTRALIAN OPEN TENNIS CHAMPIONSHIP

Stephen R. Clarke*
School of Mathematical Sciences
Swinburne University of Technology
PO Box 218, Hawthorn
Victoria 3122, Australia
sclarke@groupwise.swin.edu.au

Pam Norton
School of Mathematical Sciences
Monash University
PO Box 28M
Victoria 3800, Australia
pam.norton@sci.monash.edu.au

Abstract

This paper discusses methods used to collect statistical data at the Australian Tennis Open. The process of data entry, and the uses of the statistics are discussed, along with methods used to ensure consistency. Areas where the statistics may be misinterpreted are highlighted, and suggestions for improving the process are made.

1 Introduction

For many fans the discussion of statistics associated with their particular sport is an integral part of their enjoyment. For many events, where previously limited statistics could be gained from the newspapers or magazines the following day or week, statistics are now given real time on live broadcasts or via the web. These statistics are not only used by the fans, but by modellers such as Magnus and Klaassen [2, 3]. To satisfy the media and the public's thirst for statistics, the collection and dissemination of data concerning the event is now an important element in the organisation of many sporting events. The Australian Open Tennis Championships are no exception.

The Australian Open is unique among the grand slam championships in that statistics are collected at all courts for all events. As at 2001, the other grand slams only collected statistical data on their main courts, and only for the main events. At the Australian Open, statisticians are rostered on all 22 courts, for senior, junior and veteran events, singles and doubles. This even extends to the qualifying tournament. However radar guns are installed only on seven courts (Rod Laver Arena, Vodaphone Arena, Show Courts 1, 2 and 3, and Courts 6 and 7), so service speed is only measured on some matches.

Early in 2000, we wrote to Tennis Australia requesting some data from the 2000 Open for analysis. They replied with a request for our assistance in improving the professionalism of their "scoreboard operators". Traditionally this role had often been filled by ball persons. These applicants had passed the maximum age allowed for the position of ball person, but wanted to continue an association with the Open. While interested in tennis, these applicants had no inherent interest in statistics, and the Open organisers had received some negative feedback on their attitude and the quality of statistics collected. We were requested to publicise the position among our students. This resulted in about 15 people associated with Monash and Swinburne Universities, including the authors, forming part of the 60 strong army of tennis statisticians for the 2001 Australian Open. Some continued in this role in 2002.

In this paper we describe the methods used to collect statistical data at the tennis, and discuss some associated statistical issues.

*The authors would like to thank Chris Simpfendorfer and the staff of the Australian Open for their assistance in collecting and providing statistical data.

2 Selection of statisticians

The human resources personnel handle the selection and employment of tennis statisticians in a very professional manner. Applications are called early in the previous year, and applicants are required to fill out an application form and attach a resume. Selected applicants are interviewed, and undergo a short test collecting data on one game from a television replay. Essential skills are knowledge of tennis and some familiarity with computers. Successful statisticians are notified late in the year, and are required to attend a training evening. They were also required to attend the Nike junior tournament, at which statisticians from previous years showed beginners the ropes. While old hands usually only attended one day of this tournament, beginners could return a second day if they felt it necessary.

The position carries a small stipend, and a daily allowance for meals. In addition, a full uniform of shirts, shorts, tracksuit, runners, socks and hat is provided. The staff ID gives ground entry to all days of the tournament, and a guest ground pass is also provided. The statisticians are provided with a common room during the Open. Clearly the main motivation in performing the job for most, if not all, of the statisticians is to be a part of the Open. The position is one that suits university students, for whom the pay is reasonable and the intrusion into their holidays marginal. However few full-time employees would be willing to give up two weeks of their holidays to undertake the position. For this reason, average tenure is a few years, and turnover is great. Virtually all the statisticians were under 25, many under 20. This was in contrast to the line umpires, who were a more mature group. To be an umpire requires one to join an association, do regular training, and have a long term commitment to quality umpiring. Perhaps with the increasing number of people applying for the position, this is something tennis statisticians could investigate.

A couple of suggestions from the authors to improve the procedure were taken up by the organisers. We suggested changing the name from Scoreboard Operator to Tennis Statistician, as the task required much more than the original name implied. We also suggested removal of some sections from the application form, such as parental approval for night duty, as these implied young people were required.

3 Statistics collection at the Australian Open

The tennis statistician sits on court, usually just behind the umpire on the outside courts, with a notebook computer on a small table linked to the central IBM scoring network. The show courts usually have a purpose built stand capable of seating three statisticians. Conditions are often difficult, particularly on hot days. The screen can be hard to read in the sunlight, and umpire's chairs, television cameras and the like sometimes impair court vision. Data entry is via the keyboard, and usually requires just the arrow and the <Enter> keys. While this may sound antiquated, with little practice it is extremely quick, and can even be performed without looking at the keyboard or the screen. With the cramped conditions on court, there is often little room for mouse operation anyway. Statisticians resort to hand sheets if the central computer goes down, and must then catch up computer entry during a break in play.

The statistician works independently of the umpire, but has to ensure the score is consistent with what the umpire calls. When the statistician enters the point result, the information is relayed back to the central computer, which in turn updates the on court scoreboard, the various scoreboards around the venue, and the statistics which go live to the media and the internet. When match point is won, the computer advances the winner in the draw. Hence the statistical data entry scoring system is a real-time system, and the heart of the Australian Open. The statistics are also reported back to players and coaches in the form of a two page summary forwarded after the match.

Interestingly, in singles, the statistician does not enter the winner of the point, but the last player to make a play on the ball. Thus an entry of Agassi/forehand/forced error would result in the computer crediting the point to Agassi's opponent and advancing the score appropriately. The person serving the first game of the match has to be entered, and the computer tracks the server for the remainder of the match. For singles, each serve is entered as one of in play/fault/winner/ace (lets are not recorded), the

point conclusion is entered as player to make last play, one of forehand/backhand/overhead/volley, and one of unforced error/forced error/winner. In addition, if either or both players are at the net when the point is concluded, this is entered. While this may sound straightforward, a left-handed player with double-handed forehand and backhand requires vigilant concentration. Background noise (e.g. from Swedish fans) can make it difficult to hear line and central umpires' calls, and one cannot look at the scoreboard for confirmation (since the scoreboard is awaiting your entry). The program will not accept the entry unless all required fields are entered, and protects against certain errors. For example, if the first serve is in play, and the second is a fault, the program will flag an error, which must be corrected before the data is accepted and the next point can be entered. Some statistics of interest that are not collected include the number of strokes in a rally, the side (forehand or backhand) of any volleying winners or errors, and the side of the opponent's court to which winning shots are hit.

The main subjective element comes in with forced and unforced errors. Conceptually a forced error is the result of the opponent's good play, whereas unforced is the result of a player's poor play. The decision is usually made on the basis of whether the player has had time to prepare and position adequately to return the ball. Statisticians are regularly observed by experienced assessors to check for uniformity of application. However there remain situations which training will not have covered specifically. (While being assessed, in one of the author's matches, a player fell over while attempting to return the ball, and managed an ineffectual play on the ball while lying on his back. The player clearly did not have time to adequately prepare, but then it was his own fault that he fell.) Two statisticians are rostered for each court, and each works for an hour followed by an hour's break. Thus interpretation may change within a match. As the tournament progresses, fewer statisticians are required, and the less capable are rostered off. To ensure greater consistency for finals, five "gun" statisticians will usually be chosen. One will record the men's singles from the quarter finals onwards, another the women's singles, etc.

For doubles, data entry is simplified. Service result is as for singles, the service return is recorded as one of in play/forced error/unforced error/winner, and finally the pair to win the point is recorded. At the beginning of the first two games of each set, the statistician has to enter the server and receiver of the first point. With unknown, or similar partners, and with some teams wanting to keep their intentions from their opponents until the last second, this can be difficult. With statisticians changing midway through sets, care must be taken in written cues to player identities. One quickly learns to use colour of shorts rather than shirts for male players—the former are rarely changed during a match. Thus while doubles statistics will not give information on the individual players' statistics throughout the match, they will show their serving and return of serving statistics.

Server	Set score	Point score	1st serve	2nd serve	Last play	Point action		At net?
Agassi	5-2	0–0	in play		Clement	foreh	forc err	
Agassi	5-2	15–0	fault	in play	Clement	foreh	winner	
Agassi	5-2	15–15	in play		Clement	volley	winner	Clement
Agassi	5-2	15–30	ace					
Agassi	5-2	30–30	fault	in play	Clement	backh	unfor err	
Agassi	5-2	40–30	fault	in play	Agassi	backh	winner	

Table 1: Audit trail for the last game of the 2001 Australian Open men's singles final.

The Audit trail for the last game of the 2001 Australian Open men's singles final is shown in Table 1. The authors obtained similar data for all men's and women's singles matches at the 2000 Australian Open. Table 2 is a match summary as produced by the statistics collection package.

4 Radar

Statisticians also operate the radar facility which is installed on the main courts. This is used to collect data on the speed of service. Radar is not operated on all matches, and a separate statistician is allocated when necessary for this function. The program is mouse operated. The operator has to enter the server of the first game, and reset the radar just before each player serves. Since the speed may be registered for other movements, such as a bird or leaf, this needs to be done as late as possible. The statistician then enters the area of the court in which the ball lands. For faults, this can be inside the singles court but long, into the net, or outside the singles court. For good serves, the service box is divided into wide, middle or centre. The radar operator also enters if the serve is an ace or winner, or, at the point conclusion, the winner of the point. The program automatically keeps track of the score and hence the server. Periodically the radar will show a speed obviously incorrect, which the statistician must manually edit during a break.

Table 3 shows data as published on the World Wide Web (www.ausopen.org) by the Australian Open, which incorporates statistical and radar information. Similar data for 2001 and 2002, together with the point-by-point data from the men's and women's singles at the Australian Open in 2000, have been used by Norton and Clarke [4] to compare the server's advantage at Wimbledon, and the French and Australian Opens, and to compare singles and doubles service characteristics. Data for this paper were also obtained from such summaries for 2001 and 2002. The statistic that is missing is the percentage of points won on serve. Interestingly this is the parameter that is used in most models of tennis, such as Croucher [1] and Pollard [5]. However, it is easily calculated from the first serve percentage, winning percentage on first serve and winning percentage on second serve.

	Clement				Agassi			
Set	1	2	3	M	1	2	3	M
1st serve points	18	11	18	47	21	15	17	53
1st serve points won	15	5	8	28	15	13	10	38
2nd serve points	16	11	16	43	4	7	17	28
2nd serve points won	5	6	8	19	1	3	9	13
1st serve aces	4	1	0	5	2	4	1	7
2nd serve aces	0	0	0	0	0	0	0	0
1st serve winners	0	0	1	1	2	1	0	3
2nd serve winners	1	0	0	1	0	0	0	0
double faults	2	3	2	7	1	0	0	1
FH unforced errors	3	3	9	15	5	1	6	12
FH winners	7	2	7	16	2	1	4	7
BH unforced errors	7	5	8	20	6	4	6	16
BH winners	0	2	6	8	0	1	5	6
OH unforced errors	0	0	1	1	0	0	0	0
OH winners	0	0	1	1	0	0	0	0
Volley unforced errors	0	0	0	0	0	1	1	2
Volley winners	1	0	1	2	0	0	1	1
Net points won	1	1	3	5	2	4	5	11
Net points lost	0	1	3	4	1	1	2	4
Break opportunity	1	0	5	6	2	3	10	15
Break conversion	1	0	1	2	2	2	3	7

Table 2: Match summary for the Clement–Agassi final, Australian Open Men's Singles Championship, 2001.

Match summary	Safin	Johansson
1st serve %	90 of 142 = 63%	64 of 115 = 56%
Aces	13	16
Double faults	2	4
Unforced errors	36	43
Winning % on 1st serve	60 of 90 = 67%	55 of 64 = 86%
Winning % on 2nd serve	28 of 52 = 54%	27 of 51 = 53%
Winners (including service)	39	53
Break point conversions	3 of 6 = 50%	3 of 14 = 21%
Net approaches	29 of 53 = 55%	36 of 49 = 73%
Total points won	121	136
Fastest serve	209 kph	209 kph
Average 1st serve speed	187 kph	179 kph
Average 2nd serve speed	144 kph	140 kph

Table 3: Statistical summary of the 2002 Australian Open men's singles final.

5 Some statistical issues

In interpreting the statistics, and comparing statistics between majors, certain points should be borne in mind. Some of the statistics are very objective, while others require varying degrees of subjectivity on the part of the statistician. Since the statistician operates the scoreboard, most point results would be correct. However there may be cases where a statistician gets the score incorrect, and has to re-enter the data. This requires backtracking to the point in question, and re-entering the point and the subsequent points. If the statistician cannot remember, there may be time pressures to enter an ace, as this is the quickest to record. However such cases would be very isolated. We do not know if there is any rationalisation between the statisticians' record and the umpire's score sheet. This would be a simple way to ensure that the number of points won by each player was correct, even if the sequencing was astray. There is a little subjectivity in deciding between a winner and a forced error, as it depends on the amount of racquet the defending player gets on the ball. By far the greatest degree of judgement is needed for the forced/unforced error judgement. It can depend on the speed of the ball, the position of the player and his opponent, and the event. Thus an unforced error in men's tennis may be forced in women's or junior events. The authors feel much more group training could be put into this aspect. However many of the published statistics combine forced and unforced errors, which minimises the effects of subjectivity of classification.

Some of the definitions may seem surprising, and could easily be misinterpreted. For example, for 2002, a volley includes a half volley, defined as a "ball struck down low (at feet) just after the ball has hit the ground with minimal to no backswing", and there was sometimes confusion on the correct recording of a drive volley. When a player is at the net is also difficult to gain uniformity of application. Thus a point recorded as an Agassi winner with Rafter at the net may be interpreted as a passing shot, when in fact it was a drop shot, played by Agassi with both players at the back of the court, which Rafter didn't get his racquet to. There is also a tendency to sometimes make adjustments for the skill of a particular player. Thus the same shot classified as a forced volley error by Agassi may be recorded as unforced when hit by Rafter, since Rafter is a known better volleyer than Agassi. Of course, allowances are made for the power differences between men's and women's tennis. Thus virtually all errors off a man's first service would be classified as forced, whereas this might only apply to the strongest of the women's first serves. Prior to 2002, there was little attempt to achieve consistency of definition across major championships. This was rectified to some extent for the 2002 championships when definitions were agreed to by operators of all the major tournaments.

The authors feel there should be some study made as to the interpretation of the statistics by the

public and the media, to produce underlying definitions to be used in marginal cases. For example, a winner implies the player hit a shot that was beyond any reasonable expectation of return—the player won the point off his own racquet by hitting an un-returnable good shot. In practice, this translates into a working definition of a winner as any shot which the opponent does not get his racquet to, or which tips the racquet and goes behind. One of the authors had a player misjudge his approach to a ball on which he had a play, which then hit his foot. Under the definition, this is a winner, as the player did not get his racquet to the ball. However under an underlying definition based on the interpretation that will be made of the statistic, this is an unforced error. Similar problems arise when players completely mishit the ball.

One problem of the statistical packages used was the absence of any concept of a missing value. If a statistician missed a point, and had no idea whether it was an ace or winner or whatever, they had to enter something (and quickly!). If a radar operator missed the court area the ball landed in, the package insisted they enter something. If the facility for entering a missing value was available, the statistics could be made more reliable by promulgating a culture of not making statistics up. When operating radar, the equipment sometimes gave a speed reading obviously incorrect. Statisticians were instructed to insert the average value of the player for the appropriate serve up to that point. This clearly depends on when the correction is made, and affects both the mean and variation of the recorded speeds. A simple facility to enter a missing value would be preferable. Another interesting point was that although recorded, the speed of faults was not stored. While the programmers obviously felt they were only recording statistics of valid serves, we felt much useful information was discarded. Surely coaches and players would be interested in differences between speeds of faults and good serves, as it may give a lead to corrective action.

One of the reasons for collecting statistics is to increase our knowledge of the game of tennis. While summary statistics are available via the web, detailed modelling requires the point by point data. This is difficult to obtain. Despite the full cooperation of Australian Open staff to the authors' original request for data, it still took several months to obtain the point by point statistics for the 2000 Open through IBM, and we have not obtained any of the radar data. In addition, different years' data are in different formats. We have not attempted to obtain data from the other grand slam tournaments, but suspect these would be different again. With a little effort, the data from the tournaments could be collected in one place. The data could then be documented, checked and merged where necessary. For example, data collected by radar operators and statisticians could be compared. The data from all tournaments could then be put in the same format, archived and made available to interested parties. There may be issues of ownership and cost to be resolved, but such an archive would greatly enhance the value of the data.

6 Conclusions

It is interesting to compare the collection of statistics in tennis with other major Australian sports such as football and cricket. In AFL football, statistical data collection as undertaken by Champion Data requires two people—one caller and one recorder. The recording can be done live or from a television replay. With eight matches per weekend, there are at most two or three simultaneous matches. Furthermore, the matches are played outside normal working hours. This allows a small consistent team to be built up, which remains relatively stable over the years. Similarly, cricket statistics as shown live on television can be recorded from the television broadcast by a team of three statisticians. This means the same small team can be used for overseas matches. There is also some time between balls to discuss interpretation.

The collection of statistical data at the Australian Open requires a large team of over 60 statisticians for a short period of the year. Each statistician is generally alone on the court, and there is little time for reflection. To continually improve the quality of the collected statistics requires increased effort to reduce turnover and improve training. Perhaps an association of sports statisticians, similar to that of the tennis umpires, might be a path toward maintaining interest of statisticians past their youth,

and reduce turnover. The basic recording and reporting program could be made available as a stand-alone system for a notebook computer, so that statisticians could record matches during the year, and compare their analysis to that of a "gun". An exchange program between the major tournaments might encourage greater uniformity of interpretation between the tournaments. Perhaps some statistics courses could include a practical component of data collection. It would give students an appreciation of the commitment, scale and intensity of the data collection operation, and its limitations. Most of our students who participated certainly thought it a worthwhile exercise.

For their part, the authors gained a lot from participating in the data collection. It gave us an appreciation of the practical difficulties in collecting data, and in ensuring consistency among data recorders. The original aim in contacting Tennis Australia was to obtain some statistics for analysis, and this aim was realised. However being intimately involved with the collection gives some ownership of the data, and a feeling for where errors may not be random and statistical assumptions may not hold. While preferring to model and analyse data, we still enjoyed the experience of getting our hands dirty in the collection. With our wider experience, we could also give the younger tennis statisticians an appreciation of the uses made of the data collected. We would certainly recommend to any statistician to take the opportunity to participate in a similar exercise should the chance arise.

References

[1] J. S. Croucher, "The conditional probability of winning games of tennis", *Res. Quart. for Exercise and Sport*, **57** (1986), 23–26.

[2] J. R. Magnus and F. J. G. M. Klaassen, "On the advantage of serving first in a tennis set: four years at Wimbledon", *The Statist.*, **48** (1999), 247–256.

[3] J. R. Magnus and F. J. G. M. Klaassen, "The effect of new balls in tennis: four years at Wimbledon", *The Statist.*, **48** (1999), 239–246.

[4] P. Norton and S. R. Clarke, "Serving up some grand slam service statistics", these *Proceedings*.

[5] G. H. Pollard, "An analysis of classical and tie-breaker tennis", *Austral. J. Statist.*, **25** (1983), 496–505.

DYNAMIC EVALUATION OF CONDITIONAL PROBABILITIES OF WINNING A TENNIS MATCH

Shaun Clowes, Graeme Cohen and Ljiljana Tomljanovic*
Faculty of Information Technology
University of Technology, Sydney
PO Box 123, Broadway
NSW 2007, Australia
shaman@progsoc.org, graeme.cohen@uts.edu.au, ltomljan@socs.uts.edu.au

Abstract

Let p and q be the probabilities of players A and B, respectively, winning a point in a game of tennis ($p+q=1$). We describe a program which has been devised and implemented to give the probability of A or B winning the match ("best of three tiebreaker sets") from any stage of the match.

1 Introduction

It is possible, based on a study of earlier matches between two players under similar match conditions, to give each a probability of winning a point in any subsequent match between them. The probabilities may be subjectively altered, as it is felt warranted, and this may occur as the match is in progress. What are the respective chances of the two players winning the match, from any particular stage of the match?

A corresponding question was considered by Croucher [2], but for a fixed probability of winning a point and with regard to a single game.

Our players, A and B, will be assigned respective probabilities p and q of winning any point ($p+q=1$). In the first place, these may be determined as the proportion of points won by each in all matches played previously under similar conditions. A more detailed analysis, such as that of Carter and Crews [1], might consider separately the proportions p_A and p_B of points won by A and B, respectively, in service games, so that, for example, $q_A = 1-p_A$ is the probability of B winning a point against A's service. This level of detail is not required for our purposes.

It is crucial to our work that "the outcome of any point, game or set is independent of the outcome of any other point, game or set" (Pollard [4]). Some doubt has recently been cast on this statement by Klaassen and Magnus [3]. They concluded that winning a point in tennis has a positive effect on winning the following point and that at important points the server has a disadvantage. However, they also conclude that divergence from independence is small, particularly for strong players, and "in many practical applications concerning tennis—such as predicting the winner of the match while the match is in progress—the iid hypothesis will still provide a good approximation *if we correct for the quality of the players*" [their emphasis]. We explicitly allow this correction at any stage of the match, and so are in fact reinforced in our application of independence.

*This project began as an assignment carried out by the first and third authors in a subject taught by the second author.

Our aim is to describe the calculations in the implementation of a publicly-available program which gives the probability of A or B winning a "best of three tiebreaker sets" match, from any stage of the match, as well as the probability of winning the current game or set. This would be of interest to media commentators and to those who might be having a bet on the result. Since it assumes an overall probability of winning a point, whether serving or receiving, results produced by the program should be quoted only after an even number of games have been played.

For example, if player B is initially given a probability of only 0.44 of winning a point (based on past experience), but has just won the first four games of the first set, would you bet on B to win that set? Would you bet on B to win the match? What odds would you accept? Our calculations show that, if you still believe B's previous form to apply, then there is a 75% chance of winning that set, but only a 20% chance of winning the match. If you feel it warranted to raise B's probability of winning a point to a modest 0.5, then there is better than a 90% chance that B will win this set and better than a 70% chance of winning the match.

2 Conditional probabilities of winning a game

Let A and B have probability p, q, respectively, of winning a point ($p + q = 1$). We take the point of view always of A winning a game, set or match. If A leads 30–15, for example, then A wins the game by winning the next two points (probability p^2), or two of the next three including the third (probability $2p^2q$), or by reaching deuce (probability $3pq^2$) and winning from there. The probability d of winning from deuce satisfies

$$d = p^2 + 2pqd,$$

so that

$$d = \frac{p^2}{p^2 + q^2}. \tag{1}$$

Hence the probability of A winning the game when ahead 30–15 is given by

$$p^2 + 2p^2q + 3pq^2\frac{p^2}{p^2 + q^2}.$$

Such calculations have previously been given by Croucher [2], for example.

We need to determine the corresponding probability from any stage in the game in such a way that it may be readily programmed.

Let the current scoreline in an uncompleted game be specified as x points won by A ($x \geq 0$) and y points won by B ($y \geq 0$), including points beyond deuce. The value of x or y is incremented by 1 as each point is played, until the game is won. Notice that: if $y \leq 2$ then $x \leq 3$; if $y = 3$ then $x \leq 4$; if $y \geq 4$ then $x = y$ (the score is deuce), or $x = y + 1$ (advantage A) or $x = y - 1$ (advantage B). Let $G(x, y)$ denote the probability that A wins the game from this scoreline, and put $G_0 = G(0, 0)$.

If, initially, $y \leq 2$ and A wins the game without reaching deuce, then A must win a further $3 - x$ points as well as the final point of the game, with B winning at most a further $2 - y$ points. The probability of this is

$$G_1(x, y) = \sum_{k=0}^{2-y} \binom{3 - x + k}{k} p^{4-x} q^k.$$

If, still for $y \leq 2$, the game first reaches deuce then A wins a further $3 - x$ points and B wins a further $3 - y$ points. The probability of this happening is

$$G_2(x, y) = \binom{6 - x - y}{3 - y} p^{3-x} q^{3-y},$$

and the probability that A then wins the game is $G_2(x, y)d$.

Values of x and y	Value of $G(x,y)$
$y \leq 2$ (so $x \leq 3$)	$G_1(x,y) + G_2(x,y)d$
$y = 3$ and $x \leq 3$	$G_2(x,3)d$
$y \geq 3$ and $x = y+1$	$p + qd$
$y \geq 4$ and $x = y$	d
$y \geq 4$ and $x = y-1$	pd

Table 1: The probability $G(x,y)$ that A wins a game, given the number of points x won by A and the number y won by B.

Take $x = 2$ and $y = 1$ to reconstruct the example above in which A is leading 30–15. Take $x = y = 0$ to determine G_0.

The calculations are easier when $y \geq 3$. The complete situation is summarised in Table 1.

When the games in a set reach six all, a (twelve-point) tiebreaker game is played. For us, the only effective differences between a tiebreaker game and an ordinary game are that the tiebreaker is played to seven points, rather than four, with "deuce" reached after six points all, rather than three points all. Let $B(x,y)$ be the probability that A wins a tiebreaker game, given that A has already won x points ($x \geq 0$) and B has won y points ($y \geq 0$). Put $B_0 = B(0,0)$ and

$$B_1(x,y) = \sum_{k=0}^{5-y} \binom{6-x+k}{k} p^{7-x} q^k,$$

$$B_2(x,y) = \binom{12-x-y}{6-y} p^{6-x} q^{6-y}.$$

Then, in the same way as we obtained Table 1, the values for $B(x,y)$ are given in Table 2. The expression for d in equation (1) is used again; it is the probability that A wins the tiebreaker game from six points all.

Values of x and y	Value of $B(x,y)$
$y \leq 5$ (so $x \leq 6$)	$B_1(x,y) + B_2(x,y)d$
$y = 6$ and $x \leq 6$	$B_2(x,6)d$
$y \geq 6$ and $x = y+1$	$p + qd$
$y \geq 7$ and $x = y$	d
$y \geq 7$ and $x = y-1$	pd

Table 2: The probability $B(x,y)$ that A wins a tiebreaker game, given the number of points x won by A and the number y won by B.

3 Conditional probabilities of winning a set

Suppose, in an uncompleted set, that A has won g games ($g \geq 0$), B has won h games ($h \geq 0$), and, in the current game of that set, A has won x points and B has won y points. Let $S(g,h;x,y)$ be the probability that A wins the set. If $g = h = 6$, then the current game is a tiebreaker.

In the manner of an odometer, the point counters x and y are put back to zero with each new game and the game counters g and h are incremented by one, according as who wins the game, A or B, respectively.

Notice that: if $h \leq 4$ then $g \leq 5$; if $h = 5$ then $g \leq 6$; if $h = 6$ then $g = 5$ or 6.

Recall that the probability that A wins a game (not a tiebreaker) from love all is G_0. Put $H_0 = 1 - G_0$; this is the probability that B wins a game. The probability that A wins a tiebreaker game is B_0.

Suppose initially that $x = y = 0$.

If $h \leq 4$ and A wins the set other than 7–5 or 7–6, then A must win a further $5 - g$ games as well as the final game of the set, with B winning at most a further $4 - h$ games. The probability of this is

$$S_1(g,h) = \sum_{k=0}^{4-h} \binom{5-g+k}{k} G_0^{6-g} H_0^k.$$

If A wins the set 7–5, then A must win a further $5 - g$ games and B a further $5 - h$ games, so that the score reaches 5–5, and then A must win the next two games. The probability of this is

$$S_2(g,h) = \binom{10-g-h}{5-h} G_0^{7-g} H_0^{5-h}.$$

If A wins the set 7–6, then the score must first reach 5–5, then A and B must win one game each (there are two ways to do this), and then A must win the tiebreaker game. The probability of this is

$$S_3(g,h) = 2\binom{10-g-h}{5-h} G_0^{6-g} H_0^{6-h} B_0.$$

When $h \leq 4$, we therefore have

$$S(g,h;0,0) = S_1(g,h) + S_2(g,h) + S_3(g,h).$$

It is easy now to complete Table 3, giving the probability that A wins a set after any number of completed games in the set.

Values of g and h	Value of $S(g,h;0,0)$
$h \leq 4$ (so $g \leq 5$)	$S_1(g,h) + S_2(g,h) + S_3(g,h)$
$h = 5$ and $g \leq 5$	$S_2(g,5) + S_3(g,5)$
$h = 5$ and $g = 6$	$G_0 + H_0 B_0$
$h = 6$ and $g = 5$	$G_0 B_0$
$h = 6$ and $g = 6$	B_0

Table 3: The probability $S(g,h;0,0)$ that A wins a set, given the number of games g won by A and the number h won by B, with no points played of the subsequent game.

Now consider the more general situation in which $x + y > 0$, and put $H(x,y) = 1 - G(x,y)$. We must determine all the variations of the following basic result, which is easily seen to be true. When $h \leq 4$ and $g \leq 4$, then

$$S(g,h;x,y) = G(x,y)S(g+1,h;0,0) + H(x,y)S(g,h+1;0,0).$$

One such variation, for example, is if $h \leq 4$ and $g = 5$. In that case,

$$S(g,h;x,y) = G(x,y) + H(x,y)S(g,h+1;0,0),$$

since A wins the set if B loses the current game.

All the possibilities are given in Table 4.

4 Conditional probabilities of winning the match

Let $M(s,t;g,h;x,y)$ be the probability that A wins the "best of three tiebreaker sets" match, when A has won s sets and B t sets ($s = 0$ or 1, $t = 0$ or 1), with g, h, x and y having their previous significance

Values of g and h	Value of $S(g,h;x,y)$
$h \leq 4$ and $g \leq 4$	$G(x,y)S(g+1,h;0,0) + H(x,y)S(g,h+1;0,0)$
$h \leq 4$ and $g = 5$	$G(x,y) + H(x,y)S(5,h+1;0,0)$
$h = 5$ and $g \leq 4$	$G(x,y)(S_2(g+1,5) + S_3(g+1,5))$
$h = 5$ and $g = 5$	$G(x,y)(G_0 + H_0B_0) + H(x,y)G_0B_0$
$h = 5$ and $g = 6$	$G(x,y) + H(x,y)B_0$
$h = 6$ and $g = 5$	$G(x,y)B_0$
$h = 6$ and $g = 6$	$B(x,y)$

Table 4: The probability $S(g,h;x,y)$ that A wins a set, given the number g of games already won by A and the number h already won by B, and the number x of points won by A and the number y won by B in the current game $(x+y>0)$.

Values of s and t	Value of $M(s,t;g,h;x,y)$
$t = s = 0$	$S(g,h;x,y)(S_0 + T_0S_0) + T(g,h;x,y)S_0^2$
$t = 0$ and $s = 1$	$S(g,h;x,y) + T(g,h;x,y)S_0$
$t = 1$ and $s = 0$	$S(g,h;x,y)S_0$
$t = s = 1$	$S(g,h;x,y)$

Table 5: The probability $M(s,t;g,h;x,y)$ that A wins the match, given the number s of sets already won by A, and the number t won by B, the number g of games won by A and the number h won by B in the current set, and the number x of points won by A in the current game of the current set and the number y won by B.

for the current set. The latter are all put back to zero with each new set, and s or t is incremented by one, depending on who won the previous set, A or B, respectively.

Put $T(g,h;x,y) = 1 - S(g,h;x,y)$, the probability that B wins the current set, and put $S_0 = S(0,0;0,0)$ and $T_0 = 1 - S_0$.

It is easy to arrive at the results in Table 5. This is our final table, giving, by reference back to the earlier tables, the probability of winning a match from any stage of the match.

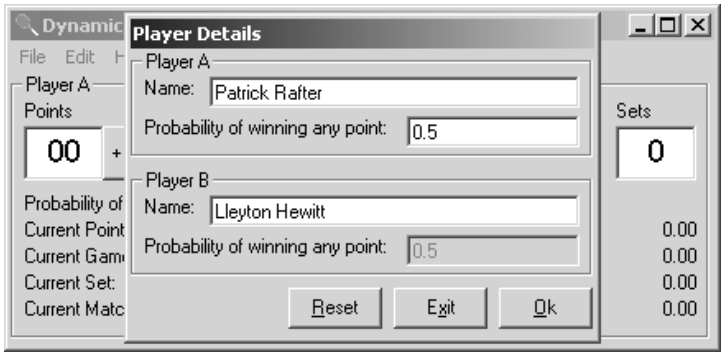


Figure 1: Player details entry screen.

5 Computer implementation

A computer application has been written in Visual Basic to accompany this paper. The program calculates the current game, set and match probabilities of winning for each player in a game of tennis ("best

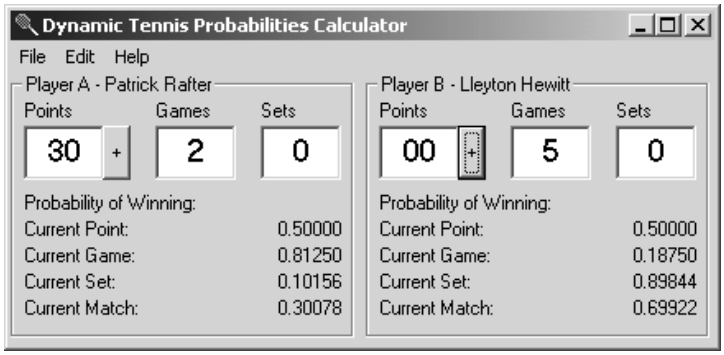


Figure 2: Main screen.

of three tiebreaker sets"), and recalculates them every time the score changes. The program may be freely downloaded from the following web site: http://www.progsoc.org/~shaman/dynamictennis.html.

When the program starts, it asks for the details of the players: the user must enter each player's name and the first player's probability of winning a point. The details screen is shown in Figure 1.

The user is then presented with the main program screen, shown in Figure 2. This shows the current score of each player, as well as each player's chance of winning the game, set and match. The user clicks on the "+" buttons to change the score, incrementing that player's point score by 1 (in the terms of the above analysis). Each time the score is altered the probabilities for each player, other than that for the current point, are recalculated.

If, at any stage of the match, the user wishes to modify the estimate of a player's chance of winning each point, the user can click on the "Edit" menu and choose "Probabilities". The "Edit" menu also provides the ability to "Undo" the last entry. This option can also be accessed using the standard Ctrl-Z keyboard shortcut.

For other forms of experimentation (or if you arrived late), it is possible for users to set the points, games and sets won by both players to values of their choosing using the "Scoreline" option in the "Edit" menu. An example of this screen is shown in Figure 3: any of the windows may be altered. If an impossible scoreline is entered, such as a games score of 6–3 for a set in progress, then an invalid score message of some form is received. There are restrictions on the use of this feature for scorelines that involve tiebreaker games, and in certain other cases.

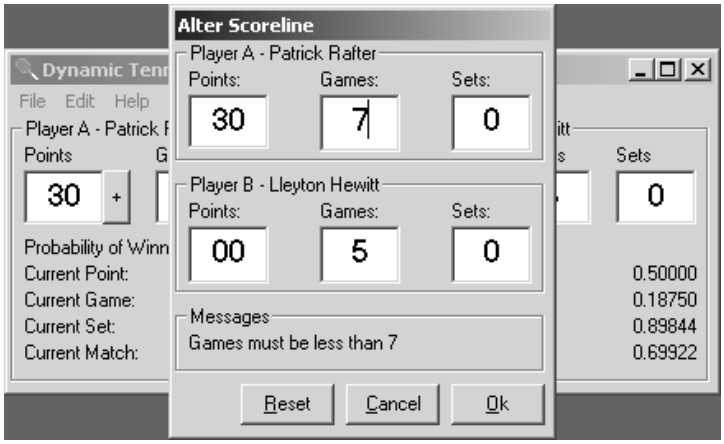


Figure 3: Alter scoreline screen.

References

[1] W. H. Carter, Jr and S. L. Crews, "An analysis of the game of tennis", *Amer. Statist.*, **28** (1974), 130–134.

[2] J. S. Croucher, "The conditional probability of winning games of tennis", *Res. Quart. for Exercise and Sport*, **57** (1986), 23–26.

[3] F. J. G. M. Klaassen and J. R. Magnus, "Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model", *J. Amer. Statist. Assoc.*, **96** (2001), 500–509.

[4] G. H. Pollard, "An analysis of classical and tie-breaker tennis", *Austral. J. Statist.*, **25** (1983), 496–505.

ANALYSING SCORES IN ENGLISH PREMIER LEAGUE SOCCER

John S. Croucher
Department of Statistics
Division of Economic and Financial Studies
Macquarie University
NSW 2109, Australia
jcrouche@efs.mq.edu.au

Abstract

Under the assumption that goals are scored at random times in soccer, using previous history it is possible to calculate the probability that a match will end in a draw or in fact with any scoreline. Common techniques include the fitting of Poisson or negative binomial distributions and a comparison is made of each using the results of the English Premier League. Theoretical scores are matched with those that occurred in practice. It is found that the negative binomial distribution yields a better fit of goal scoring than does the Poisson distribution, but requires much more data collection and calculation.

1 Introduction

One of the most popular and played sports worldwide is soccer, also known as Association Football, with great interest in the World Cup between the 32 qualifying teams in May/June 2002. The game is played between two teams each of 11 players including a goalkeeper who is the only player allowed to handle the ball.

There are a great many sources of information of soccer statistics, particularly those matches played in the UK, where examples include the yearbooks by Rothmans [14] and Playfair Football Annuals [10]. Nowadays the internet is a rich source of information regarding current and past matches with sites such as www.soccernet.com being widely accessed.

Apart from the widespread interest of the fans who follow the fortunes of individual teams, there is another important reason why many thousands of people are interested in the final scores. This involves gambling on the outcome in an activity known as *Soccer Pools*, in which punters can win vast amounts of money by using their skills to try and predict the outcome of individual matches.

In particular, gamblers in the Soccer Pools try to predict those matches that will end in a draw. Moreover, the idea is to attempt to identify those drawn matches where there will be the most goals scored since these will be the first numbers selected in the winning combination. Information regarding the precise rules of the Pools can be found at the URL http://www.nswlotteries.com.au/games/pools.html which contains the following description from the promoters:

> Six from 38 Pools is a Lotto style game where to win, you must select 6 numbers from the 38 drawn each Saturday evening. 6 from 38 Pools is known as 'The Smart Choice' because you only have to pick 6 from only 38 numbers. The winning numbers are based on the outcome of English or Australian soccer matches but you don't need to know anything about soccer to enter or to win.

> Six winning numbers are drawn first, followed by one supplementary number. You will win a prize in one of five prize divisions if you correctly select one of the following combinations in a single game panel.

In Australia, the Pools consists of trying to predict the six high-score drawn matches among thirty-eight selected from the UK FA Premiership and lower divisions 1, 2 and 3. On occasions, in the UK off-season, it will revolve around matches played in Australia. In 2000–2001, a total of around $8,000,000 was invested in the Pools across the country to identify those six matches that will have the highest score draws.

Although it is certainly true that many winners of the Pools know nothing about soccer, this does not stop keen students of the game trying to use some science to make a profit. This involves using the data available to determine those matches that are most likely to end in a draw.

From a statistical point of view, one method of achieving this aim is to fit an appropriate model to the number of goals scored by each team and use this information to select the most likely high-score draws.

2 Fitting statistical distributions

The idea of modelling sporting data is not new with some of the earlier published research being that by Wood [17] who claimed that batsmen's scores in cricket follow a geometric distribution and Moroney [8] who fitted distributions to soccer goal scoring. In soccer, goals are relatively rare with 57% of matches having a total of two goals or less scored and initial assertions that they occur at random times. For this reason it is tempting to use the Poisson distribution to fit the number of goals scored.

However, studies by Moroney [8], Pollard *et al.* [11] and Norman [9] suggest that within a match, a team's goal scoring rate may change as the match develops. That is, a team that is losing may concentrate on attack in the hope of trying to level up the score while running a greater risk of conceding even more goals. This has led to using models based on the negative binomial distribution which has proven in some cases to yield a superior fit of actual data. One example is given in Table 1. This shows a summary by Pollard [12] of the number of goals scored in 924 matches in the English First Division Football League during 1967–68. A negative binomial distribution using equation (1) was used to calculate the expected number of goals:

$$\Pr(\text{Number of goals scored} = r) = \binom{k+r-1}{k-1} p^k (1-p)^r, \qquad r = 0,\ 1,\ 2,\ 3,\ \ldots, \tag{1}$$

where $k > 0$ is an integer and $0 < p < 1$.

If the method of moments is used, where m is the sample mean and s^2 is the sample variance, then

$$p = \frac{m}{s^2} \quad \text{and} \quad k = \frac{m^2}{s^2 - m}.$$

In Table 1, the mean number of goals scored per match is $m = 1.51$ with a variance of $s^2 = 1.75$. As can be seen from this table, the negative model fitted the data extremely well, as supported by a Chi-square test with a p-value of 0.57.

Other attempts to fit statistical distributions to soccer include the works of Reep and Benjamin [13], Hill [6], Maher [7] and Croucher [4]. The evidence of a home ground advantage has been well explored by Pollard [11] and Courneya and Carron [3], while Clarke and Norman [1] and Clarke [2] provide a model based on that of Stefani, [15] and [16], that looks at winning margins taking into account both home and away performance.

Indeed, any model that tries to predict the outcome of any sporting event must properly take account of the home ground advantage since it is well known that, in the vast majority of cases, teams will exhibit a better performance when playing in front of their home crowd. The next section looks at models that use both the Poisson and negative binomial distributions to predict the combination of scores of both teams in a soccer match.

Number of goals	Observed	Expected
0	225	226.6
1	293	296.4
2	224	213.9
3	114	112.6
4	41	48.3
5	15	17.9
6	9	5.9
7+	3	2.5
Total	924	924.1

Table 1: Observed and expected number of goals scored in 924 English soccer matches in 1967–68 using a negative binomial distribution.

3 Poisson and negative binomial distribution models

The Poisson distribution typically represents events that occur randomly in a fixed time period. If this average rate is represented by the parameter λ, the event is the scoring of a goal and the time period is the duration of a match, then we have:

$$\Pr(\text{Number of goals scored} = r) = \frac{e^{-\lambda}\lambda^r}{r!}, \qquad r = 0,\ 1,\ 2,\ 3,\ \ldots, \tag{2}$$

To keep our research as current as possible, data were collected from the first 277 matches played in the 2001–2002 English Premier League up to and including Saturday, 2 March 2002. At this stage all teams had played either 27 or 28 matches. With three competition points awarded for a win, one point for a draw and no points for a loss, the race for the premiership at this stage was very close with only one point separating the top three teams of Manchester United, Arsenal and Liverpool.

To fit a Poisson distribution to the data while taking into account the home and away aspect discussed earlier, two models were constructed. The first of these was based on the overall average scoring rate of teams playing at home. In these 277 matches, there was a total of 387 goals scored by the home team for an average rate of 1.397 goals per game. At the same time, the away teams scored a total of 331 goals at an average rate of 1.195 goals per game.

The individual team average scoring rates at home (λ_1) and away (λ_2) are shown in Table 2. The correlation coefficient between home and away scoring averages of 0.556 (p-value 0.025) shows that there is a reasonable degree of association between the way teams score at home and away, with the majority as expected having a larger scoring rate for home games. Notable exceptions are Liverpool who had scored only 16 goals at home in 15 games but 31 goals away in 16 games and Bolton with 11 home goals but 20 away goals each in 14 games.

Not surprisingly, there is a strong correlation between home scoring average and competition points ($r = 0.707$, $p = 0.002$) but an even stronger one between away scoring average and competition points ($r = 0.819, p < 0.001$). Although further research is required, it seems that a team's ability when playing away from home could be the major deciding factor in determining their position on the ladder since this is their most difficult assignment.

It is interesting to note that if a team plays for a draw in *every* match then it will be in grave danger of relegation to a lower Division. The three bottom teams in any Division are relegated to the next lower Division. This is easily seen from Table 1 where a team who had drawn all their 28 matches at this stage would be only two competition points above the third bottom team Blackburn.

On the other hand, a team that won all their home games but lost all away games would be on about 42 competition points or in the top one-third of the table. Another strategy would be to try and win all home games and be content to draw away games. After 28 matches this would have yielded 56

Team	Matches played	Home scoring average λ_1	Away scoring average λ_2	Competition points
Manchester United	28	2.40	2.38	57
Arsenal	28	2.15	1.93	57
Liverpool	29	1.23	1.94	56
Newcastle	28	1.79	1.93	55
Chelsea	27	2.17	1.40	44
Leeds	27	1.54	1.21	44
Aston Villa	28	1.20	1.23	41
Tottenham	27	1.87	0.92	38
Charlton	28	1.27	1.00	37
Fulham	28	1.20	0.69	35
Southampton	28	1.08	1.33	34
Middlesborough	28	1.29	0.64	34
West Ham	28	1.38	0.87	34
Sunderland	28	0.93	0.64	31
Ipswich	27	1.27	1.33	30
Bolton	28	0.79	1.43	30
Everton	27	1.29	0.69	29
Blackburn	27	1.62	0.86	26
Derby	27	0.93	0.62	25
Leicester	28	0.62	0.67	17

Table 2: Individual team average scoring rates at home (λ_1) and away (λ_2) after 277 matches in the English Premier League 2001–2002 season.

competition points or only one point off the lead. Of all teams, Liverpool is the only one not to have lost any of their fifteen away matches while every other team has lost at least three.

Separate Poisson models for goals scored at home and away goals are shown in Table 3 using the combined scoring rate for all teams. To avoid distorting the statistical analysis, there is a category of "4 or more" goals to ensure that each has an expected frequency of at least 5. With a Chi-square value of 2.73 ($p > 0.20$) for the home data, the Poisson gives a reasonable fit, although, like the negative binomial model in Table 2, the larger relative discrepancies occur for a large number of goals scored.

The Poisson model is far less accurate for the away data with a Chi-square value of 10.82 ($p < 0.01$). In particular, there are more teams who fail to score *any* goals than might be expected but the total number that score one goal or less is not too far away from expectations.

To compare the two distributions, the same data were then modelled using a negative binomial

	Home ($\lambda = 1.397$)		Away ($\lambda = 1.195$)	
Goals scored	Observed	Expected	Observed	Expected
0	70	68.5	99	83.8
1	97	95.7	89	100.2
2	60	66.9	49	59.9
3	29	31.1	24	23.8
4 or more	20	14.8	16	9.3
Total	277	277.0	277	277.0
	$\chi^2 = 2.73$, $p > 0.20$		$\chi^2 = 10.82$, $p < 0.01$	

Table 3: Observed and expected numbers of goals scored at home and away using a Poisson model.

	Home		Away	
Goals scored	Observed	Expected	Observed	Expected
0	70	72.6	100	96.4
1	97	91.9	89	91.6
2	60	63.1	49	52.3
3	29	31.2	24	23.3
4 or more	21	18.2	15	13.4
Total	277	277.0	277	277.0
	$m = 1.419$, $s^2 = 1.592$		$m = 1.177$, $s^2 = 1.458$	
	$\chi^2 = 1.11$, $p > 0.50$		$\chi^2 = 0.63$, $p > 0.50$	

Table 4: Observed and expected numbers of goals scored at home and away using negative binomial distribution models.

distribution with parameters. The results are shown in Table 4.

A comparison of the results in Tables 3 and 4 reveals that the negative binomial distribution does indeed provide a superior fit for both home and away data. This is especially evident since the Poisson suggests that there will be far fewer away teams scoring four or more goals than actually happened. This category alone contributed nearly half of the total Chi-square test statistic value.

4 Scoring combinations

To make a reasonable prediction of the outcome of a match it is of course necessary to estimate the score of both teams. There are a number of approaches that might be made, but we will consider one based solely on the Poisson model and another based solely on the negative binomial model.

In the case of fitting a Poisson model, equation (1) will be used along with the combined parameters for all teams. This means that the probability that the home team will score x goals and the away team will score y goals is given by:

$$\Pr(\text{home goals} = x, \text{away goals} = y) = \frac{e^{-\lambda_1}\lambda_1^x}{x!}\frac{e^{-\lambda_2}\lambda_2^y}{y!} = \frac{\lambda_1^x\lambda_2^y e^{-(\lambda_1+\lambda_2)}}{x!\,y!}, \tag{3}$$

where

$$\lambda_1 = \text{the average goals per match for home teams,}$$
$$\lambda_2 = \text{the average goals per match for away teams.}$$

In particular, from equation (3) the probability that a match will end in a scoreless draw (where $x = y = 0$) is $e^{-(\lambda_1+\lambda_2)}$.

Using the data in the 277 matches played so far in 2001–2002, the parameter values for all teams combined are:

$$\lambda_1 = 1.397,$$
$$\lambda_2 = 1.195.$$

The observed and expected frequencies (shown in parentheses) using these parameters and equation (3) are shown in Table 5.

Although the individual score *totals* for both home and away are relatively close (as seen in Table 3) there are significant differences for individual combinations. In particular, there are about nine *more* 0–0 scorelines but ten *fewer* 0–1 scorelines than expected. Combined, these could be seen to "cancel each other out" since they are adjacent. It is interesting to speculate as to why away teams might find it difficult to score that single goal to break a 0–0 deadlock, while home teams (where observed frequencies seem to match almost perfectly with expected) do not seem to have that problem.

	Away					
Home	0	1	2	3	4 or more	Total
0	30	15	15	4	6	70
	(20.7)	(24.8)	(14.8)	(4.0)	(4.2)	(68.5)
1	30	36	17	12	2	97
	(29.0)	(34.6)	(20.7)	(8.2)	(3.2)	(95.7)
2	25	18	8	6	3	60
	(20.2)	(24.2)	(14.4)	(5.8)	(2.3)	(66.9)
3	6	11	9	1	2	29
	(9.4)	(11.3)	(6.7)	(2.7)	(1.0)	(31.1)
4 or more	9	9	0	1	2	21
	(4.5)	(5.3)	(3.3)	(3.1)	(0.5)	(14.8)
Total	100	89	49	24	15	277
	(83.8)	(100.2)	(59.9)	(23.8)	(9.3)	(277)

Table 5: Observed and expected (in parentheses) numbers of times given scores have appeared in 277 matches based on overall Poisson parameter values.

5 Drawn matches

Of the 277 matches in our sample there were a total of 75 (27%) drawn games, a figure that is consistent with the overall figure of 25% suggested by Haigh [5]. Table 6 shows the actual number of draws for each score along with the theoretical probability and expected values given by the Poisson distribution using the overall values of λ. The expected frequencies are based on 277 matches. There were no drawn games higher than 3–3.

Scoreline	Observed frequency	Theoretical probability	Expected frequency
0–0	30	0.0749	20.7
1–1	36	0.1250	34.6
2–2	8	0.0522	14.4
3–3	1	0.0097	2.7
4–4 or more	0	0.0010	0.3
Total	75	0.2628	72.7

Table 6: Observed and expected numbers of draws in 277 matches using a Poisson model.

Although the observed number of draws (75) matches fairly well the expected number of about 73 draws, the figures in Table 6 reveal a significantly higher number of 0–0 draws than expected. In fact, 10.8% of matches had no goals scored at all while 56% had a total of two or fewer goals at full time. This aspect has led to some groups calling for techniques to increase the goal scoring such as wider and/or higher goalposts.

To reward attacking play, the league decided in the early 1980s to award one competition point for a draw and three points for a win. Previously there were only two points awarded for a win. The idea was to promote the scoring of goals and make the game more exciting for spectators. The result of this change was investigated by Croucher [4] who found that there was little evidence of it having the desired effect, at least in the initial stages of operation. Perhaps this is because if the scores are locked at, say, 0–0 late in a match then both teams will be satisfied with a draw and play defensively rather than risk their opponents gaining three valuable competition points.

6 Actual match prediction

Suppose we wanted to predict the result of a particular match using Poisson models. On 6 March 2002, Southampton at home played Middlesborough. Using the parameters for each team shown in Table 2 ($\lambda_1 = 0.93$, $\lambda_2 = 0.64$) and equation (3), the figures in Table 7 show the probability of individual scorelines occurring.

	Middlesborough					
Southampton	0	1	2	3	4 or more	Total
0	0.208	0.133	0.043	0.009	0.002	0.395
1	0.193	0.124	0.040	0.008	0.002	0.367
2	0.090	0.058	0.018	0.004	0.001	0.171
3	0.028	0.018	0.006	0.001	0.000	0.053
4 or more	0.008	0.004	0.001	0.001	0.000	0.014
Total	0.527	0.337	0.108	0.023	0.005	1.000

Table 7: Probabilities of various scorelines when Southampton played Middlesborough.

The most likely result (the one with the highest probability) in Table 7 is a 0–0 scoreline with a 66% chance that neither team would score more than one goal. The probability of a draw was 0.351, higher than the overall chance for all matches, with 95% of this figure coming from either a 0–0 or 1–1 final score.

In fact, Sunderland and Middlesborough played a 1–1 draw which, while not being our first choice, would not be classified as a surprise result using the probabilities in Table 7.

7 Remarks

It seems by most accounts that the negative binomial distribution yields a better fit of goal scoring in soccer than does the Poisson distribution, but requires much more data collection and calculation. For example, to construct a table such as Table 7 using the negative binomial distribution requires the individual scores of each match played by the two teams involved so that the appropriate values of m and s^2 can be calculated to use in the formula.

For the average Pools punter, who most likely does not have the time or ability to use a computer to assist in undertaking copious calculations, trying to find the most likely candidates for a drawn match, it would be quite tedious to update this data for every team each week, while the Poisson simply requires the parameter of the overall scoring rate. It may not be quite as accurate but it is certainly a lot easier in practice.

The negative binomial distribution could also be used to calculate expected frequencies for every scoreline as in Table 5 and this could be compared to the values suggested by the Poisson. For example, a negative binomial distribution using overall data suggests that there would be about 25 scoreless draws, still below the observed value of 30 but much closer than the 21 suggested by Poisson.

There is also the natural question of independence when dealing with the separate scores of two teams playing in the same match. Although their scores in practice would not be strictly independent, for our purposes the assumption seems to have nevertheless yielded useful results.

Another area for further research is to analyse the data separately for teams who are in, say, the top half of the table and those who are in the bottom half. It may well be that while one group produces significant results the other does not.

References

[1] S. Clarke and J. M. Norman, "Home ground advantage of individual clubs in English soccer", *The Statist.*, **44** (1995), 509–521.

[2] S. R. Clarke, "Home advantage in balanced competitions—English soccer 1990–1996", in *Third Conference on Mathematics and Computers in Sport*, N. de Mestre (editor), Bond University, Queensland, Australia (1996), 111–116.

[3] K. S. Courneya and A. V. Carron, "The home advantage in sport competitions: a literature review", *J. Sport and Exercise Psychol.*, **14** (1992), 13–27.

[4] J. S. Croucher, "The effect of changing competition points in the English Football League", *Teaching Statistics*, **2** (1984), 39–42.

[5] J. Haigh, *Taking Chances*, Oxford University Press (1999), 239–240.

[6] I. D. Hill, "Association football and statistical inference", *J. Appl. Statist.*, **23** (1974), 203–208.

[7] M. J. Maher, "Modelling association football scores", *Statist. Neerlandica*, **36** (1982), 109–118.

[8] M. J. Moroney, *Facts from Figures*, Penguin, Harmondsworth (1951).

[9] J. M. Norman, "Soccer", in *Statistics in Sport*, J. Bennett (editor), Arnold, London (1998), 105–120.

[10] *Playfair Football Annual*, MacDonald and Jane Publishers, London.

[11] R. Pollard, "Home ground advantage in soccer: a retrospective analysis", *J. Sports Sciences*, **4** (1986), 237–248.

[12] R. Pollard, P. Benjamin and C. Reep, "Sport and the negative binomial distribution", in *Optimal Strategies in Sports*, S. P. Ladany and R. E. Machol (editors), North Holland, New York (1977), 188–195.

[13] C. Reep and P. Benjamin, "Skill and chance in association football", *J. Roy. Statist. Soc. Ser. A*, **131** (1968), 581–585.

[14] *Rothman's Football Yearbook*, J. Rollins (editor), Headline, London.

[15] R. T. Stefani, "Observed betting tendencies and suggested betting strategies for European football pools", *The Statist.*, **32** (1983), 319–329.

[16] R. T. Stefani, "Applications of statistical methods to American football", *J. Appl. Statist.*, **14** (1987), 61–73.

[17] G. H. Wood, "Cricket scores and geometrical progression", *J. Roy. Statist. Soc. Ser. A*, **108** (1945), 12–22.

REVIEW OF THE APPLICATION OF THE DUCKWORTH/LEWIS METHOD OF TARGET RESETTING IN ONE-DAY CRICKET

Frank Duckworth
"Moonrakers"
Taits Hill, Stinchcombe
Glos. GL11 6PS, UK
f.duckworth@rss.org.uk

Tony Lewis
The Business School
Oxford Brookes University
Wheatley
Oxford OX33 1HX, UK
ajlewis@brookes.ac.uk

Abstract

After its introduction to one-day cricket in 1997 the Duckworth/Lewis method has spread to become the standard rule throughout senior levels of the game for resetting the target in matches shortened after their start due to rain or other reasons. This paper reviews how the method is working following its application to around 300 known cases. This includes examining its fairness both practical and perceived and the consideration of alternatives to the method's logical way of handling interruptions within the first innings. A comparison of the Duckworth/Lewis model with several years of data from international matches is made and some suggestions are examined for improvement of the model but are shown to be at the expense of loss of simplicity and transparency. Data from actual matches are used throughout to illustrate concepts and suggestions.

1 Introduction

The Duckworth/Lewis method of target resetting in interrupted one-day cricket matches was presented to the Third Australian Conference on Mathematics and Computers in Sport [1] and published in the *Journal of the Operational Research Society* in 1998 [2]. The method was introduced into the national one-day competitions of the England and Wales Cricket Board in 1997 and into one-day internationals (ODIs) on 1 January 1997 in the match between Zimbabwe and England in Harare.

Following slight modifications to the method's implementation [6] it has gradually spread to be used in all major one-day cricket-playing countries both domestically and in internationals.

2 Summary of D/L model

The foundation of the method is the Duckworth/Lewis (D/L) model of average runs that are scored $Z(u, w)$ for the remaining overs available u when w wickets ($0 \leq w \leq 9$) have been lost:

$$Z(u, w) = Z_0 F(w)(1 - \exp(-b_0 u / F(w)). \quad (1)$$

In this model, the result of the combination of statistical and mathematical modelling, Z_0 and b_0 are positive constants and $F(w)$ is a positive, decreasing step function with $F(0) = 1$. This function is interpreted as the proportion of runs that are scored with w wickets lost compared with that with no wickets lost and, hypothetically, infinitely many overs available. That is, $F(w) = \lim_{u\to\infty} Z(u, w)/Z(u, 0)$. Fitting the model to the data includes the estimation of values for b_0, Z_0 and $F(w)$. At the time of the

initial development of the model [1, 2] little of the necessary over-by-over data were readily available from public sources and so estimation of the $F(w)$ parameters in (1) came largely from cricket commonsense.

Since u represents overs available to the fraction of one ball it can be regarded as a continuous variable and the function $Z(u, w)$ differentiable with respect to u. This function has the property that, when $u = 0$, $\partial Z/\partial u = b_0 Z_0$, a constant independent of w which models the generally sensible cricketing characteristic that at the final ball of the innings the average runs scored is more or less the same regardless of the loss of wickets. The graph of the D/L model (1), with the horizontal axis reversed to reflect the progression of an innings, is given as Figure 1.

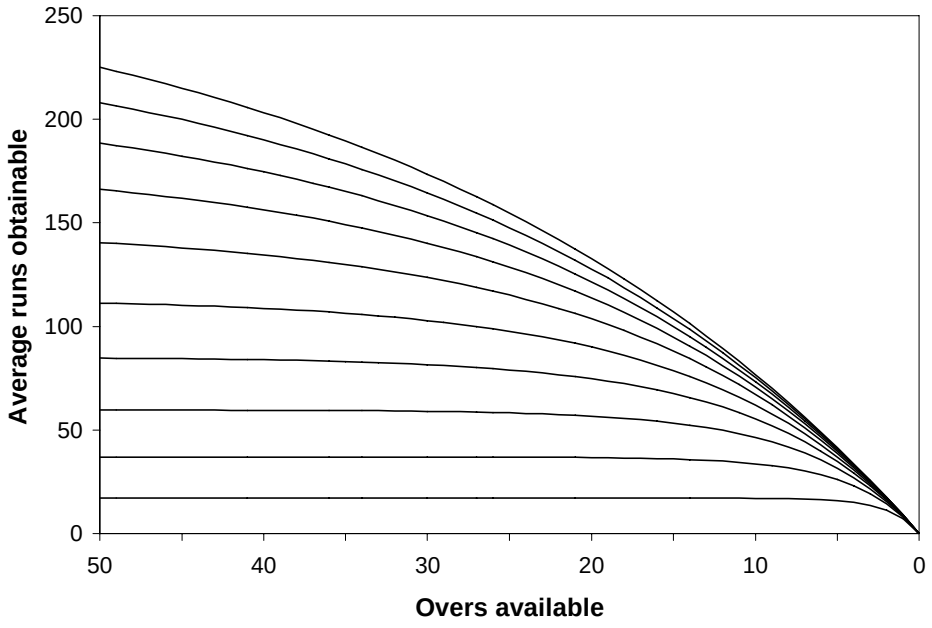


Figure 1: Average runs obtainable for overs available and wickets lost; $w = 0$ at top, $w = 9$ at bottom.

3 Summary of D/L method

In order to reset targets, the D/L method converts (1) into proportions of the average total score in 50 overs to produce a table which can be interpreted as the proportion of resources remaining, compared with those for a full 50-over innings, from any position in an innings:

$$P(u, w) = \frac{Z(u, w)}{Z(50, 0)}. \tag{2}$$

Although not all one-day matches are of 50 overs per side, these tables, as supplied in Appendix 1 but printed in terms of percentages, can be used for any length of innings up to 50 overs per side. But there is no restriction to 50 overs. Prior to 1999 some England and Wales Cricket Board (ECB) competitions consisted of matches up to 60 overs per side for which an extended table was supplied. Both over-by-over and ball-by-ball versions of the table are supplied to cricketing authorities and are printed in Duckworth and Lewis [3].

The D/L table is used to calculate each team's resources available. Denote R_1 and R_2 as the resource percentages available for the whole innings, having allowed for any stoppages. The suffixes refer respectively to Team 1, the side batting first, and Team 2. If Team 1 finished their innings with S runs

then the target for Team 2 is derived from

$$T = \frac{SR_2}{R_1}, \qquad \text{if } R_2 \leq R_1, \tag{3a}$$

$$T = S + G_{50}(R_2 - R_1), \qquad \text{if } R_2 \geq R_1. \tag{3b}$$

The target, which is the minimum total required to win, is the next integer above T. One less than this is the score required to tie. When Team 2's innings is in progress then R_2 represents the resources consumed and the equivalent of the tie is called the par score. In (3b), G_{50} is the average 50-over first innings total for the standard of cricket in which D/L is being applied. For first-class standard and international matches this has been 225. The reasons for the different formulas depending on the relative magnitudes of R_1 and R_2 are explained in Duckworth and Lewis [2, 7] and Lewis and Duckworth [6].

In using the same set of tables for different standards of matches with different values for G_{50} we are assuming that the percentages of resources remaining are the same as for ODI competitions which were the source of the data used to estimate the parameters of (1). If this were not true then we would need different sets of tables for other levels of competition. We have examined matches in the ECB domestic competitions and matches between Associate Member countries of the International Cricket Council (ICC). This has shown that although the G_{50} does indeed vary significantly between standards of competition, the variations in the resource percentages as calculated by (2) are small enough that the one set of tables, as Appendix 1, do reasonably represent matches at the different standards. Equations (3), therefore, can be applied to produce fair targets for most standards of competition. There follow two examples to illustrate how the D/L method works.

Example 1

On 25 April 1999 in Barbados, Australia scored 252 for the loss of nine wickets in their 50 overs. In reply, West Indies had scored 138 for the loss of only one wicket in 29 overs when crowd trouble caused the WI innings to be reduced by ten overs. Upon the resumption the WI target was revised to 196 in 40 overs.

For Australia, who completed their 50-over innings, $R_1 = 100\%$. West Indies started with 100% but at the stoppage, for one wicket lost and 21 overs left, they had, from Appendix 1, 58.6% of their resources remaining. Upon the resumption, with only 11 overs left, they had 36.1% resources remaining so the stoppage cost them 22.5% leaving them with $R_2 = 77.5\%$. By (3a), with $S = 252$, $T = 195.30$. They needed 196 runs to win, a further 58 in 11 overs which they achieved with three overs to spare, winning by eight wickets.

Example 2

In a match of the Carlton United Brewery one-day series on 25 January 2001 at the Adelaide Oval, West Indies had scored 6/235 in 47 of their 50 overs when rain interrupted play. When play was able to be resumed the umpires terminated the WI innings and allocated Zimbabwe 47 overs. The D/L target was an enhancement of the West Indies total to 253 runs.

West Indies commenced their innings expecting (regardless of weather forecasts) their full 50 overs and 100% resources. Rain deprived them of their last three overs when they had lost six wickets. From Appendix 1, their lost resources were 10.2% and so for their innings $R_1 = 100 - 10.2 = 89.8\%$. For their 47-over innings, Zimbabwe's resources were, from Appendix 1, $R_2 = 97.4\%$. From (3b) with $G_{50} = 225$, $T = 252.1$ with the next higher integer being the score required to win, an enhancement of 17 runs.

The enhancement is a logical and fair way of compensating West Indies for the unexpected shortening of their innings and neutralising Zimbabwe's advantage of knowing in advance of their shorter innings. In the match Zimbabwe were all out for 175 runs in 40.2 overs and lost by 77 runs.

4 Experiences of application

4.1 Reaction

Initial reaction to the introduction of the D/L method in 1997/98 is summarised in Lewis and Duckworth [6]. Since that time the method has spread to be used in all major countries [10]. Although many journalists still enjoy making disparaging comments about the method, there is general acceptance of its superiority over all other methods that have been tried. Players' attitudes appear to have been summarised by the Australian captain Steve Waugh ahead of the 1999 World Cup ("It's the best method by a long way" [8] and more recently by England captain Nasser Hussain ("I think it is probably the fairest system ..." [4]). Journalists' acceptance has been expressed by the method's initial arch-critic Martin Johnson of the *Daily Telegraph* ("The two rain-rule boffins have, by general consent, devised a fairer system ..." [5]). In countries where the method is long established, revised targets by D/L rarely raise more than a ripple of passing comment. We believe, therefore, that the method has been generally accepted although we recognise there are situations which sometimes still cause concern. We shall discuss one of these in this section. Other issues of concern are addressed later.

4.2 Enhanced targets

In cases of interruptions during Team 1's innings such as that in Example 2 the enhanced target, whilst very logical, still causes disquiet not only amongst the supporters of the team batting second but also in some sections of the media. Whilst the playing regulation that lost overs be shared as equally as possible between the two teams is in operation, the enhanced target is a logical way of neutralising the advantage to Team 2 from knowing in advance of the shorter innings after Team 1's innings has been unexpectedly shortened after its start. The alternative of reducing Team 2's wickets would be an unacceptable change to the nature of the game.

We have suggested to the Cricket Committee (Playing) of the ICC that, rather than enhance the target, playing regulations could be modified so that the lost overs would be distributed unequally between the two teams. This would need to be done in such a way that Team 2's target is one more than Team 1's final total as is traditional, but should be obtained in an appropriate number of fewer overs than that received by Team 1. The mechanism for the division would be such that R_2 is as close as possible to R_1 without exceeding it. In practical terms the closeness would be limited by the need for each team to have a whole numbers of overs available.

How this might be applied in Example 2 would be to split the six lost overs 1:5. This would allow for West Indies to bat for two more overs and limit Zimbabwe to 45 overs. Suppose WI finish on 254. Using Appendix 1, $R_1 = 100 - 10.2 + 7.2 = 97.0\%$ and $R_2 = 95.5\%$ so that, from (3a), $T = 250.07$ and Zimbabwe would require 251 to win in 45 overs. Any more even split of complete overs between the teams would provide an enhanced target for Zimbabwe.

The process of finding the appropriate division of lost overs would be iterative and would need to be performed within the D/L software, CODA [3]. In practice such a division of lost overs would reduce the chance of a viable match whereby each side needs to have a minimum number of overs available to bat. This varies from ten to 25 between the various one-day competitions around the world.

5 Analysis of match data

We have already commented on the limited data at the detail required to make accurate estimates of the parameters of the D/L model (1). Thanks to the Internet, match details are now widely available, including those from the Cricinfo database [9]. This database provides large amounts of statistical data of all international matches including the over-by-over scores. Slightly lesser detail is provided on domestic one-day matches. However, data on ECB domestic matches have been made available to us via The Press Association (Sports).

This statistical data can be used in several ways to investigate some issues related to the validity of the D/L model (1) and the fairness of the D/L method (3).

5.1 Win/Loss ratio in interrupted matches

The main method of target resetting prior to the advent of D/L was based on average run rate for the overs available [1, 2]. This method favoured the team batting second and so, anecdotally, Team 2 generally won such matches. We claim that our method is fair to both teams. Although a fair result to a match depends on particular match circumstances, it would nevertheless appear unfair if there were a long-term bias to either team from the use of D/L. Only the ECB has had sufficient case history in which to investigate our claim; other countries so far have relatively small numbers of cases partly due to their use of D/L over fewer seasons and partly due to different climatic conditions compared with the United Kingdom.

The ECB case history from 1997 to 2001 includes 126 matches that were rain interrupted and were brought to a conclusion using the D/L method [9]. Of these, 67 were won by Team 1, 56 by Team 2 and there have been three tied matches. Apportioning the ties equally between Team 1 and Team 2 ($1\frac{1}{2}$ "wins" each) gives a Team 1 win-percentage of approximately 54.4% which is not statistically significantly different from a 50:50 split, suggesting that there is no evidence to reject our claim of unbiasedness.

These data are consistent also with the win/loss ratio of 54% to 46% in uninterrupted ECB domestic one-day matches. Clearly on the basis of these pieces of evidence we can reasonably claim a lack of bias to either team.

5.2 Analysis of average runs scored for overs available and wickets lost

The graph in Figure 1 shows the D/L model of average runs scored with the overs available and wickets already lost. The model was developed in 1996 [1] and the tables of Appendix 1 provide the over-by-over version of resource percentages available that have been used for resetting targets since 1997.

The ODIs from mid-1997 until early 2002, for which over-by-over scores are available on Cricinfo, have been used to investigate the validity of the D/L model. Some 330 matches have been included in the analysis. Appendix 2(a) provides the summary of the actual average scores and Appendix 2(b) provides the numbers of cases on which these averages are based.

Excluding averages based on fewer than five data points, Figures 2(a) to 2(j) provide the comparisons between the runs expected to be scored from the D/L model (1), represented by the dotted lines, and the actual averages in Appendix 2(a), the solid lines.

Figure 2(a) shows the average runs obtained for no wickets lost for the overs still available. In 1996 when the parameters in (1) were set, the average 50-over ODI first innings total was around 225, consistent with the D/L model. There is evidence (see Appendix 2(a)) that the average score in ODIs is now significantly higher at around 233 (with a standard error of around 3.0). However the shape of the data line is fairly consistent with the D/L model for most of the data points even though sample sizes with fewer than 30 overs left are small.

Figure 2(b) is an extremely interesting graph. Whereas the general shape is consistent with the D/L model, the effect of the loss of the first wicket on average run-scoring capability appears to have been significantly over-estimated. In practice the early loss of the first wicket does not have a great effect on average total runs in an innings. Appendix 2(a) shows that close to 50% of ODIs have lost at least one wicket within the first five overs of a 50-over innings.

Similarly, Figure 2(c) shows an overestimated effect of the loss of two wickets on run-scoring capability. Both of these graphs have implication for the values of $F(1)$ and $F(2)$ in (1).

Figures 2(d) to 2(h), for wickets lost from three to seven in what might be termed the middle phase of an innings, show the D/L model fitting the data quite well.

In Figure 2(i), the graph for eight wickets lost, evidence begins to appear that the model tends to over-estimate the average further runs scored in the overs available. This suggests that $F(8)$ in (1) is

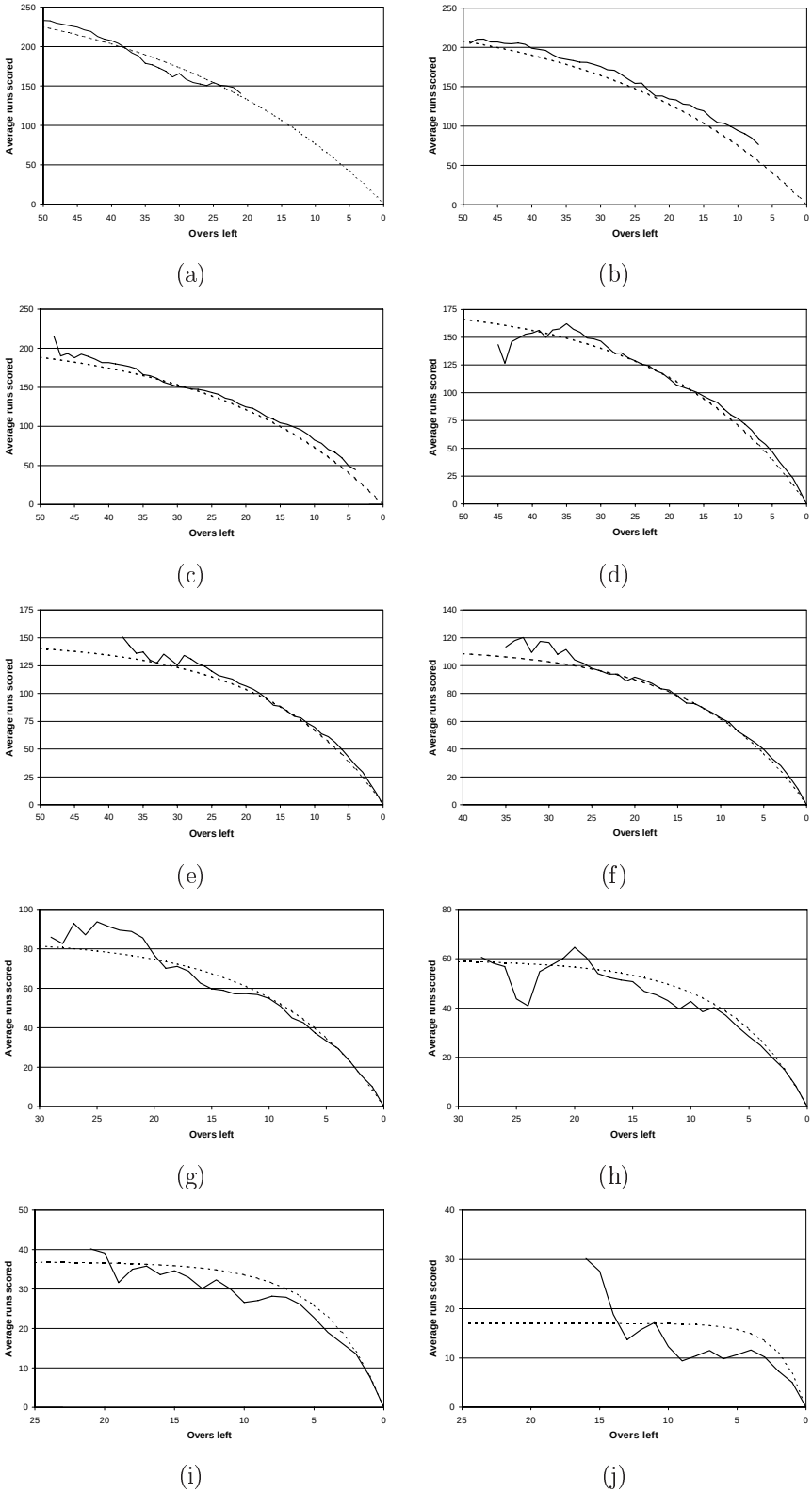


Figure 2: Average runs in ODIs for overs left: (a) no wickets lost; (b) one wicket lost; (c) two wickets lost; (d) three wickets lost; (e) four wickets lost; (f) five wickets lost; (g) six wickets lost; (h) seven wickets lost; (i) eight wickets lost; (j) nine wickets lost.

rather high attributing greater batting skill than is the reality. Figure 2(j), the graph for nine wickets lost, is based on few data points before the last eight overs available which partly explains the gross variations from the model. In the last few overs, however, the model is again observably over-estimating the contribution of the last wicket, on average.

What all this is telling us is that the D/L model is consistent in shape with the data. But there is evidence to suggest that we may need to adjust slightly the parameters of the formula to update the effects of various wicket partnerships and perhaps allow for the slightly higher average runs that are being scored in ODIs. However recent changes to playing conditions that now allow one "bouncer" per over may prove to have a reducing effect on average totals.

6 Can D/L be improved?

Whereas the D/L method is working well in practice, we are aware of several areas in which improvements may be possible. These are discussed within this section. A major disadvantage, however, in their possible implementation is the explicit need to evaluate revised targets using computer software. We believe that one of the strengths of the D/L method is that it can be implemented entirely by hand, using nothing more than the tables, as in Appendix 1, and a pocket calculator. This advantage would be lost if computers became essential and in the process barring the method's use at the grass-roots levels of the game at which a computer is unlikely to be available. Nevertheless it is important to outline areas of potential improvement if cricket authorities believe that the loss of transparency is a price worth paying for improved performance of the method.

6.1 Above average totals

When Team 1 score a well-above-average total, the D/L par scores for Team 2 as its innings progresses can seem unreasonably low. This is because the required run-scoring behaviour to achieve high totals will not normally match the average performance pattern in the D/L model.

Example 3

During the cricket World Cup of 1999, in a certain match whose teams for the moment will remain anonymous, Team 1 scored a well-above-average 329 in their 50 overs. In reply Team 2 had reached 2/139 in 27 overs, still requiring 191 in 23 overs at a rate of 8.30 per over. Most cricket observers would likely have concurred that Team 2 were behind their "asking" rate. Nevertheless, according to D/L, Team 2 were four runs ahead of par and would have won the match had it been abandoned at that point.

Whereas on average, from Appendix 1, teams score 58.9% of the average total in their remaining 23 overs when they have already lost two wickets, when the total is well above average cricketing sense suggests that teams need to sustain a scoring rate closer to their final requirement in order to stay on par. In other words, this cricket sense suggests that the D/L curves in Figure 1 need to be straighter when chasing large totals.

We can illustrate this point further with an extreme example. Suppose that in a 50-overs-per-side match, Team 1 hit six runs off every ball. If we ignore the possibility of wides and no balls (which add runs to the total and provide extra run-scoring opportunities) then, with six balls per over, the maximum score is 1800. If rain falls between the two innings reducing Team 2 to 25 overs then, from (3a) and Appendix 1, $T = 1800 \times 68.7/100 = 1236.6$ and they require 1237 to win at the impossible rate of 8.25 runs per ball. Much more logical would be a target of 901 based on half the runs in half the overs. In this limiting case the D/L curves would need to be entirely linear.

Equation (1) is such that, with the parameters in use, $Z(50, 0)$ has the value of approximately 225. Whilst retaining these values as being consistent with average ODI matches we may introduce two further parameters, λ and n, which provide the ability to shape the graph of average runs scored, as

in (4):

$$Z(u, w \mid \lambda, n) = Z_0 F(w) \lambda^n (1 - \exp(-b_0 u / \lambda^{n-1} F(w))). \tag{4}$$

Clearly, for any n, if $\lambda = 1$ then (4) reduces to (1). Note also that (4) has the property that, when $u = 0$, $\partial Z / \partial u = \lambda b_0 Z_0$, again a constant. Note that λ (≥ 1) can be interpreted as a "match factor" that adjusts the "average" target according to how much Team 1's score S is above the G_{50} average.

Using (4), with suitable choice of λ and n, we can produce graphs that more readily represent the average further runs that can be scored with overs left and wickets lost, when the final total is much larger than the overall average for the standard of the competition as represented by G_{50} in (3b). Figure 3 shows three cases. The lower dotted line refers to the average runs scored from the D/L model (1) with its 225 runs expected from 50 overs. Note that far more runs are expected when 25 overs are left (when no wickets have been lost) than are expected in the first 25 of an expected 50 overs. The top solid line refers to an appropriate scoring pattern when the overall total is an extremely high figure of around 560. At an average rate of nearly two runs per ball this high scoring rate needs to be sustained throughout the innings and so the graph of the average runs scored in remaining overs is nearly a straight line.

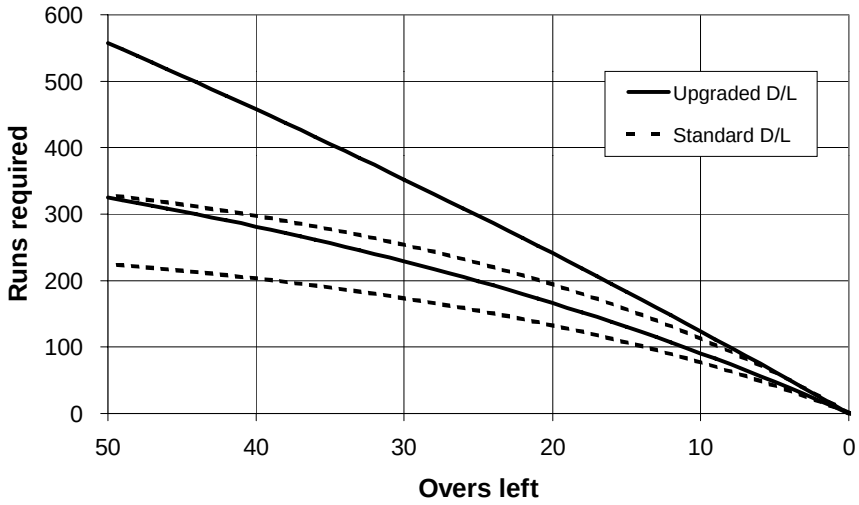


Figure 3: Upgraded D/L — a model for realistic distribution of scoring patterns.

To apply this upgraded model to target recalculations we need to produce the equivalent of (2) by dividing the right hand side of (4) by $Z(50, 0 \mid \lambda, n)$. This produces

$$P(u, w \mid \lambda, n) = \frac{F(w)(1 - \exp(-b_0 u / \lambda^{n-1} F(w)))}{1 - \exp(-50 b_0 / \lambda^{n-1})}. \tag{5}$$

Experimentation with matches such as that in Example 3 suggests that an appropriate value for n is 6. Our program CODA finds an appropriate value for λ by iteration but it may also be determined using the Microsoft® Excel tool *Goal Seek*. In Example 3 the desired value for λ is such that $Z(50, 0 \mid \lambda, 6) = 329.0$ which turns out to be 1.1178.

The curves relevant to this example appear as the central two curves in Figure 3. The heavy line represents $Z(u, 0 \mid 1.1178, 6)$ as in (4). It shows a straighter curve than the equivalent D/L curve which is the adjacent dotted line and addresses the concerns of the scenario in Example 3 in that cricket sense expects more runs from the earlier part of the innings than under standard D/L.

Using the other parameters in (5) as those used for producing the D/L tables in Appendix 1, revised tables appropriate for any first innings total S can be produced. Table 1 is an extract of such revised tables for the $S = 329$ in Example 3.

329	wickets lost				
overs left	0	2	5	7	9
50	100.0	88.9	60.2	34.8	10.1
40	86.5	78.4	56.2	34.2	10.1
30	70.3	65.2	49.9	32.6	10.1
25	61.1	57.2	45.5	31.2	10.1
23	57.2	53.8	43.5	30.4	10.1
20	51.0	48.3	40.0	29.0	10.1
10	27.8	27.0	24.4	20.4	9.6
5	14.5	14.3	13.6	12.3	7.9

Table 1: Upgraded D/L table for $S = 329$.

Referring back to Example 3, we see that the par score using Table 1, when Team 2 were 2/139 in 27 overs when chasing 329 is, from (3a), $T = 329 \times (100 - 53.8)/100 = 152.00$, so that the par score is 152. Under this upgraded model of D/L, Team 2 would have been losing by 13 runs, a situation more in keeping with cricketing instinct.

In the match being described, Team 1 were previous winners of the World Cup, India, and Team 2 were Kenya who were newcomers to the top tier of recognised one-day countries having qualified from the minor countries' one-day competition, the ICC Trophy, in 1997. Naturally D/L makes no distinction between the quality of individual teams playing in a competition of a given status. The apparent disparity of the situation of Example 3 is exaggerated by the knowledge of the teams involved and would be substantially lessened if it were India doing the chasing of 329. (As a matter of interest India, as Team 1, scored 203 in their last 23 overs [9] but this is actually irrelevant to the target resetting process.) Nevertheless cricketing instinct in such a situation would be assuaged by the use of this upgraded set of tables.

In the limiting case when Team 1 score 1800 in 50 overs then $\lambda = 4.031$. The graph for $w = 0$ in the upgraded D/L model (4) (equivalent to the top graph in Figure 3) is linear and the resource percentage for 25 overs left and no wickets lost is 50%, so that the revised target would be 901.

In principle, a different set of tables would be needed for each Team 1 total score up to totals of 400 plus. Such totals have been achieved in several recognised matches between mis-matched one-day sides.

This upgraded D/L procedure would probably have the advantage that if Team 1's innings has been interrupted then the mechanism of enhancement could be unified so that (3a) would be required to reset the target regardless of the relative magnitudes of R_1 and R_2.

There is further complexity, however, if Team 1's innings is interrupted leading to an enhanced target. The appropriate table to use is not known because Team 1's final score will not be known until the completion of their innings following an interruption. The appropriate value for λ in (5) would need to be found retrospectively and iteratively. In practice, the methodology can be included in the software CODA [3], as used by all major cricket-playing countries. Consequently targets could be obtained relatively swiftly and accurately but an interested spectator would no longer be able to perform the target calculations manually since appropriate tables would not be available.

6.2 Maintaining the win probability across a stoppage

In the D/L method the margin of Team 2's score with respect to par is maintained across a stoppage. This is totally logical as the target revision takes account only of the run scoring resources which were lost by the stoppage; the actual runs scored are irrelevant in working out the resource loss and hence the target revision. The principle that runs already scored do not affect the revised target has applied in all other target resetting methods that have ever been used [1, 2].

So long as Team 2 keep abreast of their par score, then a stoppage is unlikely to have a significant effect on their likelihood of achieving their target. However, if a team are well ahead or well behind par, the likely outcome may be substantially affected and the remainder of the match after play resumes may be a non-contest.

For instance, in Example 1, West Indies were 34 runs ahead of the D/L par score of 104 at the interruption. Their remaining task was to score 115 in 21 overs at under 5.5 runs per over. After the resumption, the par score is unchanged as is their margin ahead of it but their revised task was one of scoring 58 in 11 overs with nine wickets still in hand. Although the match was certainly not a non-contest, West Indies' task was easier after the interruption because of D/L's mechanism of maintaining the *margin* of advantage across a stoppage.

We have considered a mechanism for adjusting the target resetting process so as to maintain the probability of each team winning the match across a stoppage. In Duckworth and Lewis [7], we outline this mechanism and show that although there are seemingly useful ways of overcoming the perceived problem, there is substantial inequity in such a mechanism. It turns out that an egalitarian philosophy of taxing the run-rich to aid the run-poor becomes necessary. Also necessary are many special rules in order to overcome several anomalous situations that could arise in practice. We also show that it is impossible to maintain the "equal probability" concept in all circumstances. With all the calculations involved it would only be practical to implement such a concept by means of a computer program.

6.3 Review of D/L parameters

The D/L tables in Appendix 1 have been in use since 1 January 1997 and were devised based on data available up to mid-1996. Since that time the nature of the one-day game has continued to evolve; coaches try out different batting and/or bowling strategies, physical fitness of the players is regarded as more and more important, and there have been several minor changes to the rules of the game.

All of these factors, and the evidence from matches as summarised in Figures 2(a) to 2(j) and Appendices 2(a) and 2(b), suggest that from time to time the parameters of the model in (1) need to be reviewed. Such a review is currently in process. Although it is likely that the parameters will not change substantially nor the consequential resource percentages, nevertheless it is important for the game, and for the credibility of D/L, to keep abreast of the changing nature of the game.

7 Conclusion

The D/L method is now in use universally at international and national levels having been trialled over several years. Following a meeting of the Cricket Committee (Playing) of the International Cricket Council, the method has been adopted as the standard "rain-rule" for the three-year planning period to August 2004 [10]. The method has also been adopted at many lower levels of the game including grade cricket in Australia and village cricket in England.

Its track record over around 300 known cases shows that it generally works well. Some scenarios have seemed strange to several pundits but, upon closer inspection, the targets have been shown to be generally logical and fair. In few cases have reports suggested that the result of a match was blatantly unrealistic, unlike results when previous methods were in use.

One of D/L's strength is its simplicity in operation; interested persons can calculate the target using nothing more than a pocket calculator and the D/L tables. We know that in this format it is not perfect and shows its weakness mainly early in replies to well-above-average Team 1 totals. The methodology is in place, however, to handle all of these weaknesses. They can be implemented by the cricketing authorities if the extra precision is deemed desirable and this can be done through our computer software program CODA.

The price for this improved precision is the loss of transparency of the interested spectator. It is for the various cricket authorities to decide whether this price is worth paying but we are ready to provide the facility if it is ever wanted.

Appendix 1 The D/L (Duckworth/Lewis) method of adjusting target scores in interrupted one-day cricket matches

Table of resource percentages remaining - over by over

Overs left
50 to 0

	wickets lost										
overs left	0	1	2	3	4	5	6	7	8	9	overs left
50	100.0	92.4	83.8	73.8	62.4	49.5	37.6	26.5	16.4	7.6	50
49	99.2	91.8	83.3	73.5	62.2	49.4	37.6	26.5	16.4	7.6	49
48	98.3	91.1	82.7	73.1	62.0	49.3	37.6	26.5	16.4	7.6	48
47	97.4	90.3	82.2	72.7	61.8	49.2	37.6	26.5	16.4	7.6	47
46	96.5	89.6	81.6	72.3	61.5	49.1	37.5	26.5	16.4	7.6	46
45	95.5	88.8	81.0	71.9	61.3	49.0	37.5	26.4	16.4	7.6	45
44	94.6	88.0	80.4	71.5	61.0	48.9	37.5	26.4	16.4	7.6	44
43	93.6	87.2	79.7	71.0	60.7	48.7	37.4	26.4	16.4	7.6	43
42	92.5	86.3	79.0	70.5	60.4	48.6	37.4	26.4	16.4	7.6	42
41	91.4	85.4	78.3	70.0	60.1	48.4	37.3	26.4	16.4	7.6	41
40	90.3	84.5	77.6	69.4	59.8	48.3	37.3	26.4	16.4	7.6	40
39	89.2	83.5	76.8	68.9	59.4	48.1	37.2	26.4	16.4	7.6	39
38	88.0	82.5	76.0	68.3	59.0	47.9	37.1	26.4	16.4	7.6	38
37	86.8	81.5	75.2	67.6	58.6	47.7	37.1	26.4	16.4	7.6	37
36	85.5	80.4	74.3	67.0	58.2	47.5	37.0	26.4	16.4	7.6	36
35	84.2	79.3	73.4	66.3	57.7	47.2	36.9	26.3	16.4	7.6	35
34	82.9	78.1	72.4	65.6	57.2	47.0	36.8	26.3	16.4	7.6	34
33	81.5	76.9	71.4	64.8	56.7	46.7	36.6	26.3	16.4	7.6	33
32	80.1	75.7	70.4	64.0	56.1	46.4	36.5	26.3	16.4	7.6	32
31	78.6	74.4	69.3	63.2	55.5	46.0	36.4	26.2	16.4	7.6	31
30	77.1	73.1	68.2	62.3	54.9	45.7	36.2	26.2	16.4	7.6	30
29	75.5	71.7	67.0	61.3	54.3	45.3	36.0	26.1	16.4	7.6	29
28	73.9	70.2	65.8	60.4	53.5	44.9	35.8	26.1	16.4	7.6	28
27	72.2	68.8	64.5	59.3	52.8	44.4	35.6	26.0	16.4	7.6	27
26	70.5	67.2	63.2	58.3	52.0	43.9	35.4	25.9	16.4	7.6	26
25	68.7	65.6	61.8	57.1	51.2	43.4	35.1	25.9	16.4	7.6	25
24	66.9	64.0	60.4	55.9	50.3	42.8	34.8	25.8	16.3	7.6	24
23	65.0	62.3	58.9	54.7	49.3	42.2	34.4	25.6	16.3	7.6	23
22	63.0	60.5	57.3	53.4	48.3	41.5	34.1	25.5	16.3	7.6	22
21	61.0	58.6	55.7	52.0	47.2	40.8	33.7	25.3	16.3	7.6	21
20	58.9	56.7	54.0	50.6	46.1	40.0	33.2	25.2	16.3	7.6	20
19	56.8	54.8	52.2	49.0	44.8	39.1	32.7	24.9	16.2	7.6	19
18	54.6	52.7	50.4	47.4	43.5	38.2	32.1	24.7	16.2	7.6	18
17	52.3	50.6	48.5	45.8	42.2	37.2	31.5	24.4	16.1	7.6	17
16	49.9	48.4	46.5	44.0	40.7	36.1	30.8	24.1	16.1	7.6	16
15	47.5	46.1	44.4	42.1	39.1	35.0	30.0	23.7	16.0	7.6	15
14	45.0	43.7	42.2	40.2	37.5	33.7	29.1	23.2	15.8	7.6	14
13	42.4	41.3	39.9	38.1	35.7	32.3	28.2	22.7	15.7	7.6	13
12	39.7	38.8	37.6	36.0	33.9	30.8	27.1	22.1	15.5	7.6	12
11	36.9	36.1	35.1	33.7	31.9	29.2	25.9	21.4	15.3	7.5	11
10	34.1	33.4	32.5	31.4	29.8	27.5	24.6	20.6	14.9	7.5	10
9	31.1	30.6	29.8	28.9	27.6	25.6	23.1	19.6	14.5	7.5	9
8	28.1	27.6	27.0	26.3	25.2	23.6	21.5	18.5	14.0	7.5	8
7	25.0	24.6	24.1	23.5	22.7	21.4	19.7	17.2	13.4	7.4	7
6	21.7	21.4	21.1	20.6	20.0	19.0	17.7	15.7	12.6	7.2	6
5	18.4	18.2	17.9	17.6	17.1	16.4	15.5	14.0	11.5	7.0	5
4	14.9	14.8	14.6	14.4	14.1	13.6	13.0	11.9	10.2	6.6	4
3	11.4	11.3	11.2	11.1	10.9	10.6	10.2	9.6	8.5	6.0	3
2	7.7	7.7	7.6	7.6	7.5	7.4	7.2	6.9	6.3	4.9	2
1	3.9	3.9	3.9	3.9	3.9	3.8	3.8	3.7	3.5	3.1	1
0	0	0	0	0	0	0	0	0	0	0	0
overs left	0	1	2	3	4	5	6	7	8	9	overs left
	wickets lost										

Table of resource percentages remaining—over by over.

Appendix 2(a)

Overs left u	Wickets lost w									
	0	1	2	3	4	5	6	7	8	9
50	233.1									
49	232.6	205.1	241.0							
48	229.7	210.3	215.3							
47	228.1	210.5	190.1							
46	226.1	206.9	193.2	145.0						
45	225.0	206.7	187.8	143.2						
44	221.4	205.0	192.3	126.6	140.0					
43	219.2	204.4	189.5	146.0	150.0					
42	213.1	205.6	185.9	149.4	149.5					
41	209.5	203.8	181.7	152.7	161.0	131.0				
40	207.3	198.8	181.6	153.7	157.5	128.0	134.0			
39	204.2	197.5	180.0	156.4	156.8	140.5	69.0			
38	198.5	196.0	178.4	150.0	150.9	136.5	64.5			
37	192.2	190.6	176.6	156.6	143.4	120.3	117.0	8.0		
36	188.2	186.2	173.8	157.5	136.2	118.8	108.0	6.0		
35	178.6	184.9	166.3	162.2	137.4	113.5		104.0	1.0	
34	177.0	183.1	164.8	157.2	130.1	117.9		97.0		
33	172.9	181.1	161.6	154.4	127.4	120.3	106.0	97.0		
32	168.6	180.7	156.6	149.4	135.5	109.5	104.5	89.0		
31	161.4	178.4	153.8	148.5	130.6	117.5	86.5	90.3		
30	166.0	175.7	150.8	146.5	125.5	116.7	85.8	86.0		
29	158.4	171.4	150.0	140.8	134.3	108.2	86.0	77.0	75.0	
28	154.2	170.9	147.9	135.5	131.0	111.6	82.6	60.6	75.0	
27	152.3	165.5	147.5	135.8	126.8	104.2	92.9	58.2	73.0	
26	150.6	159.7	145.9	131.4	124.4	101.7	87.1	56.8	71.0	
25	154.3	154.2	143.3	128.8	120.0	98.1	93.7	43.7	71.0	
24	150.9	154.5	140.9	125.7	116.3	96.3	91.4	40.9	70.0	
23	149.9	145.3	136.2	124.3	114.6	93.8	89.5	54.8	63.0	
22	147.8	138.3	134.2	119.7	112.9	93.9	88.9	57.4	35.5	62.0
21	140.7	138.3	128.5	117.4	109.0	89.1	85.6	60.1	40.2	60.0
20	141.3	134.5	125.0	112.4	106.7	91.8	76.9	64.6	39.1	33.0
19	135.3	133.1	123.0	107.2	103.9	89.8	70.1	60.6	31.7	36.0
18	129.0	128.0	118.3	105.0	100.2	87.2	71.1	54.0	35.0	50.5
17	124.0	127.3	112.4	102.9	95.3	83.3	68.7	52.4	35.8	38.7
16	117.3	121.7	109.0	100.5	89.4	82.5	62.7	51.3	33.7	30.2
15	119.0	119.3	104.4	97.1	88.5	77.7	59.8	50.7	34.6	27.6
14	126.0	111.0	102.7	93.6	84.4	73.1	59.0	46.8	33.0	18.8
13	136.0	104.7	99.2	91.2	79.5	72.7	57.2	45.4	30.2	13.7
12	127.0	103.5	95.7	85.2	78.1	69.5	57.3	43.1	32.3	15.7
11	119.0	99.2	90.0	80.1	73.3	66.3	56.8	39.5	30.1	17.2
10	105.0	94.1	82.4	76.6	69.7	62.3	54.8	42.7	26.6	12.3
9	100.0	90.2	78.3	71.9	63.9	59.1	50.9	38.5	27.1	9.4
8		84.8	70.4	66.4	61.2	52.9	45.0	40.3	28.2	10.4
7		76.4	66.7	58.4	56.1	48.8	42.6	37.0	27.9	11.5
6		70.3	59.8	53.5	49.6	44.8	37.4	32.5	26.2	9.8
5		55.3	49.3	46.9	42.4	39.8	33.4	28.5	22.7	10.7
4		45.0	44.5	38.1	35.2	33.1	29.6	24.8	19.0	11.6
3		32.5	32.5	30.6	29.2	28.0	23.2	19.9	16.3	10.2
2			21.0	23.3	19.8	20.0	16.1	15.3	13.6	7.3
1			11.5	12.0	10.5	11.3	10.1	8.7	7.7	5.0

Table of average runs scored in ODI matches for overs left and wickets lost.

Appendix 2(b)

Overs left u	Wickets lost w									
	0	1	2	3	4	5	6	7	8	9
50	330									
49	290	36	4							
48	264	60	6							
47	238	78	15							
46	209	94	27	1						
45	185	104	37	5						
44	160	111	53	7	1					
43	132	130	54	14	2					
42	115	129	66	20	2					
41	100	134	70	23	3	2				
40	87	129	81	29	4	1	1			
39	71	126	91	37	4	2	2			
38	62	125	89	45	8	2	2			
37	52	120	93	50	13	3	1	1		
36	44	118	96	48	22	4	1	1		
35	32	112	102	56	22	8		1	1	
34	26	106	99	67	25	9		1		
33	23	97	101	70	29	10	2	1		
32	22	90	101	71	33	12	2	2		
31	20	85	95	72	41	15	2	3		
30	19	80	92	78	46	14	4	3		
29	17	74	88	82	42	24	5	3	1	
28	17	64	91	83	45	25	5	5	1	
27	16	58	86	84	48	30	8	5	1	
26	14	58	79	87	52	28	12	5	1	
25	10	54	77	90	56	29	13	7	1	
24	9	53	69	94	58	26	19	8	1	
23	7	45	70	94	58	34	15	13	1	
22	6	40	72	90	58	38	15	15	2	1
21	6	33	65	90	67	38	16	14	5	1
20	4	33	61	94	65	35	21	13	7	2
19	4	28	55	95	66	35	27	16	6	3
18	4	26	52	87	70	39	27	21	5	2
17	4	22	52	82	76	42	25	23	6	3
16	4	20	48	81	77	44	29	21	6	5
15	3	17	47	76	71	53	32	23	8	5
14	2	15	37	75	79	57	31	23	10	6
13	1	15	31	70	79	59	33	28	12	7
12	1	10	28	61	88	60	32	30	15	7
11	1	9	26	54	84	67	34	34	16	5
10	1	8	23	44	86	66	37	37	21	6
9	1	6	21	41	78	69	43	39	23	8
8		5	20	38	66	81	48	28	27	14
7		5	17	30	61	79	54	35	25	18
6		3	13	31	56	67	62	40	32	18
5		3	7	32	49	60	66	48	34	16
4		2	6	27	37	60	64	53	44	15
3		2	4	19	38	45	61	68	47	20
2			4	12	29	44	50	71	49	38
1			2	12	17	39	40	64	61	48

Table of number of ODI cases producing averages in Appendix 2(a).

References

[1] F. Duckworth and T. Lewis, "A fair method for resetting the target in interrupted one-day cricket matches", in *Third Conference on Mathematics and Computers in Sport*, N. de Mestre (editor), Bond University, Queensland, Australia (1996), 51–68.

[2] F. C. Duckworth and A. J. Lewis, "A fair method of resetting the target in interrupted one-day cricket matches", *J. Oper. Res. Soc.*, **49** (1998), 220–227.

[3] F. Duckworth and T. Lewis, *Your Comprehensive Guide to the Duckworth/Lewis Method*, University of the West of England, Bristol (1999).

[4] N. Hussain, "Captain's diary", *Sunday Telegraph* (24 February 2002).

[5] M. Johnson, "West Indies show madness in the method", *Daily Telegraph* (11 May 1999).

[6] T. Lewis and F. Duckworth, "Developments in the Duckworth–Lewis (D/L) method of target-resetting in one-day cricket matches", in *Fourth Conference on Mathematics and Computers in Sport*, N. de Mestre and K. Kumar (editors), Bond University, Queensland, Australia (1998), 131–151.

[7] F. Duckworth and T. Lewis, "Can the probability of winning a one-day cricket match be maintained across a stoppage?", these *Proceedings*.

[8] A. Longmore, "The odd couple getting even with the weather", *The Independent on Sunday* (16 May 1999).

[9] Web site, Cricinfo: http://www.cricket.org/link_to_database/ARCHIVE.

[10] Web site: http://www.cricket.org/link_to_database/ARCHIVE/CRICKET_NEWS/2001/MAY/128586_ICC_25MAY2001.html.

Copyright

CAN THE PROBABILITY OF WINNING A ONE-DAY CRICKET MATCH BE MAINTAINED ACROSS A STOPPAGE?

Frank Duckworth
"Moonrakers"
Taits Hill, Stinchcombe
Glos. GL11 6PS, UK
f.duckworth@rss.org.uk

Tony Lewis
The Business School
Oxford Brookes University
Wheatley
Oxford OX33 1HX, UK
ajlewis@brookes.ac.uk

Abstract

Substantial improvement in the equity of target resetting in interrupted one-day cricket matches has been brought about by the now universally adopted Duckworth/Lewis method. Because of one of its principles, namely that of maintaining the margin of advantage, it can sometimes happen that what was a difficult task at a stoppage becomes a virtually impossible one upon the resumption. Conversely, if the batting side were heading comfortably for victory, they may already have won when playing conditions would have allowed a resumption. In the interests of maintaining an exciting contest on resumption, we have therefore examined a variation in the method of applying the D/L method whereby the probability of winning is maintained across a stoppage. In this paper we describe how such a method might be implemented and how adjustments would be necessary to avoid anomalous situations. The paper also shows there is considerable inequity in the equal probability concept and that it cannot be sustained in all situations. As a computer program would be necessary to perform the calculations, the transparency of the basic Duckworth/Lewis method is lost. We conclude that a change to an equal probability concept would create more problems than it would solve.

1 Introduction

Ever since the inception of one-day cricket in the 1960s there has been a need for a method of resetting the target in interrupted matches. The Duckworth/Lewis (D/L) method has recently become the international standard method and is used in most countries playing cricket under the auspices of the International Cricket Council. The D/L method is discussed in several papers including Duckworth and Lewis [4, 5]. These papers also outline the mechanisms of several other target-resetting methods that have been used in the past.

All of these methods have based the revised target in various ways on the final score obtained by the team batting first (Team 1). D/L has become accepted as the fairest system around as it yields fair targets in the vast majority of the varied situations in which stoppages can occur in one-day cricket. One of its principles is that it maintains the margin of advantage one team may have established at the time of a stoppage. The effect of this can sometimes be that upon resumption in play a team batting second (Team 2) which were in a weak position at the stoppage find themselves with an impossible task on the resumption. Conversely, if Team 2 were in a strong position at a stoppage, upon resumption the remaining task can be very easy or it may be the case that they have already achieved the revised target without any resumption.

Such scenarios, as described below, have prompted some critics to suggest that the target resetting process should maintain the probability of the win across a stoppage. If this could be achieved then there would always be a meaningful contest for the shortened innings remaining with some interest left for spectators who may have braved the inclement weather hoping at some stage for a restart. This paper discusses the viability of using "probability" to set the target, however this might be defined.

2 The standard D/L procedure

The D/L method of target resetting uses resource percentages of the two teams in order to reset the target [6]. We have

$$T = \frac{SR_2}{R_1}, \qquad \text{if } R_2 \leq R_1, \tag{1a}$$

$$T = S + G_{50}(R_2 - R_1), \qquad \text{if } R_2 \geq R_1, \tag{1b}$$

where S is Team 1's final total, R_1, R_2 are the resource percentages available to Teams 1 and 2 respectively and G_{50} is the average first innings total for 50 overs in the standard of competition in which D/L is being used. The next integer above T is the minimum number of runs needed to win. This is commonly known as the "target", and one run below this is the score required to tie.

Resource percentages for the overs remaining and wickets lost are read from the D/L tables, the over-by-over version of which can be seen in Appendix 1. Some of the examples discussed below need the ball-by-ball version of the tables for the precision required to obtain the correct target for the matches concerned. Our booklet [6] contains these tables as used in official matches run under the auspices of the International Cricket Council (ICC).

As an innings progresses the heavy lines of Figures 1 and 5 represent the required target, under the two respective scenarios of (1a) and (1b). They also represent the par score for the innings in progress, which is used as the basis for deciding the winner if a game is abandoned and also a benchmark against which Team 2's progress towards achieving their target can be assessed.

Example 1

In an English Norwich Union National League (NUNL) match between Northamptonshire and Somerset on 30 August 2001, Northants scored 214 in their full allotment of 45 overs. Somerset were 1/155 in 24 overs and four balls of the six balls permitted per over (written as 24.4 overs) when it rained with 20.2 overs remaining. In the event the match was washed out with no further play. Under the D/L method and from the D/L resource percentage tables [6], in (1a) $S = 214$, $R_1 = 95.5\%$, $R_2 = 95.5 - 57.4 = 38.1\%$ and $T = 85.48$. The par score was 85 and so Somerset were declared the winners by their margin of advantage of 70 runs. But if it had been possible to resume for a further 8.2 overs losing 12 overs of play then $R_2 = 38.1 + 28.6 = 66.7$ and, from (1a), $T = 149.46$ so that the revised target would have been 150, which they had already achieved. Consequently it would have had to be announced that Somerset had already won and no further play would have been necessary. Could an equal probability approach have provided a fair target for Somerset to chase?

Example 2

In another NUNL match on 12 August 2001, Durham scored 8/269 in their 45 overs. In reply, Derbyshire had reached 5/134 in 28.5 overs when rain interrupted play and initially 14 overs were lost. At the stoppage 16.1 overs remained. In (1a) with $S = 269$, $R_1 = 95.5$, $R_2 = 95.5 - 36.3 = 59.2\%$ so that $T = 166.75$ with a par score of 166. Derbyshire were 32 runs behind par. On a restart for the remaining 2.1 overs they would resume 32 runs behind par and the revised target, with $R_2 = 59.2 + 7.9 = 67.1$, $T = 189.00$, would be 190. Their requirement had changed from the difficult one of making 136 in 16.1 overs to the virtually impossible one of making 56 in 13 balls. In the event just as play was about

to resume further rain caused the match to be abandoned with no further play possible and Durham were the victors by 32 runs. But could an equal probability approach have provided a fairer target for Derbyshire had play resumed?

3 Defining probability

In the two examples, if the scores of the separate Teams 2 at their stoppages were respectively 85 and 166 for the corresponding number of wickets lost, then they would both be level par. This suggests that such matches would then be evenly balanced representing the probability 0.5 of each team winning in each of these matches. Case history in the application of the use of D/L is not yet sufficient to test whether the distribution of winners is reasonably balanced from such situations. Some, as yet unpublished, analysis of appropriate matches which have not been interrupted, however, is suggesting that such a proposition is not unreasonable.

Whereas a definition of probability in balanced matches can be determined, defining the probability of victory when Team 2 are ahead or behind par is a more difficult concept.

Clarke [1] used dynamic programming first in establishing optimum batting strategy for Team 1 to maximise their expected final total and then to determine Team 2's optimal strategy to maximise the probability of achieving their target. He suggests that the model could be developed to determine revised targets in interrupted matches to preserve Team 2's probability of achieving their target.

Preston and Thomas [8] have developed a methodology of defining probability of victory based on their dynamic programming model of optimum batting strategy. They go on to illustrate how the target could be adjusted based on cricketing information that is not readily available in practice.

4 "Probability" based on D/L resources

If "degree of difficulty" is used as a surrogate for probability then the D/L resource percentages can be used to model the maintenance of the chance of a Team 2 victory across a stoppage. We first consider the situations in which $R_2 \leq R_1$. Later, we will consider the other situation in which $R_2 > R_1$.

Once Team 1's innings has been completed, S and R_1 in (1a) are constants for the remainder of the game. Now T is a linear function of R_2 so that the heavy line in Figure 1 depicts the score that Team 2 need to beat for the resources they have available.

Suppose the target is adjusted so that the remaining tasks before and after the stoppage are in proportion to the resources remaining before and after. We shall refer to this as our Equal Probability (EP) approach.

Assume that before a stoppage Team 2's score to beat was S_b and that they had scored Y runs with the resources consumed X. Let R_b denote their total resources available before the stoppage. Denote Team 2's resources available, after allowing for the resources lost by the stoppage, by R_a. Then the score to beat after the stoppage S'_b would be such that

$$\frac{S'_b - Y}{S_b - Y} = \frac{R_a - X}{R_b - X},$$

so that

$$S'_b = Y + (S_b - Y)\frac{R_a - X}{R_b - X}. \tag{2}$$

Figure 1 shows Team 2's score Y, at a point H, below the par score for the resources they have consumed, X by the time of the stoppage. The revised score to beat, S'_b, is where the line from H to the original target intersects the ordinate at R_a. This is the value of the ordinate at G in Figure 1. Point E would represent the standard D/L target reset according to (1a). The length of the line segment EG represents the difference in the revised scores to beat between the standard D/L and an EP approach.

The same logic applies if Team 1 are ahead of the par score. A separate diagram illustrating this is felt not to be necessary although some examples will relate to such a position.

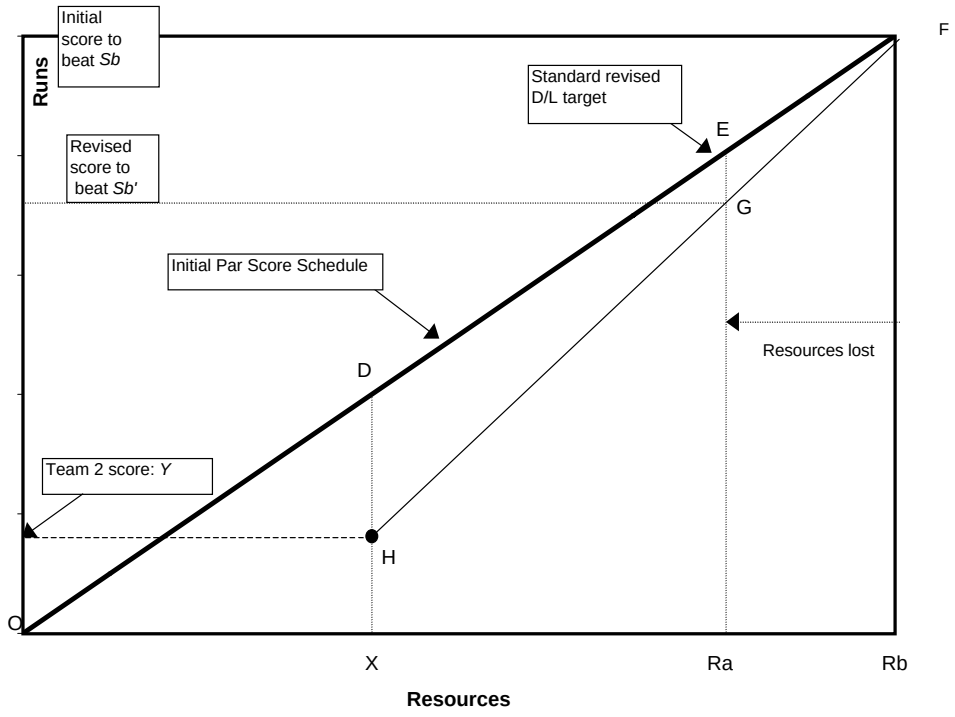


Figure 1: Runs/resources plot.

Examples

In Example 1, $S_b = 214$, $Y = 155$, $X = 95.5 - 57.4 = 38.1\%$, $R_b = 95.5\%$, $R_a = 66.7\%$ so that $S'_b = 184.39$ and instead of the match being all over without a need for a resumption, the revised target would be 185, a further 30 in the 8.2 overs. This would still be a relatively easy task for Somerset, but arguably fairly so given the strength of their position at the stoppage and would provide some interest for the spectators had play been able to be resumed.

In Example 2, $S_b = 269$, $Y = 134$, $X = 95.5 - 36.3 = 59.2\%$, $R_b = 95.5\%$, $R_a = 67.1\%$ so that $S'_b = 163.38$ and the revised target would be 164. The remaining task would be the still difficult, but not now impossible, one of making 30 runs in 13 balls.

These examples illustrate that there could be considerable merit in resetting the target along EP lines since some entertainment appears to be guaranteed after a stoppage. Since cricket is not just in the sporting industry but also in the entertainment industry this is a very important consideration provided equity to teams can be maintained. Unfortunately this is not always the case, as subsequent discussion will explain.

5 Taxing the rich to aid the poor

In (2), the revised score to beat is clearly a function of the runs that Team 2 have already scored at the interruption. Since $R_b \geq R_a$, then $(R_a - X)/(R_b - X) \leq 1$ and it follows that the better a team have been performing the greater the task remaining following an interruption; their good performance is "taxed" in order to make an interesting game at the resumption. On the contrary a team performing

poorly before a stoppage is subsidised in order to create an entertaining match. This characteristic is not present in any of the previous methods of target resetting; the adjustment is made entirely on the overs lost. We consider this egalitarian approach to target resetting as alien to the competitive nature of cricket. It can also lead to unreasonable inconsistency as Example 3 illustrates.

Example 3

In two matches played on adjacent grounds A and B, Team 1 score 250 in 50 overs on each ground. Team 2 face 20 overs and have lost three wickets on both grounds when it rains and ten overs are lost at each match. On ground A, Team 2 had scored 3/120 whereas on ground B they had only scored 3/50. By the standard D/L method, the target on both grounds would be 221. However, under EP teams are treated differently because of their comparative performances up to the time of the stoppage. Team 2's target on ground A will be 226 whereas the target for Team 2 on ground B will be only 213. Although such inconsistency would rarely, if ever, become apparent, its existence demonstrates a fundamental unfairness in the EP concept.

6 Further stoppages

Although the adjustment to the final target using resources appears to offer a more equitable mechanism one must consider what would happen if a match suffers further interruptions or subsequent abandonment.

In Figure 1, the line HG provides the remaining task for Team 2. Suppose that after a few more overs their subsequent score for the corresponding resources further consumed places them above this line and then the match has to be abandoned. How should the winner of the match be decided? If we consider the revised score to beat represented by point G in Figure 1 then Team 2 are ahead of their revised task. If Team 2 are to be declared the winners on this basis then the target resetting process under EP is tantamount to starting a new match after the interruption and laying aside any advantage Team 1 had established at the stoppage. This is clearly unfair to Team 1 in this circumstance and to Team 2 if, at the stoppage, they were considerably ahead of the par score, represented by point D in Figure 1.

It is clear, therefore, that the margin of advantage at the stoppage should not be ignored, and yet, reducing this margin would appear to be an appropriate corollary to maintaining the probability of the win across the stoppage. Figure 2 suggests a mechanism of how this can be achieved.

7 Adjusting the margin of advantage

At the stoppage, Team 1's margin of advantage is represented by the line-segment DH in Figure 2. Equal probability suggests that this margin should be reduced commensurate with the EP revised target. This could be achieved by resetting the par score on the notional total S^* that Team 1 would have scored to produce the revised target at G using the standard D/L formula (1a). Substituting, in (1a), S'_b for T, R_a for R_2 and rearranging provides this notional Team 1 score:

$$S^* = \frac{R_1}{R_a} S'_b. \tag{3}$$

This is represented, in Figure 2, by extending a line OG to intersect the ordinate at R_b. The revised par score is shown at D′, which, from (1a) and (3), has the value

$$\text{Par score} = \frac{S'_b X}{R_a} \tag{4}$$

and the line segment D′G represents the revised par-score schedule.

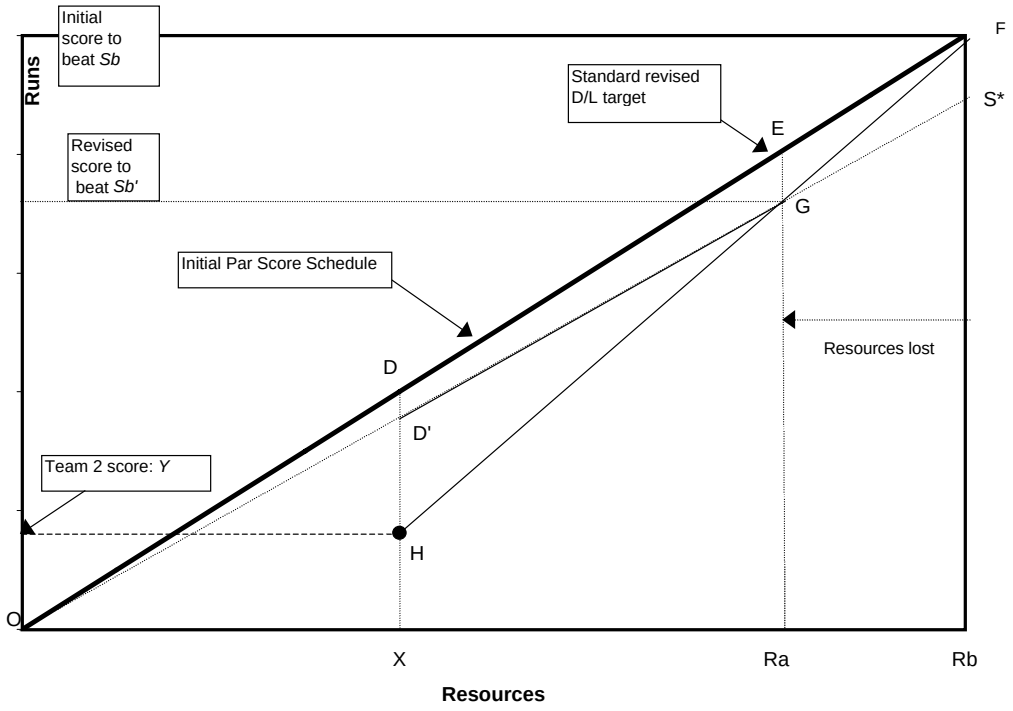


Figure 2: Runs/resources plot — possible adjusted par score schedule.

At the stoppage in Example 1, the par score, from (1a) with X replacing R_2, was 85, rounded down from 85.48. That is, Somerset were 70 runs ahead of par. Under standard D/L, the margin would still be +70 on the restart, but under EP the revised par score, from (4), would be 105.33. The revised par-score schedule represented by the equivalent of the line-segment D′G in Figure 2 would then be used to decide the winners if further rain subsequently abandoned the match.

In Example 2, from (4) with $S_b' = 163.38$, $X = 59.2\%$ and $R_a = 67.1\%$, the revised par score would be 144 so that Derbyshire, at 5/134, requiring 30 in 13 balls, would be ten below the revised EP par score instead of 32 below the standard D/L par of 166.

These two examples suggest that (4) appears to offer a compromise between the harshness of standard D/L to teams well behind par and its generosity to teams well ahead of par in the provision of a "reasonable" par schedule.

8 Very short resumptions

Suppose that in Example 2, Derbyshire have resumed play wanting those 30 in 13 balls. The first ball is a no ball (which normally adds one further run to those scored in any other way) and is hit for six, as is the next ball, which is legitimate. Derbyshire's score is now 5/147 with two overs left. But before any more balls are possible, more rain comes down and the match has to be abandoned. Who has won?

Their aggressive, some might say lucky, approach has made their remaining task one of 17 runs in 12 balls. In (4), inputs to the right-hand side remain fixed except for X which, after that one legitimate ball, becomes 59.7% and the par score is 145 from rounding down the 145.36, so that Derbyshire, now at 5/147, would have won by two runs.

Durham might well feel aggrieved at such a result. Prior to the stoppage their position was extremely strong from a sustained piece of good cricket over the 28.5 overs of Derbyshire's innings. After the stoppage, they had just two bad deliveries and events have somehow conspired to deprive them of the victory that was justly theirs. And this has come about through reducing the margin of advantage that one team had established.

These cases highlight the issue concerning the par score, which is such a useful concept in standard D/L. It would be unjust for the margin of advantage before a stoppage to be totally negated. But by using the line segment HG as the revised par score schedule this is exactly what is being done. Consequently, in the interests of fairness, the "equal probability" concept needs to make due allowance for the whole of the margin of advantage established by one team over another. This is important also if the margin of victory were to be used in round-robin one-day competitions for ranking teams on equal points [2, 7, 3].

9 Maintaining the difference from par

From discussions so far, it is becoming clear that in order to avoid unfair and anomalous situations the appropriate mechanism for establishing the revised par score schedule is to maintain the margin of advantage that either team has established at a stoppage. This is consistent with the current standard D/L procedure.

In the EP environment, the par score schedule in Figure 2 should be represented by the line DG and have the equation

$$\text{Par score} = P + (S_b' - P)\frac{R_2 - X}{R_a - X}, \qquad X \le R_2 \le R_a, \tag{5}$$

where P is the unrounded D/L par score at the stoppage and is calculated through (1) with R_2 replaced by X, and where R_2 ($\ge X$) is the total resource consumed by Team 2 including the play after the stoppage. Other symbols are as already defined.

By this mechanism, the margin of advantage over the par score is maintained across the stoppage. In Figure 3, the line segment DG represents the subsequent par score schedule as defined in (5).

For matches described so far the criterion summarised in (5) could overcome the difficulties over the margin of advantage and yet capture the concept of equal probability. But the issue of egalitarianism in cricket illustrated in Example 3 is still present and its unfairness can be underlined by revisiting that example.

Example 3 (continued)

Suppose in this hypothetical example that after the resumptions in the two parallel matches the fortunes of the two teams batting second are reversed. In ten more overs of play, Team 2 on ground A suffer a minor collapse and advance their score to 6/160. On the parallel ground B, Team 2 throw caution to the wind, throw their bats but although losing a few wickets advance their score to 6/155. There are now ten overs left in the match (after the ten overs stoppage from the first rain interruption). Suppose now that further rain causes the two matches to be abandoned. According to (5), the par score on ground A will be 161 and on ground B it will be 154. Although at the end of 30 overs of play the two teams have consumed exactly the same resources, Team 2 on ground A with the higher total would lose whereas Team 2 on ground B with the lower total would win! Again such an anomaly would be unlikely to be manifest in public, but it is there nevertheless and illustrates again the fundamental unfairness of the EP approach.

10 Long stoppage with short resumption

In Example 2, the requirement to maintain the margin of advantage yields an anomalous situation. The D/L par score at the stoppage was 166 yet the EP revised target would be 163.38, which is rounded up

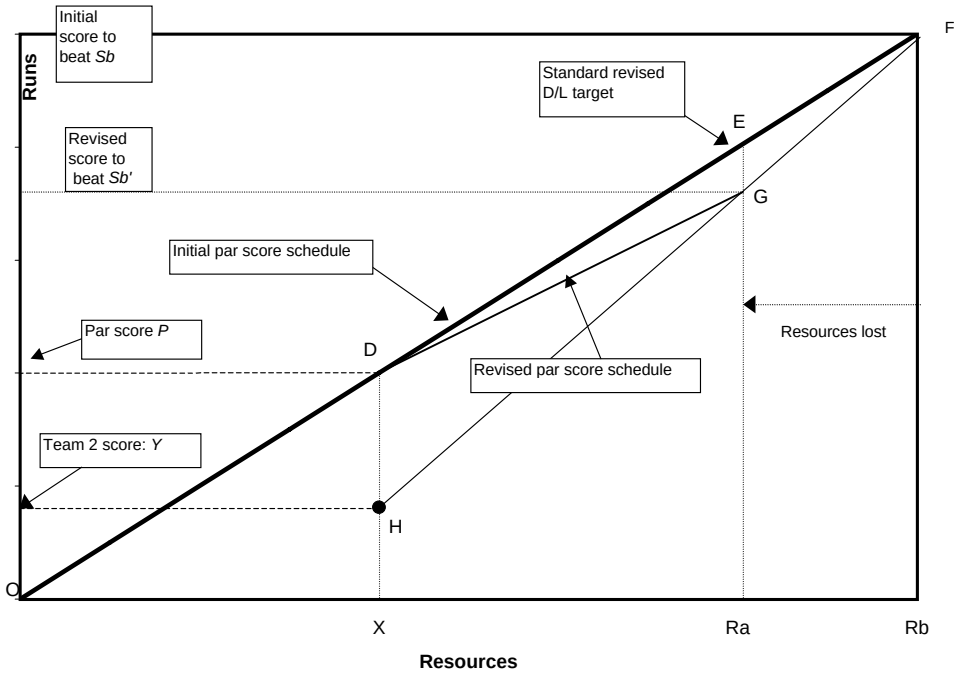


Figure 3: Runs/resources plot — EP revised par score schedule.

to 164 with 163 required to tie. As Derbyshire's innings would have progressed, the par score would not have risen as is normal with the consumption of resources, but would decline!

Figure 4 illustrates such a scenario depicted by the line-segment DG, the equivalent to that in Figure 3. The absurdity of this would be manifested in Example 2 if Derbyshire just blocked out the last 13 balls without making any attempt to score any runs. Their margin of defeat would *decline* from the 32 runs that they were behind par at the stoppage to a margin of 29 runs.

In order to eliminate such inconsistency, it would be necessary to constrain the par score in such situations so that the par score can never drop below the target. This means that the revised score to beat for any further interruptions would not change to allow for the further loss of resources. Consequently, the equal probability principle through D/L resources could not be sustained in resetting the target in these types of circumstances.

11 Stoppages to Team 1's innings

So far discussion has concentrated on the simpler case of stoppages reducing the duration of only Team 2's innings. A major strength of the standard D/L method is that, unlike earlier methods of target adjustment, it allows in a fair and logical way for the disadvantage that Team 1 usually suffer when their innings is unexpectedly shortened after it has started or is prematurely terminated. The measure of disadvantage to Team 1 is modelled in D/L by the excess resource that Team 2 subsequently enjoy over Team 1. Rather than adjusting targets in the same way as (1a), D/L recognises the implications of a possibly unrealistic assumption of extrapolation of performance per unit of resource. Instead "average performance" is used as the model of expected further runs with the excess run-scoring resources. Equation (1b) summarises the application in target resetting in this circumstance through use of G_{50}, the average score in 50 overs for the relevant standard of cricket.

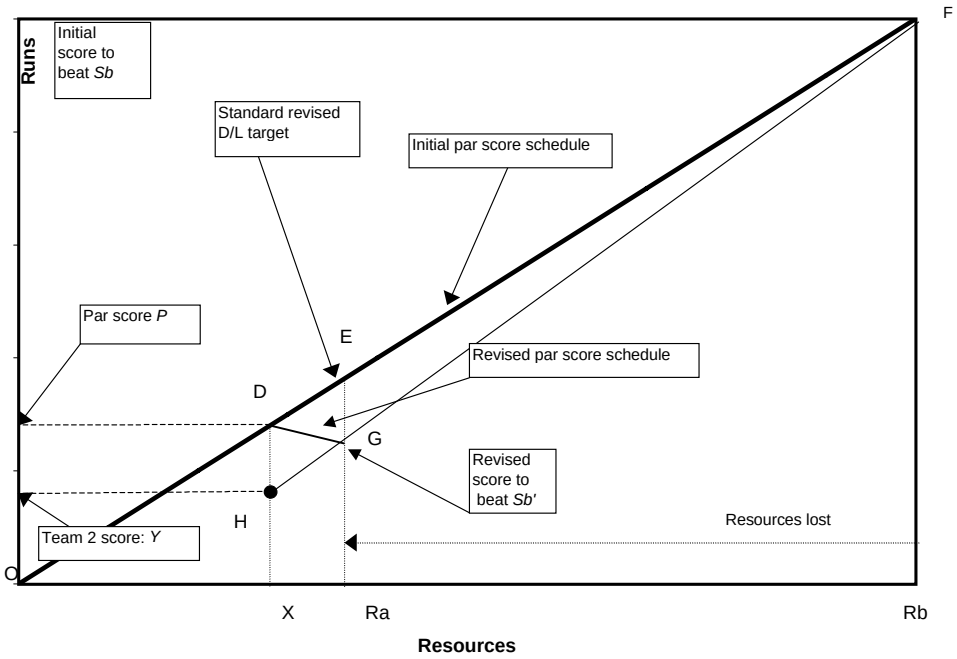


Figure 4: Runs/resources plot for a long stoppage.

Example 4

In a hypothetical one-day international (ODI), Team 1 score 4/196 in 40 of their 50 overs. Rain causes their innings to be terminated and Team 2 are given 36 overs. They reach 4/119 in 23 overs when more rain shortens their innings by a further five overs to 31. Following the restart, they have reached 6/166 after 27 of their 31 overs when yet more rain causes the match to be abandoned. Which team have won the match?

At the termination of Team 1's innings, ten overs remained and four wickets were down. Thus resources remaining from the D/L table are 29.8%. These were lost, so that the resources available to Team 1 were $100 - 29.8 = 70.2\%$ and so $R_1 = 70.2\%$. Team 2 are given 36 overs; so $R_2 = 85.5\%$. From (1b) with $G_{50} = 225$, the current ODI standard, $T = 230.43$ so that Team 1 are compensated for the premature termination of their innings and Team 2's advantage of knowing in advance of the shorter innings is neutralised by the provision of the enhanced target of 231 in 36 overs (230 to tie).

At the stoppage during Team 2's innings, standard D/L has $X = 85.5 - 35.7 = 49.8\%$ and from (1a), $T = 139.04$. Team 2 are 20 runs behind par and would lose by this amount if the match were abandoned (assuming 23 overs in an innings constituted a viable match). However, after a stoppage deducts five overs there are only eight more overs to play. From (1b), standard D/L provides the revised target of $T = 206.8$ using $R_2 = 85.5 - 35.7 + 25.2 = 75.0\%$ with the other inputs unchanged. Team 2 are still 20 runs behind the par score and have the difficult task of scoring a further 88 runs in eight overs to win.

The procedure for calculating the target under EP is depicted in Figure 5. In principle it is the same as that finalised for when $R_2 \leq R_1$. In this example, $S_b = 230.43$, $Y = 119$, $X = 85.5 - 35.7 = 49.8\%$, $R_b = 85.5\%$, $R_a = 75.0\%$ so that $S'_b = 197.66$ and the revised target would be 198, a further 79 in the eight overs. This would be a slightly easier and a more achievable task than the 88 required under standard D/L.

However, even further anomalous scenarios are possible because of the "kink", at K in Figure 5, in

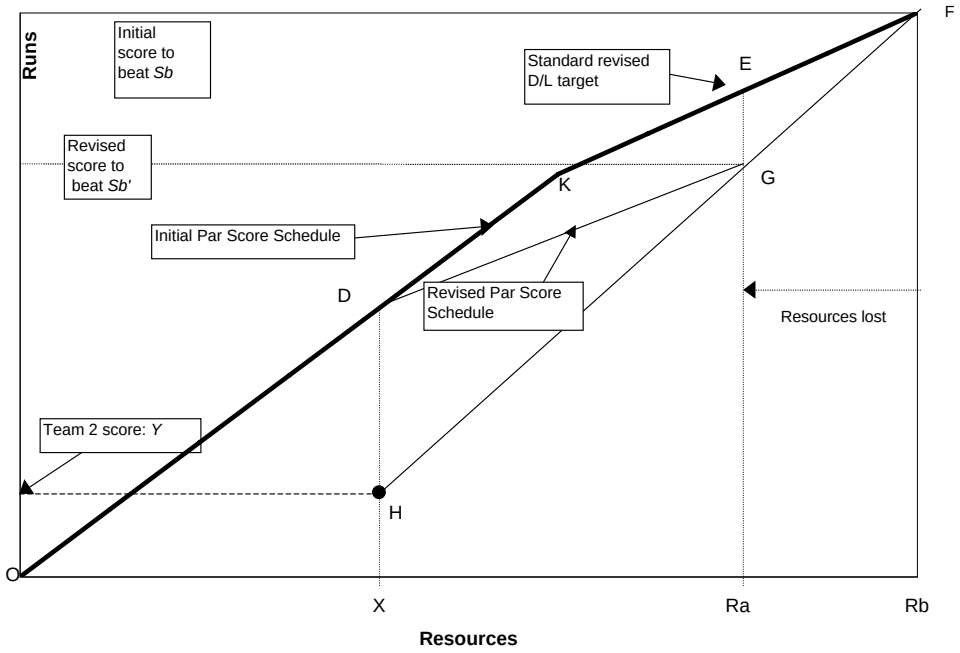


Figure 5: Team 1 interruption — run/resources plot.

the standard D/L par score curve which arises from the modelling of D/L.

Suppose, in this example, Team 2, have scarcely begun their reply and there is an interruption resulting in a loss of $R_b - R_a$ resources. Figure 6 represents this scenario with Team 2's score again represented by the point H after the small resource consumption X but somewhat exaggerated for clarity. Because of the lack of proportionality in the D/L target calculation for such scenarios with $R_2 > R_1$, the revised D/L target would be significantly different to that obtained via EP. This would seem unsatisfactory because there would be a negligible difference in cricketing terms but the different rules required for EP after a stoppage would yield a significantly different revised target.

In truth, if D/L set the target always using (1a) regardless of the relative magnitudes of R_2 and R_1, then this anomaly would not exist. But the reasons for the existence of the D/L "kink", already outlined, would require an EP approach to make due allowance for it. One "solution" to this anomalous situation would be to reduce the target along standard D/L lines for the loss of resource down from R_b to the value of the abscissa at K. Resource lost in excess of this could then be allowed for using EP principles already outlined. But if this were included in the methodology it would be another situation where the rationale of the EP approach could not be applied. It must therefore be concluded that EP is incompatible with $R_2 > R_1$.

12 A nonlinear model of maintaining probability

The way we have suggested of establishing equal probability targets is by scaling down the runs required for a tied match by the ratio of the resources remaining after and before the stoppage as in (2). This would assume, of course, that the probability of achieving these runs is directly proportional to the resources in hand.

But in reality this assumption breaks down when the task before the stoppage is either very difficult or very easy, its failure being most manifest when an innings is resumed for a very short period. For

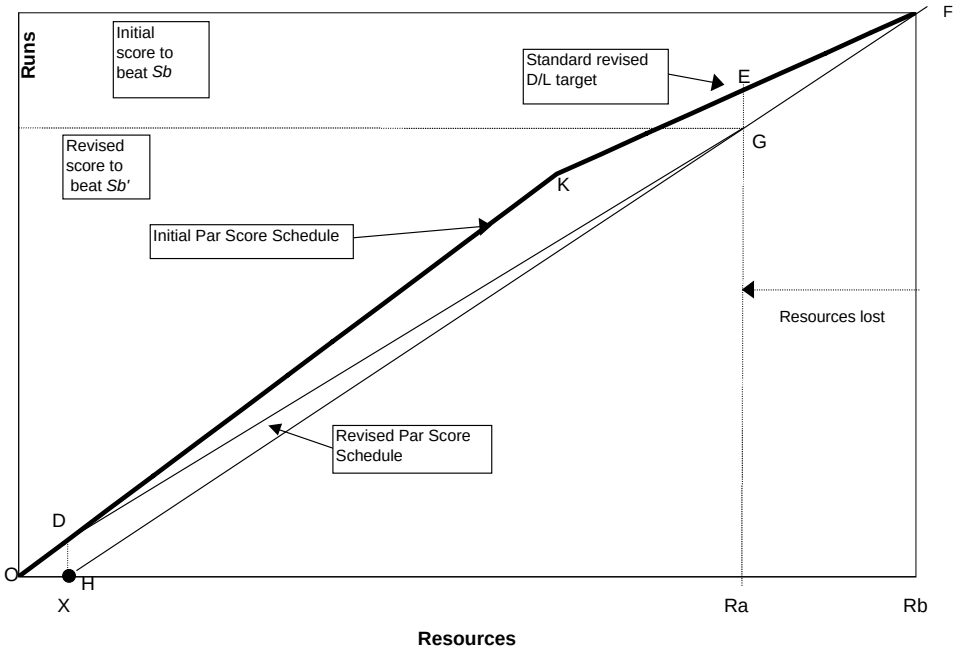


Figure 6: Early stoppage following Team 1 interruption.

instance, to score 400 in a 50-over innings is so big a requirement as to be a very rare event, yet the required rate of 4 runs per percentage resource (a 50-over innings constituting 100% resource, by definition) could quite possibly be achieved if only one over (about 4%) remained after the stoppage, 16 runs being needed.

Conversely, a virtual certainty of winning could become a by-no-means-certain task; for instance if just 50 runs were required from 50% resource (e.g. 24 overs to go with only four wickets down), a resumption for one over would still leave the team with the demanding task of making four runs (five to win).

To avoid such distortions, any realistic consideration of an equal probability approach would have to be based on the assumption that an additional regulation would be introduced setting a minimum length of play, say one or perhaps two overs, for a resumption to be permitted. However, in some circumstances such a regulation could prevent an exciting finish to a balanced match, which would be contrary to the very purpose for which the equal probability option would have been chosen.

A way of circumventing the problem would be to use a more realistic mathematical model of the way the number of runs obtainable with equal probability varies with resources available. This would add even further complexity to the operation of the method.

13 Summary

There are significant advantages to adopting an EP approach to target resetting but there are also significant disadvantages. These can be summarised as follows.

Advantages

1. No matter how far Team 2 may be ahead of par, the revised target will always be higher than the runs already scored, so the likelihood is much reduced of a situation where the rain stops and play is again possible but it is found that the revised target has already been achieved.

2. No matter how far Team 2 may be behind par, their task is unlikely to become totally impossible on resumption.

Disadvantages

1. Instead of one procedure, as in the standard D/L method, several different procedures would be necessary to calculate the revised targets. Standard D/L would still be applicable for any target revision that may be necessary before Team 2 start their innings and for a termination of Team 2's innings without previous interruptions within Team 2's innings. But for subsequent interruptions within Team 2's innings a different procedure would usually be required. Furthermore, after such an interruption, another procedure would be required to determine the par score schedule. And an even further procedure would be required to ensure that the EP target was never lower than the par score (+1) at the stoppage. Thus the unified standard D/L system, which was relatively easy to explain and operate manually, would be replaced by four different procedures.

2. This additional complexity would mean that EP could only be operated in practice by computer and the transparency would be lost. It would be possible to write revised regulations explaining how the calculations could be done by hand/calculator, but these would not be easily comprehensible by scorers, match officials or the general public.

3. EP means that a team's target is higher the more runs it has scored and lower the less it has scored. Subsequently the outcome of a match may be determined by the way Team 2's performance varied throughout its innings. This could well lead to anomalies and to basic unfairness.

4. If play could resume only for a very short period, then Team 1 could be presented with a chance of winning that was not a realistic possibility before the interruption; conversely a very difficult target could become a possibility, e.g. six off one ball. To reduce such distortion, it would be necessary either to introduce a regulation that play could only resume for a minimum of, say, two overs (and such a "minimum overs" rule could easily backfire denying the possibility of a result in a close match) or to use a more complex model for maintaining the win-probability across the stoppage.

5. EP is incompatible with situations where Team 2's resources exceed those of Team 1. It would therefore be necessary to change the current playing regulations whereby lost overs during Team 1's innings are divided equally between the two sides, thus reducing the probability of completion of a viable match, especially where further interruptions in play are liable to occur.

6. For matches resuming with few overs left, the margin of victory, which could be used for ranking purposes, would be unfairly affected under EP, resulting in unfair margins.

14 Conclusion

The equal probability option would usually ensure that there is a meaningful game after a resumption of play. However, it has many disadvantages, in particular its basic unfairness to teams doing well and the additional complexity of operation and consequent loss of transparency. The problems created by the equal probability approach seem to be greater than the problems it solves.

Appendix 1 The D/L (Duckworth/Lewis) method of adjusting target scores in interrupted one-day cricket matches

Overs left
50 to 0

	wickets lost										
overs left	0	1	2	3	4	5	6	7	8	9	overs left
50	100.0	92.4	83.8	73.8	62.4	49.5	37.6	26.5	16.4	7.6	50
49	99.2	91.8	83.3	73.5	62.2	49.4	37.6	26.5	16.4	7.6	49
48	98.3	91.1	82.7	73.1	62.0	49.3	37.6	26.5	16.4	7.6	48
47	97.4	90.3	82.2	72.7	61.8	49.2	37.6	26.5	16.4	7.6	47
46	96.5	89.6	81.6	72.3	61.5	49.1	37.5	26.5	16.4	7.6	46
45	95.5	88.8	81.0	71.9	61.3	49.0	37.5	26.4	16.4	7.6	45
44	94.6	88.0	80.4	71.5	61.0	48.9	37.5	26.4	16.4	7.6	44
43	93.6	87.2	79.7	71.0	60.7	48.7	37.4	26.4	16.4	7.6	43
42	92.5	86.3	79.0	70.5	60.4	48.6	37.4	26.4	16.4	7.6	42
41	91.4	85.4	78.3	70.0	60.1	48.4	37.3	26.4	16.4	7.6	41
40	90.3	84.5	77.6	69.4	59.8	48.3	37.3	26.4	16.4	7.6	40
39	89.2	83.5	76.8	68.9	59.4	48.1	37.2	26.4	16.4	7.6	39
38	88.0	82.5	76.0	68.3	59.0	47.9	37.1	26.4	16.4	7.6	38
37	86.8	81.5	75.2	67.6	58.6	47.7	37.1	26.4	16.4	7.6	37
36	85.5	80.4	74.3	67.0	58.2	47.5	37.0	26.4	16.4	7.6	36
35	84.2	79.3	73.4	66.3	57.7	47.2	36.9	26.3	16.4	7.6	35
34	82.9	78.1	72.4	65.6	57.2	47.0	36.8	26.3	16.4	7.6	34
33	81.5	76.9	71.4	64.8	56.7	46.7	36.6	26.3	16.4	7.6	33
32	80.1	75.7	70.4	64.0	56.1	46.4	36.5	26.3	16.4	7.6	32
31	78.6	74.4	69.3	63.2	55.5	46.0	36.4	26.2	16.4	7.6	31
30	77.1	73.1	68.2	62.3	54.9	45.7	36.2	26.2	16.4	7.6	30
29	75.5	71.7	67.0	61.3	54.3	45.3	36.0	26.1	16.4	7.6	29
28	73.9	70.2	65.8	60.4	53.5	44.9	35.8	26.1	16.4	7.6	28
27	72.2	68.8	64.5	59.3	52.8	44.4	35.6	26.0	16.4	7.6	27
26	70.5	67.2	63.2	58.3	52.0	43.9	35.4	25.9	16.4	7.6	26
25	68.7	65.6	61.8	57.1	51.2	43.4	35.1	25.9	16.4	7.6	25
24	66.9	64.0	60.4	55.9	50.3	42.8	34.8	25.8	16.3	7.6	24
23	65.0	62.3	58.9	54.7	49.3	42.2	34.4	25.6	16.3	7.6	23
22	63.0	60.5	57.3	53.4	48.3	41.5	34.1	25.5	16.3	7.6	22
21	61.0	58.6	55.7	52.0	47.2	40.8	33.7	25.3	16.3	7.6	21
20	58.9	56.7	54.0	50.6	46.1	40.0	33.2	25.2	16.3	7.6	20
19	56.8	54.8	52.2	49.0	44.8	39.1	32.7	24.9	16.2	7.6	19
18	54.6	52.7	50.4	47.4	43.5	38.2	32.1	24.7	16.2	7.6	18
17	52.3	50.6	48.5	45.8	42.2	37.2	31.5	24.4	16.1	7.6	17
16	49.9	48.4	46.5	44.0	40.7	36.1	30.8	24.1	16.1	7.6	16
15	47.5	46.1	44.4	42.1	39.1	35.0	30.0	23.7	16.0	7.6	15
14	45.0	43.7	42.2	40.2	37.5	33.7	29.1	23.2	15.8	7.6	14
13	42.4	41.3	39.9	38.1	35.7	32.3	28.2	22.7	15.7	7.6	13
12	39.7	38.8	37.6	36.0	33.9	30.8	27.1	22.1	15.5	7.6	12
11	36.9	36.1	35.1	33.7	31.9	29.2	25.9	21.4	15.3	7.5	11
10	34.1	33.4	32.5	31.4	29.8	27.5	24.6	20.6	14.9	7.5	10
9	31.1	30.6	29.8	28.9	27.6	25.6	23.1	19.6	14.5	7.5	9
8	28.1	27.6	27.0	26.3	25.2	23.6	21.5	18.5	14.0	7.5	8
7	25.0	24.6	24.1	23.5	22.7	21.4	19.7	17.2	13.4	7.4	7
6	21.7	21.4	21.1	20.6	20.0	19.0	17.7	15.7	12.6	7.2	6
5	18.4	18.2	17.9	17.6	17.1	16.4	15.5	14.0	11.5	7.0	5
4	14.9	14.8	14.6	14.4	14.1	13.6	13.0	11.9	10.2	6.6	4
3	11.4	11.3	11.2	11.1	10.9	10.6	10.2	9.6	8.5	6.0	3
2	7.7	7.7	7.6	7.6	7.5	7.4	7.2	6.9	6.3	4.9	2
1	3.9	3.9	3.9	3.9	3.9	3.8	3.8	3.7	3.5	3.1	1
0	0	0	0	0	0	0	0	0	0	0	0
overs left	0	1	2	3	4	5	6	7	8	9	overs left
	wickets lost										

Table of resource percentages remaining—over by over.

References

[1] S. R. Clarke, "Dynamic programming in one-day cricket — optimal scoring rates", *J. Oper. Res. Soc.*, **39** (1988), 331–337.

[2] S. R. Clarke and P. Allsopp, "Fair measures of performance: The World Cup of Cricket", *J. Oper. Res. Soc.*, **52** (2001), 471–479.

[3] S. R. Clarke and P. Allsopp, "Response to Comment on Clarke SR and Allsopp P (2001) 'Fair measures of performance: The World Cup of Cricket' ", *J. Oper. Res. Soc.*, to appear.

[4] F. Duckworth and T. Lewis, "A fair method for resetting the target in interrupted one-day cricket matches", in *Third Conference on Mathematics and Computers in Sport*, N. de Mestre (editor), Bond University, Queensland, Australia (1996), 51–68.

[5] F. C. Duckworth and A. J. Lewis, "A fair method of resetting the target in interrupted one-day cricket matches", *J. Oper. Res. Soc.*, **49** (1998), 220–227.

[6] F. Duckworth and T. Lewis, *Your Comprehensive Guide to the Duckworth/Lewis Method*, University of the West of England, Bristol (1999).

[7] F. C. Duckworth and A. J. Lewis, "Comment on Clarke SR and Allsopp P (2001) 'Fair measures of performance: The World Cup of Cricket' ", *J. Oper. Res. Soc.*, to appear.

[8] I. P. Preston and J. M. Thomas, "Rain rules for limited overs cricket and probabilities of victory", *J. Roy. Statist. Soc. Ser. D*, to appear.

Copyright

A BIOMECHANICAL POWER MODEL FOR WORLD-CLASS 400 METRE RUNNING

Chris Harman
Department of Mathematics and Computing
University of Southern Queensland
Toowoomba
Queensland 4350, Australia
harman@usq.edu.au

Abstract

A model is developed here for computation of the effects of power allocation strategies during the 400 metre track event. The model incorporates four mechanisms for consumption of energy: (i) speed/acceleration, (ii) heat generation in the body, (iii) air drag, and (iv) resistance to centrifugal effects in the curved portions of the track. This consumption is balanced in the model by aerobic and anaerobic energy supplies to the muscles. Considered in the anaerobic system are effects of the three important processes: (i) ATP cleavage, (ii) phosphocreatine cleavage (the phosphagen system), and (iii) glycolysis (the lactic acid system). Certain parameters in the model are estimated using averaged data from the 50 metre time splits of the eight finalists in each of the men's and women's 400 metre final in the 1999 Seville World Championships. The model is then tested against 50 metre time splits of Michael Johnson and Cathy Freeman. There is a good match with actual data. Using the developed model, variations to running strategies are then investigated. The model verifies the common wisdom that running at maximum available power for the entire race is a poor choice. Too much energy is wasted on increased heat, drag, and centrifugal effects. The model computes faster times for actual strategies commonly used by world-class runners. Alternative variations in strategy are then tested which clearly indicate possible improvements in race times.

1 Introduction

The task of accurately modelling the 400 metre track event is interesting. Unlike the case for the 100 m and 200 m events, runners in the 400 m do not exert maximum available power throughout, but use a strategy of conserving power in order to be able to maintain a reasonably high speed for as long as possible. In this paper, a model is developed which enables any power allocation strategy to be tested against available data. It is shown that running at maximum power throughout is a poor strategy. In particular the model tests the effects of accelerating for a short time at maximum power and then maintaining a lower than expected constant speed for the middle section of the race. It is demonstrated that world-class runners use strategies which can be approximated by such a constant speed phase and that potential improvements in race times are predicted by optimising the strategy.

The early models of running were based upon a force formulation using Newton's Law. The early work inspired by the biologist Hill [4] was developed by Keller [8, 9] who developed an analysis using optimal control. An excellent commentary on Keller's method is given by Pritchard [12]. One of the problems in Keller's approach is the assumption of a simple form for the athlete's energy profile. Keller's approach has been extended and applied more recently by Woodside [18], Tibshirani [13] and Murieka [11].

Ward-Smith [14, 15, 16] has developed an alternative energy/power formulation of a mathematical model for running. He includes developing physiological theories of energy supply and usage by athletes.

There are some crucial advantages in this approach. Current understandings on anaerobic and aerobic energy supplies to the muscles can be immediately incorporated into the model. Also, the effects of fatigue are an intrinsic component of the model, enabling more realistic matches with time-split data for sprint events.

In this paper a modification of the energy/power formulation of Ward-Smith is to be used. It is found that such an approach can be adapted to the difficult task of modelling curvature-induced retardation effects for accelerating motion on a curve. It is demonstrated that such curvature effects are important for the 400 m event where the runner initially accelerates on the curve and in fact runs less than half the total distance on the straights.

For their energy/power model, Ward-Smith and Radford [17], and Ward-Smith [16] have developed and refined values of relevant physical constants and parameters associated with sprinting the 100 m under maximum power. Most of these constants and parameters are also relevant to the 400 m event. However there are certain differences. In order to estimate variations in their values, 50 m time-split data are to be used from the 400 m finals of the 1999 Seville World Championships (see Ferro *et al.* [3]). In each of the men's and women's final, averaged data is used from all eight competitors. A more detailed examination is then made of individual data for the respective winners, Michael Johnson (world record) and Cathy Freeman.

Ward-Smith [16] tested a strategy for inserting a constant-speed phase into the 100 m sprint model. It was found that there was nothing to be gained and that running at maximum available power was optimal for this event. In this paper, it is shown that insertion of a constant speed phase not only approximates the well-known strategy used by world-class athletes in the 400 m, but also reduces times significantly from those expected for maximum available power exertion. Optimising the constant speed phase indicates potential time improvements for Johnson and Freeman.

2 Energy and power formulation

Following Ward-Smith [14, 15, 16], chemical energy from the muscles is transformed into two main forms, heat and external work. Chemical energy C at the muscles is produced by fairly well understood anaerobic and aerobic mechanisms, and is transformed into heat production H in the body and external work W done on the centre of mass of the runner. In this paper, all terms used will be assumed to be normalised per kilogram of body mass of the runner and all quantities will be measured in excess of the body's resting metabolic rate. Conservation of energy/power then gives the basic power equation

$$\frac{dC}{dt} = \frac{dH}{dt} + \frac{dW}{dt}, \tag{1}$$

where t is the running time from the start.

Some assumptions are required. External work is done in providing kinetic energy to the centre of mass. It is also done against air drag. It will be assumed here that there is no appreciable wind. Raising the runner's centre of mass from the starting position is a factor, but is ignored here. It is generally acknowledged that vertically upward losses of kinetic energy during a stride are recovered during the downwards phase. Hence, it is to be assumed that vertical motion of the centre of mass during a typical stride can be ignored. The work done in retarding centrifugal effects due to curvature is postponed until the next section. So, for the moment, straight line motion is assumed.

If v represents the runner's horizontal speed, kinetic energy (per kilogram) is $\frac{1}{2}v^2$ and the force acting due to air drag is $D = \frac{1}{2}\rho C_D S v^2$ where ρ is air density, C_D is the drag coefficient for a runner, and S is the effective frontal area of the runner. Hence,

$$\begin{aligned}\frac{dW}{dt} &= \text{drag power} + \text{kinetic power} \\ &= Dv + \frac{d}{dt}0.5v^2 \\ &= Kv^3 + v\frac{dv}{dt},\end{aligned} \tag{2}$$

where $K = \rho S C_D/2m$, m the runner's mass.

The rate of transfer of mechanical energy into body heat is well modelled by

$$\frac{dH}{dt} = Av, \tag{3}$$

where A is a constant, see Ward-Smith [14]. If we represent the power contributions from anaerobic and aerobic energy sources as dC_{an}/dt and dC_{aer}/dt, respectively, equation (1) becomes

$$\frac{dC_{an}}{dt} + \frac{dC_{aer}}{dt} = Av + Kv^3 + v\frac{dv}{dt}, \tag{4}$$

where A is the heat parameter and K is the drag parameter.

It remains to specify the aerobic and anaerobic energy supply mechanisms. It is well known that the rate of aerobic energy release, the power P_{aer}, is modelled by the equation

$$P_{aer}(t) = \frac{dC_{aer}}{dt} = R(1 - e^{-\lambda t}), \tag{5}$$

where R is the limiting aerobic power and λ is a parameter determining the rate of aerobic energy release, see Ward-Smith [14, 16]. The form of equation (5) indicates that the aerobic energy supply is low in the early stages of exercise and builds up towards the maximum R as time progresses. The principal energy source in the early stages is anaerobic and is provided in the muscles by three fundamental mechanisms: (i) the rapid utilisation of adenosine triphosphate (ATP), (ii) phosphocreatine cleavage (the phosphagen system), and (iii) glycolysis (the lactic acid system). Ward-Smith [15, 16] has successfully modelled each of these mechanisms with the product of exponential functions which represent the rates of rise and decline of each of the energy supply processes. In fact, the power equations determined by Ward-Smith are

$$P_n(t) = (P_{\max})_n r_n^{1/r_n} \left(\frac{1 + r_n}{r_n}\right)^{(1+r_n)/r_n} (1 - e^{-\psi_n t})e^{-\lambda_n t}, \tag{6}$$

where the subscript n refers to each of the three anaerobic mechanisms, with $n = 1$ for the ATP system, with $n = 2$ for the phosphagen system, and with $n = 3$ for the glycolysis system. In (6), $P_n(t)$ is the corresponding available power at time t and $(P_{\max})_n$ is the maximum attainable power by mechanism n. The parameters ψ_n and λ_n determine each power growth and decline rates respectively and r_n is defined by

$$r_n = \frac{\psi_n}{\lambda_n}.$$

Hence the total anaerobic power P_{an} is given by

$$P_{an}(t) = P_1(t) + P_2(t) + P_3(t), \tag{7}$$

and the total available power $P(t)$ is given by

$$P(t) = P_{an}(t) + P_{aer}(t) = \frac{dC_{an}}{dt} + \frac{dC_{aer}}{dt} = P_1(t) + P_2(t) + P_3(t) + R(1 - e^{-\lambda t}), \tag{8}$$

where P_1, P_2, and P_3 are given by equation (6).

Figure 1 shows a typical representation of the three anaerobic power mechanisms and the somewhat slower aerobic power supply for the duration of the men's 400 metre event. Note that the initial ATP supply is largely depleted within the first two or three seconds and in fact all three anaerobic mechanisms have peaked within ten seconds. The total power available to the runner is also displayed.

Combining equations (4) and (8) results in the following nonlinear differential equation in v:

$$P_{an}(t) + P_{aer}(t) = Av + Kv^3 + v\frac{dv}{dt}. \tag{9}$$

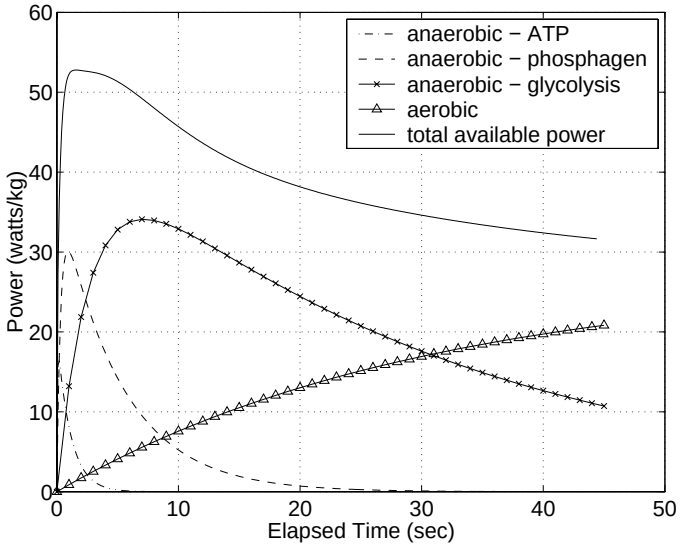


Figure 1: Typical available power in the men's 400 m event.

This equation forms the basis of the model and can be solved numerically using an initial value solver such as a Runge–Kutta scheme.

It should also be noted that the amount of aerobic energy used in time t can easily be found by integrating equation (5) to get

$$C_{aer} = Rt - \frac{R}{\lambda}(1 - e^{-\lambda t}), \tag{10}$$

and, similarly, by integrating equation (6) the amount of anaerobic energy used in time t is

$$C_{an} = \sum_{n=1,2,3} (C_{an})_n,$$

where

$$(C_{an})_n = (P_{\max})_n r_n^{1/r_n} \left(\frac{1+r_n}{r_n}\right)^{(1+r_n)/r_n} \left(\frac{\psi_n}{\lambda_n(\lambda_n+\psi_n)} - e^{-\lambda_n t}\frac{\psi_n + \lambda_n(1-e^{-\psi_n t})}{\lambda_n(\lambda_n+\psi_n)}\right). \tag{11}$$

3 Track curvature effects

The Seville 1999 World Championship 400 m track consisted of two 80 m straights and two semicircular sections with the inside track circumference being 120 m. Eight lanes of approximate width 1.3 m are used. The inside lane runner starts and ends at the end of the straight, with the other seven runners starting at set positions around the semicircle such that they all finish at the end of the straight after running exactly 400 m. For example the runner in lane 8 starts from about 57 m ahead around the semicircle. The IAAF standard for an outdoor stadium does allow some variations to this layout and so not all 400 m tracks are exactly to this specification. For the purposes of this paper, the Seville track measures are to be used.

When running on a curve, there is a resulting centrifugal effect which must be counteracted in order to maintain motion along the curve. Obviously runners on the inside lanes are more affected. This phenomenon has been studied by Greene [5], whose model has been applied to baserunning models in softball and baseball by Harman [6], and also applied to specific case studies of 200 m and 400 m results data, Harman [7]. Greene's model takes account of the runner's speed (assumed constant), foot contact

time, stride length, stride time, and ballistic air time. The model predicts the retarding effect on speed, of running at various constant speeds, on circles of defined radii. The results are supported by experimental data. More complex models have subsequently been formulated, see for example Behncke [1] and Mureika [11], but with some difficulty and uncertainty about parameter values. In fact, Behncke commented that "the simplicity of Greene's result and the apparent ease of its derivation make it ... an ideal candidate for the analysis of the track and field situation."

Greene found that a runner, with constant straight line speed v_0, has that speed reduced to v on a circular curve of radius r, where v can be determined as the solution of the polynomial equation

$$v^6 + (rg)^2 v^2 - (rgv_0)^2 = 0, \tag{12}$$

where g is the acceleration due to gravity.

There is a problem in applying Greene's speed modification formula since it assumes constant speed. During the initial stages of the 400 m event, the runner accelerates rapidly from rest while on the curve. Hence an additional formulation is needed which allows Greene's modification to be applied to accelerating motion. An innovation here is to formulate the curvature retardation effect in terms of the power required to overcome the centrifugal effect. It will then be seen that this will enable Greene's formula, given by equation (12), to be used during accelerating motion.

Denote by $P_{cen}(v)$ the additional power (per kilogram) required to overcome centrifugal effects when running on a curve of radius r at speed v. Combining with heat, drag, and kinetic effects, and equating to the available power supply, equation (9) must then be modified to

$$P_{an}(t) + P_{aer}(t) = Av + Kv^3 + v\frac{dv}{dt} + P_{cen}. \tag{13}$$

Without the curve, and using the same power supply, the runner's speed would be v_0 and so the power equation is just equation (9), in this case

$$P_{an}(t) + P_{aer}(t) = Av_0 + Kv_0^3 + v_0\frac{dv_0}{dt}, \tag{14}$$

where, from equation (12) it follows that

$$v_0 = \sqrt{\frac{v^6}{r^2g^2} + v^2}. \tag{15}$$

From equations (13) and (14), it follows that

$$P_{cen} = A(v_0 - v) + K(v_0^3 - v^3) + v_0\frac{dv_0}{dt} - v\frac{dv}{dt}. \tag{16}$$

Differentiating equation (15),

$$\frac{dv_0}{dt} = \frac{rg(3v^4 + r^2g^2)}{\sqrt{v^4 + r^2g^2}}\frac{dv}{dt}. \tag{17}$$

Equation (16), after some simplification, becomes

$$P_{cen} = Av\left(\sqrt{w+1} - 1\right) + Kv^3\left(\sqrt{(w+1)^3} - 1\right) + v\frac{dv}{dt}(3w), \tag{18}$$

where

$$w = \frac{v^4}{r^2g^2}, \tag{19}$$

and P_{cen} is measured in watts per kilogram. Substituting back into equation (13) and simplifying, gives the power balance equation

$$P_{an}(t) + P_{aer}(t) = Av\sqrt{1+w} + Kv^3\sqrt{(1+w)^3} + v\frac{dv}{dt}\left(1 + 3w\right), \tag{20}$$

where w is zero on the straight and is given by equation (19) on the curve. This is a very general power balance equation applying to any running situation on an arbitrary curve or straight line. The actual outcomes of the retarding effect of curvature in the 400 m event are quite significant and will be discussed in the concluding section.

4 The model

Equation (20) can be rewritten as

$$\frac{dv}{dt} = \frac{1}{v(1+3w)}\left(P_{an} + P_{aer} - Av\sqrt{1+w} - Kv^3\sqrt{(1+w)^3}\right), \tag{21}$$

where $w = v^4/(rg)^2$. Using appropriate values of the parameters and constants (normalised per kilogram), this was then solved for v, x at times up to $x(t) = 400$ using a fourth order Runge–Kutta scheme with step size $\Delta t = 0.0025$. The initial conditions are: (i) $t_0 = t_r$, the runner's reaction time in seconds, (ii) $v_0 = 0.00001\,\mathrm{ms}^{-1}$, to bypass the singular value at $v = 0$, and (iii) $x_0 = -0.17$ metres, the average position in metres of a runner's centre of mass at the start, see Ward-Smith [16]. Such a solution represents motion under maximum available power supplied by P_{aer} and P_{an}.

As mentioned in the Introduction, Ward-Smith [16] tested a strategy for inserting a lower constant-speed phase into the 100 m sprint model and found that there was nothing to be gained. Running at maximum available power is optimal for this event. However, for the 400 m event things are very different. Curvature effects are significant and increase with speed. In addition, as Figure 1 illustrates, the anaerobic power supplies are well in decline after ten seconds with the aerobic supply becoming more prominent.

The strategy to be used in this model is to accelerate at maximum available power until a chosen speed v_c is reached at time t_1. The athlete then continues at this constant speed until time t_2, the point where the total energy used is equal to the total available energy given by equations (10) and (11). The athlete's speed then declines as the race is finished at reducing maximum available power. During the constant speed phase $v = v_c$, the amount of energy used can easily be calculated by accurate numerical integration of the appropriate power function. From equation (20), and since $dv_c/dt = 0$, it follows that the energy used during this phase, denoted by E_c, is given by

$$E_c(t) = \int_{t_1}^{t} \left(Av_c\sqrt{1+w} + Kv_c^3\sqrt{(1+w)^3}\right) dt,$$

where $w = v_c^4/(rg)^2$.

It should be noted that if the value of v_c is chosen sufficiently small, the race can finish at 400 m before the point where the total energy used matches the total available energy. However, it is found that such a low value of v_c is always a poor strategy, resulting in slower times for a 400 m race.

It remains to estimate the values of the heat coefficient A, the drag constant K, the anaerobic parameters $(P_{\max})_{1,2,3}$, $\lambda_{1,2,3}$ and $\psi_{1,2,3}$, and the aerobic parameters R and λ. The values of A and K are well known and Ward-Smith has used world-class time splits for the 100 m event to estimate the anaerobic and aerobic parameters, see [16, 17]. Table 1 summarises the values used by Ward-Smith for the men's 100 m event and Table 2 gives the corresponding values for the women's event.

$A = 3.96$ J kg^{-1} m^{-1}	$K = 0.0029$ m^{-1}	
$(P_{\max})_1 = 16.6$ W kg^{-1}	$\lambda_1 = 0.9$ s^{-1}	$\psi_1 = 20$ s^{-1}
$(P_{\max})_2 = 30.1$ W kg^{-1}	$\lambda_2 = 0.20$ s^{-1}	$\psi_2 = 3.0$ s^{-1}
$(P_{\max})_3 = 34.1$ W kg^{-1}	$\lambda_3 = 0.033$ s^{-1}	$\psi_3 = 0.34$ s^{-1}
$R = 25$ W kg^{-1}	$\lambda = 0.033$ s^{-1}	

Table 1: Ward-Smith's constants and parameters for the men's 100 m sprint.

$A = 3.94$ J kg^{-1} m^{-1}	$K = 0.0031$ m^{-1}	
$(P_{\max})_1 = 15.8$ W kg^{-1}	$\lambda_1 = 0.9$ s^{-1}	$\psi_1 = 20$ s^{-1}
$(P_{\max})_2 = 21.9$ W kg^{-1}	$\lambda_2 = 0.21$ s^{-1}	$\psi_2 = 3.0$ s^{-1}
$(P_{\max})_3 = 32.4$ W kg^{-1}	$\lambda_3 = 0.041$ s^{-1}	$\psi_3 = 0.44$ s^{-1}
$R = 21.5$ W kg^{-1}	$\lambda = 0.041$ s^{-1}	

Table 2: Ward-Smith's constants and parameters for the women's 100 m sprint.

It is known that the build up of aerobic power is proportional to the depletion of the glycogen store in anaerobic glycolysis. As a consequence, Ward-Smith [15] argues that there are good reasons for assuming that λ can be approximated by λ_3. He supports this notion using theoretical considerations and experimental evidence. Ward-Smith used parameter estimation of λ_3 and thus deduced the value of λ.

Ward-Smith's estimated constants and parameters listed in Tables 1 and 2 are obtained from 100 m time-split data where the anaerobic supply mechanism is predominant. For the 400 m event, it is to be assumed in this paper that world-class runners have similar characteristic constants and parameter values, with the notable exception of the value of the aerobic parameter R, which represents the limiting aerobic power. In the 100 m event, aerobic energy supply has barely commenced (see Figure 1) and so any estimation of R from associated time-split data must be subject to question. Estimates of R from longer events have been made, see for example Ward-Smith [14], where a men's value of $R = 23.5$ watts per kilogram was obtained, averaged over a range of event distances from 100 to 10000 metres. However, this estimate did not include the effects of track curvature on energy, which are significant in the 400 m event. The associated aerobic parameter λ will be retained for the 400 m model, since as mentioned above, there are compelling reasons for its approximation to the anaerobic parameter λ_3.

5 Results for averaged time splits

The model for the 400 m event will now be matched with averaged 50 m time splits of the eight finalists from the Seville final. First the men's event. The parameter values from Table 1 are initially used, including a starting value of $R = 25$. The running strategy to be used is assumed to include the constant speed phase, so that maximum available power is used until speed v_c is reached at time t_1. This speed is maintained until time t_2, the point at which total energy used equals total energy available. Using maximum available power, the speed then declines until 400 m is reached. It is assumed that the averaged runner runs in a lane with averaged curvature (i.e. halfway between lanes 4 and 5). The average reaction time was 0.18 seconds.

Using starting values $(R, v_c) = (25, 9.5)$, the Nelder–Mead Simplex Method was then used to search for the optimum value of (R, v_c). The optimal value is to be taken in the sense of minimum value of the sum of absolute differences between the computed model results and the averaged time-split data at the 50 m intervals. The results are summarised in Table 3, with optimal values $R = 26.71$, $v_c = 9.78$, sum of absolute differences $(SAD) = 0.82$, and race time 44.45 seconds which is identical with the actual recorded time.

It should be noted at this stage that averaging the time splits at each 50 m interval mark is bound to be rough. Each runner is in a defined lane with individual curvature characteristics. Also, each runner is using an individual race plan which need not approximate a constant speed phase. Mindful of these reservations, the optimal values produce strong support for the model.

It is interesting to compute the race time for the case where the runner's strategy is to use maximum available power for the entire race. In this case there is no constant speed phase. The model computes a time of 44.59 seconds, a significantly slower time and thus a poor strategy.

The above optimisation is an attempt to match the model closely with the actual time splits used in the race. It is possible to check alternative strategies by using the model to simulate the corresponding times. Using the value $R = 26.71$ from above, the Nelder–Mead Simplex Method was again used to

Distance (m)	0	50	100	150	200	250	300	350	400
Time splits (s)	0	6.11	11.05	16.05	21.23	26.59	32.12	37.99	44.45
Computed times (s)	0	5.94	11.05	16.16	21.28	26.39	31.88	38.05	44.45
Differences (Δt)	0	0.17	0.00	−0.11	−0.05	0.20	0.24	−0.06	0.00

Table 3: The best constant speed phase fit for the averaged time splits of the eight finalists in the men's 400 metre final in Seville 1999. Optimal aerobic $R = 26.71\ \mathrm{W\,kg^{-1}}$, optimal constant speed $v_c = 9.78\,\mathrm{ms^{-1}}$; measure of fit, $SAD = 0.82$ seconds.

optimise the value of v_c to minimise the race time for this averaged athlete. It was found that $v_c = 9.26\,\mathrm{ms^{-1}}$ yielded a minimum race time of 44.36 seconds.

The analysis was repeated for the women's 400 m event which yielded an optimal value of $(R, v_c) = (24.12, 8.51)$ for the best fit to the data with a race time of 50.24 seconds. Optimising v_c for the fastest race time gave a value of $v_c = 8.19$ for a time of 50.15 seconds. The resulting times are summarised for the men's and the women's averaged results in Table 4.

	Max. avail. power	Best fit to data	Fastest time fit
Averaged men's times (s)	44.59	44.45	44.36
Averaged women's times (s)	50.47	50.24	50.15

Table 4: Time comparison using (i) maximum available power, (ii) best R and v_c to fit the given data, and (iii) race time for optimal choice of v_c.

6 Models for Michael Johnson and Cathy Freeman

In the 1999 Seville World Championships, Michael Johnson, running in lane 5, set a world record of 43.18 seconds for the 400 metre event. His reaction time was 0.15 seconds. Based on the above averaged results, starting values $(R, v_c) = (26.71, 9.78)$ were then used to search for optimal results for Johnson's data fit to the model, again using the Nelder–Mead Simplex Method and the values of the constants and parameters from Table 1. The optimal values were $(R, v_c) = (29.15, 9.76)$. The corresponding computed race time was 43.18 seconds with the sum of absolute differences between the computed model results and the averaged 50 m time-split data $SAD = 0.57$. The results for Johnson are summarised in Table 5.

Distance (m)	0	50	100	150	200	250	300	350	400
Time splits (s)	0	6.14	11.10	16.10	21.22	26.42	31.66	37.18	43.18
Computed times (s)	0	5.92	11.05	16.17	21.29	26.42	31.54	37.14	43.18
Differences (Δt)	0	0.22	0.05	−0.07	−0.08	−0.00	0.12	0.04	−0.00

Table 5: The best constant speed phase fit for Michael Johnson's 50 m time splits in the men's 400 metre final in Seville 1999. Optimal aerobic $R = 29.15\ \mathrm{W\,kg^{-1}}$, optimal constant speed $v_c = 9.76\,\mathrm{ms^{-1}}$; measure of fit, $SAD = 0.57$ seconds.

Figure 2 illustrates the closeness of the fit of the computed model to Johnson's actual 50 m time-split data. Hence an athlete having all the attributes of the estimated constants and parameters used here, and running using the defined strategy of a constant speed phase $v_c = 9.76\,\mathrm{ms^{-1}}$, would closely match the Johnson time-splits and would finish with the same time of 43.18 seconds.

Running under maximum available power, the model computes a race time of 43.30 second which is significantly slower.

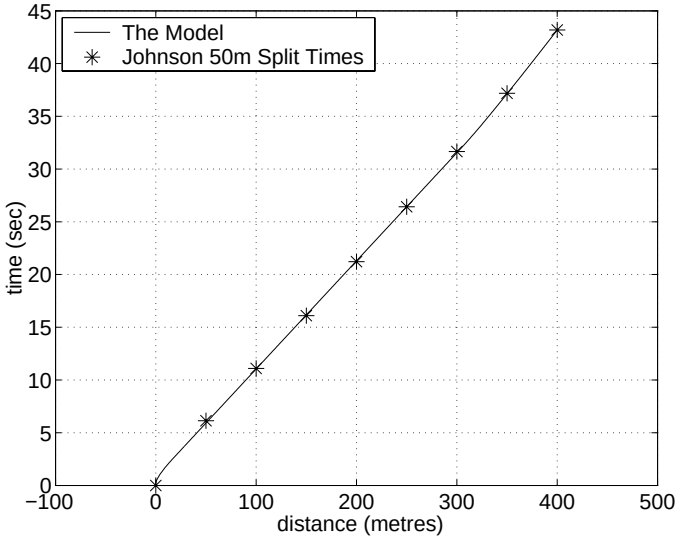


Figure 2: Michael Johnson's Seville 1999 50 m time splits and the model fit using constant speed phase $v_c = 9.78\,\mathrm{ms}^{-1}$.

	Max. avail. power	Best fit to data	fastest time fit
Johnson's times (s)	43.30	43.18	43.14
Freeman's times (s)	49.88	49.67	49.59

Table 6: Time comparison using (i) maximum available power, (ii) best R and v_c to fit the given data, and (iii) race time for optimal choice of v_c.

Using the estimated value of $R = 29.15$, the Nelder–Mead Simplex Method was then used again to find the value of v_c which gives minimum race time for the computed model. In fact the optimal value was $v_c = 9.52$ with corresponding race time 43.14 seconds. This represents a small but not insignificant improvement in time.

Figure 3 shows the corresponding three velocity profiles which are computed by the model based upon Johnson's estimated parameters. The maximum available power profile clearly shows the effect of the change from curved to straight track and vice versa.

The Seville women's 400 m event was won by Cathy Freeman in a time of 49.67 seconds. She ran in lane 5 and had a reaction time of 0.193 seconds. A similar model analysis was carried out based upon the 50 m time splits. The optimal fit to the data corresponded to $(R, v_c) = (24.77, 8.59)$ with a computed time of 49.67 seconds. A fastest time of 49.59 seconds corresponded to an optimal value of the constant speed phase given by $v_c = 8.28\,\mathrm{ms}^{-1}$. The time corresponding to usage of maximum available power was 49.88 seconds.

Figure 4 illustrates the fit of the computed model to Freeman's actual 50 m time-split data. With an SAD of 1.2, the fit is not quite as good as Johnson's, but an athlete with the assumed characteristics of the model, and running using the defined strategy of a constant speed phase $v_c = 8.59\,\mathrm{ms}^{-1}$, would closely match Cathy Freeman's time splits and would finish with the same time of 49.67 seconds.

Figure 5 shows the corresponding three velocity profiles computed by the model based upon Freeman's estimated parameters.

A summary of the times for these model computations based upon Johnson's and Freeman's profiles is given in Table 6.

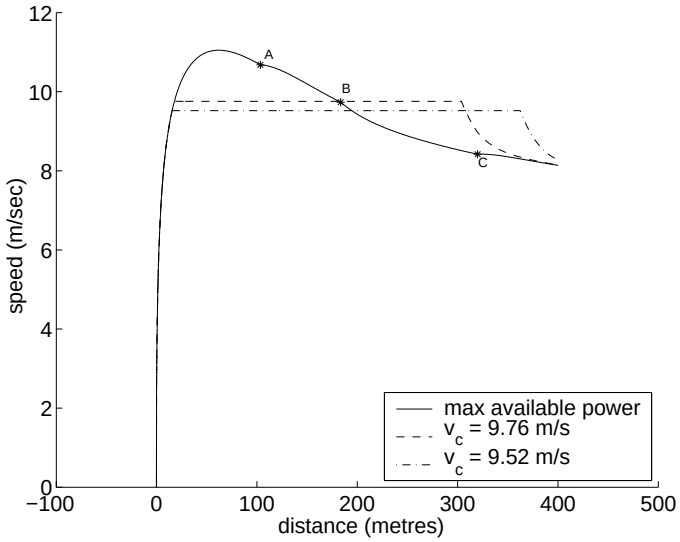


Figure 3: Computed model profiles based on Michael Johnson; A marks the end of the first curve and B, C mark the start and end of the second curve in the track.

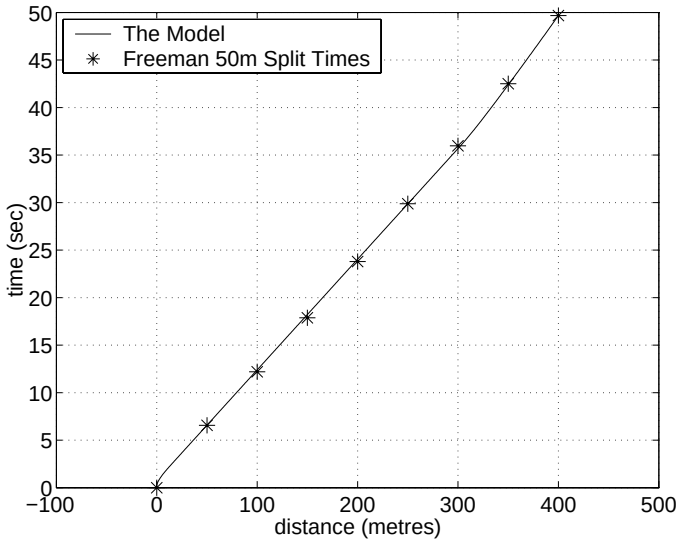


Figure 4: Cathy Freeman's Seville 1999 50 m time splits and the model fit using constant speed phase $v_c = 8.59\,\mathrm{ms}^{-1}$.

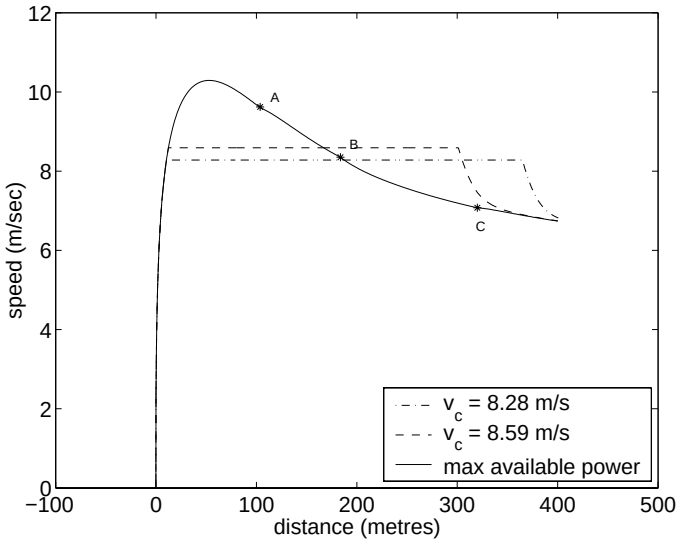


Figure 5: Computed model profiles based on Cathy Freeman; A marks the end of the first curve and B, C mark the start and end of the second curve in the track.

7 Discussion

Values of the energy supply parameters used in the model obviously require further investigation. The assumption of the applicability of the majority of parameters from Tables 1 and 2 to the 400 m is only a start. It has been argued above that it is reasonable to so do. However, it is quite possible that this results in a compensating biased estimate of the aerobic R. Additional estimates of the predominant anaerobic glycolysis parameters would be very useful. Ideally these would be for averaged results and for individual athletes. This would require more physiological measurement data and more track time splits.

Alternative approaches to the energy supply mechanism might be possible. For example, Morton [10] has developed a model incorporating three important parameters, anaerobic work capacity, critical power and maximal instantaneous power, and di Prampero [2] has a model for energy/power consumption, but this applies only up to running speeds of about six metres per second.

One of the difficulties faced is the lack of time-split data. At the present time, the only available data are the 50 m time splits. There is no readily available information on the most interesting phase of the race, the first 50 metres, where the predominant acceleration occurs. Film coverage of finals might yield useful information. In the above model fits to the data, it is interesting to note that the fit is less accurate at the 50 metre mark. Runners seem to take longer to reach this mark than predicted by the model. It has been assumed that runners accelerate at maximum available power until the constant speed phase $v = v_c$ is reached. Perhaps this is not quite so. More data is needed.

Curvature has a significant retarding effect on expected race times for the 400 m event. Computations were carried out using the model, with and without the curvature effect. For Johnson's profile, there was a 1.7% difference and for Freeman there was a smaller difference of 1.1% in race time. This compares with the model's computed air drag effects of 6.6% and 6.0% for Johnson and Freeman respectively.

Of course, 400 m runners do not run exactly with a constant speed phase as assumed in the model. However the model's match with time splits is sufficiently good to enable alternative running strategies to be tested for effectiveness. In this paper, it has been shown that strategies with an optimally selected slower constant speed phase lead to improvements in race time. Any alternative strategy could be tested with the model developed here.

References

[1] H. Behncke, "Small effects in running", *J. Appl. Biomech.*, **10** (1994), 270–290.

[2] P. E. di Prampero, "The energy cost of human locomotion on land and water", *Int. J. Sports Med.*, **7** (1986), 55–72.

[3] A. Ferro, A. Rivera, I. Pagola, M. Ferreruela, A. Martin and V. Rocandio, "Biomechanical analysis of the 7th world championships in athletics Seville 1999", *New Studies In Athletics*, **16** (2001), 25–60.

[4] K. Furusawa, A. V. Hill and J. L. Parkinson, "The dynamics of sprint running", *Proc. Roy. Soc. London Ser. B*, **102** (1927), 29–42.

[5] P. R. Greene, "Running on flat turns: experiments, theory, and applications", *J. Biomech. Eng.*, **107** (1985), 96–103.

[6] C. J. Harman, "Who's on first! What? What's on second! And how What got there on an optimal baserunning path", in *Fourth Conference on Mathematics and Computers in Sport*, N. de Mestre and K. Kumar (editors), Bond University, Queensland, Australia (1998), 217–223.

[7] C. J. Harman, "Curvature effects in running the 200 and 400 metre sprints", in *Proceedings of the Fifth Australian Conference on Mathematics and Computers in Sport*, G. Cohen and T. Langtry (editors), University of Technology, Sydney (2000), 118–123.

[8] J. B. Keller, "A theory of competitive running", *Physics Today*, **26** (1973), 43–47.

[9] J. B. Keller, "Optimal velocity in a race", *Amer. Math. Monthly*, **81** (1974), 474–480.

[10] R. H. Morton, "A 3-parameter critical power model", *Ergonomics*, **39** (1996), 611–619.

[11] J. R. Mureika, "A simple model for predicting sprint race times accounting for energy loss on the curve", *Can. J. Physics*, **75** (1997), 837–851.

[12] W. G. Pritchard, "Mathematical models of running", *SIAM Rev.*, **35** (1993), 359–379.

[13] R. Tibshirani, "Who is the fastest man in the world?", *Amer. Statist.*, **51** (1997), 106–111.

[14] A. J. Ward-Smith, "A mathematical theory of running, based on the first law of thermodynamics, and its application to the performance of world-class athletes", *J. Biomech.*, **18** (1985), 337–349.

[15] A. J. Ward-Smith, "The kinetics of anaerobic metabolism following the initiation of high-intensity exercise", *Math. Biosciences*, **159** (1999), 33–45.

[16] A. J. Ward-Smith, "Energy conversion strategies during 100 m sprinting", *J. Sports Sci.*, **19** (2001), 701–710.

[17] A. J. Ward-Smith and P. F. Radford, "Investigation of the kinetics of anaerobic metabolism by analysis of the performance of elite sprinters", *J. Biomech.*, **33** (2000), 997–1004.

[18] W. Woodside, "The optimal strategy for running a race (a mathematical model for world records from 50 m to 275 km)", *Math. Comput. Modeling*, **15** (1991), 1–12.

ANALYSIS OF A TWISTING DIVE

K. L. Hogarth* and D. M. Stump
Mathematics Department
University of Queensland
Queensland 4072, Australia
kateh@maths.uq.edu.au

M. R. Yeadon
Sports Biomechanics Laboratory
Loughborough University
Ashby Road
Loughborough LE11 3TU, UK
m.r.yeadon@lboro.ac.uk

Abstract

An 11-segment mathematical model of a diver is presented where the human body is described as a series of nested levels. The diver's orientation is defined in terms of Euler parameters instead of Euler angles. From this model, equations of motion are formulated using a Hamiltonian approach where the resultant equations describe the diver's displacement, orientation and momentum. When compared with known solutions of a rigid body, the simulation model is highly accurate. Comparisons with experimental data are available.

1 Introduction

Although the physical laws governing motion have been known for centuries, research aimed at making mathematical predictions of a twisting dive has been minimal. Traditionally, theoretical studies of twisting somersaults have described movement in terms of a mathematical simulation model. To date, all diving models have derived their equations of motion from angular momentum formulas calculated about the diver's centre of mass.

One of the earliest models was introduced by Dapena [1] who employed a 15-segment model where each segment was represented by a thin rod. Using the fact that a diver's angular momentum is conserved, he derived angular velocity equations for the upper trunk. Dapena found that his simulation model was reasonably accurate for 0.6–0.8 seconds. Errors resulted from the fact that each segment had no inertia about its longitudinal axis.

As an extension of his cinematographical study [7], Van Gheluwe [6] combined a mathematical simulation model with experimental data. Using a six-segment model, Van Gheluwe derived differential equations for angular velocity of the trunk and used numerical techniques to find the time evolution of the whole body angular velocity. From this analysis, Van Gheluwe's model gave reasonable estimates of a backward twisting somersault. However, errors became apparent toward the end of the motion.

Perhaps the most detailed analysis of twisting dives can be found in two series of papers published by Yeadon [9–17]. In his 1990 series of papers, Yeadon developed an 11-segment geometric inertia model. Combining this inertia model with whole body angular momentum equations, he calculated time histories of somersault, tilt and twist angles and solved them analytically.

Yeadon evaluated his model with film of twisting somersaults and found a good agreement with simulation estimates. However, he noted some differences between simulated and filmed performances in the second half of the simulations. This model also showed a close match with mean angular momentum

*This paper was researched at the University of Queensland, Australia, and at the University of Loughborough, UK. The author is grateful for the hospitality and guidance given by the members of the Sports Biomechanics Laboratory at Loughborough University.

estimates, but displayed higher discrepancies for momentum time histories [10]. These discrepancies were attributed to measurement and inertia value errors.

One of the most recent additions to twist analysis has been computer animation. Wooten and Hodgins [8] simulated the motion of a platform diver using forward dynamics with a control system. The equations of motion were generated from a commercially available computer package, SD/FAST. This package uses Kane's method and was coded for a 15-segment rigid body model. While the model did provide a good simulation of simple somersaulting dives, discrepancies were noted for simple twisting dives. These discrepancies were heightened when dives of increased complexity were modelled.

In each of the aforementioned studies, Euler angles were used to describe the diver's orientation. Describing orientation in terms of a triad of Euler angles is tempting as there are the same number of parameters as degrees of freedom. However, the problem with using this description is that no three-parameter set can be both global and non-singular [5]. To date, researchers have had to proceed with extreme caution to avoid singularities. Not only does this mean writing lengthy code, but it also means that results must be interpreted with great care.

The model presented in this paper represents a novel approach to modelling twisting somersaults. Instead of using the diver's centre of mass to describe flight, it is assumed that the diver's reference frame lies at a predetermined site within his body. Thus, the motion of a physical point is generated rather than the imaginary centre of mass. The use of a physical point makes video analysis and physical interpretation easier. It will also ensure that applications are relevant for coaches.

The diver's orientation will be described by Euler parameters instead of Euler angles. The most powerful advantage of using a four-parameter set is that it does not contain any geometric singularities. Moreover, the resultant equations of motion are universally nonsingular.

2 Equations of motion

To solve the problem of modelling a diver, a nested coordinate system has been employed. This approach was chosen as it resembles the human body; distal limbs are most naturally described in terms of proximal limbs, and proximal limbs are most naturally described in terms of the torso.

By defining a system of rigid bodies in terms of a reference frame and two nested levels (see Figure 1), the diver can be described as a system of 11 segments. These segments represent the diver's head and chest, middle torso, lower torso, upper arms, lower arms, upper legs and lower legs.

For this simulation model, each nested level segment is prescribed with its own mass, inertia, angular velocity, linear momentum, orientation and displacement. For simplicity, the coordinate axes of each level have their origin at the segment's centre of mass.

It is easier to calculate Euler angles than Euler parameters from film data. Given the Euler angles of each coordinate system, the Euler parameters, b_t, are defined as

$$
\begin{aligned}
b_0 &= \cos\frac{\phi}{2}\cos\frac{\theta}{2}\cos\frac{\psi}{2} - \sin\frac{\phi}{2}\sin\frac{\theta}{2}\sin\frac{\psi}{2},\\
b_1 &= \sin\frac{\phi}{2}\cos\frac{\theta}{2}\cos\frac{\psi}{2} + \cos\frac{\phi}{2}\sin\frac{\theta}{2}\sin\frac{\psi}{2},\\
b_2 &= \cos\frac{\phi}{2}\sin\frac{\theta}{2}\cos\frac{\psi}{2} - \sin\frac{\phi}{2}\cos\frac{\theta}{2}\sin\frac{\psi}{2},\\
b_3 &= \cos\frac{\phi}{2}\cos\frac{\theta}{2}\sin\frac{\psi}{2} + \sin\frac{\phi}{2}\sin\frac{\theta}{2}\cos\frac{\psi}{2}.
\end{aligned}
$$

Here, we have chosen the 1–2–3 set of Euler angles as they represent the diver's somersault, tilt and twist. However, the definition of Euler parameters may be parallelled for all 12 sets of Euler angle sequences [3].

Another powerful advantage of using Euler parameters is that the relationship between their deriva-

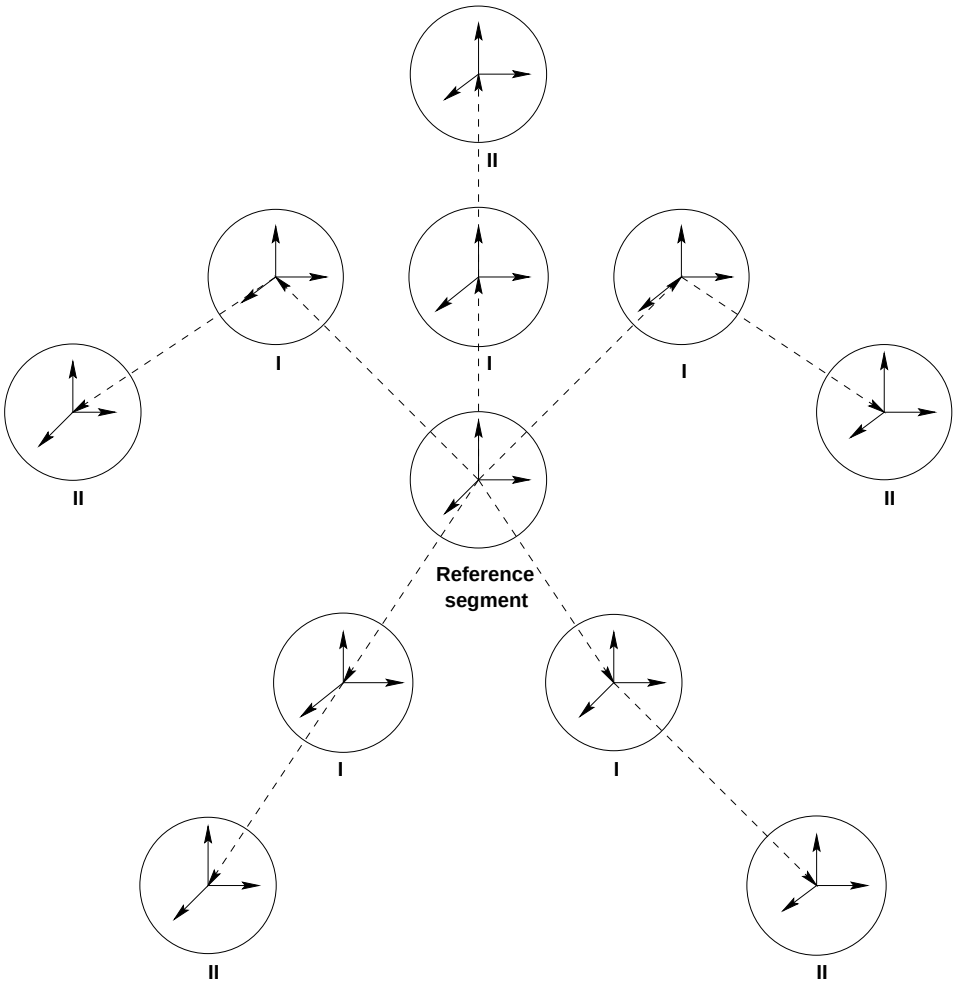


Figure 1: Eleven-segment model of the human body.

tives and angular velocity is extremely elegant:

$$\begin{pmatrix}\Omega_1\\ \Omega_2\\ \Omega_3\end{pmatrix} = 2\begin{pmatrix}-b_1 & b_0 & b_3 & -b_2\\ -b_2 & -b_3 & b_0 & b_1\\ -b_3 & b_2 & -b_1 & b_0\end{pmatrix}\begin{pmatrix}\dot{b}_0\\ \dot{b}_1\\ \dot{b}_2\\ \dot{b}_3\end{pmatrix}. \tag{1}$$

From equation (1) it can be seen that the transformation matrix relating $\dot{b}_t$ to Ω_s is orthogonal. This relationship ensures our equations of motion will be universally nonsingular. The corresponding kinematic equations for any definition of Euler angles are transcendental, nonlinear and contain a 0/0 singularity.

The equations of motion have been formulated by finding the kinetic and potential energy of each level, calculating the Lagrangian for the system of rigid bodies and then applying the Legendre transform to find the Hamiltonian (see Appendix). By deriving differential equations for the reference segment, the motion of the whole diver can be analysed and practical applications, throughout the dive, can be made. Using this approach we are left with four sets of nonlinear differential equations to describe the diver's flight. These equations describe the linear momentum, displacement and orientation of the diver.

3 Evaluation of method

In order to ascertain the accuracy of this approach the equations of motion were coded into fortran and integrated numerically using the Merson form of Runge–Kutta. Firstly, results were compared with known rigid body solutions and then they were evaluated against experimental data to assess their accuracy.

3.1 Rigid body motion

As shown in Appendix A4, the differential equations for linear momentum, displacement and orientation are consistent with the laws of motion. However, as the generalised momentum for the Euler parameters has no physical significance, the model can only be validated by how closely it replicates motion.

To check the model's angular velocity estimates, Euler's equations of rigid body motion were coded. The body's total inertia was calculated using the parallel axis theorem.

Results were compared for a symmetric and an asymmetric rigid body. The symmetric case was chosen as it depicts a simple somersaulting movement. The whole body inertia for this example was

$$\mathcal{I}_{sk} = \begin{pmatrix} 5.275 & 0 & 0 \\ 0 & 5.509 & 0 \\ 0 & 0 & 0.732 \end{pmatrix}.$$

When given initial angular velocity about the x-axis, initiating the somersault, both Euler's equations and the simulation model showed that angular velocity remained constant throughout the simulation; as expected. For the asymmetric case, the diver's inertia was calculated to be

$$\mathcal{I}_{sk} = \begin{pmatrix} 5.418 & 0 & -0.067 \\ 0 & 5.683 & 0 \\ -0.067 & 0 & 0.763 \end{pmatrix}.$$

This slightly asymmetric case was chosen as it is a good indicator of how well the simulation model will somersault and twist at the same time. The diver was given the same initial conditions as for the symmetric case. The angular velocity comparisons can be seen in Figure 2.

Figure 2 shows that the simulation model closely approximates Euler's equations of motion. Further, it can be seen that Euler parameters handle the singularities, inherent in Euler angles, very well. The slight discrepancies in the output can be attributed to the fact that the whole body inertia was given for Euler's equations where segmental inertias were prescribed for the simulation model.

3.2 Experimental analysis

As the simulation model has been shown to accurately compute known rigid body solutions, the accuracy of the proposed theoretical model was evaluated using video data of divers performing various twisting somersaults.

For this analysis, body segment parameters (BSPs) were calculated using de Leva's adjustments to Zatsiorsky's data [2]. The estimates are the most comprehensive data set currently available for females and non-Caucasian males. Data from an *in vivo* method was chosen as the validity of mathematically determined BSPs is contingent upon how well geometric solids can describe body segments.

Fifteen joint centres were digitised from the film data. These points were smoothed with a quintic spline and each segment's orientation was calculated. Segmental centre of mass estimates were based on segment length and angle calculations to limit the effects of digitising error. Each segment's linear and angular velocities were determined from quintic splines.

The orientation, velocities, mass and inertia of each nested level segment were input into the simulation model. The diver's somersault, tilt and twist were based upon Yeadon's [10] design where the whole body orientation in space is specified by the orientation of an orthogonal reference frame $\mathbf{f}$ in the body relative to the inertial frame. The rotating body frame $\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3$ is not fixed to one particular

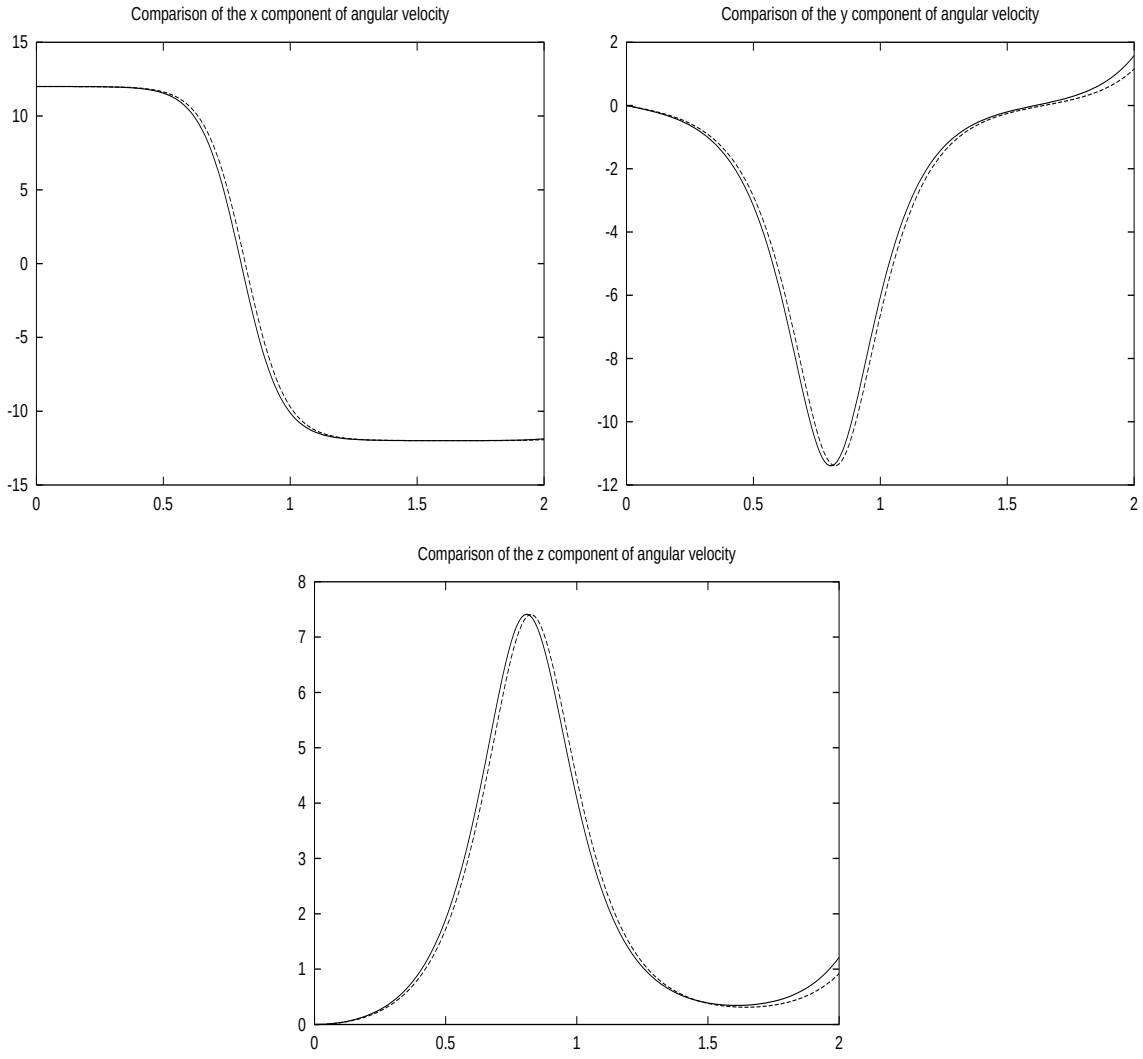


Figure 2: Angular velocity comparisons for Euler's equations and the 11-segment rigid body model. The solid line shows the solution for Euler's equations of motion and the dashed line represents the simulation model's solution.

segment of the diver, but representative of the whole body orientation. For this model, $\mathbf{f}_3$ extends from the mid-shoulders to the mid-thighs, $\mathbf{f}_1$ is parallel to the vector connecting the right hip joint centre to the left hip joint centre and $\mathbf{f}_2$ extends from the front of the body to the back. The body frame axes were calculated from simulation estimates.

By combining the whole body orientation with the location and orientation of the diver's lower torso, a complete picture of the dive is painted. A detailed analysis of these simulations is available.

4 Conclusion

A novel approach to modelling a twisting dive has been described. By defining the diver's orientation in terms of Euler parameters, universally nonsingular equations of motion were derived. Further, by employing a Hamiltonian approach, it was noted that physical points may be followed with ease.

The simulation model was tested against Euler's equations of motion for a rigid body. The simulation was shown to accurately replicate somersaulting and twisting movements. It was also shown to handle the singularities inherent in Euler angles very well.

A method for testing the accuracy of the simulation model for actual performances has been outlined.

There are many possibilities for the application of this approach. With the accuracy of the model for filmed performances determined, the simulation model can be used to examine the effect of different twisting techniques on overall performances. Further, it may be used to create new dives without incurring injury to divers.

Appendix

Formulating rigid body dynamics by directly applying Newton's laws can be difficult. For a complex system of bodies, accurately determining each segment's acceleration and force may be problematic. In order to overcome this difficulty, a general approach is useful.

The principle of least action is the most general form of the physical laws governing motion. This principle states that every mechanical system can be described in terms of a function, $L(q, \dot{q}, t)$, where q are generalised coordinates and $\dot{q}$ are their velocities [4].

For a closed system (that is, a system of bodies which interact with each other but do not interact with any other bodies), the Lagrangian is simply

$$L = T - U, \tag{2}$$

where T is the kinetic energy of the system and U is its potential energy.

A1 Kinetic energy

The kinetic energy of a system of rigid bodies is given by

$$T = \sum \frac{1}{2} m v_s^2 + \frac{1}{2} \Omega_s \mathcal{I}_{sk} \Omega_k,$$

where the summation is taken over all segments in the body [4].

As the velocity v_s of each nested level segment is not naturally expressed in terms of the inertial frame, we need to determine the motion of a moving body in terms of a fixed coordinate system [4]. Assume that dA_s/dt is the rate of change of an arbitrary vector A_s with respect to an inertial coordinate system. If the position of A_s changes in the moving system,

$$\frac{dA_s}{dt} = \frac{d\,^m A_s}{dt} + \epsilon_{skm} \Omega_k A_m,$$

where $d\,^m A_s/dt$ is the rate of change of A_s with respect to the moving system [4]. Thus, for the reference segment,

$$T_r = \frac{1}{2} m |\dot{R}_s|^2 + \frac{1}{2} \Omega_s \mathcal{I}_{sk} \Omega_k,$$

where m is the mass of the reference segment, R_s is the displacement vector from the inertial origin to the segment's centre of mass, Ω_s is the segment's angular velocity and $\mathcal{I}_{sn}$ is its inertia. For the first nested level,

$$T_I = \frac{1}{2}\,^i m |\dot{R}_s + {}^i\dot{\rho}_s + \mathcal{C}_{sk} \epsilon_{kln} \Omega_l\, {}^i\rho_n|^2 + \frac{1}{2} (\Omega_s + {}^i C_{sk}\, {}^i\omega_k)\, {}^i C_{sl}\, {}^i I_{lm}\, {}^i C_{nm} (\Omega_n + {}^i C_{np}\, {}^i\omega_p),$$

where ${}^i m$ is the mass of the ith body connected to the reference segment, $\mathcal{C}_{sk}$ is the direction cosine matrix which defines vectors from the reference frame in terms of the inertial frame, ${}^i C_{sk}$ is the direction cosine matrix which defines vectors from the frame of the ith segment in nested level I in terms of the

reference frame, ${}^i\omega_s$ is the angular velocity of the ith segment and ${}^i\rho_n$ is the displacement vector connecting the reference segment to the ith segment of nested level I.

For the second nested level,

$$\begin{aligned} T_{II} =& \frac{1}{2}{}^i\mu|\dot{R}_s + {}^i\dot{\rho}_s + {}^i\dot{\tau}_s + {}^iC_{sk}\epsilon_{kln}{}^i\omega_l{}^i\tau_n + \mathcal{C}_{sk}\epsilon_{kln}\Omega_l({}^i\rho_n + {}^iC_{nm}{}^i\tau_m)|^2 \\ &+ \frac{1}{2}(\Omega_s + {}^iC_{sk}({}^i\omega_k + {}^i\mathbb{C}_{kw}{}^i\xi_w)){}^iC_{sl}{}^i\mathbb{C}_{lm}{}^i\mathbb{I}_{mn}{}^i\mathbb{C}_{pn}{}^iC_{qp}(\Omega_q + {}^iC_{qu}({}^i\omega_u + {}^i\mathbb{C}_{uv}{}^i\xi_v)), \end{aligned}$$

where ${}^i\mu$ is the mass of the ith segment of nested level II connected to the ith segment of nested level I, ${}^i\xi_s$ is the angular velocity of this segment, ${}^i\mathbb{I}_{sk}$ is its inertia, ${}^i\tau_s$ is the displacement vector connecting the ith segment of the first nested level to the ith segment of the second nested level and ${}^i\mathbb{C}_{sk}$ is the direction cosine matrix defining the coordinates of the ith segment of nested level II in terms of the coordinates of the ith segment of nested level I.

A2 Potential energy

The potential energy of a rigid body is somewhat easier to compute than the kinetic energy. In a uniform field, the potential energy of a body is given by

$$U = -F_s r_s,$$

where the same external force F_s acts on each body at any point in the field and r_s is the displacement vector of each body from the inertial origin.

For the reference segment it can be seen that

$$U_1 = m\delta_{s3}gR_s,$$

where g is gravity. For the ith segment of nested level I we have

$$U_2 = {}^im\delta_{s3}g(R_s + \mathcal{C}_{sn}{}^i\rho_n),$$

and finally, for the ith segment of nested level II connected to the ith segment of nested level I, we find that

$$U_3 = {}^i\mu\delta_{s3}g(R_s + \mathcal{C}_{sn}({}^i\rho_n + {}^iC_{nk}{}^i\tau_k)).$$

A3 The Lagrangian

Using the equations for the kinetic and potential energies of each level and substituting them into (2) the Lagrangian for this system of rigid bodies is

$$\begin{aligned} L =& \frac{1}{2}m|\dot{R}_s|^2 + \frac{1}{2}{}^im|\dot{R}_s + {}^i\dot{\rho}_s + \mathcal{C}_{sk}\epsilon_{kln}\Omega_l{}^i\rho_n|^2 \\ &+ \frac{1}{2}{}^i\mu|\dot{R}_s + {}^i\dot{\rho}_s + {}^i\dot{\tau}_s + {}^iC_{sk}\epsilon_{kln}{}^i\omega_l{}^i\tau_n + \mathcal{C}_{sk}\epsilon_{kln}\Omega_l({}^i\rho_n + {}^iC_{nm}{}^i\tau_m)|^2 \\ &+ \frac{1}{2}\Omega_s\mathcal{I}_{sn}\Omega_n + \frac{1}{2}(\Omega_s + {}^iC_{sk}{}^i\omega_k){}^iC_{sl}{}^iI_{lm}{}^iC_{nm}(\Omega_n + {}^iC_{np}{}^i\omega_p) \\ &+ \frac{1}{2}(\Omega_s + {}^iC_{sk}({}^i\omega_k + {}^i\mathbb{C}_{kw}{}^i\xi_w)){}^iC_{sl}{}^i\mathbb{C}_{lm}{}^i\mathbb{I}_{mn}{}^i\mathbb{C}_{pn}{}^iC_{qp}(\Omega_q + {}^iC_{qu}({}^i\omega_u + {}^i\mathbb{C}_{uv}{}^i\xi_v)) \\ &- m\delta_{s3}gR_s - {}^im\delta_{s3}g(R_s + \mathcal{C}_{sn}{}^i\rho_n) - {}^i\mu\delta_{s3}g(R_s + \mathcal{C}_{sn}({}^i\rho_n + {}^iC_{nk}{}^i\tau_k)). \end{aligned} \tag{3}$$

A4 The Hamiltonian

The primary disadvantage of describing the motion of a diver in terms of the Lagrangian is that, for a system with n degrees of freedom, we are left with n second-order equations. These equations can be rather difficult to solve. However, if we express the dynamical equations of motion in terms of a Hamiltonian, we are left with $2n$ first-order equations.

In turn, these equations are much simpler. Furthermore, the generalised momenta from the Legendre transform have physical significance. For our system, the generalised momentum of R_s is the diver's linear momentum.

The solution to each of the resultant equations of motion gives us the displacement, orientation and momentum of the reference segment as a function of time. Thus we can analyse the complete dynamical behaviour of a diver in flight.

Given the Lagrangian of our system (see equation (3)), we can use the Legendre transform to find the generalised momenta. These momenta are given by

$$P_s = \frac{\partial L}{\partial \dot{R}_s},$$
$$\beta_t = \frac{\partial L}{\partial \dot{b}_t} = \frac{\partial L}{\partial \Omega_s}\frac{\partial \Omega_s}{\partial \dot{b}_t},$$

where

$$\Omega_s = \alpha_{st}\dot{b}_t \tag{4}$$

and

$$\alpha_{st} = \frac{1}{2}\begin{pmatrix} -b_1 & -b_2 & -b_3 \\ b_0 & -b_3 & b_2 \\ b_3 & b_0 & -b_1 \\ -b_2 & b_1 & b_0 \end{pmatrix}.$$

Thus, given

$$^iA_s = \mathcal{C}_{sk}\epsilon_{klm}\Omega_l\,{}^i\rho_m,$$
$$^iB_s = \mathcal{C}_{sk}\epsilon_{klm}\Omega_l\,{}^iC_{mn}\,{}^i\tau_n,$$
$$^iD_s = {}^iC_{sk}\epsilon_{klm}\,{}^i\omega_l\,{}^i\tau_m,$$

the resultant equations of motion are

$$\dot{P}_s = -\delta_{s3}g(m + {}^im + {}^i\mu), \tag{5}$$
$$\dot{R}_s = \frac{P_s - {}^im({}^i\dot{\rho}_s + {}^iA_s) - {}^i\mu({}^i\dot{\rho}_s + {}^i\dot{\tau}_s + {}^iA_s + {}^iB_s + {}^iD_s)}{m + {}^im + {}^i\mu},$$
$$\dot{b}_t = \gamma_{ts}\Omega_s,$$
$$\dot{\beta}_t = -L_{2,st}I_{sm}\Omega_m - {}^im\delta_{s3}gS_{skt}\,{}^i\rho_k - {}^i\mu\delta_{s3}gS_{skt}({}^i\rho_k + {}^iC_{kn}\,{}^i\tau_n),$$

where

$$L_{2,st} = \frac{\partial \Omega_s}{\partial \beta_t},$$
$$S_{snt} = \frac{\partial \mathcal{C}_{sn}}{\partial b_t}.$$

It is evident that the equation for $\dot{P}_s$ is correct. By integrating equation (5), we get

$$P_s = -\delta_{s3}(m + {}^im + {}^i\mu)gt.$$

Now, the velocity of the body is $-\delta_{s3}gt$, which satisfies $P_s = mv_s$.

Similarly, the equation for $\dot{R}_s$ can be proved true. For a system of rigid bodies, the whole body centre of mass is defined as

$$\mathrm{cm}_s = \frac{1}{M}\sum_{i=1}^{n} {}^i m\, {}^i r_s$$

where n is the number of segments in the body, M is the whole body mass, ${}^i m$ is the mass of segment i and ${}^i r_s$ is the displacement of segment i from the inertial origin. Thus, for this model

$$\begin{aligned}\mathrm{cm}_s &= \frac{mR_s + {}^i m(R_s + {}^i\rho_s) + {}^i\mu(R_s + {}^i\rho_s + {}^i\tau_s)}{m + {}^i m + {}^i\mu} \\ &= R_s + \frac{{}^i m\, {}^i\rho_s + {}^i\mu({}^i\rho_s + {}^i\tau_s)}{m + {}^i m + {}^i\mu}. \end{aligned} \tag{6}$$

Now, the velocity of the centre of mass of a body, falling under gravity, is $-\delta_{s3}gt$. By differentiating equation (6) and noting that $P_s = -(m + {}^i m + {}^i\mu)\delta_{s3}gt$, we get

$$\dot{R}_s = \frac{P_s - {}^i m({}^i\dot{\rho}_s + {}^i A_s) - {}^i\mu({}^i\dot{\rho}_s + {}^i\dot{\tau}_s + {}^i A_s + {}^i B_s + {}^i D_s)}{m + {}^i m + {}^i\mu}.$$

Finally, it is evident that the equation for $\dot{b}_t$ is correct as it is the definition of angular velocity given in equation (4), where $\gamma_{ts} = \alpha_{st}^{-1}$.

There is no direct way to check that the equation for $\dot{\beta}_t$ is correct as it has no physical meaning. Thus the accuracy of this equation will be determined by how closely the model replicates human movement.

References

[1] J. Dapena, "Simulation of modified human airborne movements", *J. Biomech.*, **14** (1981), 81–89.

[2] P. De Leva, "Adjustments to Zatsiorshy–Seluyanov's segment inertia parameters", *J. Biomech.*, **29** (1996), 1223–1230.

[3] J. L. Junkins and J. D. Turner, *Optimal Spacecraft Rotational Maneuvers*, Elsevier, Oxford (1986).

[4] L. D. Landau and E. M. Lifshitz, *Mechanics: Course of Theoretical Physics* Volume 1, Pergamon, Sydney (1976).

[5] J. Stuelpnagel, "On the parameterization of the three-dimensional rotation group", *SIAM Rev.*, **6** (1964), 422–430.

[6] B. Van Gheluwe, "A biomechanical simulation model for airborne twist in backward somersaults", *J. Human Movement Studies*, **7** (1981), 1–22.

[7] B. Van Gheluwe and W. Duquet, "A cinematographic evaluation of two twisting theories in the backward somersaults", *J. Human Movement Studies*, **3** (1977), 5–20.

[8] W. L. Wooten and J. K. Hodgins, "Animation of human diving", *Computer Graphics Forum*, **15** (1996), 3–13.

[9] M. R. Yeadon, "The simulation of aerial movement—II, A mathematical inertia model of the human body", *J. Biomech.*, **23** (1990), 67–74.

[10] M. R. Yeadon, " The simulation of aerial movement—III, The determination of the angular momentum of the human body", *J. Biomech.*, **23** (1990), 75–83.

[11] M. R. Yeadon, "The simulation of aerial movement—IV, A computer simulation model", *J. Biomech.*, **23** (1990), 85–89.

[12] M. R. Yeadon, J. Atha and F. D. Hales, "The simulation of aerial movement—I, The determination of orientation angles from film data", *J. Biomech.*, **23** (1990), 59–66.

[13] M. R. Yeadon, "The biomechanics of twisting somersaults. Part I: Rigid body motions", *J. Biomech.*, **11** (1993), 187–198.

[14] M. R. Yeadon, "The biomechanics of twisting somersaults. Part II: Contact twist, *J. Biomech.*, **11** (1993), 199–208.

[15] M. R. Yeadon, "The biomechanics of twisting somersaults. Part III: Aerial twist, *J. Biomech.*, **11** (1993), 209–218.

[16] M. R. Yeadon, "The biomechanics of twisting somersaults. Part IV: Partitioning performances using tilt angles", *J. Biomech.*, **11** (1993), 219–225.

[17] M. R. Yeadon, "Twisting techniques used by competitive divers", *J. Biomech.*, **11** (1993), 337–342.

TENNIS SERVING STRATEGIES

Graham McMahon and Neville de Mestre
School of Information Technology
Bond University
Gold Coast
Queensland 4229, Australia
gmcmahon@bond.edu.au, neville_de_mestre@bond.edu.au

Abstract

The use of a fast serve followed by a slower serve is almost universal in tennis. Data taken from grand slam tournaments show that, in many cases, one or both players would benefit from departing from this strategy.

1 Introduction

There is substantial research literature on the game of tennis, and several papers are concerned with the appropriate use of the two services which are permitted to the player when serving for a point. The principal references are Croucher [1], George [2], and King and Baker [3]. It is well known that the normal pattern is a fast serve followed by a slower serve. It has been suggested that the case for this traditional pattern is not particularly strong in all matches, and that some players might achieve better results by using their fast service twice.

We have taken data from all singles matches in the four grand slams of the year 2000, and show that in a surprising number of cases, one or both players would apparently benefit from departing from the standard pattern. In particular, we show that the use of two slower serves is often preferable. This extends the work of earlier investigators who considered only a small number of matches.

2 A model of tennis tournament play

A model for the probabilities involved in winning a service point is as follows. We assume that each player has two serving techniques. These are a fast service (F), which has a higher probability of being a fault, and a slower service (S) which is faulted less frequently. The fast service is usually more difficult to return than the slower service, but in a particular match this may not be the case. Our model assumes that the server places his serve in the most advantageous fashion. This may be to the opponent's weaker side or, more commonly, may be varied in direction to make the receiver's task more difficult. The strategy of varying the direction of the serve is not part of this study.

We have four probabilities for each player which are assumed to remain constant throughout the match. These are the probabilities of being not faulted on each type of serve (P_1 for F and P_2 for S) and the conditional probabilities of winning the rally when each type of serve is not faulted (W_1 for F and W_2 for S). The fast serve is defined by the property $P_1 < P_2$. Four strategies are possible, based on the four possible combinations of fast and slow serves, which we label FF, FS, SF and SS. The conventional method is to use FS.

Note that although we describe these probabilities in terms of the server, the values of W_1 and W_2 are partially determined by the skill of the person in returning the serve, and the skill of each player in

the subsequent rally. The probabilities of success using each of the four strategies are given in Table 1, which is essentially the same as a table in Croucher [1].

Strategy	Probability of success
FF	$P_1W_1(2-P_1)$
FS	$P_1W_1+(1-P_1)P_2W_2$
SS	$P_2W_2(2-P_2)$
SF	$P_2W_2+(1-P_2)P_1W_1$

Table 1

The relative values of the strategies are determined by two expressions:

$$D = P_1W_1 - P_2W_2,$$
$$E = P_2W_2(P_1 - P_2).$$

In these, D measures the advantage of a single fast serve compared with a single slow serve (usually assumed to be negative), whereas E has no simple interpretation.

If $D > 0$, then

$$P(\text{FF}) > P(\text{FS}) > P(\text{SF}) > P(\text{SS}),$$

where $P(\text{FF})$ is the probability of winning a point with the strategy FF, etc.

However, if $0 > D > E$, then

$$P(\text{FS}) > P(\text{FF})$$

and

$$P(\text{FS}) > P(\text{SS}) > P(\text{SF}),$$

while if $0 > E > D$, then

$$P(\text{SS}) > P(\text{SF})$$

and

$$P(\text{SS}) > P(\text{FS}) > P(\text{FF}).$$

Note that although SF is not always the worst strategy, there is no region in which it is the best. Hence its use cannot be justified except perhaps as part of a mixed strategy where surprise is considered to be important. On the other hand strategies FF or SS are superior to FS for certain combinations of the probabilities P_1, P_2, W_1, W_2.

3 Results in four grand slam tournaments

A large number of matches from the singles matches in all grand slam tournaments for the year 2000 have been analysed, and the probabilities P_1, P_2, W_1, W_2 obtained for each match, since the relevant data were available from the official web sites.

Table 2 shows raw data for play in the quarter-finals, semi-finals and final in the women's US championships in the year 2000. However, in all, data was collected for 858 completed men's and women's singles matches from the grand slam tournaments of 2000, covering the US Open, the Australian Open, the French Open and Wimbledon (see web site references [4]). The first player in each pairing was the winner. Note that the player who won the higher proportion of points on serve always won in those seven matches. Over all matches recorded, this was not true in 6.5% of cases. Presumably, in those cases, the winner was more successful on the important points.

We assume that in all matches the normal strategy FS was used. The final two columns of Table 2 show projected results if either player moved to strategies FF or SS. These are based on the following

Player	First serves	First serves in	Won on 1st	Second serves	Second serves in	Won on 2nd	Double faults	% with FS	% with FF	% with SS
Hingis	56	39	29	17	17	7	0	0.643	0.675	0.412
Seles	63	35	21	28	22	7	6	0.444	0.481	0.304
V. Williams	77	40	30	37	27	15	10	0.584	0.577	0.515
Tauziat	79	41	24	38	36	17	2	0.519	0.450	0.471
Dementieva	78	50	33	28	22	14	6	0.603	0.575	0.607
Huber	80	45	29	35	26	13	9	0.525	0.521	0.467
Davenport	69	33	24	36	33	21	3	0.652	0.529	0.632
S. Williams	82	54	34	28	26	12	2	0.561	0.556	0.459
V.Williams	89	52	38	37	33	18	4	0.629	0.604	0.539
Hingis	105	83	48	22	22	15	0	0.600	0.553	0.682
Davenport	63	32	29	31	30	15	1	0.698	0.687	0.499
Dementieva	84	53	33	31	25	12	6	0.536	0.538	0.462
V. Williams	81	48	33	33	26	14	7	0.580	0.573	0.514
Davenport	70	39	27	31	24	10	7	0.529	0.557	0.395

Table 2: US women's championship, 2000 (quarter finals, semi-finals and final).

formulas, which enable P_1, W_1, P_2, W_2 (defined earlier) to be calculated from the published data for each match:

$$\begin{aligned} P_1 &= 1 - (\text{number of second serves / number of points}),\\ W_1 &= \text{number won on first serve / number of first serves made},\\ P_2 &= 1 - (\text{number of double faults / number of second serves}),\\ W_2 &= \text{number won on second serve / number of second serves made}. \end{aligned}$$

Given these ratios, we can obtain the strategy probabilities for the last three columns of Table 2 using the formulas from Table 1.

We are interested in the strategy which gives the highest probability of winning a point on serve, as shown in Table 2. A strategy change may not change the result of the match, but it could either make the match easier for the winner or bring the loser closer to victory.

In two of the seven matches shown in Table 2, both players were correct to use the conventional strategy; in four matches, one player should have varied; while in one game both players should have varied. In just one game, the semi-final between Hingis and Venus Williams, there is strong evidence for a different result if Hingis had used her second service throughout.

This sample of seven matches is not atypical. Results over the whole set of data are given in Table 3 below.

4 Discussion

The number of situations where we have prima-facie evidence of inferiority of the conventional strategy is surprisingly large. In more than 50% of men's games and 80% of women's games, one or both players would gain by changing strategy. Of course, the evidence is of the hindsight variety. It may not be obvious early in a match that a shift in strategy is advisable. However, given the data above an astute coach should at least look for opportunities to shift to FF or SS. Table 3 shows that the most common beneficial change for women is to an SS strategy, whereas an FF strategy works more often for men, in spite of the additional double faults. Part of the motivation for this study was the appearance of

Event	Total matches available	Matches suggesting strategy change	Winner to FF	Winner to SS	Loser to FF	Loser to SS	Result change
Women's							
Aust. Open	114	84	21	30	30	39	18
French Open	114	96	32	38	32	46	16
US Open	62	49	15	13	20	20	5
Wimbledon	124	94	27	29	40	33	14
Men's							
Aust. Open	126	67	17	17	30	14	15
French Open	111	59	12	18	23	16	5
US Open	84	42	7	12	16	11	4
Wimbledon	123	70	23	16	34	11	12

Table 3

information on a television screen during a Wimbledon match between Agassi and Rafter which showed that Rafter was winning the same percentage of points on his second slower service as on his first faster service. In that case SS was indicated to save energy and reduce the possibility of a double fault.

The number of situations where it appears that the result of the match would change due to a change in serving strategy by the loser is also surprisingly higher than expected. Even where the winner also changes to optimum strategy, Table 3 predicts that 53 of the 414 results would change for the women, while 36 of the 444 results in the men's matches would change. Additional results would be reversed if the winner did not change strategy. Comparison of the proportion of points won on serve is an excellent predictor of a match result, but not perfect. Hence the number of changed results might be slightly less or greater than those presented.

The results clearly show that at least the coaches of women tennis professionals should be more aware of the possibilities of a strategy change for their protégées.

We have found cases where a player should use the FF strategy against one player and SS against another. Such strategies may vary depending on the court surface. These aspects would need further analysis.

It is possible to argue that the real gain in a match from shifting strategy might be less than the above figures indicate:

(a) The opponent realises what you are doing and takes some counter action.

This argument is without merit. The results shown above are not based on the element of surprise. In conventional play, the receiver prepares for a first or second serve, knowing what to expect. There is uncertainty in the placement of the serve (backhand, forehand or body) but not in the type of serve. If a player decides to change, say to SS, he or she is free to announce this to the receiver, and the above probabilities should still apply.

(b) The opponent gains practice in receiving the same service continually. This may be true.

(c) It may be more tiring to shift to FF if the fast service takes more energy. True, but there may be compensation in the fact that rallies following a fast serve are usually shorter.

There are however somewhat stronger arguments for the view that an unconventional strategy may be better than the figures suggest.

(a) If, as predicted, a player changes strategy and wins a higher proportion of service points, then some service games will be shorter, with attendant physical and psychological advantages.

(b) A shift to SS saves energy, not just because the service effort may be less, but mainly because there is rarely a need for a second serve.

5 Mixed strategy

A mixed strategy may sometimes be more effective than a shift to an unconventional pure strategy. There are several ways in which this could be implemented. An unconventional service strategy could be used at random times during the match. Alternatively, the unconventional strategy could be used on important points such as break points or just in the tiebreaker.

If the receiver prepares in a very different way for the fast and slow serves, then we have a two-person game in the mathematical sense. Even if the preparation is almost identical, the additional uncertainty might be useful. These cases can be modelled and analysed, but we have no data to make this analysis useful.

References

[1] J. Croucher, "Developing strategies in tennis", in *Statistics in Sport*, J. Bennett (editor), London, Arnold (1998).

[2] S. L. George, "Optimal strategies in tennis", *J. Roy. Statist. Soc. Ser. C*, **2** (1973), 97–104.

[3] H. A. King and J. A. Baker, "Statistical analysis of service and match-play strategies in tennis", *Canad. J. Appl. Sports Sci.*, **4** (1979), 298–301.

[4] Web sites: http://www.ausopen.com, http://www.frenchopen.org, http://www.usopen.org and http://www.wimbledon.org.

TEAM RATINGS AND HOME ADVANTAGE IN SANZAR RUGBY UNION, 2000–2001

R. Hugh Morton
Institute of Food, Nutrition and Human Health
Massey University
Private Bag 11-222
Palmerston North, New Zealand
h.morton@massey.ac.nz

Abstract

This paper examines team ratings and home advantages of both national (Super 12) and international (Tri-nations) rugby union teams from South Africa, New Zealand and Australia in the years 2000 and 2001. Linear modelling methodology is used for parameter estimation. High variability is a significant feature of the estimation process. Team ratings vary from zero (by convention the lowest) to over 20 points. Home advantages, which average around eight points, range from low negative values (a home disadvantage) to positive values in excess of 20 points. Depending on the schedule of matches, an unusually high percentage of home wins (50 of 69 or 72.5%) occurred in the Super 12 in 2001. On the contrary, seven of the first twelve matches in 2002 have been away wins. The high variability makes any attempt at arbitrage a risky venture despite a positive expectation.

1 Introduction

There have been many studies of team (or individual) ratings and home advantage in sport. Literature reviews in the area include those by Courneya and Carron [6], Nevill and Holder [16] and Stefani [22, 23]. Specific examples include the summer [3] and winter [1] Olympic Games, Australian Rules football [24], basketball [10, 21, 25], tennis and golf [11, 17], cricket [4, 7], ice hockey [8], alpine skiing [2], baseball [14], athletics [15], netball [18], soccer [5], rugby league [9, 13]; and many more.

These studies have almost invariably found significant rating differences between teams or individuals, and the existence of significant home advantage, either on average or to most teams in a competition. However, there is very little literature available on rugby union. Only Pretorius *et al.* [19] investigated the 1997 domestic Currie Cup season in South Africa, later [20] relating this to altitude, but not weather conditions.

The purpose of this paper is to investigate team ratings and home advantages at national and international level rugby union in competitions involving sides from South Africa, New Zealand and Australia (SANZAR). The years 2000 and 2001 are considered and analysed using linear modelling methodology.

2 Linear modelling methodology

In the game of rugby union, teams can score five points for a try (touchdown), two further points for converting a try (a successful kick of the ball between the goal posts, above the bar), and three points for a penalty or dropped goal (also a successful kick). Both teams in the match accumulate points by

one or more of these means, the match winner being that team scoring the most points at the end of the game.

The simplest linear model in such a situation incorporates a neutral ability rating $r.$ for each of the m teams in the competition, together with a common home advantage h applicable to all teams. It models the difference in point scores d_{ij}, over a match between the home team i (by convention named first) and away team j, by the difference between their respective ratings adjusted for the home advantage:

$$d_{ij} = h + r_i - r_j + e, \tag{1}$$

where r_i and r_j are the respective ratings and e is an error term (white noise).

Expressed more fully over a set of n matches in a competition between m teams,

$$D = hI + JR + E,$$

where D is an $n \times 1$ matrix of score differences; I is an $n \times 1$ unit matrix; J is an $n \times (m-1)$ indicator matrix such that, for any row, column entries are $+1$ for the home team, -1 for the away team and zeros elsewhere; R is an $(m-1) \times 1$ matrix of ratings; and E is an $n \times 1$ matrix of random errors.

Note that an mth column for J is linearly dependent on the previous $m-1$ columns, and so is omitted to avoid singularity. The mth team rating is therefore by default zero.

This model can most easily be fitted by ordinary least squares using any common linear regression procedure. Estimates of h and the $r.$ together with their standard errors, and an overall error of estimation are obtained. Ancillary output such as goodness-of-fit statistics is also available.

A second model simply removes the restriction that all teams share a common home advantage, introducing parameters $h.$, one for each team. Thus:

$$d_{ij} = h_i + r_i - r_j + e. \tag{2}$$

The full matrix formulation adds in similar fashion to above, a second indicator matrix, with units in the columns of the home team for each match, and zeros elsewhere. Again this model can be fitted by ordinary least squares, though the "No constant" option should be selected in the regression command.

A third model considers the set of 2n scores, $s_{.(.)}$ or $s_{(.).}$ themselves, and widens the concept of a single rating of team ability, into two ratings. One represents the offensive strength of a team, $o.$, reflecting its ability to score points. The second represents the defensive strength of a team, $d.$, reflecting its ability to prevent its opponents scoring. Specifically:

$$s_{i(j)} = h_i + o_i - d_j + e \qquad \text{and} \qquad s_{(i)j} = o_j - d_i + e \tag{3}$$

represent the modelling of the pair of scores obtained in the match between team i at home and team j away, respectively. Three indicator matrices, of appropriately located $+1$'s, -1's and zeros are required, and the whole model can be fitted by ordinary least squares once again.

The data to hand consists of the records of all full-time scores by both sides in all the Super 12 and Tri-nations rugby union matches played in the years 2000 and 2001. It is these data that are to be modelled. Data for 2002 (Super 12 only) are available only up to the time of writing.

3 Results and discussion

With three models and two competitions over two years (which may or may not be combined, weighted or otherwise) there are a large number of possible applications to choose between. I shall consider only a few for the purposes of illustration.

3.1 The Tri-nations

This competition involving South Africa, New Zealand and Australia is a neat example of a small balanced home and away league, in which each team plays each other team once at home and once

Year	Ability ratings South Africa	New Zealand	Australia	Home advantage
2000	0.0	6.3	7.7	5.5
2001	0.0	4.0	3.5	0.3
Combined	0.0	5.2	5.6	2.9

Table 1: Team ability ratings in the Tri-nations competition, 2000 and 2001.

away. Model (1) applied to 2000 and 2001, and to both years combined (unweighted), yields the results given in Table 1.

South Africa has clearly been the weakest team, and the honours seem to be only very slightly in Australia's favour, despite Australia winning the series in both years.

Now consider model (3), and, for the purposes of making a prediction, exclude only the final match of 2001, played between Australia and New Zealand. The estimates in Table 2 are obtained based on the preceding eleven matches.

Country	Offensive rating	Defensive rating	Home advantage
Australia	16.0	−2.7	6.2
New Zealand	29.4	−8.5	−8.5
South Africa	6.2	0.0	10.0

Table 2: Offensive and defensive ratings, Tri-nations 2000 and 2001.

That final match, played in Sydney would have had a predicted scoreline of Australia 31 ($16.0 + 6.2 - (-8.5) = 30.7$), New Zealand 32 ($29.4 - (-2.7) = 32.1$). In fact, Australia, in scoring a last gasp try, won by 29 to 26.

3.2 The Super 12

The Super 12 competition is played between twelve regional teams; four from South Africa, five from New Zealand and three from Australia. Each team plays each other team once only, with a home/away split of five/six or six/five. Pool play consists therefore of 66 matches, and thereafter there are two semi-finals and one final; 69 matches in total.

In order to investigate both differences between teams and between years in a larger competition, I illustrate by considering the application of model (2) to the 69 matches in each of 2000 and 2001. The estimates of ratings and home advantages in Table 3 are obtained.

The dominance of the Brumbies in both years, both at home and away is clearly evident. Likewise is the poor performance of the Bulls. Big downward movers between the years were the Blues, Crusaders and Hurricanes. The Crusaders dropped only in their rating, but the Blues and Hurricanes dropped both in their ratings and home advantages. There are no obvious big upward movers, so it can be concluded that the Sharks (who lost in the final in 2001) made the most of their opportunities. Given that the competition is not balanced home and away, and noting some of the differences observed in this table, it would be fair to say that the outcome of the competition is to some extent influenced by which opponents a team plays at home, and which it plays away.

The overall standard, as represented by the range of the ratings (the worst team is conventionally zero rated), appears to be more even in 2001 than in 2000. This is true both at home and away as the average and range of home advantages have changed little.

The leading four teams after the completion of pool play in the 2000 season were the Brumbies, Cats, Crusaders and Highlanders. The Crusaders beat the Brumbies by one point in the final in Canberra.

	Away rating		Home advantage		Home rating	
Team	2000	2001	2000	2001	2000	2001
Blues	33.9	13.7	−8.6	−2.6	25.3	11.1
Brumbies	36.6	18.8	5.8	20.7	42.4	39.5
Bulls	0.0	7.7	18.7	−3.8	18.7	3.9
Cats	12.9	13.6	30.6	2.6	43.5	16.2
Chiefs	14.4	7.8	6.5	13.0	20.9	20.8
Crusaders	29.5	7.8	9.3	7.7	38.8	15.5
Highlanders	20.3	0.9	16.4	21.5	36.7	22.4
Hurricanes	19.3	8.7	11.8	2.0	31.1	10.7
Reds	23.1	16.1	9.5	0.6	32.6	16.7
Sharks	19.9	18.6	−4.1	2.3	15.8	20.9
Stormers	28.6	17.2	−3.2	−2.0	25.4	15.2
Waratahs	24.1	0.0	4.3	30.5	28.4	30.5
Mean	21.9	10.9	8.1	7.7	30.0	18.6

Table 3: Home and away ratings in Super 12 rugby, 2000 and 2001.

A prediction would have had a Brumbies win by 13 points ($42.4 - 29.5 = 12.9$). Likewise in 2001, the Brumbies, Cats, Reds and Sharks comprised the semi-finalists, with the Brumbies clear winners by 30 points over the Sharks in the final in Canberra. A prediction would have had them winning by 21 points ($39.5 - 18.6 = 20.9$).

3.3 A statistical note

Of course not all the estimates presented above are significant in a strictly statistical sense, but they are presented for comparative purposes rather than in any attempt at parsimonious modelling. One particular feature of the modelling not mentioned previously is a high degree of variability. This manifests through a large standard error of estimation (around ±13 points in a Super 12 year and ±20 points when two years are combined) and low to moderate goodness-of-fit statistics (R^2 in the 35–60% range). It goes without saying that this makes any attempt at arbitrage a risky venture despite the ability to identify bets with positive expectation.

To illustrate the variability inherent in such activity, Figure 1 shows the ups and downs of the following three betting strategies over the 69 matches of the 2001 Super 12 season starting with a pool of $750.00 (thirty investors each contributing $25):

(a) Bet proportional to the Kelly criterion [12] using model (1) but limiting the total outlay in any one week to 60% of the current pool (the real $ strategy adopted).

(b) Bet 10% of the current pool on the predicted winners each week, using model (1) (virtual $).

(c) Bet 10% of the current pool on the home team in each match each week (virtual $).

It is clear that a profit can be made, but the ups and downs and different strategy outcomes are enough to unnerve most investors.

References

[1] N. J. Balmer, A. M. Nevill and A. M. Williams, "Home advantage in the Winter Olympics (1908–1998)", *J. Sport Sci.*, **19** (2001), 128–139.

[2] S. Bray and A. Carron, "The home advantage in alpine skiing", *Aust. J. Sci. Med. Sport*, **25** (1993), 76–81.

[3] S. R. Clarke, "Home advantage in the Olympic games", in *Proceedings of the Fifth Australian Conference on Mathematics and Computers in Sport*, G. Cohen and T. Langtry (editors), University of Technology, Sydney (2000), 76–85.

[4] S. R. Clarke and P. Allsopp, "Fair measures of performance: The World Cup of Cricket", *J. Oper. Res. Soc.*, **52** (2001), 471–479.

[5] S. R. Clarke and J. M. Norman, "Home ground advantage of individual clubs in English soccer", *The Statist.*, **44** (1995), 509–521.

[6] K. S. Courneya and A. V. Carron, "The home advantage in sport competitions: a literature review", *J. Sport. Exer. Psychol.*, **14** (1992), 13–27.

[7] J. S. Croucher, "Player ratings in one-day cricket", in *Proceedings of the Fifth Australian Conference on Mathematics and Computers in Sport*, G. Cohen and T. Langtry (editors), University of Technology, Sydney (2000), 95–106.

[8] W. F. Gayton, G. R. Matthews and C. J. Nickless, "The home field disadvantage in sports championships: does it exist in hockey?", *J. Sport Psychol.*, **9** (1987), 183–185.

[9] W. Griffiths and D. Chotikapanich, "A Bayesian decision model for point spreads in rugby league", *ASA Proc. Bayesian Statist.*, (1993), 258–263.

[10] D. A. Harville and M. H. Smith, "The home-court advantage: How large is it, and does it vary from team to team?", *Amer. Statist.*, **48** (1994), 22–28.

[11] R. L. Holder and A. M. Nevill, "Modelling performance at international tennis and golf tournaments: is there a home advantage?", *The Statist.*, **46** (1997), 551–559.

[12] J. Kelly, "A new interpretation of information rate", *Bell Syst. Tech. J.*, **35** (1956), 917–926.

[13] A. J. Lee, "Modelling rugby league data via bivariate negative binomial regression", *Aust. N. Z. J. Statist.*, **41** (1999), 141–152.

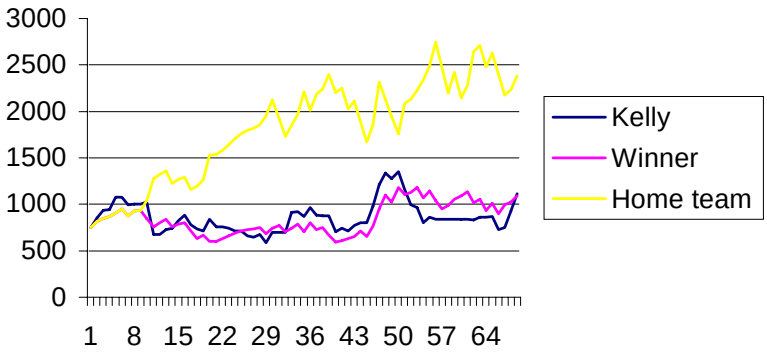


Figure 1: Comparison of various betting strategies.

[14] W. M. Leonard, "Specification of the home advantage: the case of the World Series", *J. Sport Behav.*, **21** (1998), 41–52.

[15] L. E. McCutcheon, "The home advantage in high school athletics", *J. Sport Behav.*, **7** (1984), 135–138.

[16] A. M. Nevill and R. L. Holder, "Home advantage in sport: an overview of studies on the advantage of playing at home", *Sports Med.*, **28** (1999), 221–236.

[17] A. M. Nevill, R. L. Holder, A. Bardsley, H. Calvert and S. Jones, "Identifying home advantage in international tennis and golf tournaments", *J. Sport Sci.*, **15** (1997), 437–443.

[18] P. Norton and S. R. Clarke, "Home ground advantage in the Australian netball league", in *Proceedings of the Fifth Australian Conference on Mathematics and Computers in Sport*, G. Cohen and T. Langtry (editors), University of Technology, Sydney (2000), 168–173.

[19] B. Pretorius, I. N. Litvine, A. M. Nevill and L. Terblanche, "An analysis and estimation of home and away game performance of South African rugby teams: is a home field advantage present?", *South African J. Res. Sport Phys. Ed. Rec.*, **21** (1998/1999), 69–80.

[20] B. Pretorius, M. W. Pierce and I. N. Litvine, "The possible effect of climate and altitude on home advantage and game performance in South African rugby teams", *South African J. Res. Sport Phys. Ed. Rec.*, **22** (2000), 37–48.

[21] E. E. Snyder and D. A. Purdy, "The home advantage in collegiate basketball", *Sociol. Sport J.*, **2** (1985), 352–356.

[22] R. T. Stefani, "Survey of the major world sports rating systems", *J. Appl. Statist.*, **24** (1997), 635–646.

[23] R. T. Stefani, "Predicting outcomes", in *Statistics in Sport*, J. Bennett (editor), London, Arnold (1998), 249–275.

[24] R. T. Stefani and S. R. Clarke, "Predictions and home advantage for Australian Rules football", *J. Appl. Statist.*, **19** (1992), 251–261.

[25] P. E. Varca, "Analysis of home and away game performance of male college basketball teams", *J. Sport Psychol.*, **2** (1980), 245–257.

SOCCER: FROM MATCH TAPING TO ANALYSIS

Emil Muresan, Jelle Geerits and Bart De Moor
ESAT-SCD, KULeuven
Kasteelpark Arenberg 10
3001 Heverlee (Leuven)
Belgium
emil-mihai.muresan@esat.kuleuven.ac.be, jelle.geerits@esat.kuleuven.ac.be,
bart.demoor@esat.kuleuven.ac.be

Abstract

Soccer games analysis has so far been done only by watching the recording of the game and making remarks about the interesting elements in it. There is no exact data to use, let alone detailed and quantitative analysis of player and team performances and tactics.

This paper proposes a different approach to match analysis, using cameras and computers to extract player coordinates throughout the whole match. Once we have the coordinates database, any statistics can be inferred, e.g. on individual player performance (position, speed, acceleration, field coverage, etc.) as well as on team tactics (players' complementarity, cohesion, field coverage, etc.).

1 Overview

In this paper we propose a full technological process for match analysis. As it is a proposal, not all the steps are implemented and there are still problems to be tackled. But the potential of such an analysis method is huge, as will be pointed out.

The process by which we obtain the desired analysis and statistics of a soccer match has several steps: match taping, digitising the video, coordinate extraction, and full analysis on the coordinates database. Each of these steps will be described in the following. The whole process is depicted in Figure 1.

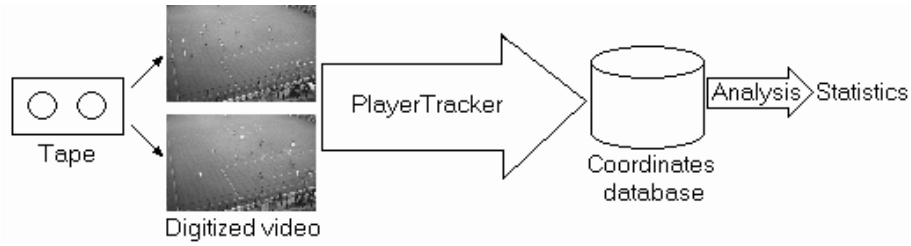


Figure 1: The whole analysis process.

2 Match taping and digitisation

We start from the video recording of a soccer match on tape. Taping of the match is done with four fixed cameras from the four corners of the field (Figure 2), each surveying a quarter (in order to get sufficient video quality for the next step). The cameras are fixed, static ones. This is required in order to calculate the mapping between the position of the player in the image and its real position on the field.

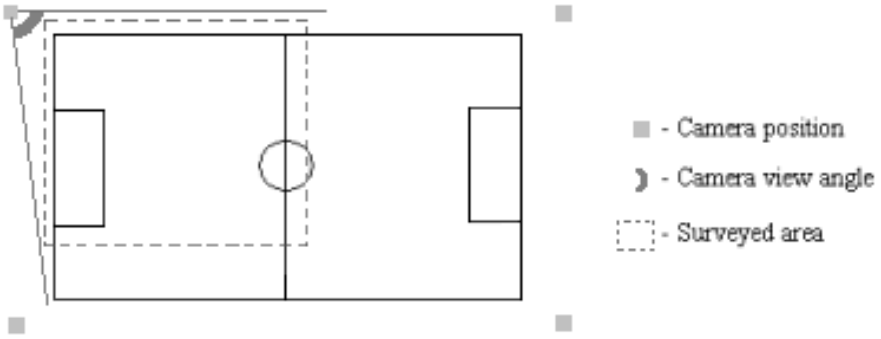


Figure 2: Camera positions for match taping.

After the whole match is recorded on tape, it has to be digitised and stored as a video file (avi). Once we have the digital recording of the match, the coordinate extraction tool can be applied.

The digitised videos can have any resolution as far as the extraction tool is concerned, but there are limitations due to the minimum image quality needed for being able to identify the players in the images. For digitisation we used Dazzle MovieStar (www.dazzle.com), a very easy-to-use multimedia tool that uses some dedicated hardware to do digitisation and video compression in real time. Some parameters that we used in the digitisation process are: MPEG-2 video encoder, image size not bigger than 720×480 pixels, a sampling rate of 25 frames per second, and colour depth of 24 bits per pixel.

The processing of these images does not run in real time. The application is run on a computer with a 500 MHz processor and 512 MB of RAM, using an image processing algorithm that achieves a compromise between accuracy and speed, resulting in a processing rate of ten frames per second, that is, 2.5 times slower than real time.

3 Coordinate extraction

PlayerTracker is a tool that we developed especially for this step of the process, i.e. coordinates extraction. It uses the images digitised in the previous step. Those images are processed and the players are automatically tracked throughout the match. The coordinates of the players are saved in a database for exploitation in the next step (analysis).

The PlayerTracker's objective is to obtain the coordinates of the players, the referee and the ball during the match. This tool makes use of image processing techniques to track each player during the whole match and compute their positions on the field in every frame of the video recording.

The coordinate extraction process is a cycle. A frame is taken from the video, players are detected in this frame, and the real coordinates of these players on the real field are computed and saved in the database. After this we move on to the next frame.

To make the detection easier and faster, player position information from the previous frame is used to detect the players in the current frame. We make use of two rectangles to track each player: one rectangle has the dimensions of the player and is centred on the player, the second bigger rectangle is given such dimensions that the player cannot exit that rectangle in between two consecutive frames (because of physical limitations).

In the beginning, some user interaction is needed for initialisation. The user must specify the mapping between the field in the perspective view of the video and the real field. This mapping will be used to translate the detected player position from image coordinates to real physical coordinates on the field. Next, because information from a previous frame is not available, the user has to initialise the system by interactively placing (e.g. mouse click and drag in the image) the tracking rectangles on the players. Then, the detection is done automatically (apart from some exceptions that will be discussed later in this paper). The system after the initialisation phase is shown in Figure 3. The three players to be tracked (left image) are indicated by the tracking rectangles. Their physical positions on the field are computed and visualised in the "virtual" top view of the field (right image).

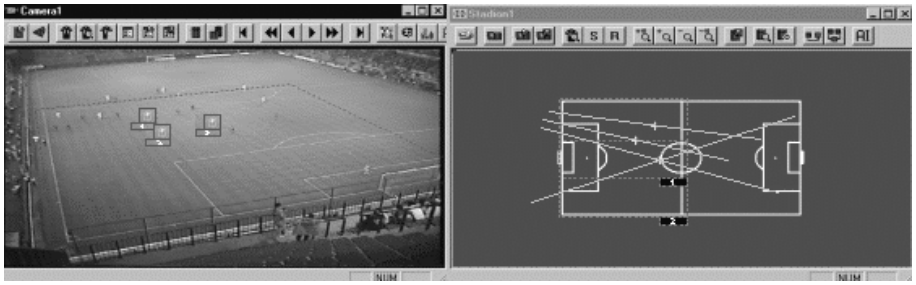


Figure 3: The detection system after initialisation.

Using the tracking rectangles makes the detection faster and more robust because each player is searched for only inside their own rectangle, and not in the whole frame image.

For each player the same detection process is applied. The image inside the bigger tracking rectangle is preprocessed (smoothing, contrast enhancement). Then image processing and object detection algorithms are used for the detection of the player.

The algorithms used to detect a player in a rectangle employ different image processing techniques for contrast enhancement, smoothing, edge detection (Sobel and Prewitt operators) and segmentation [2]. The actual detection algorithms are based on pattern matching, histogram analysis, dynamic thresholding, etc. Their performance differs and is dependent on the image quality. The images on which the tests were made are not of very high quality, because one of the objectives of our project is to keep the whole procedure at low cost, so the images were taken with ordinary cameras.

The detection can be performed either on gray-scale or on colour images. One of the algorithms for gray-scale images with its subsequent steps (a. initial image; b. noise elimination with a Max kernel and enlargement of the object; c. gray scale stretching, contrast enhancement; d. binarisation, object detection is straightforward) is exemplified in Figure 4.

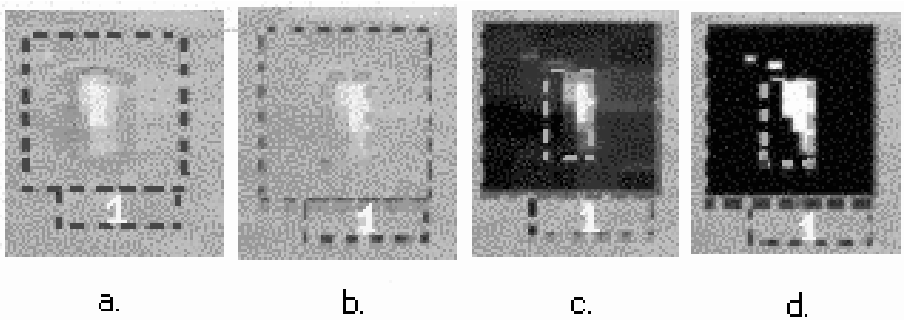


Figure 4: Player detection in gray-scale images (a. initial image; b. noise elimination with a Max kernel and enlargement of the object; c. gray scale stretching, contrast enhancement; d. binarisation).

An example of player detection algorithm for colour images is the detection by histogram comparison.

The small detection rectangle parses the bigger rectangle, and for each position its colour histogram is computed and it is compared with the reference histogram of the empty field. The position for which the histogram is the most distant from the reference is considered to be the player.

In Figure 5 an example of the histograms for the reference and player rectangles is given. Histograms are computed on the inner tracking rectangle, and displayed separately for each of the three colour components (R, G, B). Reference image histograms are more peak-like because they are quite uniform (green); player histograms are less uniform due to equipment color.

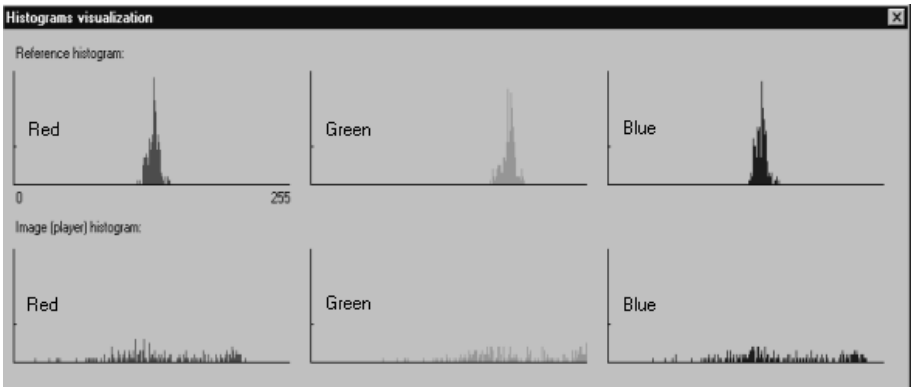


Figure 5: Histogram comparison between the reference image (first row) and the player image (second row).

Several metrics to compute the "distance" between two colour histograms were taken into consideration. One is the Kullback–Leibner distance [1] that is widely used in probability distribution problems. This is

$$D(p,q) = \sum_{n \in X} p(n) * \log \frac{p(n)}{q(n)},$$

where p and q are two discrete distributions over the domain X.

In this case, the two distributions are the concatenated colour histograms (R,G,B) of the reference (grass) and the player image at the current position of the detection rectangle.

However, because of the computational complexity of this metric, a simpler one was chosen:

$$D(RH, CH) = \sum_{i \in [0,255]} (|RHr[i] - CHr[i] + |RHg[i] - CHg[i]| + |RHb[i] - CHb[i]|)$$

where RHr, RHg, RHb are the R, G, B histograms of the reference (grass) image, and CHr, CHg, CHb are the R, G, B histograms of the detection rectangle at the current position.

After having obtained the position of the player in the current frame in the video, the mapping to the real field coordinates is used to compute the player's real position on the field. These real field coordinates are then saved in the database.

The mapping function maps the perspective polygonal view of the scaled field onto the real rectangular field. The user is required to graphically make the correspondence between the field lines in the image and field projection during the initialisation phase.

Of course there are also some problems in the process and not everything goes smoothly from one whistle to another. Player collisions make this task more difficult than it seems. Some algorithms were developed to deal with several of the collision situations, but for some cases user interaction is still needed. Figure 6 shows an example of how the software deals with the collision problem. The application detects when a collision is about to occur (tracked players are lost) and displays a warning (Figure 6a). If a collision has occurred (Figure 6b) the detection is stopped with the tracking rectangles in the last known positions and the user has to reinitialise the colliding players.

Figure 6: Player collision solving.

The algorithms for player collision solving are based on previous player positions, multiple cameras information, player equipment colour. But if the colliding players overlap completely in the image, the detection process cannot discriminate between them.

Collisions are detected from the intersections of the tracking rectangles. If the inner rectangles of two players intersect more than a certain percentage (a threshold), a collision is signalled and the corresponding solving algorithms are enabled.

One type of collision solving is to use two cameras surveying the same portion of the field. While for one camera two players appear to collide because they are on the same line with the camera, the other camera from a different angle can clearly differentiate between them.

One such situation is shown in Figure 7, as well as the way in which two cameras with different orientations can solve this problem. Two players are tracked in this examples (each player's number is displayed at the bottom of the tracking rectangle; in Figure 7 the players with numbers 1 and 2 are tracked). Figure 7a shows the two players from the viewpoint of the first camera (the two players are colliding). Viewing the same scene from the viewpoint of the second camera (Figure 7b), the two players can clearly be discriminated. Figure 7c shows the real field positions of the two players, computed using information from both cameras (here numbers in the figure specify the two cameras surveying the corresponding regions; the two crosses show the player positions).

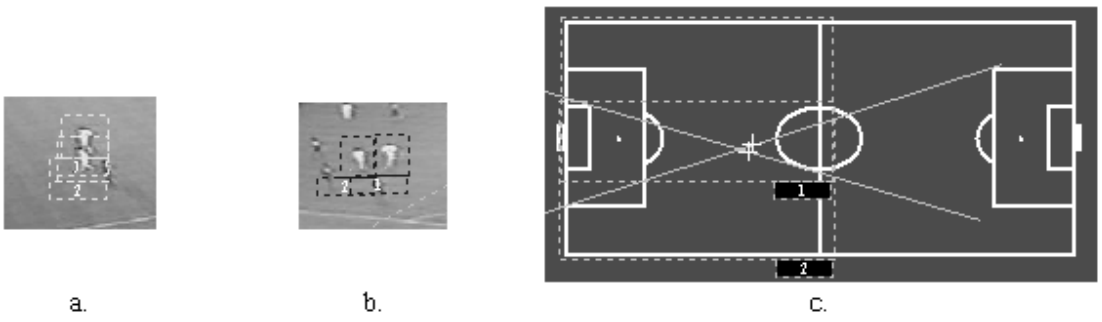


Figure 7: Collision solving with two cameras; a. first camera; b. second camera.

Even if players are not colliding, the two cameras are useful for a more precise computation of the real player position on the field (player positions on the field are displayed in Figure 7c). Normally, there are small errors in the mapping from image to real field, and these can be corrected a certain amount by averaging in between the coordinates computed from the two different cameras (sensor fusion).

4 Analysis based on the coordinates database

Once the coordinates are extracted, quantitative game analysis can be performed (different analysing methods that can be priceless to soccer coaches), and a virtual reconstruction of the match can be made for analysis, entertainment or commercial purposes.

All kinds of statistics can be calculated such as: field coverage, total distance run by a player, velocity, acceleration, ball possession percentage, ball losses, etc.

Even tactics can be built in. There is, for instance, a claim that when you take the polygon that wraps the defensive players (the "convex hull") and calculate its centre of gravity, and you do the same thing for the offensive players of the other team, you can determine, by the relative position of the body mass point if there is danger for a counter-attack [3]. The coach can then be alerted about this situation and give the team directions to be careful in the future. The player polygon can easily be drawn from the player coordinates at each moment of the game, and displayed as in Figure 8.

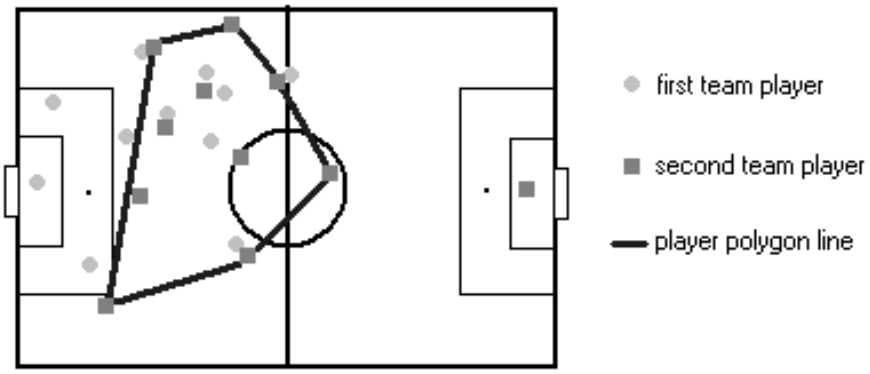


Figure 8: The player polygon ("convex hull") is the smallest polygon that encloses all players of a team.

Another possible use of the coordinates is 3D replay. A situation can be replayed and the user can have a look at the scenario from all angles (e.g. how did the goal-keeper see the ball).

A first scheme of what kind of analysis can be performed once the coordinates are extracted is given below to emphasise the importance of this process. This is a kind of knowledge tree with respect to possible kinds of analysis:

Analysis scope:
- — per player
- — per group of players
  - — attack
  - — midfield
  - — defence
- —per team
- —per club

Time horizon:
- — per phase
- — per *a priori* defined game period
- — per game
- — over different games

Statistics:
- — ball possession per team
- — time/position tensor
- — tracking the line-up through a match (4:4:2 or 3:2:5 or 5:2:3)

Markov model:
- — gives the *a posteriori* statistics of passes
- — by deriving the Markov model of both teams:
  - — ball possession, loss and recuperation can be quantified
  - — per group of players a passing matrix can be made to show their complementarity

Goal keeper:
- — visualise the goal attempts on his goal
- — distant or close shots / good or bad keeper reactions
- — preferential directions

Per individual player:
- — physical condition
- — run distance
- — speed profiles
  - — double acceleration
  - — explosivity
- — run lines
- — position per unit of time
- — strategy and tactical insight of the game:
  - — goes deep in the field
  - — works hard
  - — can anticipate well (e.g. positioning with respect to the others)

For the referee:
- — track/monitor the referee performance
- — position on the field in relation to field situations

This is only a coarse categorisation of possible analysis methods, and the elements mentioned above will be detailed and completed.

5 Future work

In the future there will be two distinct approaches for the tracking. The use of fixed cameras has proven to be useful and will be improved, but also moving cameras will enter the scene. The use of the moving cameras is already tested in reality for ball tracking purposes. Here the pan and tilt of the camera is measured using the optical mouse principle.

Another use for the moving camera is to determine the exact identity of a player after a collision that may have caused a switch of identity. Therefore, the moving camera scans the field and if a player whose identity is uncertain is found, the camera follows him until the back number can be read.

Acknowledgments

Dr Bart De Moor is a full professor at the Katholieke Universiteit Leuven, Belgium. Our research is supported by grants from several funding agencies and sources: Research Council KUL: Concerted Research Action GOA–Mefisto 666 (Mathematical Engineering), IDO (IOTA Oncology, Genetic networks), several

PhD/postdoctoral and fellowship grants; Flemish Government: Fund for Scientific Research Flanders (several PhD/postdoctoral grants, projects G.0256.97 (subspace), G.0115.01 (bio-i and microarrays), G.0240.99 (multilinear algebra), G.0197.02 (power islands), G.0407.02 (support vector machines), research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary/Poland), IWT (Soft4s (softsensors), STWW–Genprom (gene promotor prediction), GBOU–McKnow (Knowledge management algorithms), Eureka–Impact (MPC–control), Eureka–FLiTE (flutter modeling), several PhD grants); Belgian Federal Government: DWTC (IUAP IV-02 (1996–2001) and IUAP V-10-29 (2002–2006): Dynamical Systems and Control: Computation, Identification & Modelling), Program Sustainable Development PODO-II (CP-TR-18: Sustainability effects of Traffic Management Systems); Direct contract research: Verhaert, Electrabel, Elia, Data4s, IPCOS.

References

[1] S. Kullback, *Information Theory*, Dover, New York (1968).

[2] V. S. Nalwa, *A Guided Tour of Computer Vision*, Addison–Wesley (1993).

[3] F. Masson, *Football +: Comment gerer une equipe?* manual from the Belgian Heizel Trainer School, Brussels (1998).

VIDEO ANALYSIS IN SPORTS: VIDEOCOACH©

Emil Muresan, Jelle Geerits and Bart De Moor
ESAT-SCD, KULeuven
Kasteelpark Arenberg 10
3001 Heverlee (Leuven)
Belgium
emil-mihai.muresan@esat.kuleuven.ac.be, jelle.geerits@esat.kuleuven.ac.be,
bart.demoor@esat.kuleuven.ac.be

Abstract

The traditional way of analysing soccer matches by coaches or whoever is interested in getting statistics about them, is to take notes on a piece of paper or to make remarks about what happens while watching the game.

This paper proposes a different approach to match analysis, utilising computers and the appropriate software tools. VideoCoach© allows the user to annotate the match through easy-to-use interfaces. The annotations are saved in a database. Match analysis and statistics are automatically retrieved from the annotation database, video compilations can be made and analysis can be done. This kind of analysis is not strictly bound to soccer, but the examples will be from this sport.

1 Introduction

The Canadian researcher Ian Franks proved that coaches only remember 30% of the game content correctly. Even after going through a special training they do not do any better than 50%. As a consequence, a trainer often misjudges certain situations. A quantitative and objective analysis of soccer games, with the aid of computers, could therefore be a priceless advantage.

VideoCoach© is a tool we specially developed to assist coaches. Starting from a video of the game and the software package VideoCoach©, an annotation of the game can be done after some initialisation. From the annotation, statistics can be derived, compilations can be made and video analysis can be done. The flow chart of the entire process is shown in Figure 1.

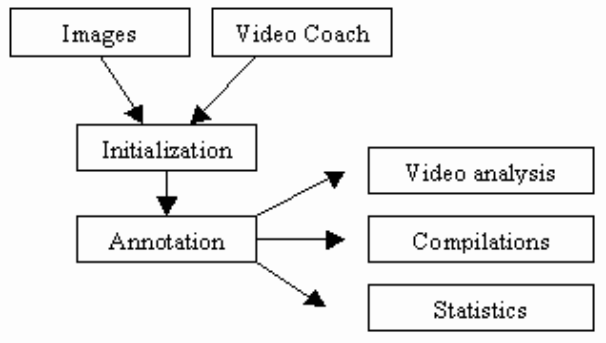


Figure 1: Flow chart of VideoCoach© process.

2 Images

The images of the game can be recorded with an ordinary camera or directly taped from television. They have to be digitised. We use Dazzle MovieStar [1] to deal with this problem. It can digitise input from a VCR, directly from a camera or even from television. It writes the input to mpeg format (real time) and the user can specify the bit rate. A quality of 12MB/minute is ideal because this is high enough and one half of the play can be burned onto one CD-ROM for distribution. We tape 55 minutes for each half to deal with extra time and camera set-up.

3 Initialisation

Before the annotation can start in VideoCoach© there is an initialisation phase. The purpose is that the coach has to decide what he or she wants to annotate (which actions, players, etc.) and how to annotate (with the keyboard interface or the touch screen interface). An absolute advantage of VideoCoach© is that the labels can be chosen freely by the user, that is, the labels are not predefined by the program.

The first step is to define the teams. Therefore a team database is set-up and saved on the hard disk. The team information can contain only the name of the team or can be more detailed including the names of the players, the shirt numbers and the labels attached to the players.

What are those labels and why are they needed? As stated above, the purpose of VideoCoach© is to help the trainer to remember the critical situations and to make it easy to reconstruct the situations when analysing the match with the players. Using this tool, a coach can mark each situation that is worth remembering. This marking is done by pressing a key (keyboard interface) or touching a button on the screen (touch screen interface). The label behind the key or button is attached to the phase. This can be "free kick", "goal", "fault", "corner", "throw in" or any other label that was defined by the user. These labels thus define actions and are called action labels. There are also quotation labels (to give a quantitative appreciation to an action), time labels (to identify the moment at which the action occurred, these are automatically filled in by the program) and player labels (usually the player's name, to indicate which player did the action).

A coach can define different label sets. Standard label sets for each kind of annotation (e.g. team or player) will be available for download in the near future. When the coach wants to analyse an individual player he or she will need other action labels than when wanting to analyse a whole team at once. The coach can combine the labels in any fashion, but before starting the annotation will need to select at most one action label set and one quotation label set.

After defining the label sets, matches can be created. Just the two teams that are playing against each other can be defined as a match, but we may also define the number of spectators, team line-up, date and hour, as well as a whole collection of notes that are important to remember for later use.

4 The annotation

Once a game is created, it is possible to annotate it. This can be done while watching the game in real time (the video is then linked afterwards) or otherwise. A label is attached to every interesting phase, using a touch screen or the keyboard.

The use of the touch screen is developed for a quick and user-friendly annotation. Figure 2 shows the interface. The yellow buttons are the labels. The label is written to the annotation list at the time the button is pressed and a new line is started (one click thus results in one label). On the left there is the video that is annotated and there are buttons to navigate through the video, start the first half, second half, extra time and penalties. In the left corner the annotation list is shown.

By using the keyboard a combination of the different label sets can be made if wanted. So one line in the annotation list can contain the time label, the player label, the actions made and the quotation. Each label has a unique key attached to it. By pressing this key, the label is displayed on the screen.

Figure 2: The touch screen interface.

The user specifies which label set defines a new line. When a label of this set is used, the annotation line is written in the database and a new line is started.

The annotation results in a list containing all the labels, as shown in Figure 3. The list can be ordered on time, action, player or quotation. The video fragment attached to every phase can be shown by double clicking on the corresponding line. A compilation can be made from this list. For example, in the screen dump all the actions of Crasson are selected. Figure 4 shows the corresponding compilation.

Figure 3: The annotation list showing the labelled actions.

5 Video analysis and compilation

By just double clicking on the label in the label list, the video of the action behind each label can be shown. By selecting the desired actions a compilation can be made. This way, an individual player can get a video of his or her actions, ordered by action, time or appreciation. In Figure 3 all the actions of

Bertrand Crasson are selected. This player's actions are grouped in a compilation displayed in Figure 4. The compilation can be played, deleted or written to a new mpeg file of the same quality as the original one. The compilation can also be ordered on time, action, player or quotation. Extra information can be added manually or automatically if wanted.

Figure 4: The compilation from the selection made in Figure 3.

6 Statistics

Statistics are calculated automatically after the annotation is completed (Figure 5). The number of times that every action occurred is calculated. From these occurrences the positively and negatively evaluated ones are counted and displayed separately. The analysis can be done for every player, for the whole team or for both teams. In Figure 5, it is done for the whole team.

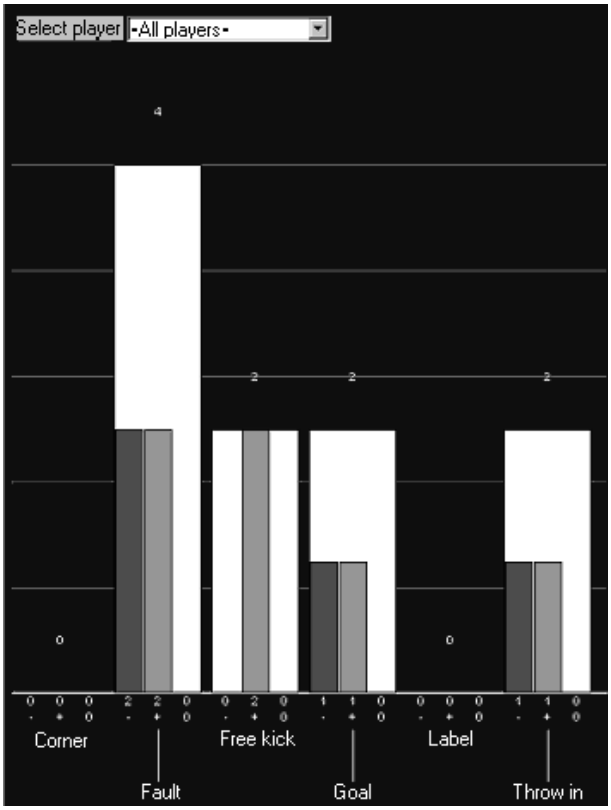


Figure 5: Statistics calculated automatically from the annotation.

Besides these statistics, sequence statistics can be calculated; for example, one can calculate how many times action x was followed by action y (Figure 6) and this for several time intervals. This can be useful, for example, to see how long the ball is kept in possession by the team after a free kick or a throw-in. Figure 6 shows how many times a throw-in was followed by a free kick.

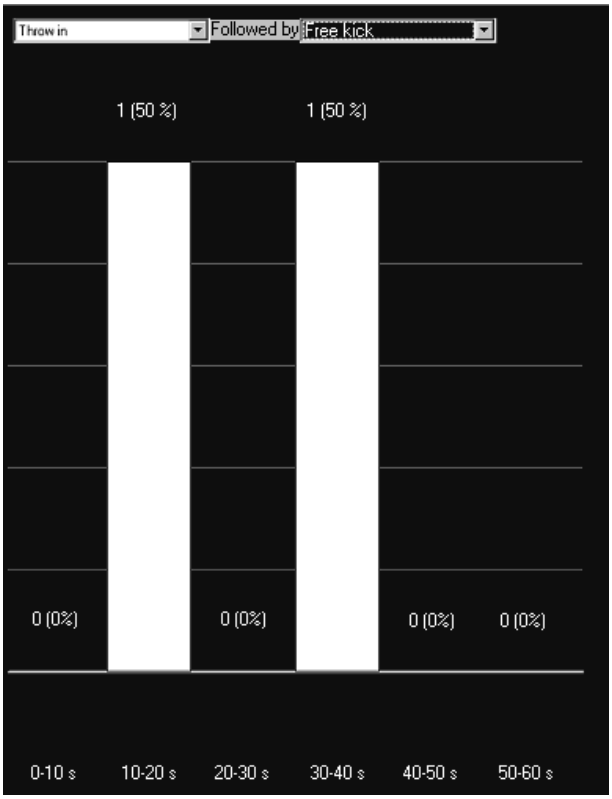


Figure 6: Sequence statistics show how many actions x are followed by actions y in different time intervals.

7 Future work

In the future, the statistics and data collected during several games will be used to make predictions, search for patterns and make a complete player database. Also a LUI (Language User Interface) for labelling will be developed.

A downloadable version will be available soon. The downloadable version can be obtained from www.sport4s.com (not yet online). Those concerned can write to the authors to obtain VideoCoach©. Predefined VideoCoach© label sets for soccer and other sports and in several languages will also be downloadable from the website.

Acknowledgments

Dr Bart De Moor is a full professor at the Katholieke Universiteit Leuven, Belgium. Our research is supported by grants from several funding agencies and sources: Research Council KUL: Concerted Research Action GOA–Mefisto 666 (Mathematical Engineering), IDO (IOTA Oncology, Genetic networks), several PhD/postdoctoral and fellowship grants; Flemish Government: Fund for Scientific Research Flanders

(several PhD/postdoctoral grants, projects G.0256.97 (subspace), G.0115.01 (bio-i and microarrays), G.0240.99 (multilinear algebra), G.0197.02 (power islands), G.0407.02 (support vector machines), research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary/ Poland), IWT (Soft4s (softsensors), STWW-Genprom (gene promotor prediction), GBOU–McKnow (Knowledge management algorithms), Eureka–Impact (MPC–control), Eureka–FLiTE (flutter modeling), several PhD grants); Belgian Federal Government: DWTC (IUAP IV-02 (1996–2001) and IUAP V-10-29 (2002–2006): Dynamical Systems and Control: Computation, Identification & Modelling), Program Sustainable Development PODO-II (CP-TR-18: Sustainibility effects of Traffic Management Systems); Direct contract research: Verhaert, Electrabel, Elia, Data4s, IPCOS.

Reference

[1] Website: www.dazzle.com.

SERVING UP SOME GRAND SLAM TENNIS STATISTICS

Pam Norton*
School of Mathematical Sciences
Monash University
PO Box 28M
Victoria 3800, Australia
pam.norton@sci.monash.edu.au

Stephen R. Clarke*
School of Mathematical Sciences
Swinburne University
PO Box 218, Hawthorn
Victoria 3122, Australia
sclarke@groupwise.swin.edu.au

Abstract

Data from the 2000 Australian Tennis Open are used to measure the advantage of serving with new balls in both men's and women's singles. We discuss the main features of the men's, women's and mixed doubles at the Australian Open in 2001. Data from a variety of sources are used to compare the server's advantage at Wimbledon, the French and the Australian Open.

1 Introduction

The Australian Open Tennis Championships are the first Grand Slam tournament of each calendar year, and are held on "Rebound Ace" hardcourts in January at Melbourne Park, Melbourne.

The first Australasian Tennis Championships were played in 1905 at the Warehouseman's Cricket Ground (now Albert Reserve) in Melbourne. Men's events only were played. Women's events, mixed doubles and junior boys' events were first held in 1922, and junior girls' events in 1930. In 1969, the championships were open to amateurs and professionals. The championships were held each year at a different venue throughout Australia and New Zealand, up until 1972, when Kooyong became the home of the Australian Open. The championships moved to Flinders Park (now Melbourne Park) in 1988.

The other grand slam events are the French Open, Wimbledon and the US Open. There is often discussion about the differences in court surface and their effects on matches. In recent years, point-by-point playing statistics have been collected and disseminated. The availability of such statistics allows for many of the myths of tennis to be tested. For instance, Magnus and Klaassen [2, 3] use four years of Wimbledon data to test some beliefs about tennis.

Point-by-point data is collected at the Australian Open on every court for every match (unless technical difficulties occur). The point-by-point data for men's and women's singles matches in 2000 and the results for all matches were obtained in electronic form from the Australian Open organisers. The authors were actively engaged in the collection of similar data in 2001 and 2002, and these activities are discussed in Clarke and Norton [1].

Using this point-by-point data, the effects of new balls at the Australian Open, 2000, will be examined. A comparison is then made of service characteristics for non-tiebreak games and for tiebreak games. Using data from a range of sources, a comparison of points won on serve at Wimbledon, 1992–1995 and 2001, the French Open 2001 and the Australian Open 2000–2002 is given. Finally, doubles (including mixed doubles) have rarely been looked at closely from a statistical point of view. An overview of the main features of men's, women's and mixed doubles at the Australian Open in 2001 is given.

*The authors would like to thank Chris Simpfendorfer and the staff of the Australian Open for their assistance in providing statistical data.

2 Is serving with new balls an advantage at the Australian Open?

"Serving with new balls . . . " is a familiar saying by tennis commentators today. The implication is that they are an advantage to the server, but there is no discussion as to what this means. During a tennis match, new balls are provided at the start of the warm up to the match, after the first seven games of a match, and then after every subsequent nine games. Tiebreaks are counted as one game, even though there are usually more points played in a tiebreak. At the Australian Open 2000, in the men's singles there were an average of 6.1 points played in an ordinary game, and 11.8 in a tiebreak, whereas for women, there were an average of 6.5 points played in an ordinary game and 11.6 in a tiebreak. (This compares favourably to the standard theoretical calculations by Pollard [4] of 6.5 (6.8) points played in an ordinary game where the probability of winning a point on serve is 0.6 (0.5), and 11.7 in a tiebreak when the probability of winning a point on serve is 0.5.)

Magnus and Klaassen [2] used point-by-point data from some of the matches at The Championships, Wimbledon, and examined the effect of new balls. They had previously showed that the probability of winning a point on service is the best statistic to measure service quality or dominance. If new balls were an advantage to the server, then it would be expected that the dominance of service, measured by the probability of winning a point on service, would decrease with the age of the balls. They used point-by-point data from Wimbledon over the period 1992–1995, and were able to show that there was no decrease in the probability of winning a point on service with the age of the balls. They suggested that new balls may affect the way that points are won. There was evidence that the chance of getting a first serve in increases with age of the balls, but this is counteracted by the chance of winning a point given a good serve decreasing. They were hampered to some extent by not having data on all singles matches.

	Percentages of the characteristics for the following ages of balls:									
Characteristic	1	2	3	4	5	6	7	8	9	All
First serves in	60.4	57.5	57.3	55.8	58.2	57.9	57.3	56.6	58.0	57.6
	(0.9)	(0.9)	(0.8)	(0.8)	(0.8)	(0.8)	(0.9)	(0.9)	(0.9)	(0.3)
Second serves in	88.9	91.3	90.5	88.1	89.8	89.3	88.2	89.5	87.6	89.2
	(0.9)	(0.8)	(0.7)	(0.8)	(0.8)	(0.8)	(0.9)	(0.8)	(0.9)	(0.3)
Games won on serve	81.3	81.5	79.2	78.5	80.4	78.9	79.4	81.3	78.9	79.9
	(1.8)	(1.8)	(1.7)	(1.7)	(1.7)	(1.8)	(1.8)	(1.8)	(1.9)	(0.6)
Aces	8.9	8.4	8.6	8.5	7.7	8.3	8.9	7.8	9.0	8.4
	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.2)
Double faults	4.4	3.7	4.0	5.3	4.3	4.5	5.1	4.6	5.2	4.6
	(0.4)	(0.4)	(0.3)	(0.4)	(0.3)	(0.4)	(0.4)	(0.4)	(0.4)	(0.1)
Points won if first serve in	73.5	74.3	72.3	73.4	71.7	73.5	74.2	73.3	74.7	73.4
	(1.1)	(1.1)	(1.0)	(1.0)	(1.0)	(1.0)	(1.0)	(1.1)	(1.1)	(0.3)
Points won on first serve	44.4	42.7	41.5	41.0	41.7	42.6	42.5	41.4	43.3	42.3
	(0.9)	(0.9)	(0.8)	(0.8)	(0.8)	(0.8)	(0.9)	(0.9)	(0.9)	(0.3)
Points won on serve	64.4	64.9	63.2	63.0	63.3	63.6	63.8	64.5	64.1	63.8
	(0.9)	(0.9)	(0.8)	(0.8)	(0.8)	(0.8)	(0.9)	(0.9)	(0.9)	(0.3)

Table 1: Service characteristics for the 2000 Australian Men's Singles Championships depending on the age of the balls.

The data in Table 1 give the summary statistics for service characteristics for the men's singles at the 2000 Australian Open based on all points played, including points played during tiebreaks. The age of the balls in games are from 1 (new balls) to 9 (old balls). New balls are taken for the warm up (five minutes in total) to each match, and are then used for the first seven games of the match. So the age

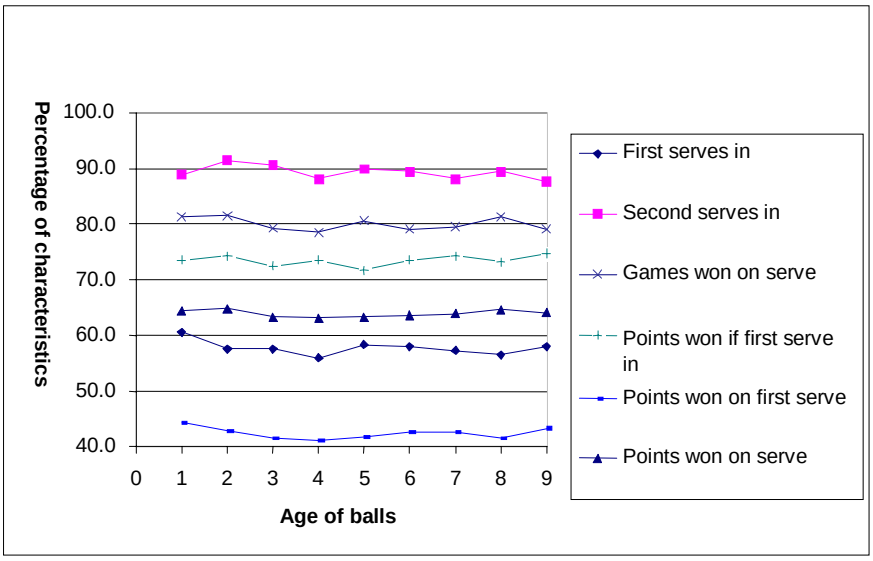


Figure 1: Service characteristics for the 2000 Australian Men's Singles Championships depending on the age of the balls.

of the balls in the first game of a match is set to three. Standard errors are given in parentheses. These have been obtained by assuming all points are independent. The proportion of points won if the first serve is in is the ratio of the number of points won on the first serve divided by the number of first serves which are in. The proportion of points won on first serve is the ratio of the number of points won when the first serve is in to the total number of points.

If new balls provide an advantage to the server, then we would expect the probability of winning a point on service would decrease with increasing age of the balls. The percentages from Table 1 have been graphed in Figure 1. There appears to be no consistent trend in any of the characteristics. In particular, the percentage of points won on serve does not decrease with the age of the balls. The percentage of first serves in and the percentage of points won on first serve are highest with new balls, but the percentage of points won if the first serve is in is highest with the oldest balls.

Similar data and graph for the women's singles are shown in Table 2 and Figure 2. Again no clear trend appears. There is little evidence that age of ball affects any of the serving statistics. The percentages of first serves in, games won on serve, points won if first serve is in, points won on first serve and points won on serve is greatest, but not significantly, with balls being used in their first game.

Of interest is the fact that points won on first serve, points won on serve and games won on serve start to decline as the age of the balls increases, but then rise again when the balls are at about age six. This suggests a possible confounding effect due to tiebreaks. When points from tiebreaks were removed from the data, there were still no trends evident.

It is impossible to compare statistics in tiebreaks for different ages of balls, as very few tiebreaks are played with balls of certain ages. If a tiebreak is played in the first set, then the balls are of age 6. If a tiebreak is played in the second set, then the balls could be of any age, except 2 and 8, depending on the score in the first set. Of the 92 tiebreaks played in the men's singles, 41 were with balls of age 6, and only two with balls of age 1. For the women's singles, 11 of the 24 tiebreak games were played with balls of age 6, and no tiebreaks were played with balls of age 2, 3 or 8. Figures 3 and 4 show the distribution of tiebreaks against age of balls for the men's and women's singles events.

Characteristic	Percentages of the characteristics for the following ages of balls: 1	2	3	4	5	6	7	8	9	All
First serves in	63.0	63.5	61.9	62.5	61.9	62.5	61.0	61.3	61.6	62.1
	(1.2)	(1.2)	(1.0)	(1.0)	(1.0)	(1.0)	(1.1)	(1.1)	(1.1)	(0.4)
Second serves in	84.8	88.8	87.0	84.9	86.9	87.1	85.2	86.8	86.3	86.4
	(1.4)	(1.3)	(1.1)	(1.2)	(1.2)	(1.2)	(1.3)	(1.3)	(1.3)	(0.4)
Games won on serve	69.9	67.7	65.4	64.0	62.5	69.0	67.3	63.6	65.4	65.9
	(2.8)	(2.9)	(2.5)	(2.6)	(2.6)	(2.6)	(2.7)	(2.8)	(2.8)	(0.9)
Aces	3.5	2.7	3.3	2.8	3.7	3.8	3.5	3.6	2.8	3.3
	(0.4)	(0.4)	(0.4)	(0.3)	(0.4)	(0.4)	(0.4)	(0.4)	(0.4)	(0.1)
Double faults	5.6	4.1	5.0	5.6	5.0	4.8	5.8	5.1	5.2	5.2
	(0.6)	(0.5)	(0.4)	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)	(0.2)
Points won if first serve in	66.0	62.4	62.3	62.5	61.4	64.2	65.0	62.9	62.7	63.2
	(1.4)	(1.5)	(1.2)	(1.3)	(1.3)	(1.3)	(1.4)	(1.4)	(1.4)	(0.5)
Points won on first serve	41.6	39.6	38.5	39.1	38.0	40.1	39.7	38.5	38.6	39.3
	(1.2)	(1.2)	(1.0)	(1.0)	(1.0)	(1.1)	(1.1)	(1.1)	(1.1)	(0.4)
Points won on serve	58.9	57.1	56.5	55.7	55.9	58.6	57.3	56.2	57.4	57.0
	(1.2)	(1.2)	(1.0)	(1.0)	(1.1)	(1.1)	(1.1)	(1.1)	(1.1)	(0.4)

Table 2: Service characteristics for the 2000 Australian Women's Singles Championships depending on the age of the balls.

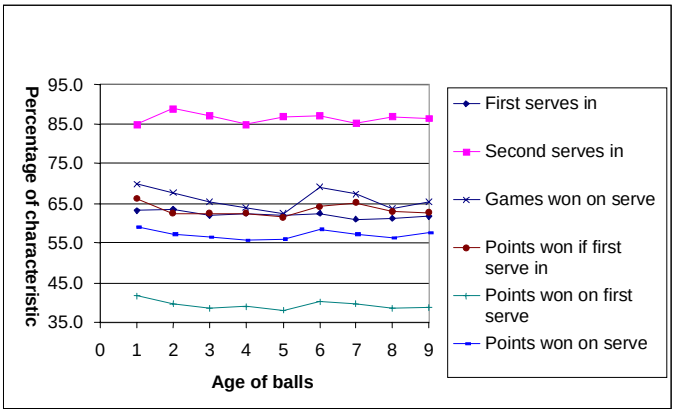


Figure 2: Service characteristics for the 2000 Australian Women's Singles Championships depending on the age of the balls.

3 Comparison of tiebreaks and non-tiebreak games

Tiebreaks were introduced to reduce the chance of extremely long matches. All the grand slam events have tiebreaks in the non-final sets, and the US open has a tiebreak in the final set. At the Australian Open in 2001, a single tiebreak game replaced the third set in the Open Mixed Doubles. (In 2002, this was extended to a super tiebreak—first pair to win ten points by two clear.) While tiebreaks remove the possibility of a long advantage set, they increase the element of luck, and increase the importance of points played. At 5–5 in a tiebreak, both players are within two points of winning the set, and every second point will be a set point. This may cause players to play differently to a normal game. In addition, players are serving in groups of two points, rather than complete games. This may also change their usual serving pattern.

Table 3 compares the serving characteristics of men's and women's tennis in normal games and

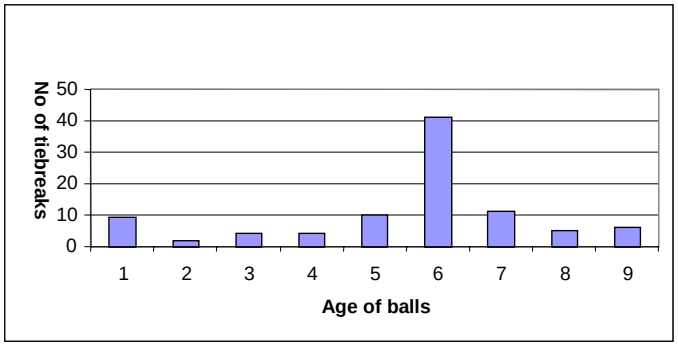


Figure 3: Age of balls in men's singles tiebreak games, Australian Open, 2000.

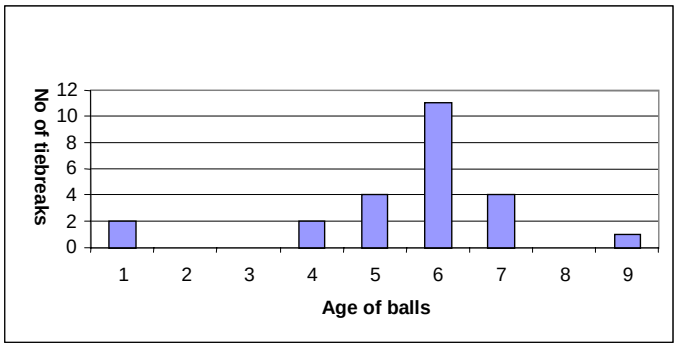


Figure 4: Age of balls in women's singles tiebreak games, Australian Open, 2000.

tiebreaks.

Both men and women appear to perform worse on service in tiebreaks than in normal games. Men appear to serve more conservatively in the tiebreak, with a greater proportion of first and second serves in, a smaller proportion of points won if the first serve is in (p-value < 0.001), a smaller proportion of aces (p-value 0.013) and fewer double faults than for normal games. But the overall result is that the percentage of points won on serve goes down, although not significantly. The picture is similar for women, although the percentage of first serves in actually decreases in the tiebreak. Again, overall the

	Percentage of the characteristics			
	Men		Women	
Characteristic	Normal games	Tiebreaks	Normal games	Tiebreaks
First serve in	57.6 (0.3)	58.6 (1.5)	62.2 (0.4)	58.3 (3.0)
Second serves in	89.2 (0.3)	90.5 (1.4)	86.3 (0.4)	90.5 (2.7)
Aces	8.5 (0.2)	6.2 (0.7)	3.3 (0.1)	1.1 (0.6)
Double faults	4.6 (0.1)	3.9 (0.6)	5.2 (0.2	4.0 (1.2)
Points won if first serve in	73.5 (0.3)	69.4 (1.8)	63.2 (0.5)	60.5 (3.8)
Points won on first serve	42.4 (0.3)	40.7 (1.5)	39.4 (0.4)	35.3 (2.9)
Points won on serve	63.9 (0.3)	62.1 (1.5)	57.0 (0.4)	55.4 (3.0)

Table 3: Serving statistics for Men's and Women's Singles, Australian Open, 2000.

Tournament	Percentage of points won on serve	
	Men	Women
Wimbledon 1992	64.9 (0.4)	57.0 (0.6)
Wimbledon 1993	64.9 (0.4)	56.6 (0.6)
Wimbledon 1994	63.9 (0.4)	55.4 (0.6)
Wimbledon 1995	64.0 (0.4)	55.4 (0.6)
Wimbledon 2001	64.5 (0.3)	57.1 (0.4)
Australian Open 2000	63.8 (0.3)	57.0 (0.4)
Australian Open 2001	61.9 (0.3)	54.9 (0.4)
Australian Open 2002	61.7 (0.3)	54.4 (0.4)
French Open 2001	60.1 (0.3)	54.1 (0.4)

Table 4: A comparison of points won on service in singles at selected grand slam tournaments.

percentage of points won on serve goes down, although not significantly.

4 Statistical comparison of grand slam tournaments

Tennis player's styles usually lie between two extremes. The serve volleyer attacks the net following a big serve, and likes a fast court. Wimbledon, played on grass, is said to suit a serve–volleyer. The baseliner stays back, has long rallies, and is suited to a slower court, where the serve is not such a big advantage. The French Open, played on clay, is usually won by baseline players. The Australian Open is played on hardcourt, and is said to be relatively even handed to baseliners and serve–volleyers. A common point of discussion is the varying surfaces of the grand slam tournaments and the relative advantages they give to the server.

Table 4 summarises the percentage of points won on service at three grand slam tournaments. The data for Wimbledon 1992–1995 have been taken from Magnus and Klaassen [1], and the other data were extracted from the websites www.wimbledon.com, www.ausopen.org, and www.rolandgarros.com. The US Open does not take statistics on every court, and comparable data are not available.

The figures for Wimbledon are fairly consistent, although there appears to be a smaller percentage of points won on serve in 1994 and 1995. Figures for the Australian Open in 2000 were similar to those for Wimbledon, but there has been a significant drop in the percentage of points won on serve since 2000. As would be expected, the percentage of points won on serve at the French Open is less than that for the other two grand slams, although the figures for women at the Australian Open in both 2001 and 2002 are not significantly different from the figure for the French Open in 2001.

Table 5 gives the serving characteristics for Wimbledon, 1992–1995, and the Australian Open, 2000. For the men's singles, at Wimbledon, a significantly larger proportion of first serves were in, a smaller proportion of second serves were in, a larger proportion of double faults were served and a larger proportion of points were won on first serves, than at the Australian Open. For the women, almost the opposite is true. At Wimbledon, a smaller proportion of first serves were in, and a smaller proportion of points were won on the first serve than at the Australian Open.

5 A first look at doubles

Table 6 gives an overview of some statistics related to doubles for the Australian Open Tennis Championships 2001, obtained from data available from the Australian Open website www.ausopen.org. Figures in brackets are standard errors. Matches in men's and women's doubles were the best of three sets, first two tiebreak and third advantage, except for the men's doubles final, which was the best of five sets. Matches in mixed doubles were the best of three sets, first two tiebreak, with the third set consisting

Characteristic	Percentage of the characteristics Men Wimbledon 1992–1995	Men Australian Open 2000	Women Wimbledon 1992–1995	Women Australian Open 2000
First serves in	59.4 (0.2)	57.6 (0.3)	60.8 (0.3)	62.1 (0.4)
Second serves in	86.4 (0.2)	89.2 (0.3)	86.0 (0.3)	86.4 (0.4)
Games won on serve	80.8 (0.4)	79.9 (0.6)	63.4 (0.7)	65.9 (0.9)
Aces	8.2 (0.1)	8.4 (0.2)	3.1 (0.1)	3.3 (0.1)
Double faults	5.5 (0.1)	4.6 (0.1)	5.5 (0.1)	5.2 (0.2)
Points won if first serve in	73.3 (0.2)	73.4 (0.3)	62.2 (0.4)	63.2 (0.5)
Points won on first serve	43.6 (0.2)	42.3 (0.3)	37.8 (0.3)	39.3 (0.4)
Points won on serve	64.4 (0.2)	63.8 (0.3)	56.1 (0.3)	57.0 (0.4)

Table 5: Summary of men's and women's singles service characteristics for Wimbledon 1992–1995 and the Australian Open 2000.

of a single tiebreak game. Unfortunately the statistics of seventeen of the ladies doubles matches are missing, but an estimate of the overall statistics based on simple proportions has been given. There is a fairly good agreement between men's and mixed doubles in terms of statistics. There are significant differences between these and women's doubles in the percentage of aces, percentage of first serve points won, percentage of second serve points won and overall percentage of points won on serve. However, the percentage of tiebreak sets played in the first two sets is similar in women's doubles and mixed doubles. Generally, the percentage of points won on serve is higher in the mixed, which leads to a larger average number of points played in a tiebreak and a smaller number of points in a normal game.

In comparison with singles events (2000 data), the men get significantly more first serves in in doubles, but serve half the number of aces, and win less points if the first serve is in. The same differences show between women's singles and doubles, but are not significant. In both men's and women's doubles (in 2001), there is a higher proportion of points won on serve compared to the corresponding singles events (in 2001), but this difference is not significant.

6 Conclusion

There is now a range of data available for analysing tennis. This allows some of the folklore of tennis to be placed under the statistical microscope. Here we have shown the supposed beneficial effects of new balls to the server are largely illusory. However, there is evidence to suggest players alter their serving patterns during tiebreaks. The data can also be used to compare the different tournaments. Such comparisons highlight the difference in men's and women's tennis, particularly in relation to the performance on serve. When comparing Wimbledon and the Australian Open, the percentage of first serves in, games won on service, points won on serve and points won on first serve all move the opposite way for men than for women in going from one tournament to the other. As expected, the percentage of points won on serve is lower in the French Open tournament than the Australian Open, but the magnitude of the difference is not great, particularly for women's tennis. We have also looked at some doubles data. As the less glamorous events, doubles has received little attention in the literature, yet there is much potential gain from its study. Mixed doubles data in particular provides the opportunity to calibrate some of the differences in standard between the men's and women's game.

However, there is still some progress to be made in ensuring a greater degree of uniformity of data collection and dissemination at the various grand slam tournaments, and even between events at the same tournament. It would also improve the standard of analysis if more of the point-by-point data were released as a matter of course. This would allow a deeper analysis. Topics such as differences in the backhand and forehand sides, independence of points, could then be investigated.

Characteristic	Men's doubles[1]	Women's doubles	Mixed doubles[2]
Number of matches	63	63	31
Number of 4-set matches	1		
Number of 3-set matches	29	25	9
Number of 2-set matches	32	38	21
Number of sets	155	151	60**
Number of tiebreaks	28	9	11 (first two sets) 9 (third set)
Total games	1574	1380	599
Average no. of games per set (excluding third set of mixed)	10.2	9.1	9.8
Average no. of points in tiebreak	11.9	11.2	13.0
Average no. of points in a game (excluding tiebreaks)	6.3	6.6	5.9
Server points won	6295	3841*	2271
Total points	10002	6936*	3603
First serves in	6452	4588*	2314
% first serves in	64.5	66.1*	64.2
Total aces	375	159*	151
% aces	3.7 (0.2)	2.3 (0.2)*	4.2 (0.3)
Total double faults	429	309*	158
% double faults	4.3 (0.2)	4.5 (0.2)*	4.4 (0.3)
1st serve points won	4508	2805*	1624
% 1st serve points won	69.9 (0.6)	61.1 (0.7)*	70.2 (1.0)
% 2nd serve points won	50.3 (0.8)	44.1 (1.0)*	50.2 (1.4)
% points won on serve	62.9 (0.5)	55.4 (0.6)*	63.0 (0.8)

[1]one walkover. [2]one walkover, two matches missing statistics.
*Estimates only: 17 matches missing statistics. **Does not include third set tiebreaks.

Table 6: A comparison of doubles events at the 2001 Australian Open Tennis Championships.

References

[1] S. R. Clarke and P. Norton, "Collecting statistics at the Australian Open Tennis Championships", these *Proceedings*.

[2] J. R. Magnus and J. G. M. Klaassen, "The effect of new balls in tennis: four years at Wimbledon", *The Statist.*, **48** (1999), 239–246.

[3] J. R. Magnus and J. G. M. Klaassen, "On the advantage of serving first in a tennis set: four years at Wimbledon", *The Statist.*, **48** (1999), 247–256.

[4] G. H. Pollard, "An analysis of classical and tie-breaker tennis", *Austral. J. Statist.*, **25** (1983), 496–505.

OPTIMISATION TOOLS FOR ROUND-ROBIN AND PARTIAL ROUND-ROBIN SPORTING FIXTURES

David Panton and Kylie Bryant
Centre for Industrial and Applied Mathematics
University of South Australia
Mawson Lakes
South Australia 5095, Australia
david.panton@unisa.edu.au

Jan Schreuder
Department of Applied Mathematics
University of Twente
The Netherlands
j.a.m.schreuder@math.utwente.nl

Abstract

The creation of good match schedules for a sporting league is a combinatorially complex task. Constraints relating to equity, the media, crowd safety, and team and ground requirements make this a difficult task. The process we will discuss in this paper for the creation of round-robin and partial round-robin match schedules involves three steps. The first is to create home and away pattern sets providing a pattern of home and away games for each team. Next is the creation of basic match schedules, where games are assigned to the pattern sets. Finally teams are assigned to the rows in the basic match schedule. This creates several possible solutions, one of which is selected by the league. Integer programming optimisation models and solution processes for each step will be discussed. Examples of applications to the National Soccer League, the Australian Football League and the South Australian National Football league will also be discussed.

1 Introduction

Schedules for sporting leagues come under a lot of scrutiny from teams and fans. It is important that the schedules are fair for all teams. Teams and the leagues they are a part of also have many constraints on the schedules, such as maintaining or increasing gate receipts, and having traditional games always played in the same round each year. The media can also have an impact on schedules since their objective is to have a large audience which requires games between popular teams. There are many different groups that place constraints on the schedule and the problem of finding a good schedule can be difficult.

At present, many sporting schedules are created by hand. The advantage of this is that many constraints can be met by dealing with them first. Traditional games can be fixed initially as well as other team preferences. The problem comes when the fairness and quality of the schedule is analysed. Situations such as a team having to visit three different states in three consecutive rounds, or playing a majority of their away games in the last half of the season arise and are inequitable. These situations would not be so bad if all teams had similar schedules. However, it often happens that while one team

has a situation such as the ones described above, other teams may always alternate their home and away games, or only have to travel interstate once or twice during the entire season. Unfortunately, a fair schedule is difficult to create by hand. It can also be difficult to meet all constraints.

Many different mathematical approaches to the problem of constructing match schedules have been explored. It has generally been accepted that a good way to solve the problem is by splitting it into smaller parts. An example of this is to decide first whether each team plays at home or away in each round and assign games later. Methods such as this have been used to schedule the Dutch Football League [9], the Atlantic Coast Conference for basketball in the USA [6] and major [1] and minor [8] league baseball. However there are also other approaches that have been successful, [2]. The connection between timetables and graph theory has provided some useful theorems on the existence of timetables, [5]. Constraint programming has been used more recently to solve the problem and has proved successful, [4]. Due to the diversity of leagues and competitions, one method may not be suitable for all. This paper describes a three step process to generate high quality sporting schedules. In Section 2 we discuss the generation of patterns and pattern sets which are used as a basis for the schedule. Models for generation of quality pattern sets are discussed in this section. In Section 3 we discuss models for assigning team numbers to the home and away games specified in the pattern sets. This results in a Basic Match Schedule (BMS) or timetable. In Section 4 we present a model for the assignment of teams to the team numbers, based on team preferences and requirements. Section 5 discusses the concept of carry over which impinges on the quality of a match schedule, and finally in Section 6 we discuss examples taken from the National Soccer League, the Australian Football League (AFL) and the South Australian National Football league (SANFL). Issues relating to the creation of double and partial round-robin schedules are also discussed in this section.

2 Generation of pattern sets

2.1 Home and away patterns

The first step in the process of creating a match fixture schedule is to generate appropriate home and away patterns for each team over the entire season. Initially teams are regarded as anonymous and it is not until the last step in the process (see Section 4) that actual teams are assigned to a number. We assume that there are an even number of teams given by $2n$. When this is not the case in practice, an extra team is created, and matches against this team represent a bye. A *round-robin* tournament is one in which $2n-1$ rounds are played with each team playing each other team exactly once either at home or away. In a *double round-robin* competition the $2n-1$ rounds are repeated but with the home or away venues reversed, giving a total of $4n-2$ rounds altogether. The actual number of rounds played in a competition is dependent on a number of factors, which include seasonal time limits, revenue generation, and the physical nature of the sport. For example in a 16-team soccer competition a double round-robin tournament consisting of 30 rounds may be appropriate. In the first half of the season each team plays each other team once (15 rounds) in a series of home and away games. In the second half of the season these games are repeated but with the home and way components reversed. Such a schedule of 30 home and way games would not be feasible in certain other sports however. For example in the AFL competition, the physical nature of the sport would prevent this number of games being played, and in their 16-team competition only 22 rounds are played. We call such a competition in which only a fraction of one complete set of rounds is played a *partial round-robin* competition.

A *home and away pattern* involving $2n$ teams is a sequence of $2n-1$ 0's or 1's, where a 0 represents an away game and a 1 represents a home game. Certain restrictions apply on the structure of a feasible pattern. For a round-robin competition involving $2n$ teams it is usual to require either $n-1$ or n home games in the pattern. This will then balance out to exactly $2n-1$ home games in a double round robin competition. In a partial round-robin competition the total number of rounds is normally chosen as an even number to enable an exact balance of home and away games. Appropriate patterns must be generated to ensure this. In addition to these restrictions, patterns will be subject to certain quality constraints. We call a *break* an unbroken sequence of either home or away games, for example 1 1 1

or 0 0. In the first sequence the break is of length 2, whereas in the second sequence the break is of length 1. Too many breaks, or breaks that are too long are seen as undesirable. Note that a break of home games of length 2 which might be seen as desirable in the first half of the schedule becomes a break of away games of length 2 in the second half of the schedule for a double round-robin competition. Breaks are unavoidable for any competition of realistic size in which an even number of teams are involved. When an odd number of teams are involved it is possible to construct pattern sets in which no breaks occur ([5]). For an even number of teams only two distinct patterns are possible without breaks. To illustrate, consider a 16 team competition for the first 15 rounds. The only two possible patterns without breaks are

$$1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1 \quad \text{and} \quad 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1\ 0.$$

It is a straightforward process to generate all patterns with only one break of length 1, i.e. either 1 1 or 0 0. For a double round-robin competition we need only generate patterns for the first $2n-1$ rounds. Patterns with only one break of length 1 in the first $2n-1$ rounds will always result in three breaks of length 1 when put together as a double round-robin pattern however, since an additional break is unavoidable at the interface. In fact certain patterns, notably those commencing with either 1 1 or 0 0 must be avoided to prevent breaks of length 2 at the interface, since in these cases patterns will either end in 0 or 1 respectively. Thus patterns in a full double round-robin competition will have either no breaks or at least 3 breaks of length 1 over the $4n-2$ rounds. While it is desirable to keep the number and size of the breaks as small as possible, other factors may influence what patterns are admissible. These will be discussed in a later section.

For partial round-robin competitions the situation is slightly more interesting since balancing the number of home and away games is not guaranteed. We will again use the 16 team, 22 round competition to illustrate. If the first seven rounds in the 15 round pattern are to be used again in their complementary form to make up rounds 16 to 22, each 0 and 1 in this section must be counted as a home game. Thus to balance the competition with exactly 11 home games we must ensure that there are exactly four home games in rounds 8 to 15. Patterns with no breaks in rounds 8 to 15 will automatically ensure this. Patterns with only one break of length 1 which occur in rounds 8 to 15 however will either have too few (10) or too many (12) home games. This can be avoided by allowing a break of length 1 to commence only in the first seven rounds. These patterns will as before have three breaks of length 1 over 22 rounds. An alternative and potentially better quality option is available in this case however. If no breaks are allowed in the first seven rounds but two breaks of length 1 are allowed in rounds 8 to 15 then patterns with only two breaks in 22 rounds can be generated. It is straightforward to generalise to any partial round-robin competition which contains an even number of rounds.

2.2 Pattern sets

Patterns are used as building blocks for creating pattern sets. A pattern set is the framework of home and away matches on which the complete match fixture schedule is built. For a competition having $2n$ teams a pattern set is a $2n \times (2n-1)$ array in which each row in the array is a feasible pattern as determined by the competition requirements. Necessary conditions for pattern sets are:

(i) each row in the array must contain either $n-1$ or n 1's;

(ii) each column in the array must have exactly n 1's and therefore exactly n 0's since each column represents a round in which a balanced number of home and away games are played;

(iii) patterns cannot be duplicated since this prevents the two teams associated with these patterns from playing each other at all (either at home or away) in any round.

In addition, other desirable features of the pattern set are:

(iv) there are exactly two rows with no breaks;

(v) there are at least two pairs of rows which are complementary.

Condition (iv) may be important when scheduling matches between pairs of teams who cannot both play at home in the same round. This situation will be revisited in a later section.

A graph-theoretic analysis of pattern sets can be found in [5]. Optimisation models are well documented for creating pattern sets (see for example [6]). We describe a set partitioning model which can be dynamically used to generate all feasible pattern sets for a given basis set of patterns P.

Let

$$x_i = \begin{cases} 1, & \text{if pattern } i \text{ is chosen,} \\ 0, & \text{otherwise,} \end{cases}$$

where $i \in P$. Let c_i be the cost of each pattern in P. Costing can be based on the number and length of breaks in each pattern. Then we wish to

$$\text{minimise} \quad \sum_i c_i x_i \tag{1}$$

$$\text{subject to} \quad \sum_i a_{ki} x_i = n, \quad \text{for all } k = 1, \ \ldots, \ 2n-1, \tag{2}$$

$$\text{and} \quad \sum_i x_i = 2n, \tag{3}$$

where k is an index for the rounds and $a_{ki} = 1$ if pattern i has a 1 in round k. The objective is to create the best quality pattern set. Constraint (2) ensures that each column in the pattern set has exactly n home games, while constraint (3) ensures that there are exactly $2n$ rows chosen in the pattern set. In general there are several possible pattern sets for a given set of patterns P. If we let S' be the current pattern set, then addition of the constraint

$$\sum_{j \in S'} x_j < 2n \tag{4}$$

will prevent this pattern set from recurring. These constraints are progressively added to the model until no feasible solution is found, in which case all feasible pattern sets have been generated.

An example of a pattern set for six teams and five rounds is shown in Table 1.

Team	Round				
1	1	0	1	1	0
2	0	1	0	0	1
3	1	0	1	0	1
4	0	0	1	0	1
5	1	1	0	1	0
6	0	1	0	1	0

Table 1: Home and away pattern set for six teams and five rounds.

3 Creating the basic match schedule

Stage 2 in the generation of round-robin tournaments is to construct the Basic Match Schedule (BMS), often referred to as the *timetable*, in which team numbers (but not team names) are assigned to a pattern set, together with all matching home and way games. An example of a BMS for the pattern set given in Table 1 is shown in Table 2.

Note that in this BMS there are exactly three home and three away games in each round, and that for each placeholder team there are either two or three games played at home in five rounds.

Team	Round				
1	+2	−4	+6	+3	−5
2	−1	+3	−5	−6	+4
3	+5	−2	+4	−1	+6
4	−3	−6	+2	−5	+1
5	+6	+1	−3	+4	−2
6	−4	+5	−1	+2	−3

Table 2: A Basic Match Schedule based on the pattern set in Table 1. A + denotes a home game and a − denotes an away game.

The process of withholding the identity of the teams at this stage has been used in several cases (see for example [6], [9]). It is argued that this approach allows for a more flexible assignment of teams in the last stage, taking into account factors which are hard to define in terms of hard constraints. A contrary view is expressed by Henz in [4] who argues that many such constraints may already be required to generate suitable pattern sets and basic match schedules and that the identity of the teams relating to these constraints may be lost in the final stage.

We create the basic match schedule using the optimisation model described below. Consider a pattern set derived from the first stage, having $2n$ rows and $2n-1$ columns. This pattern set is stored as the array (p_{ik}) where i is an index for each pattern or placeholder team and k is an index for each round. We now define

$$a_{ijk} = \begin{cases} 1, & \text{if } p_{ik} = 1 \text{ and } p_{jk} = 0,\ i \neq j, \\ 0, & \text{otherwise,} \end{cases}$$

with

$$x_{ijk} = \begin{cases} 1, & \text{if } a_{ijk} = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Then

$$\text{minimise} \quad \sum_{i,j,k} a_{ijk}\, x_{ijk}$$

$$\text{subject to} \quad \sum_{k} (a_{ijk}\, x_{ijk} + a_{jik}\ x_{jik}) = 1, \quad \text{for all } i < j, \tag{5}$$

$$\sum_{i} a_{ijk}\, x_{ijk} = 1, \quad \text{for all } j,\ k, \tag{6}$$

$$\sum_{j} a_{ijk}\, x_{ijk} = 1, \quad \text{for all } i,\ k. \tag{7}$$

This model provides solutions for the first $2n-1$ rounds. Their mirror or complementary solutions may be used for the next $2n-1$ rounds in a double round-robin tournament. Constraint (5) ensures that in the first k rounds, teams i and j play each other once, either at home or away. Constraint (6) applies only to those teams j which have a home game in round k. For these teams, this constraint ensures that team j plays only one home game in round k. Constraint (7) applies only to those teams which have an away game in round k. For these teams, this constraint ensures that team i plays only one away game in round k. The objective is superfluous in this model since we are seeking as many feasible solutions as possible. For each pattern set, once a feasible solution is found, we add a constraint which prevents this solution from recurring, and continue until all feasible solutions have been found. It is interesting to note that a feasible BMS cannot always be found for every pattern set. Necessary and sufficient conditions for the existence of a BMS are unknown at this stage. There are a relatively

small number of pattern sets admitting feasible solutions, as shown in Table 3. However, in practice, the number of alternative schedules for each feasible pattern set provides a large number of BMS's for all practical sized problems.

Number of teams	Number of pattern sets	Number of feasible pattern sets
4	0	0
6	1	1
8	4	2
10	15	4
12	56	10
14	207	17
16	752	38

Table 3: The number of pattern sets and those which admit feasible BMS's, under the assumption that there are no breaks of length 3 or longer and at most two breaks of length 2 in a double round-robin competition.

4 Assigning teams to the Basic Match Schedule

At this stage one or more Basic Match Schedules have been chosen from the many that have been generated by the process described above. These are typically chosen by the competition organisers and are then given to the individual teams who are asked to rank the team numbers (identified by the rows in the BMS) in their order of preference. Team preferences come into account at this point since a team will not choose a team number which is in conflict with their preferences. For example, team A may wish to play at home in a certain round and hence will not rank a team number which plays away in that round. It may not be possible for every team's preferences to be accommodated and it may be necessary for some teams to have only some preferences met. In addition, it may be possible to meet more preferences by using a different BMS.

A factor that might influence team ranking is safety constraints that dictate when certain teams can or cannot play each other. This is more of a factor with soccer schedules in Europe. It may also be necessary for pairs of teams to be matched in a complementary sense, where team A and team B cannot both play at home in the same round. This may occur for example if the two teams use the same ground, or if gate receipts are an issue. Thus for example teams 1 and 2, 3 and 6, or 4 and 5 are complementary in Table 2. It may be necessary to have more than one complementary pair, and not every BMS will accommodate this.

At this point we have at least three possible ways of assigning teams to the row numbers in the BMS. A common choice is to do this manually since in many situations there is a complex set of constraints that are to be met. An alternative used in [6] is to enumerate all possible permutations of teams and to check each one against the constraints. While this may be possible with ten or fewer teams, it becomes impractical beyond this number. The method we describe employs an optimisation model. All preferences are set up as shown in Table 4.

Team preferences are grouped in clusters. For example team A prefers only lines 2 or 4, while teams B and C together must have either lines 1 and 3 or lines 2 and 4. These may be complementary lines, for example. Finally teams D and A together must have either lines 1 and 2 or lines 1 and 3. Each column represents a possible assignment and can be costed according to preferences. We use a cost of zero for the most preferred team number(s) with increasing cost for the rest. It is desirable to minimise the cost, giving the objective

$$\text{minimise} \qquad \sum_i c_i x_i \tag{8}$$

	j	Cluster						
	i	A		B + C		D + A		Sum
BMS	1			1		1	1	1
	2	1			1	1		1
	3			1			1	1
	4		1		1			1
Club	A	1	1			1	1	1
	B			1	1			1
	C			1	1			1
	D					1	1	1

Table 4: Clusters for team assignment

where c_i are the costs of each team number for each team. Clearly each team can receive only one team number, and each team number can only be assigned to one team. This means that as shown in Table 4 we want each row to sum to 1, giving the constraint

$$\sum_j a_{ij} x_j = 1, \quad \text{for all } i, \tag{9}$$

where a_{ij} is 1 if cluster j uses team number i and 0 otherwise. In order to ensure that each team is assigned to only one team number the lower part of Table 4 includes a row for each team (A, B, C, D) with a 1 in each column representing a team number combination. Again, these must sum to 1 as shown in the lower half of this table. This can be ensured using the constraint

$$\sum_j b_{ij} x_j = 1, \quad \text{for all } i, \tag{10}$$

where b_{ij} is 1 if team i is in cluster j and x_j is 1 if column j is chosen and 0 otherwise.

5 Minimising carry over

Each team has an effect on each other team by playing against that team. It is reasonable to expect that a team can be demoralised by playing against a hard team or spurred on by playing against an easy team. The effect on each team is balanced out in a round-robin tournament since each team plays each other team the same number of times. However, there is also an effect that is passed on to the team which is played the following week. For example if team A plays against team B in round 1 and then team A plays against team C in round 2, team C will receive an effect from team B. If team B is a hard team, it is likely that team A would lose that game. Thus, when team A plays against team C, team A may have a diminished chance of winning. On the other hand if team B is an easy team, team A may have an enhanced chance of winning against team C. Some competitions require that no team will play two hard teams in a row, and will decide which teams are rated as hard. Nemhauser and Trick [6] faced this situation while devising a schedule for the Atlantic Coast Conference Basketball competition.

The carry over effect that is of concern here occurs when more than one team plays two particular teams consecutively (see [3]). Once again, if team A plays against teams B and C consecutively and team D also plays teams B and C consecutively, then team C receives two effects from team B. Carry over effects of this sort are most obvious when there is an odd number of teams and therefore there is a bye each round. If team A has a bye and then plays team C, and team D has a bye and then plays team C, then team C may experience some carry over effect (although there may be some debate as to whether this is positive or negative). The carry over effect due to a team is not quite so obvious but still

effective. Table 5 shows a six-team five-round competition with teams 1, 2 and 6 all playing teams 5 and 3 consecutively. The effect in this case is on team 3. It has to play three different teams immediately after these teams have played team 5. If team 5 is hard, team 3 has an advantage, whereas if team 5 is easy, team 3 has a disadvantage.

Team	Round				
1	−6	**+5**	**−3**	−4	+2
2	−4	+6	**−5**	**+3**	−1
3	−5	+4	+1	−2	+6
4	+2	−3	−6	+1	−5
5	+3	−1	+2	−6	+4
6	+1	−2	+4	**+5**	**−3**

Table 5: A Basic Match Schedule showing the carry over effect in bold for team 3.

It is theoretically possible to add constraints to the BMS model to prevent this carry over effect from occurring. However, for a 14-team competition this requires the addition of the order of 10^6 linear constraints. It is very likely that no feasible solution would be found, requiring the weakening of these constraints in a progressive manner. A far more practical approach is to identify carry overs in the generation of alternative BMS's and use only those BMS's which have a relatively low total carry over score. It has been proven by Russel [7] that a BMS with zero carry over is possible when the number of teams is a power of 2. Russel also gives a method for construction of a schedule in this case.

6 Application to round-robin and partial round-robin tournaments

Three examples will be discussed in this section. The first is a competition containing 14 teams in a double round-robin. In this case the application of the processes discussed above is relatively straightforward. The second and third examples are partial round-robins in which local rules and historical constraints play a major role in the match schedule construction process. These examples are based on preliminary work done for Soccer Australia, and the SANFL. This work precipitated interest in AFL match schedules; however, no official contact has been made with this organisation.

6.1 Double round-robin schedule

Consider a competition with 14 teams in which each team plays each other twice, giving a total of 26 rounds. A number of basic match schedules can be generated in this case. One example is given in Table 6 with only the first 13 rounds displayed. The second half of this schedule is a mirror image of the first with home and away games reversed, and taking this into account there are a total of 18 away breaks and 18 home breaks. There are no breaks of length 2 or longer. The number of carry overs in this case is 76.

The assignment of teams to the team placeholder positions could be done manually, or using the model described above.

6.2 A partial round-robin with an even number of teams

In the official current (2002) season match schedule for the AFL there are a total of 47 away breaks; 49 home breaks; ten breaks of length 2 and a reasonable number of carry overs. We have attempted to create an alternative AFL schedule. Of the 50 pattern sets generated, not all gave feasible BMS's, but a sufficient number of quality usable BMS's were available. To be usable it is necessary that there

Team	1	2	3	4	5	6	7	8	9	10	11	12	13
1	−9	+8	−11	+12	−7	+6	**−5**	**−2**	+10	−4	+3	−13	**+14**
2	−14	+13	−12	+11	−10	+7	−6	+1	**−5**	**−3**	+9	−4	**+8**
3	−13	+9	−14	+10	−11	+12	−7	+5	−6	+2	**−1**	**−8**	**+4**
4	−8	+10	−9	+14	−12	+13	−11	+6	−7	+1	−5	+2	−3
5	−12	+14	−13	+7	−6	**+11**	**+1**	−3	+2	−8	+4	−10	**+9**
6	−10	+11	−7	**+13**	**+5**	−1	+2	−4	+3	−9	+8	−14	**+12**
7	−11	**+12**	**+6**	−5	+1	−2	+3	−8	+4	−14	+10	−9	**+13**
8	+4	**−1**	**−10**	+9	−13	+14	−12	+7	−11	+5	−6	+3	**−2**
9	+1	−3	+4	**−8**	**−14**	+10	−13	+11	−12	+6	−2	+7	**−5**
10	+6	−4	+8	−3	+2	**−9**	**−14**	+13	−1	+12	−7	+5	**−11**
11	+7	−6	+1	−2	+3	−5	+4	−9	+8	−13	+14	−12	+10
12	+5	−7	+2	−1	+4	−3	+8	−14	+9	−10	**+13**	**+11**	**−6**
13	+3	−2	+5	−6	+8	−4	+9	−10	**+14**	**+11**	−12	+1	**−7**
14	+2	−5	+3	−4	+9	−8	**+10**	**+12**	−13	+7	−11	+6	**−1**

Table 6: First half of a double round-robin competition for 14 teams. Breaks are shown in bold.

are at least two complementary pairs of teams to accommodate the Adelaide Crows and Port Power, and the West Coast Eagles and Fremantle. This requirement relates to revenue generation, since it is not desirable for both local teams to play at home in the same round. In addition it is essential that these pairs of teams play each other in the first seven rounds since these rounds must be repeated (in a complementary sense) for the last seven rounds. Compliance with these conditions alone, at the same time maintaining a quality schedule, is a nontrivial task if a manual match schedule construction process is being used. Using the techniques discussed in Section 2, we have generated the match schedule shown in Tables 7 and 8. This schedule has only 16 away breaks; 16 home breaks, and no breaks of length 2 or higher. The total number of carry overs is 87, which is relatively good. Thus the total number of breaks has been dramatically reduced. It should be noted, however, that we have not had access to team preferences and other possible constraints which may affect the quality of the finished schedule.

Team	1	2	3	4	5	6	7	8	9	10	11
1	−7	+9	−4	−2	+15	−13	**+14**	**+11**	−6	+12	−3
2	−10	+11	−6	+1	−9	+4	−7	+3	−16	+15	−14
3	−5	+8	−10	+11	−6	+9	**−12**	**−2**	+7	−13	+1
4	−9	**+7**	**+1**	−14	+13	−2	**+15**	**+6**	−11	+8	−12
5	+3	**−15**	**−8**	+6	−10	+12	−11	**−16**	+13	−7	+9
6	+14	−13	+2	−5	+3	−16	+8	−4	+1	−10	+7
7	+1	−4	+13	−15	+16	−14	**+2**	**+12**	−3	+5	−6
8	+16	−3	**+5**	**+10**	−12	+11	**−6**	**−13**	+9	−4	+15
9	+4	−1	+14	−13	+2	−3	+16	**−15**	**−8**	+11	−5
10	+2	−16	+3	−8	+5	−15	+13	**−14**	**−12**	**+6**	**+11**
11	+15	−2	+16	−3	+14	−8	+5	−1	+4	**−9**	**−10**
12	+13	−14	+15	−16	+8	−5	+3	−7	+10	−1	+4
13	−12	+6	−7	+9	−4	+1	−10	+8	−5	+3	−16
14	−6	+12	−9	+4	−11	+7	−1	+10	−15	**+16**	**+2**
15	−11	+5	−12	+7	−1	+10	−4	**+9**	**+14**	**−2**	**−8**
16	−8	+10	−11	+12	−7	+6	−9	**+5**	**+2**	−14	+13

Table 7: First eleven games of a proposed AFL match schedule for 2002. Breaks are shown in bold.

Team	12	13	14	15	16	17	18	19	20	21	22
1	+10	−8	+16	−5	+7	−9	**+4**	**+2**	−15	+13	−14
2	+5	−13	+8	−12	+10	−11	+6	−1	+9	−4	+7
3	−15	+4	−14	**+16**	**+5**	−8	+10	−11	+6	−9	+12
4	+16	−3	+5	−10	+9	**−7**	**−1**	+14	−13	+2	−15
5	−2	+14	−4	+1	−3	**+15**	**+8**	−6	+10	−12	+11
6	−11	+12	−9	+15	−14	+13	−2	+5	−3	+16	−8
7	+8	−11	+10	**−9**	**−1**	+4	−13	+15	−16	+14	−2
8	−7	+1	−2	+14	−16	+3	**−5**	**−10**	+12	−11	+6
9	+12	−10	**+6**	**+7**	−4	+1	−14	+13	−2	+3	−16
10	−1	+9	−7	+4	−2	+16	−3	+8	−5	+15	−13
11	**+6**	**+7**	−12	+13	−15	+2	−16	+3	−14	+8	−5
12	**−9**	**−6**	**+11**	**+2**	−13	+14	−15	+16	−8	+5	−3
13	**+14**	**+2**	**−15**	**−11**	+12	−6	+7	−9	+4	−1	+10
14	**−13**	**−5**	+3	−8	+6	−12	+9	−4	+11	−7	+1
15	+3	−16	+13	−6	+11	−5	+12	−7	+1	−10	+4
16	−4	+15	**−1**	**−3**	+8	−10	+11	−12	+7	−6	+9

Table 8: Second eleven games of a proposed AFL match schedule for 2002. Breaks are shown in bold.

6.3 A partial round-robin with an odd number of teams

In the SANFL, nine teams play each other at least twice in a season. The SANFL program is a partial round-robin competition of 23 rounds, with the first 18 rounds making up a double round-robin. The additional five rounds create complications which do not occur in a simple double round-robin. For equity purposes it is essential that each team has the same number of byes in a season, and for this reason an additional four byes must be scheduled so that each team has exactly three. This is done by creating two extra byes in each of two rounds within the program, giving 27 byes in total between the nine teams. These must be strategically placed so that byes for teams do not occur too close together, and multiple bye rounds are not too close to either the start or end of the season. A further major constraint relates to the additional games played in the last five rounds since this requires that there are four teams that each team must play three times in a season. This must be done in such a way that a two year rotational schedule is satisfied, while at the same time balancing home and away conditions. Many other issues must also be addressed, for example the use of neutral grounds, availability of grounds and special team requests. In particular the Port Adelaide Magpies request that as far as possible, when Port Power are at home their games are away, and vice versa. Thus while the generation of a BMS using the techniques discussed in this paper provides a quality basis for the match program, a great deal of manual manipulation is required to create the final result.

7 Conclusions

In this paper we have discussed many of the major practical issues involved in generating a match program for a sporting competition where round-robin rules apply. A three step process is described in which IP optimisation models are used to create sporting schedules. While these optimisation models play an important part in the schedule generation process, our examples illustrate that there are many other factors which are difficult to formalise in a mathematical model which impinge on this process. Nevertheless we have shown that IP optimisation models can be very effective in creating basic pattern sets and Basic Match Schedules from which quality schedules can be developed.

References

[1] W. O. Cain, "A computer assisted heuristic approach used to schedule the major league baseball clubs", in *Optimal Strategies in Sports*, S. P. Ladany and R. E. Machol (editors), North Holland, Amsterdam (1977), 32–41.

[2] J. Cross, "Scheduling", in *Fourth Conference on Mathematics and Computers in Sport*, N. de Mestre and K. Kumar (editors), Bond University, Queensland, Australia (1998), 125–130.

[3] M. Henz, T. Muller, S. Thiel, and M. van Brandenburg, "Global constraints for round robin tournament scheduling", unpublished (2001).

[4] M. Henz, "Scheduling a major college basketball conference—revisited", *Oper. Res.*, **49** (2001), 163–168.

[5] D. de Werra, "Some models of graphs for scheduling sports competitions", *Discrete Appl. Math.*, **21** (1988), 47–65.

[6] G. L. Nemhauser and M. A. Trick, "Scheduling a major college basketball conference", *Oper. Res.*, **46** (1998), 1–8.

[7] K. G. Russel, "Balancing carry-over effects in round robin tournaments", *Biometrika*, **67** (1980), 127–131.

[8] R. A. Russell and J. M. Leung, "Devising a cost effective schedule for a baseball league", *Oper. Res.*, **42** (1994), 614–625.

[9] J. A. M. Schreuder, "Combinatorial aspects of construction of competition Dutch professional football leagues", *Discrete Appl. Math.*, **35** (1992), 301–312.

THE CHARACTERISTICS OF SOME NEW SCORING SYSTEMS IN TENNIS

Graham Pollard*
Division of Management and Technology
University of Canberra
ACT 2601, Australia
graham.pollard@ise.canberra.edu.au

Ken Noble*
Australian Bureau of Agricultural
and Resource Economics
GPO Box 1563, Canberra
ACT 2601, Australia
knoble@abare.gov.au

Abstract

The International Tennis Federation has recently approved a range of new tennis scoring systems. In this paper we report the effect of these changes on various measures such as the average number of points played in a match, the standard deviation of the number of points played, and the probability that each player wins. Other possible scoring systems are also considered. We identify those best-of-five short sets systems that are a reasonable alternative to the present best-of-three sets system. Further, the substantial reduction in the standard deviation of the number of points played by replacing the present third set by a tiebreak or super tiebreak game is reported.

1 Introduction

Recently the International Tennis Federation (ITF) approved several new tennis scoring systems for use in professional and amateur competition. This approval followed the belief that there may be some merit in replacing the present best-of-three sets system with a best-of-five short sets system. In these short sets the winner wins by 4 games to 0 games, 4–1, 4–2, 5–3, or 5–4 following a (first to seven points leading by at least two points) tiebreak game if the game score reaches 4–4. It can be seen that these short sets are equivalent to starting at 2–2 within the present set structure. Thus, the less important games at the beginning of the (present) set are removed, so play progresses more quickly to the most exciting part of the set (i.e. the end). Also, playing (possibly) five sets rather than three provides extra opportunities for fresh starts and additional opportunities for a swap in the lead. These were seen as attractive characteristics in moving to a best-of-five short sets structure.

The ITF also approved an alternative within the game structure. This alternative allows for the playing of only one point at deuce to determine the winner of the game (called "no ad" games). It was felt that this would reduce the number of points in a match and the variability in match duration. Also, an increase in the percentage of service breaks may not be a bad thing for the game.

Other ITF approved alternatives relate to the final set within a best-of-five or best-of-three sets structure. The final set can be a fifth or third set of the same type as the earlier ones, a regular tiebreak or super tiebreak game to replace the fifth or third set, or a "half-set" or "sudden death" set with possible game scores of 2–0, 3–1, or 3–2 following a tiebreak game at 2–2.

*This research was supported in part by the International Tennis Federation.

	Scoring system	Mean, μ	Standard deviation, σ	Probability better player loses, Q
	3S6	149.8	40.7	0.153
A1	5S4	164.2	45.0	0.140
A2	5S4, "no ad"	146.2	38.7	0.161
A3	5S4, "0–15"	144.2	40.6	0.143
A4	5S4, "0–15", "no ad"	123.5	33.3	0.169
A5	5S4, "40–0, 0–40"	150.1	42.1	0.146
A6	5S4, "0–15, 40–15"	134.6	38.5	0.147
A7	5S4, "30–0, 40–15"	128.8	36.7	0.158
B1	4S4, TB	156.7	36.9	0.162
B2	4S4, TB, "no ad"	139.6	31.7	0.182
B3	4S4, TB, "0–15"	137.9	34.0	0.165
B4	4S4, TB, "0–15", "no ad"	118.4	27.8	0.188
B5	4S4, TB, "40–0, 0–40"	143.5	35.0	0.169
B6	4S4, TB, "0–15, 40–15"	128.7	32.5	0.167
B7	4S4, TB, "30–0, 40–15"	123.3	31.0	0.177
C1	4S4, STB	157.8	37.9	0.156
C2	4S4, STB, "no ad"	140.8	32.7	0.175
C3	4S4, STB, "0–15"	139.1	35.1	0.158
C4	4S4, STB, "0–15", "no ad"	119.9	29.0	0.182
C5	4S4, STB, "40–0, 0–40"	144.8	36.0	0.164
C6	4S4, STB, "0–15, 40–15"	129.9	33.4	0.162
C7	4S4, STB, "30–0, 40–15"	124.6	32.2	0.174
D1	4S4, HS	159.9	40.2	0.152
D2	4S4, HS, "no ad"	142.3	34.6	0.173
D3	4S4, HS, "0–15"	140.1	36.3	0.155
D4	4S4, HS, "0–15", "no ad"	119.8	29.5	0.181
D5	4S4, HS, "40–0, 0–40"	146.2	37.8	0.156
D6	4S4, HS, "0–15, 40–15"	130.8	34.5	0.157
D7	4S4, HS, "30–0, 40–15"	125.5	33.5	0.167

Table 1: The characteristics of several alternative tennis scoring systems.

2 Best-of-five short sets alternatives to the present best-of-three sets system

A primary purpose of providing these alternatives within the general tennis scoring system framework was to provide a best-of-five short sets scoring system with a similar (or a little smaller) expected number of points played to the present best-of-three sets system. Other characteristics of the best-of-five short sets structure should be comparable to (or better than) the present best-of-three sets system. For example, the probability, Q, that the better player, player A, loses the match should be similar to its present value, certainly not much larger. Also, the standard deviation of the number of points played, should be no larger than at present, and preferably smaller. This is because we would ideally like the duration of tennis matches to be more predictable. Hence, the main characteristics of interest when considering a new tennis scoring system, are the expected number of points played, μ, the standard deviation of the number of points played, σ, and the probability that the better player wins, P. Note that the probability that the better player loses, Q, is equal to $1 - P$.

It is clear that the ITF approvals provide a range of best-of-five short sets alternatives to the present

best-of-three sets system. In this paper we consider the various scoring system alternatives with the view to seeing which have characteristics (μ, σ, Q) that are close to, or better than, the present system. We also consider some additional systems that may be of interest in the future.

These additional scoring systems make use of the following constructs.

(i) "0–15" tennis, in which the server starts each game at 0–15, or one point down. This concept is due to Miles [1]. It can be seen that, on the one hand, the server has the advantage of serving, while, on the other hand, has the disadvantage of being down a point. Miles demonstrated the efficiency that could be achieved by using such a construct. Note that it can be argued that "0-15" tennis is a smaller change to the scoring system than "no ad" tennis, because the score 0–15 regularly occurs at present, whereas players did not previously play only one point at deuce.

(ii) "40–0, 0–40" tennis, in which the server is declared to have won (lost) the game when the score reaches 40–0 (0–40).

(iii) "0–15, 40–15" tennis, in which the server starts each game at 0–15, and is declared the winner of the game if the point score reaches 40–15.

(iv) "30–0, 40–15" tennis, where the server is declared the winner of the game if the score reaches 30–0 or 40–15. We have considered such a system even though it is not suggested that it would be of interest to players or spectators.

These various constructs have been selected for consideration as they represent alternative ways of removing the less "important" points within a game of tennis (in which serving is a definite advantage; see Morris [2]). The efficiency of a scoring system is enhanced by making the importance of points less variable (see Pollard [4]). This can often be achieved by removing the less important points.

In order to find reasonable best-of-five short sets alternatives to the present best-of-three sets system, we need to make some assumption about the probability that a player wins a point when serving. The ITF noted that, over a large number of tournaments and court surfaces, touring men professionals won 76% of their service games in singles. We note that a player who has a probability of 0.612 of winning a point on service, has a probability of 0.76 of winning a service game, assuming points are independent. In order to compare the probability of each player winning a match for each of the various scoring systems, it is appropriate to assume one player (the better player, player A) has a point-probability on service greater than 0.612, and the other player, player B, a point-probability on service less than 0.612. We have assumed the values $p_a = 0.612 + 0.04 = 0.652$ and $p_b = 0.612 - 0.04 = 0.572$, respectively.

Table 1 gives the values for μ, σ and Q for various scoring systems when the probability player A wins a point on service is $p_a = 0.652$ and the probability player B wins a point on service is $p_b = 0.572$.

	Scoring system	Mean, μ	Standard deviation, σ	Probability better player loses, Q
	3S6	149.8	40.7	0.153
E1	2S6, TB	130.6	23.6	0.200
E2	2S6, TB, "no ad"	116.3	19.7	0.218
E3	2S6, TB, "0–15"	117.0	22.8	0.202
E6	2S6, TB, "0–15, 40–15"	109.0	22.1	0.202
F1	2S6, STB	132.7	24.5	0.194
F2	2S6, STB, "no ad"	118.4	20.9	0.207
F3	2S6, STB, "0–15"	119.0	24.0	0.193
F6	2S6, STB, "0–15, 40–15"	111.0	23.2	0.196

Table 2: The characteristics of several best-of-two sets tennis scoring systems.

Success rates on serve for players A, B: 0.6520, 0.5720.
Number of matches won by players A, B: 84693, 15307.
Proportion of service games won by players A, B: 0.833, 0.674.
Proportion of tiebreak games won by players A, B: 0.626, 0.374.
Two sets — number of occurrences: 62746.
Three sets — number of occurrences: 37254.

Duration in points has mean 149.80, standard deviation 40.67, median 142, mode 121.

skew measure 1 = 3 × (mean − median)/standard deviation = 0.58,
skew measure 2 = (mean − mode)/standard deviation = 0.71.

Frequency distribution for the 100,000 matches:

Number of points	Frequency
1–59	0
60–69	18
70–79	342
80–89	1791
90–99	5347
100–109	9243
110–119	11210
120–129	10880
130–139	9183
140–149	7781
150–159	6801
160–169	6262
170–179	5886
180–189	5797
190–199	5292
200–209	4493
210–219	3571
220–229	2601
230–239	1589
240–249	914
250–259	534
260–269	278
270–279	115
280–289	45
290–299	15
300–309	9
310–319	2
320–329	1
over 330	0

Table 3: Distributional characteristics of current best-of-three sets scoring system (3S6).

The notation 3S6 represents the present scoring system (the best-of-three tiebreak sets); 5S4 represents the best-of-five short sets (described earlier); 4S4, TB represents four short sets followed by the tiebreak game (first-to-seven points, leading by two points) if necessary; STB represents the super tiebreak game which has a first-to-ten points, leading by two points structure; and HS is the half-set described earlier.

The value of each (μ, σ, Q) in Table 1 is based on a simulation of 100,000 matches for each scoring system 3S6, A1, A2, ... By comparison with the exact results for 3S6 (see Pollard [3]), and by comparison with some simulations of 1,000,000 matches, it was ascertained that simulations of 100,000 matches gave sufficiently accurate results. Table 3 demonstrates the typical output of the simulations and provides additional detail for the current scoring system. The positive skew of the present system is noted. v

Success rates on serve for players A, B: 0.6520, 0.5720.
Number of matches won by players A, B: 80613,19387.
Proportion of service games won by players A, B: 0.832, 0.674.
Proportion of tiebreak games won by players A, B: 0.625, 0.375.
Proportion of super-tiebreak games won by players A, B: 0.646, 0.354.
Two sets — number of occurrences: 62662.
Three sets — number of occurrences: 37338.

Duration in points has mean 132.72, standard deviation 24.54, median 131, mode 130.

skew measure 1 = 3 × (mean − median)/standard deviation = 0.21,
skew measure 2 = (mean − mode)/standard deviation = 0.11.

Frequency distribution for the 100,000 matches:

Number of points	Frequency
1–59	0
60–69	17
70–79	332
80–89	1741
90–99	5422
100–109	10373
110–119	14271
120–129	15682
130–139	15220
140–149	12670
150–159	9717
160–169	6737
170–179	3874
180–189	2235
190–199	1056
200–209	446
210–219	151
220–229	41
230–239	14
240–249	1
over 250	0

Table 4: Distributional characteristics of best-of-two sets scoring system F1 (2S6, STB).

Some general observations can be made from Table 1.

(i) The use of the "no ad" construct reduces the expected number of points played by about 17 or 18 points, reduces the standard deviation by about 5 or 6 points, but increases Q by about 0.02 which is not an unsubstantial amount (about 13%).

(ii) The use of the "0–15" construct reduces μ by about 19 or 20 points, reduces σ by about three or four points, and increases Q only marginally by about 0.002 or 0.003. As this increase in Q is quite small compared to that in (i) above, the "0–15" construct would appear to be more useful than the "no ad" construct.

(iii) The use of the STB game rather than the TB game as the fifth set, increases both μ and σ by a little over one point. The STB game however, has a smaller increase in Q relative to the current system 3S6 (0.003 compared to 0.009). Thus the STB game would appear to be preferable to the TB game when used as a fifth set.

(iv) The use of the "40–15" stopping rule as an addition to the "0–15" starting rule decreases μ by about a further nine points and σ by about a further two points. The increase in Q is relatively small (0.002 to 0.004). Thus, in terms of μ, σ and Q considerations, this "0–15, 40–15" combination is a worthy contender.

In summary, the Table 1 scoring systems that have a decrease (or only quite a small increase) in Q as well as decreases in μ and σ (relative to 3S6) are those listed as A3, A6, A7, C3, D3 and D6 (D5 has only a small decrease in μ). If consideration was to be given to adopting any additional scoring systems, these could be included in those considered.

3 The best-of-three sets system

We now briefly consider a system which some people have called the "best-of-two sets" system and which has been used at important tournaments. This system involves playing two (tiebreak) sets, and if the set score reaches 1–1, a TB or STB game is played as the final set. This system has been used in mixed doubles at the Australian Open and in the Hopman Cup. In each case it replaced the best-of-three tiebreak sets system. In Table 2 we compare, for the same parameter values as in Table 1, these best-of-two sets systems within the current 3S6 system. Table 4 gives more detail for the (2S6, STB) structure. It can be seen from Table 2 that a substantial reduction in the standard deviation is achieved by using this best-of-two sets structure. There is however a substantial increase in Q, the probability that the better team loses. Similar observations to (i) to (iv) above can be made from Table 2. The reduction in the standard deviation is seen as attractive to the players in the mixed doubles events as they are typically also involved in the men's or ladies' doubles and/or the men's or ladies' single events, and would prefer not to have to play a mixed doubles match with a "long" duration. The best-of-two sets system achieves this elimination of long matches. This reduction in standard deviation in the best-of-two sets system is also very useful when two teams of four doubles players each play on adjacent courts and then swap over to play the alternate pair. Greater synchronisation is achieved across the two courts, with less time spent waiting for the second court to complete their longer match.

References

[1] R. E. Miles, "Symmetric sequential analysis: the efficiencies of sports scoring systems (with particular reference to those of tennis)", *J. Roy. Statist. Soc. Ser. B*, **46** (1984), 93–108.

[2] C. Morris, "The most important points in tennis", in *Optimal Strategies in Sports* (S. P. Ladany and R. E. Machol, editors), North–Holland, Amsterdam (1977), 131–140.

[3] G. H. Pollard, "An analysis of classical and tie-breaker tennis", *Austral. J. Statist.*, **25** (1983), 496–505.

[4] G. H. Pollard, "The optimal test for selecting the greater of two binomial probabilities", *Austral. J. Statist.*, **34** (1992), 273–284.

THE EFFECT OF A VARIATION TO THE ASSUMPTION THAT THE PROBABILITY OF WINNING A POINT IN TENNIS IS CONSTANT

Graham Pollard
Division of Management and Technology
University of Canberra
ACT 2601, Australia
graham.pollard@ise.canberra.edu.au

Abstract

A server has a range of options when starting a point. Some of these options are better than others when playing a single point. In order to avoid becoming predictable, the server will typically vary the option chosen, making use of all options, even those with the least favourable outcomes. This paper shows the considerable advantage of using the better options on the more important points.

1 Introduction

Tennis singles is usually modelled mathematically by assuming that the probability that the better player wins a point when serving is p_a, and that the probability that the other player wins a point when serving is p_b ($p_a > p_b$). Typically p_a and p_b are assumed to be constant. This constant probabilities model is a reasonable first approximation to the practical situation and probably gives reasonably accurate values for the mean μ and standard deviation σ of the number of points played in a match. Non-constant values for p_a and p_b, particularly for the situation in which one player increases his or her p-value on the more important points, can affect the probability that each player wins the match. The difference between the number of points won by each player is typically quite small, and sometimes the winner wins fewer points than the loser. In the tennis world, there is general acceptance of the importance of being able to play the "big points" well.

In this paper we consider the situation in which the server does not play a single type of point when serving, but selects from a range of options on both the first and second serve. Some of these options are more likely than others to lead to winning the point. The player however believes it is important to use the full range of options so that he or she keeps the opponents guessing as to which option is to be used next. A method for evaluating the benefit from using the better options on the more important points is demonstrated.

2 The analysis

Let us consider, for example, a right-handed player, player A, serving to a right-handed receiver in the first, or forehand, court. Let us assume that player A has four service options on the first serve. In particular we assume:

1. the service down the middle of this first court is the most likely (fast) first service to go in (that is, not be a fault), and player A has a probability 0.6 of serving in such a service, and a probability of 0.7 of winning such a point if the service does go in;

2. the service down the centre of the court to the receiver's backhand is the second most likely (fast) first service to go in, and player A has a probability 0.55 of serving in such a service, and a probability of 0.75 of winning such a point if the service does go in;

3. the service to the receiver's forehand is the third most likely (fast) first service to go in (the net is higher for this service than for 1 and 2 above), and player A has a probability 0.5 of serving in such a service, and a probability 0.8 of winning such a point if the service goes in;

4. the wide and swinging-away service to the forehand is the least likely service to go in, and player A has a probability 0.45 of serving in such a service, and a probability 0.85 of winning such a point if the service goes in.

We assume that, in order to keep the opponent guessing, player A must serve each of these first service options at least 20% of the time, and must not serve any one of the options more than 35% of the time.

It is assumed that player A has three service options for the second serve. In particular we assume:

1. the service down the centre of the service court has a probability 0.9 of going in, and player A has a probability 0.5 of winning the point if it goes in;

2. the service down the centre of the court to the receiver's backhand has a probability 0.85 of going in, and player A has a probability 0.55 of winning such a point if it goes in;

3. the service to the player's forehand has probability 0.8 of going in, and player A has a probability of 0.55 of winning the point if it goes in.

We assume player A must serve each of these second service options at least 25% of the time, and must not serve any one of them more than 40% of the time.

This range of options and probabilities realistically represents the practical solution in many matches in men's singles. Also, the range of options and probabilities to the second, or backhand, court is similar, and is assumed to be the same. Further, although the situation above has been described for right-handed players, the situation is similar for left-handed players.

Considering firstly the second service situation, player A has a probability $0.9 \times 0.5 = 0.45$ of winning the point under option 1, 0.4675 under option 2, and 0.44 under option 3. By using the best option, option 2, on 40% of occasions, the worst option, option 3, on only 25% of occasions, and option 1 the remainder of the time, each at random, the probability of player A winning a point on second service is given by $0.4 \times 0.4675 + 0.25 \times 0.44 + 0.35 \times 0.45 = 0.4545$.

We now consider the probability player A wins a point using the first of the four options for the first service, assuming the second service is a random selection as given by the above paragraph. The probability player A wins a point under first serve option 1 is equal to $0.6 \times 0.7 + (1 - 0.6) \times 0.4545 = 0.6018$. The corresponding probabilities under first service options 2, 3 and 4 are 0.6170, 0.6273 and 0.6325. By using option 4 on 35% of occasions, options 1 and 2 each on 20% of occasions, and option 3 on 25% of occasions, at random, player A's probability of winning a point on service is 0.6219. Given this probability of winning a single point, player A's probability of winning a service game can be shown to be 0.7796.

We now consider the increased probability of player A winning a service game by using first service option 4 on the 35% of occasions which are the most important, option 3 on the 25% of occasions which are the next most important, option 2 on the 20% of occasions which are the next most important, and option 1 on the 20% of occasions which are the least important. It is assumed that second serves are at random as described above. The importance of a point within a game (see Morris (1977)) is defined as the probability of winning the game given the point is won minus the probability of winning the game given the point is lost. Table 1 lists the various points (or states) within a game from the most important (when the point-probability is 0.6219) to the least. It also lists the importance of each point, I, and the expected number of times each point (or state) is played (or visited), E. Note that the sum of the E's

for states 10a and 10b equals 1 as the score 0–0 occurs once in a game. States 10a and 10b have been created so that the entry for the cumulative E column for 10a, 3.8186, is 60% (that is, 35% + 25%) of the expected number of points in the game, 6.3644. The other states 7a and 7b, and 12a and 12b, have been created for corresponding reasons. The increase in player A's probability of winning a point in states 1, 2, ..., 7a is $0.6325 - 0.6219 = 0.0105$; 0.0053 in states 7b, 8, 9 and 10a; -0.0049 in states 10b, 11 and 12a; and -0.0201 in states 12b, 13, 14 and 15. The increase (above 0.7796) in player A's probability of winning the game is given (approximately) by the sum of the products $IE\Delta p$ (where Δp is the increase in point-probability for that state), and equals 0.0090 in this case.

		I	E	$\sum E$	
1.	30–40, ad receiver	0.7302	0.3946	0.3946	
2.	15–40	0.4541	0.1344	0.5290	
3.	15–30	0.4477	0.2667	0.7957	
4.	30–30, deuce	0.4439	0.8225	1.6182	
5.	0–30	0.3853	0.1429	1.7611	
6.	0–15	0.3404	0.3781	2.1391	
7a.	15–15	0.3131	0.0884	2.2275	(35%)
7b.	15–15	0.3131	0.3819	2.6094	
8.	0–40	0.2824	0.0540	2.6634	
9.	40–30, ad server	0.2698	0.6491	3.3125	
10a.	0–0	0.2454	0.5061	3.8186	(60%)
10b.	0–0	0.2454	0.4939	4.3125	
11.	30–15	0.2312	0.4387	4.7512	
12a.	15–0	0.1877	0.3403	5.0915	(80%)
12b.	15–0	0.1877	0.2817	5.3732	
13.	30–0	0.1114	0.3868	5.7600	
14.	40–15	0.1020	0.3638	6.1238	
15.	40–0	0.0386	0.2406	6.3644	
			6.3644		

Table 1: The importance I of the various points in a game of tennis, and the expected number E of times each point is played, during one play of a game (when $p = 0.6219$).

We now suppose the player in the previous paragraph also optimises on the second service, rather than randomising on it. Thus, the best second service option, option 2, is used on the 40% of occasions which are the most important, and the worst second service option, option 3, is used on the 25% of occasions which are the least important, with option 1 being used the remainder of the time. The increase (above 0.7796) in player A's probability of winning the game can now be shown to be 0.0161, using a similar analysis to that in previous paragraph.

3 Summary

We have considered a player who has four first service and three second service options, and made what are believed to be reasonable probability assumptions concerning these options. Under these assumptions, it can be shown that a player who serves each of the service options in a random and equally likely fashion, has a probability 0.7733 of winning the game. By using the better options more often (although subject to constraints on how often) and still randomising, this probability increases to 0.7796. By randomising on the second service and using the better first service options on the more important points, this probability increases to 0.7886. Further, by using the better first and second service options on the more important points, this probability increases to 0.7957.

The benefits of "lifting your game", or playing the best options on the more important points, is clear. The probability of winning the game is increased somewhat, the probability of winning the set is increased by even more, and the probability of winning the match is increased by a substantial amount.

Players may find it useful to collect the relevant statistics for their own play, so that the approach taken in this paper can be put into practice, at least partially. One challenge in putting the approach into practice is to avoid the danger of becoming too predictable.

Reference

C. Morris (1977), "The most important points in tennis", in *Optimal Strategies in Sports* (S. P. Ladany and R. E. Machol, editors), North–Holland, Amsterdam, 131–140.

A SOLUTION TO THE UNFAIRNESS OF THE TIEBREAK GAME WHEN USED IN TENNIS DOUBLES

Graham Pollard
Division of Management and Technology
University of Canberra
ACT 2601, Australia
graham.pollard@ise.canberra.edu.au

Ken Noble
Australian Bureau of Agricultural
and Resource Economics
GPO Box 1563, Canberra
ACT 2601, Australia
knoble@abare.gov.au

Abstract

In 2000 the International Tennis Federation approved several new tennis scoring systems. One of these new scoring systems, the "best-of-two sets" system has been used for mixed doubles in the major tournament, the Australian Open. Within this best-of-two sets scoring system, the tiebreak game plays a substantially enhanced role. Questions such as "Should a mixed doubles pair elect to use the stronger server to serve first or second in the tiebreak game?" and "Is the tiebreak game fair in this mixed doubles situation?" become considerably more important than before the new systems were approved. In this paper we show that the tiebreak game, although fair in the singles context (under basic assumptions), can become unfair in the doubles context. We also outline a solution to this unfairness.

1 Introduction

In 2000 the governing body of tennis, the International Tennis Federation approved several new tennis scoring systems for use in professional and amateur competition. One of these new scoring systems is the "best-of-two sets" system. It consists of playing two (normal) sets, and if one player (or doubles pair) has not won the match by two sets to nil, the tiebreak game is then played to determine the winner. It can be seen that, in the case in which the match reaches one set all, the tiebreak game plays a particularly important and considerably enhanced role.

This new best-of-two sets system was used in 2001 in the Australian Open Mixed Doubles Championships. In this tournament the "12-point" or "first-to-seven and lead by two points" tiebreak game was used if the sets score reached 1–1. Further, the best-of-two sets system with a super tiebreak game ("first-to-ten and lead by two points") was used in the 2002 Australian Open Mixed Doubles Championship and in the recent Hopman Cup for dead rubbers.

The use of the tiebreak game as the (sudden-death) decider in the best-of-two sets mixed doubles highlights several interesting and important questions. Should a mixed doubles pair elect to use the stronger server to serve first or second in the tiebreak? Indeed, is the tiebreak game fair in this doubles situation? We now consider these two questions and outline a method for modifying the tiebreak game so that it is fair for doubles.

2 The 12-point tiebreak game

Consider two doubles teams A and B, with players A1 and A2, and B1 and B2. Suppose team A has a probability P_{A1} of winning the point when player A1 serves, and probability P_{A2} when A2 serves. Similarly, suppose team B has a probability P_{B1} of winning the point when B1 serves, and probability P_{B2} when B2 serves.

Consider the case of two equal teams when $P_{A1} = P_{B1} = 0.7$ and $P_{A2} = P_{B2} = 0.5$. For two equal teams, the probability of each team winning under a fair scoring system should equal 0.5. Assuming without loss of generality that Team A serves first in the tiebreak game, there are four cases for the order of serving.

(a) A1, B1 B1, A2 A2, B2 B2, A1 A1, B1 B1, A2 ...

(b) A2, B2 B2, A1 A1, B1 B1, A2 A2, B2 B2, A1 ...

(c) A1, B2 B2, A2 A2, B1 B1, A1 A1, B2 B2, A2 ...

(d) A2, B1 B1, A1 A1, B2 B2, A2 A2, B1 B1, A1 ...

For cases (a), (b), (c) and (d), respectively:

Prob (Team A wins in 12 points or less) is 0.3348, 0.4303, 0.4303, 0.3348,
Prob (the tiebreak game reaches 6–6) is 0.2360, 0.2323, 0.2323, 0.2360,
Prob (Team A wins the tiebreak game) is 0.4618, 0.5356, 0.5287, 0.4719.

The results (here and following) have been evaluated using a method similar to that described in the Appendix, and have been rounded to four decimal points.

It can be seen that:

(i) Team B should always elect to use its stronger server, player B1, first (cases (a) and (d)). When Team B does this, player B1 serves four points out of the first 12 points of the tiebreak game, player B2 serves only two points, and players A1 and A2 each serve for three points, leading to the observed unfairness.

(ii) Interestingly, Team A should use its weaker server, player A2, on the first point (cases (b) and (d)). When Team A does this, its stronger server, player A1, serves three points which are (on average) more important (see Morris (1977)) than the three points that player A2 serves.

(iii) It follows from (i) and (ii) that case (d) is the case of practical relevance, and in this case the tiebreak game is unfair. The probability that team A wins the tiebreak game is only 0.4719, which is less than 0.5.

Correspondingly, it can be shown that the first to ten points (and lead by at least two points) super tiebreak game is also unfair.

3 Designing a tiebreak game which is fair for doubles

To design a tiebreak game which is fair, we need to think in multiples of eight points. Consider a 16-point tiebreak game (first to nine points and leading by at least two points). Again we consider the four cases (a) to (d) above for the order of serving.

With a 16-point tiebreak, each player serves four of the first 16 points, and so one unfair aspect of the 12-point tiebreak is removed.

Now, for each of the cases (a), (b), (c) and (d), respectively:

Prob (Team A wins in 16 points or less) is 0.3974, 0.3974, 0.3974, 0.3974,
Prob (this tiebreak game reaches 8–8) is 0.2052, 0.2052, 0.2052, 0.2052,
Prob (Team A wins the tiebreak game) is 0.4904, 0.5078, 0.5166, 0.4843.

It can be seen that both teams would now elect to use their stronger server first, and that case (a) is therefore the case of practical relevance.

It is interesting to note that in each of these four cases the probability of Team A winning in 16 points or less is 0.3974, and the probability of the tiebreak game reaching 8–8 is 0.2052.

Thus, it can be seen that this tiebreak game becomes unfair *after* the 16th point is played. This is because the players are serving two points at a time. The solution is for the players to serve alternately only one point at a time at this stage. Thus, noting case (a) is the relevant one, the 17th, 18th, 19th, 20th, ... points would be served by players A1, B1, A2, B2, respectively. With this minor modification in service sequencing, Prob (Team A wins the tiebreak game) becomes 0.5, and this modified 16-point tiebreak game is fair. An analysis of this modified fair 16-point tiebreak game is given in the Appendix.

(It might be noted that the teams should change ends after eight points (not six points) during this 16-point tiebreak game. It can also be seen that for the first, second, third, ... groups of eight points, each team has exactly two out of four serves, or half of the serves, from the same end of the court as that used prior to the tiebreak game. By comparison, in the 12-point tiebreak, for each of the first, second, third, ... groups of six points, the proportion of serves from the same end is either one-third or two-thirds for each team, and is sometimes different for the two teams, leading to an additional component of unfairness in some playing conditions.)

So far we have considered two equal doubles pairs, A and B, with win-on-serve probabilities $P_{A1} = P_{B1} = 0.7$ and $P_{A2} = P_{B2} = 0.5$, and have demonstrated a method of designing a fair tiebreak game when $P_{A1} = P_{B1}$ and $P_{A2} = P_{B2}$. We now consider a doubles pair C with $P_{C1} = P_{C2} = 0.6$. Doubles pair C is equal in standard to pair A or pair B (since $P_{C1} + P_{C2} = P_{A1} + P_{A2} = P_{B1} + P_{B2}$). We now consider the 16-point tiebreak game between pair A and pair C, and there are four cases for the order of serving to consider.

(i) A1, C1 C1, A2 A2, C2 C2, A1 A1, C1 C1, A2 ...

(ii) A2, C1 C1, A1 A1, C2 C2, A2 A2, C1 C1, A1 ...

(iii) C1, A1 A1, C2 C2, A2 A2, C1 C1, A1 A1, C2 ...

(iv) C1, A2 A2, C2 C2, A1 A1, C1 C1, A2 A2, C2 ...

For each of these cases (i), (ii), (iii) and (iv), respectively:

Prob (Team A wins in 16 points or less) is 0.3981, 0.3981, 0.3981, 0.3981,
Prob (Team C wins in 16 points or less) is 0.3990, 0.3990, 0.3990, 0.3990,
Prob (this tiebreak game reaches 8–8) is 0.2029, 0.2029, 0.2029, 0.2029,
Prob (Team A wins the tiebreak game) is 0.5059, 0.4926, 0.5059, 0.4926.

It can be seen that pair C has a slightly greater (than pair A) chance of winning the tiebreak game in 16 points or less. Also, provided pair A selects player A1 to serve first, pair C has a probability less than 0.5 of winning the tiebreak game. This is because of the advantage pair A has on the points-pairs played on points numbered 17 and 18, 21 and 22, 25 and 26, ... This advantage can be removed by insisting that multiples of four points are played after 8–8 is reached, so that, if 8–8 is reached, a total of 20, 24, 28, ... points is played. The leading by two rule would still apply after each of these multiples of four points. (Note that in practice the last of each of these four points, the 20th, 24th, 28th , ..., may not be required when one pair is three points ahead). Under this modification of playing multiples of

four points once 8–8 is reached, Prob (Team A wins the tiebreak game) is 0.4992 in all cases (i) to (iv). Thus, under this modification, pair C has a slightly greater than 0.5 probability of winning the 16-point tiebreak. However, it can be seen that with this modification we have a scoring system which is much closer to fair than the one without it. Hence, this modification for the 16-point tiebreak is recommended. (Note that it would be incredibly exciting when a pair comes from two points down to win!)

The system outlined above goes a very long way to achieving fairness in the tiebreak game. However, it would appear that there is no way (in a typical scoring system) of achieving exact fairness for the case in which two equal pairs have different variances in their p-values. For typical scoring systems (indeed all reasonable scoring systems), the (equal) pair with the smaller variance in p-values has a (very slightly) greater than 0.5 probability of winning. This would appear to be due to the concave downward nature of the relationship between the probability of winning a single point and the probability of winning overall under the particular scoring system.

There would appear to be a good case for introducing this modified 16-point tiebreak game into mixed (and women's and men's) doubles in order to resolve the unfairness of the present tiebreak structure. Matches will, of course, have a slightly larger average duration, and the probability of the better team winning will be slightly higher than when the present 12-point tiebreak is used. This is seen to be an advantage for the cases in which a tiebreak is used to replace a set.

A corresponding 24-point tiebreak game, with players serving alternately if the score reaches 12–12, would also be fair (for two equal pairs with the same variances in their p-values). It is clear that such an extended tiebreak would address the concerns of those who claim that the 12-point tiebreak is a "bit of a lottery", and could be particularly relevant for grand slam mixed doubles.

Appendix — An analysis of the modified fair $8n$-point tiebreak game for doubles

Suppose that a tiebreak game of doubles is the best of $2b$ points ($b = 4, 8, 12, 16, \ldots$), and so is won by the team that first reaches a score of $b + 1$ points, provided that the opposing team has a score of $b - 1$ points, or fewer. To simplify the notation in this appendix, we refer to the players as 1, 2, 3, 4 rather than using A1, B1, A2, B2. We suppose that team A comprises players 1 and 3, and that team B comprises players 2 and 4. Player 1 (team A) serves the first point of the tiebreak, then player 2 (team B) serves for two consecutive points, followed by player 3 (team A) serving for two consecutive points, then player 4 (team B) serving for two consecutive points, etc. Thus during the first 16 points of the tiebreak, the sequence of serving by the players is as follows:

$$1, 2, 2, 3, 3, 4, 4, 1, 1, 2, 2, 3, 3, 4, 4, 1 \ldots$$

This continues until either one team reaches $b + 1$ points, leading by at least two and thereby wins the tiebreak game, or the score reaches b points all.

In the event that the score reaches b points all, the players alternate service, with the tiebreak game decided when one team leads the other by two points. Thus points $2b + 1$, $2b + 2$, ... of the tiebreak are served by the players as follows:

$$1, 2, 3, 4, 1, 2, \ldots$$

Let p_i denote the probability that player i wins a point when serving, and let $q_i = 1 - p_i$ for $i = 1, \ldots, 4$.

Given numeric values for p_1, p_2, p_3, p_4 and the value of b, it is a simple matter to program a computer to determine:

$$P(i, j) = \text{probability that the tiebreak score reaches } i \text{ points to Team A and } j \text{ points to Team B,}$$

where $i + j \leq 2b$. Then

$$\text{Prob (Team A wins the tiebreak in the first } 2b \text{ points or less)} = \sum_{j=0}^{b-1} P(b + 1, j),$$

and

$$\text{Prob (Team B wins the tiebreak in the first } 2b \text{ points or less)} = \sum_{j=0}^{b-1} P(i, b+1),$$

and

$$\text{Prob (point scores are level after the first } 2b \text{ points)} = P(b, b).$$

For the score to move from (b, b) to $(b+2, b)$, player 1 must win service at point $2b+1$, and player 2 must lose service at point $2b+2$. Thus $P(b+2, b) = P(b, b)p_1q_2$. Likewise, $P(b, b+2) = P(b, b)q_1p_2$.

Also, if either player 1 and 2 both hold their services, or both lose their services, the score moves to $(b+1, b+1)$, so that

$$P(b+1, b+1) = P(b, b)(p_1p_2 + q_1q_2).$$

If the score reaches $(b+1, b+1)$, then player 3 serves at point $2b+3$ and player 4 serves at point $2b+4$, and

$$\begin{aligned}
P(b+3, b+1) &= P(b+1, b+1)p_3q_4, \\
P(b+1, b+3) &= P(b+1, b+1)q_3p_4, \\
P(b+2, b+2) &= P(b+1, b+1)(p_3p_4 + q_3q_4).
\end{aligned}$$

If the score should reach $(b+2, b+2)$ then the cycle of players 1, 2, 3, 4 serving is repeated. Hence it is easy to see that, for $k = 0, 1, 2, \ldots,$

$$\begin{aligned}
P(b+2k+2, b+2k) &= P(b, b)((p_1p_2 + q_1q_2)(p_3p_4 + q_3q_4))^k p_1q_2, \\
P(b+2k, b+2k+2) &= P(b, b)((p_1p_2 + q_1q_2)(p_3p_4 + q_3q_4))^k q_1p_2, \\
P(b+2k+1, b+2k+1) &= P(b, b)((p_1p_2 + q_1q_2)(p_3p_4 + q_3q_4))^k (p_1p_2 + q_1q_2), \\
P(b+2k+3, b+2k+1) &= P(b, b)((p_1p_2 + q_1q_2)(p_3p_4 + q_3q_4))^k (p_1p_2 + q_1q_2)p_3q_4, \\
P(b+2k+1, b+2k+3) &= P(b, b)((p_1p_2 + q_1q_2)(p_3p_4 + q_3q_4))^k (p_1p_2 + q_1q_2)q_3p_4, \\
P(b+2k+2, b+2k+2) &= P(b, b)((p_1p_2 + q_1q_2)(p_3p_4 + q_3q_4))^{k+1}.
\end{aligned}$$

Consequently,

$$\begin{aligned}
&\text{Prob (Team A wins the tiebreak at point } 2b+2 \text{ or subsequently)} \\
&= \sum_{k=0}^{\infty} (P(b+2k+2, b+2k) + P(b+2k+3, b+2k+1)) \\
&= P(b, b)(p_1q_2 + (p_1p_2 + q_1q_2)p_3q_4) \sum_{k=0}^{\infty} ((p_1p_2 + q_1q_2)(p_3p_4 + q_3q_4))^k \\
&= P(b, b)\frac{p_1q_2 + (p_1p_2 + q_1q_2)p_3q_4}{1 - (p_1p_2 + q_1q_2)(p_3p_4 + q_3q_4)}
\end{aligned}$$

and

$$\begin{aligned}
&\text{Prob (Team B wins the tiebreak at point } 2b+2 \text{ or subsequently)} \\
&= P(b, b)\frac{q_1p_2 + (p_1p_2 + q_1q_2)q_3p_4}{1 - (p_1p_2 + q_1q_2)(p_3p_4 + q_3q_4)}.
\end{aligned}$$

Reference

C. Morris (1977), "The most important points in tennis", in *Optimal Strategies in Sports* (S. P. Ladany and R. E. Machol, editors), North–Holland, Amsterdam (1977), 131–140.

DARTBOARD ARRANGEMENTS WITH A CONCAVE PENALTY FUNCTION

E. Tonkes
Department of Mathematics
University of Queensland
Queensland 4072, Australia
ejt@maths.uq.edu.au

Abstract

This paper investigates combinatorial arrangements of the dartboard to maximise a penalty function derived from the differences of adjacent sectors. The particular penalty function is constructed by summing the absolute differences of neighbouring sectors raised to a power between zero and one. The arrangement to give the maximum penalty is found.

1 Introduction

Let $\mathcal{A}$ be the set of all permutations of $\{1, 2, \ldots, n\}$ and write

$$D_p(A) = \sum_{j=1}^{n} |a_j - a_{j+1}|^p,$$

where $A = (a_1, a_2, \ldots, a_n) \in \mathcal{A}$ and $a_{n+1} = a_1$.

We consider the combinatorial optimisation problem:

$$L_p: \qquad \text{Find } \widehat{A} \in \mathcal{A} \text{ such that } D_p(\widehat{A}) = \max_{A \in \mathcal{A}} D_p(A).$$

This problem has been studied in the context of a generalised dartboard with n sectors numbered 1, 2, ..., n. Problem L_p can then be interpreted as arranging the sectors to maximise a penalty function. Two penalty functions have received most attention: L_1, which maximises the sum of absolute differences of all pairs of adjacent numbers, and L_2, which maximises the sum of the squares of these differences.

Figure 1 shows the arrangement of the standard dartboard (also termed the "London" or the "Clock" board). This arrangement is by far the most common and along with the current rules of the game was developed in 1896 in Lancashire, [6]. It is of interest to note that, in general, the neighbours of relatively high scoring sectors are relatively low scoring sectors. Competitors aiming for a high score (say 20) will have inaccurate throws punished by the low scoring neighbours (5 and 1).

Alternative dartboard arrangements are still in use (particularly in the Manchester area of Britain, according to Selkirk [7]) and various penalty functions suggest yet more arrangements. This paper continues with the theme in the literature and extends the study of L_p to the case $0 < p < 1$.

For $n = 20$, it was demonstrated by Cohen and Tonkes [3] that the standard arrangement shown does not solve L_p for $p \geq 1$. It is well known that the standard dartboard arrangement is *close* to optimal for $p = 1$: $D_1(A_{dartboard})$ is 198 compared to the optimal value of 200.

Everson and Bassom [5] give a direct solution of L_1. Selkirk [7] and Eiselt and Laporte [4] considered L_1 and L_2. Cohen and Tonkes [3] used a string reversal algorithm to solve the L_p problem for all $p \geq 1$.

The intention here is to solve L_p for $0 < p < 1$.

More general versions of the problem have been attacked by Chao and Liang [1], where circular permutations of arbitrary numbers are considered (the differences between neighbours are not necessarily integer). Their approach involves a graphical representation of the solution, and deals with an arbitrary concave penalty function in the formulation of D_p. While they present a solution with the same ordering as ours for n even, their result cannot be correct and we have produced a counterexample in Remark 2. If the numbers a_i are arbitrary, it appears that, for n even, the optimal arrangement $\widehat{A}$ depends on the particular values of a_i and the form of the penalty function. Our Theorem 2 demonstrates that the optimal arrangement is essentially unique (when $0 < p < 1$) in the case that a_i are all consecutive integers and D_p takes the specific form above.

We refer to the permutation A above as an arrangement, and write it more usually as the sequence $A = a_1, a_2, \ldots, a_n$. We call an arrangement $B \in \mathcal{A}$ equivalent to A if $B = a_q, a_{q+1}, \ldots, a_n, a_1 \ldots, a_{q-1}$ for some q, $1 \leq q \leq n$ (cyclic permutation of A), or $B = a_r, a_{r-1}, \ldots, a_1, a_n, \ldots, a_{r+1}$ for some r, $1 \leq r \leq n$ (reversed cyclic permutation of A). For an actual dartboard, these imply that it does not matter which number is uppermost or whether the board is reflected in a diameter.

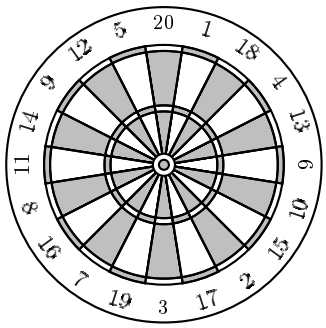


Figure 1: The traditional dartboard arrangement.

Define the left and right neighbours of elements in A in the obvious way, taking the left neighbour of a_1 to be a_n and the right neighbour of a_n to be a_1.

For n even, let $s_1, s_2, \ldots, s_{n/2}$ be a permutation of $\{1, 2, \ldots, n/2\}$ and let $l_1, l_2, \ldots, l_{n/2}$ be a permutation of $\{n/2 + 1, n/2 + 2, \ldots, n\}$. We call A an *alternating* arrangement if it can be written as $s_1, l_1, s_2, l_2, \ldots, s_{n/2}, l_{n/2}$. For n odd, let $m = (n + 1)/2$, let $s_1, s_2, \ldots, s_{(n-1)/2}$ be a permutation of $\{1, 2, \ldots, (n - 1)/2\}$ and let $l_1, l_2, \ldots, l_{(n-1)/2}$ be a permutation of $\{(n+3)/2, (n+5)/2, \ldots, n\}$. In this case, A is an alternating arrangement if it is equivalent to $m, s_1, l_1, s_2, l_2, \ldots, s_{(n-1)/2}, l_{(n-1)/2}$.

It has been shown (for instance by Selkirk [7]) that A solves L_1 if and only if A is alternating.

When $p > 1$, it is found [3] that an alternating arrangement is necessary, but not sufficient to solve L_p. Amongst all of the alternating arrangements, the solution to L_p is achieved by the arrangement which heuristically gives the most variation in the differences of adjacent pairs.

When $p < 1$, again we find that an alternating arrangement is necessary to solve L_p. However, this time uniformity amongst the differences yields a larger value of the penalty function D_p.

2 Main result

Our main result gives the explicit arrangement to solve L_p:

Theorem 1 *For $0 < p < 1$, the solution of L_p is given by*

$$\widetilde{A} = \begin{cases} \frac{n}{2}+1, 2, \frac{n}{2}+3, 4, \frac{n}{2}+5, 6, \ldots, n-2, \frac{n}{2}-1, n, \frac{n}{2}, n-1, \frac{n}{2}-2, n-3, \ldots, 3, \frac{n}{2}+2, 1, & \frac{n}{2} \text{ odd}, \\ \frac{n}{2}+1, 2, \frac{n}{2}+3, 4, \frac{n}{2}+5, 6, \ldots, n-3, \frac{n}{2}-2, n-1, \frac{n}{2}, n, \frac{n}{2}-1, n-2, \ldots, 3, \frac{n}{2}+2, 1, & \frac{n}{2} \text{ even}, \\ 1, \frac{n+3}{2}, 2, \frac{n+5}{2}, 3, \ldots, n-1, \frac{n-1}{2}, n, \frac{n+1}{2}, & n \text{ odd}. \end{cases}$$

The solution is unique up to equivalence.

3 Proof of the main result

The primary tool that we will use is Jensen's inequality for concave functions (see, for example, [2]). Recall that a real-valued function $f(x)$ is strictly concave on interval (a, b) if, for any $x, y \in (a, b)$ it holds that $\lambda f(x) + (1-\lambda) f(y) < f(\lambda x + (1-\lambda) y)$ for all $0 < \lambda < 1$. The function $f(x) = x^p$ is an example of a strictly concave function on $(0, \infty)$ when $0 < p < 1$. A smooth concave function $f(x)$ has the property that $f'(x)$ is always decreasing.

Lemma 2 (*Jensen's concave inequality*) *Let $f(x)$ be a continuous, real valued function which is strictly concave on (a, b) and let $x_1, \ldots, x_m$ be in (a, b). Let $C_1, \ldots, C_m$ all be nonnegative with the property that $\sum_{i=1}^{m} C_i = 1$. Then*

$$f\left(\sum_{i=1}^{m} C_i x_i\right) \geq \sum_{i=1}^{m} C_i f\left(x_i\right).$$

Equality holds only if x_i are identical for all $1 \leq i \leq m$ with $C_i > 0$.

For an arrangement $A = a_1, a_2, \ldots, a_n$, define the set of differences $d_i(A) = |a_{i+1} - a_i|$ for $1 \leq i \leq n$, (where $a_{n+1} \equiv a_1$). If the arrangement A is understood, we will abbreviate $d_i(A)$ to d_i. If n is even then $\{d_i(\widetilde{A})\}_{i=1}^{n}$ contains precisely $n/2 - 1$ instances of $n/2 - 1$, $n/2 - 1$ instances of $n/2 + 1$ and two instances of $n/2$. For n odd, $\{d_i(\widetilde{A})\}_{i=1}^{n}$ contains $(n-1)/2$ instances of $(n+1)/2$ and $(n+1)/2$ instances of $(n-1)/2$.

The proof of Theorem 3 is contained in [3].

Theorem 3 *An arrangement $\overline{A} \in \mathcal{A}$ is a solution of L_1 if and only if it is alternating. Furthermore, $D_1(\overline{A}) = n^2/2$ for n even and $D_1(\overline{A}) = (n^2 - 1)/2$ for n odd.*

To prove Theorem 1, we apply slightly different approaches for n even and n odd. Henceforth, we assume that $0 < p < 1$.

Lemma 4 *Let n be odd. Suppose that $\widehat{A}$ solves L_p. Then $d_i(\widehat{A}) \in \{(n-1)/2, (n+1)/2\}$ for all $1 \leq i \leq n$.*

Proof: Suppose that arrangement A solves L_p and $d_k(A) \notin \{(n-1)/2, (n+1)/2\}$ for some $1 \leq k \leq n$. Let $\{\lambda_i\}_{i=1}^{n}$ satisfy:

$$0 \leq \lambda_i \leq 1 \tag{1}$$

and

$$\sum_{i=1}^{n} \lambda_i = \frac{n+1}{2}. \tag{2}$$

Either (i) we can find some $\{\lambda_i\}_{i=1}^{n}$ fulfilling (1) and (2) and satisfying

$$\sum_{i=1}^{n} \lambda_i d_i = \frac{n^2 - 1}{4}, \tag{3}$$

or (ii) $\sum_{i=1}^n \lambda_i d_i > (n^2-1)/4$ for all $\{\lambda_i\}_{i=1}^n$ satisfying (1) and (2), or (iii) $\sum_{i=1}^n \lambda_i d_i < (n^2-1)/4$ for all $\{\lambda_i\}_{i=1}^n$ satisfying (1) and (2).

Assuming case (i), expression (3) can be written:

$$\sum_{i=1}^n \frac{2\lambda_i}{n+1} d_i = \frac{n-1}{2}.$$

Since $\sum_{i=1}^n 2\lambda_i/(n+1) = 1$, we can apply Jensen's inequality with $f(x) = x^p$ to obtain

$$\begin{aligned} &\sum_{i=1}^n \frac{2\lambda_i}{n+1} d_i^p \le \left(\frac{n-1}{2}\right)^p \\ &\Rightarrow \sum_{i=1}^n \lambda_i d_i^p \le \left(\frac{n+1}{2}\right)\left(\frac{n-1}{2}\right)^p. \end{aligned} \tag{4}$$

By Theorem 3, $\sum_{i=1}^n d_i \le (n^2-1)2$, and combined with (3)

$$\sum_{i=1}^n \frac{2(1-\lambda_i)}{n-1} d_i \le \frac{n+1}{2}. \tag{5}$$

Noting that $\sum_{i=1}^n 2(1-\lambda_i)/(n-1) = 1$, we can again apply Jensen's inequality to obtain

$$\begin{aligned} &\sum_{i=1}^n \frac{2(1-\lambda_i)}{n-1} d_i^p \le \left(\frac{n+1}{2}\right)^p \\ &\Rightarrow \sum_{i=1}^n (1-\lambda_i) d_i^p \le \frac{n-1}{2}\left(\frac{n+1}{2}\right)^p. \end{aligned} \tag{6}$$

We now claim that (4) or (6) must be a strict inequality. This statement follows trivially if (5) is a strict inequality, so assume that (5) is an equality. According to Jensen's inequality, the only way that (4) can be an equality is if d_j are identical (call them $\underline{d}$) for all $j \in \{1, 2, \ldots, n\}$ for which $\lambda_j > 0$. Consequently, expression (3) implies that $\underline{d} = (n-1)/2$. Similarly, assuming equality in (5), the only way that (6) can be an equality is if d_j are identical (call them $\overline{d}$) for all $j \in \{1, 2, \ldots, n\}$ for which $\lambda_j < 1$. Consequently expression (5) implies that $\overline{d} = (n+1)/2$. However, we have assumed in the statement of the theorem that $d_k \notin \{(n-1)/2, (n+1)/2\}$ for some $1 \le k \le n$ and this contradiction shows that (4) or (6) must be a strict inequality.

Adding (4) and (6) gives

$$\sum_{i=1}^n d_i(A)^p < \frac{n+1}{2}\left(\frac{n-1}{2}\right)^p + \frac{n-1}{2}\left(\frac{n+1}{2}\right)^p = D_p(\widetilde{A})$$

and so A cannot solve L_p.

We will show next that case (ii) is impossible. Let $I = \{i_1, i_2 \ldots, i_{(n+1)/2}\}$ be a subindex corresponding to the smallest $(n+1)/2$ elements of $\{d_i\}_{i=1}^n$. For any index $j \notin I$ it follows that $d_j(A) \ge \max_{i \in I} d_i(A)$.

Let

$$\lambda_i = \begin{cases} 1, & \text{if } i \in I, \\ 0, & \text{if } i \notin I. \end{cases}$$

Then

$$\sum_{i=1}^n \lambda_i d_i = \sum_{i \in I} d_i > \frac{n^2-1}{4} \tag{7}$$

by assumption. Let $d^* = \max_{i \in I} d_i$. We claim that $d^* \geq (n+1)/2$. Suppose to the contrary that $d^* < (n+1)/2$, or in other words (since d_i are integer), $d^* \leq (n-1)/2$. Since I consists of $(n+1)/2$ elements,

$$\sum_I d_i \leq d^* \frac{n+1}{2} \leq \frac{n^2-1}{4},$$

contradicting (7).

Consequently, $\sum_{i \notin I} d_i \geq (n-1)d^*/2 \geq (n^2-1)/2$. Thus,

$$\sum_{i=1}^{n} d_i = \sum_{i \in I} d_i + \sum_{i \notin I} d_i > \frac{n^2-1}{4} + \frac{n^2-1}{4} = \frac{n^2-1}{2}$$

and this contradicts Theorem 3.

Thus, we are left with the case (iii) where $\sum_{i=1}^{n} \lambda_i d_i < (n^2-1)/4$ for all $\{\lambda_i\}_{i=1}^{n}$ satisfying (1) and (2).

By applying Jensen's inequality as per (4), we have

$$\sum_{i=1}^{n} \lambda_i d_i^p < \frac{n+1}{2} \left(\frac{n-1}{2} \right)^p. \tag{8}$$

Certainly, if $\sum_{i=1}^{n} \lambda_i d_i < (n^2-1)/4$ for all $\{\lambda_i\}_{i=1}^{n}$ satisfying (1) and (2), then $\sum_{i=1}^{n} \alpha_i d_i < (n^2-1)/4$ for all $\{\alpha_i\}_{i=1}^{n}$ satisfying $0 \leq \alpha_i \leq 1$ and $\sum_{i=1}^{n} \alpha_i = (n-1)/2$. We note that $\sum_{i=1}^{n} (1-\lambda_i) = (n-1)/2$, so by applying Jensen's inequality once more,

$$\begin{aligned} \sum_{i=1}^{n} (1-\lambda_i) d_i &< \frac{n^2-1}{4} \\ \Rightarrow \sum_{i=1}^{n} \frac{2}{n-1} (1-\lambda_i) d_i &< \frac{n+1}{2} \\ \Rightarrow \sum_{i=1}^{n} (1-\lambda_i) d_i^p &< \frac{n-1}{2} \left(\frac{n+1}{2} \right)^p. \end{aligned} \tag{9}$$

Adding (8) and (9), we see that $\sum_{i=1}^{n} d_i(A)^p < D_p(\widetilde{A})$ and so A cannot solve L_p. $\square$

Lemma 5 *Suppose that n is even and $\widehat{A}$ solves L_p. Then $\widehat{A}$ is an alternating arrangement.*

Proof: If $n = 2$, then all arrangements are equivalent, so we take $n \geq 4$. Let A be an arrangement which is not alternating, so $D_1(A) < n^2/2$ by Theorem 3. We claim that

$$\sum_{i=1}^{n} d_i^p \leq (n-1) \left(\frac{n}{2} \right)^p + \left(\frac{n}{2} - 1 \right)^p.$$

If $\max_{1 \leq j \leq n} d_j \leq n/2 - 1$, then $\sum_{i=1}^{n} d_i^p \leq n(n/2-1)^p$ and the claim follows.

Since $\sum_{i=1}^{n} d_i < n^2/2$, it is impossible for $\min_{1 \leq i \leq n} d_i \geq n/2$. Thus we consider an arrangement which possesses $\min_{1 \leq j \leq n} d_j \leq n/2 - 1$ and $\max_{1 \leq j \leq n} d_j \geq n/2$. Since all d_i are integer, we can write $\sum_{i=1}^{n} d_i \leq n^2/2 - 1$.

For such an arrangement we take $d_r \geq n/2$ and $d_s \leq n/2 - 1$ for some $1 \leq r, s \leq n$. It follows that for

$$\lambda \equiv \frac{(n/2-1) - d_s}{d_r - d_s} \in [0, 1)$$

we have

$$\lambda d_r + (1-\lambda)d_s = \frac{n}{2} - 1 \tag{10}$$

$$d_1 + d_2 + \cdots + (1-\lambda)d_r + \cdots + \lambda d_s + \cdots + d_n \le (n-1)\frac{n}{2}. \tag{11}$$

Apply Jensen's inequality on (10) with $f(x) = x^p$ to get

$$\left(\frac{n}{2} - 1\right)^p \ge \lambda d_r^p + (1-\lambda)d_s^p. \tag{12}$$

Similarly, using Jensen's inequality twice on (11) reveals

$$\begin{aligned}\left(\frac{n}{2}\right)^p &\ge \left(\frac{1}{n-1}\left(d_1 + d_2 + \cdots + ((1-\lambda)d_r + \lambda d_s) + \cdots + d_n\right)\right)^p \\ &\ge \frac{1}{n-1}\left(d_1^p + d_2^p + \cdots + ((1-\lambda)d_r + \lambda d_s)^p + \cdots + d_n^p\right) \\ &\ge \frac{1}{n-1}\left(d_1^p + d_2^p + \cdots + (1-\lambda)d_r^p + \cdots + \lambda d_s^p + \cdots + d_n^p\right). \end{aligned} \tag{13}$$

Combining (12) and (13) yields

$$(n-1)\left(\frac{n}{2}\right)^p + \left(\frac{n}{2} - 1\right)^p \ge d_1^p + d_2^p + \cdots + d_r^p + \cdots + d_s^p + \cdots + d_n^p,$$

and the claim is confirmed.

We now verify that $D_p(\widetilde{A}) - D_p(A) > 0$ (and thus A cannot be the solution to L_p). We have

$$\begin{aligned} D_p(\widetilde{A}) - D_p(A) &\ge \left(\frac{n}{2} - 1\right)\left(\frac{n}{2} - 1\right)^p + 2\left(\frac{n}{2}\right)^p + \left(\frac{n}{2} - 1\right)\left(\frac{n}{2} + 1\right)^p \\ &\quad - (n-1)\left(\frac{n}{2}\right)^p - \left(\frac{n}{2} - 1\right)^p \\ &= 2^{-p-1}\left((n-2)((n+2)^p - n^p) - (n-4)(n^p - (n-2)^p)\right). \end{aligned} \tag{14}$$

For $f(x) = x^p$, $0 < p < 1$, one can easily check that $f''(x)$ is negative and increasing on $(0, \infty)$. Thus, for any interval $[a, a+c] \subset (0, \infty)$,

$$f''(a + \xi) \ge f''(a) \quad \text{for } 0 \le \xi \le c.$$

Integrating this with respect to ξ over an interval $[0, h]$ yields

$$f'(a+h) - f'(a) \ge f''(a)h \quad \text{for } 0 \le h \le c.$$

Integrating once more with respect to h over the interval $[0, c]$,

$$f(a+c) - f(a) - f'(a)c \ge \frac{1}{2}f''(a)c^2.$$

Substituting $a = n$ and $c = 2$, it follows that

$$(n+2)^p \ge n^p + 2pn^{p-1} + 2p(p-1)n^{p-2}. \tag{15}$$

Similarly, over an interval $[a-c, a] \subset (0, \infty)$,

$$f''(a - c + \xi) \ge f''(a - c) \quad \text{for } 0 \le \xi \le c.$$

Integrating with respect to ξ over $[h, c]$ and then with respect to h over $[0, c]$ gives

$$f'(a)c - f(a) + f(a-c) \ge \frac{1}{2}f''(a-c)c^2.$$

Substituting $a = n$ and $c = 2$, it follows that

$$(n-2)^p \geq n^p - 2pn^{p-1} + 2p(p-1)(n-2)^{p-2}. \tag{16}$$

Substituting (15) and (16) into (14) gives

$$\begin{aligned}
D_p(\widetilde{A}) - D_p(A) &\geq 2^{-p-1}\left((n-2)\left(2pn^{p-1} + 2p(p-1)n^{p-2}\right)\right. \\
&\qquad \left. - (n-4)\left(2pn^{p-1} - 2p(p-1)(n-2)^{p-2}\right)\right) \\
&= 2^{-p}p\left(2n^{p-1} + (p-1)\left((n-2)n^{p-2} + (n-4)(n-2)^{p-2}\right)\right) \\
&> 2^{-p}p\left(2n^{p-1} - (n-2)n^{p-2} - (n-4)(n-2)^{p-2}\right) \\
&> 2^{-p}p\left(n^{p-1} - (n-4)(n-2)^{p-2}\right) \\
&= 2^{-p}pn(n-2)^{p-2}\left(\left(1-\frac{2}{n}\right)^{2-p} - \left(1-\frac{4}{n}\right)\right) \\
&> 2^{-p}pn(n-2)^{p-2}\left(\left(1-\frac{2}{n}\right)^{2} - \left(1-\frac{4}{n}\right)\right) \\
&= 2^{-p}pn(n-2)^{p-2}\frac{4}{n^2} > 0.
\end{aligned}$$

Thus if A is not alternating, then $D_p(A) < D_p(\widetilde{A})$ and A cannot solve L_p. □

Lemma 6 *If A is equivalent to an arrangement $w, x, a_1, a_2, \ldots, a_k, y, z, a_{k+1}, \ldots$ where $w < z < y < x$, then $D_p(A) < D_p(\widehat{A})$.*

Proof: Transform the original arrangement A to A' by reversing the string between the brackets:

$$\begin{aligned}
A &= w, [x, a_1, a_2, \ldots, a_k, y], z, a_{k+1}, \ldots, \\
A' &= w, [y, a_k, \ldots, a_2, a_1, x], z, a_{k+1}, \ldots.
\end{aligned}$$

The calculations for the penalty functions take the form:

$$\begin{aligned}
D_p(A) &= \Delta + (x-w)^p + (y-z)^p \\
D_p(A') &= \Delta + (y-w)^p + (x-z)^p
\end{aligned}$$

where Δ is precisely the same in both sums. Letting

$$\lambda = \frac{x-y}{x-y+z-w} \in (0,1)$$

we see that

$$\begin{aligned}
y - w &= \lambda(y-z) + (1-\lambda)(x-w), \\
x - z &= (1-\lambda)(y-z) + \lambda(x-w).
\end{aligned}$$

Using Jensen's inequality with $f(x) = x^p$ on each,

$$\begin{aligned}
(y-w)^p &> \lambda(y-z)^p + (1-\lambda)(x-w)^p, \\
(x-z)^p &> (1-\lambda)(y-z)^p + \lambda(x-w)^p.
\end{aligned}$$

Summing we see that $(y-w)^p + (x-z)^p > (y-z)^p + (x-w)^p$ and so $D_p(A) < D_p(A')$ and A cannot solve L_p. □

Lemma 7 *Suppose that n is even and that A is an alternating arrangement. Suppose that $d_i(A) \notin \{n/2-1, n/2, n/2+1\}$ for some $1 \le i \le n$. Then $D_p(A) < D_p(\widehat{A})$.*

Proof: Suppose that $\max_i d_i(A) \ge n/2+2$. The case $\min_i d_i \le n/2-2$ follows in a symmetrical manner. Let A contain adjacent elements s and l where $l-s = \max_i d_i$. Then A is equivalent to an arrangement with $a_1 = s$ and $a_2 = l$. The right neighbours in A of each element of $Z = \{n/2+1, n/2+2, \ldots, l-1\}$ cannot lie in $Y = \{s+1, s+2, \ldots, n/2\}$. Otherwise A can be written $s, l, a_3, a_4, \ldots, c, b, \ldots$ where $c \in Z$ and $b \in Y$ with $s < b < c < l$. It follows by Lemma 6 that $D_p(A) < D_p(\widehat{A})$. Consequently, the right neighbours of Z must lie in $X = \{1, 2, \ldots, s\}$. But comparing cardinalities, $|X| < |Z|$ yielding a contradiction. □

Lemma 8 *Suppose that n is even, A is alternating and $d_i(A) \in \{n/2-1, n/2, n/2+1\}$ for every $1 \le i \le n$. It follows that $A = \widetilde{A}$.*

Proof: The proof uses induction. We start with a subarrangement and at each step, we add a left neighbour and a right neighbour to build up A. Initially, the subarrangement A_0 consists of only the element 1. At the first step, to maintain an alternating arrangement and with the restrictions on $d_i(A)$, we note the neighbours of 1 must be $n/2+1$ and $n/2+2$, creating the subarrangement $A_1 = n/2+1, 1, n/2+2$.

At step $k \le n/2-1$, we claim that for k even,

$$A_k = k, k+\frac{n}{2}-1, k-2, k+\frac{n}{2}-3, \ldots, \frac{n}{2}+1, 1, \frac{n}{2}+2, \ldots, k-1, k+\frac{n}{2}, k+1, \tag{17}$$

while for k odd,

$$A_k = k+\frac{n}{2}, k-1, k+\frac{n}{2}-2, k-3, \ldots, \frac{n}{2}+1, 1, \frac{n}{2}+2, \ldots, k-2, k+\frac{n}{2}-1, k, k+\frac{n}{2}+1. \tag{18}$$

At step $k+1$, we add an element to each end of A_k.

If k is even then the neighbours of k must come from $\{k+n/2-1, k+n/2, k+n/2+1\}$. From the inductive hypothesis, the right neighbour of k is $k+n/2-1$. The left neighbour cannot be $k+n/2$, since this already has two neighbours distinct from k in the subarrangement. Hence the left neighbour of k must be $k+n/2+1$. At the right end of A_k, the neighbours of $k+1$ must come from $\{k+n/2, k+n/2+1, k+n/2+2\}$. From the inductive hypothesis, the left neighbour of $k+1$ is $k+n/2$. The right neighbour of $k+1$ cannot be $k+n/2+1$ because this has just been added to the left of A_k. Thus,

$$A_{k+1} = k+\frac{n}{2}+1, k, k+\frac{n}{2}-1, k-2, k+\frac{n}{2}-3 \ldots, \frac{n}{2}+1, 1, \frac{n}{2}+2, \ldots, k-1, k+\frac{n}{2}, k+1, k+\frac{n}{2}+2$$

corresponding with the form (18).

If k is odd, then the left neighbour of $k+n/2$ must come from $\{k-1, k, k+1\}$. From the inductive hypothesis, both $k-1$ and k are already contained in A_k and hence the left neighbour of $k+n/2$ must be $k+1$. The right neighbour of $k+n/2+1$ must come from $\{k, k+1, k+2\}$. However, k is already contained in A_k and $k+1$ is the left neighbour of $k+n/2$. Consequently

$$A_{k+1} = k+1, k+\frac{n}{2}, k-1, k+\frac{n}{2}-2, k-3, \ldots, \frac{n}{2}+1, 1, \frac{n}{2}+2, \ldots, k-2, k+\frac{n}{2}-1, k, k+\frac{n}{2}+1, k+2$$

corresponding with the form (17).

To conclude the construction, when $k = n/2$ the final element is inserted to form $A_{n/2} = \widetilde{A}$. □

We are now in a position to complete the proof of the main result.

Case of n odd: By Lemma 4, each element of $\widehat{A}$ can be uniquely identified with two neighbours. The two neighbours of $1 \le j \le (n-1)/2$ must be $j+(n-1)/2$ and $j+(n+1)/2$. The two neighbours of $(n+1)/2$ must be 1 and n. The two neighbours of $(n+3)/2 \le j \le n$ must be $j-(n-1)/2$ and $j-(n+1)/2$. This yields $\widehat{A} = \widetilde{A}$.

Case of n even: By Lemma 5, $\widehat{A}$ is alternating. By Lemma 7, $d_i(\widehat{A}) \in \{n/2-1, n/2, n/2+1\}$ for each $1 \le i \le n$. By Lemma 8, $\widetilde{A}$ is the only arrangement which has these properties and so $\widehat{A} = \widetilde{A}$. □

4 Further remarks

(1) We note that the sum can easily be evaluated:

$$D_p(\widehat{A}) = \begin{cases} \left(\frac{n}{2}-1\right)\left(\frac{n}{2}-1\right)^p + 2\left(\frac{n}{2}\right)^p + \left(\frac{n}{2}-1\right)\left(\frac{n}{2}+1\right)^p, & n \text{ even}, \\ \frac{n-1}{2}\left(\frac{n+1}{2}\right)^p + \frac{n+1}{2}\left(\frac{n-1}{2}\right)^p, & n \text{ odd}. \end{cases}$$

For a dartboard with 20 sectors, the optimal permutation is $\widehat{A}$ = 1, 11, 2, 13, 4, 15, 6, 17, 8, 19, 10, 20, 9, 18, 7, 16, 5, 14, 3, 12. This gives $D_{1/2}(\widehat{A}) \approx 63.174$, compared with the penalty for the actual dartboard arrangement, $D_{1/2}(A_{dartboard}) \approx 60.569$.

The arrangement is alternating, which means that it solves L_1 and interlaces "small" numbers from $\{1, 2, \ldots, 10\}$ with "large" numbers from $\{11, 12, \ldots, 20\}$. However, a drawback can be immediately observed as the cluster 19, 10, 20, 9, 18 creates an attractive target for high scores.

(2) We remark that this approach is not valid for permutations of arbitrary distinct numbers (not necessarily consecutive integers) and a counterexample for Theorem 2 in that case is readily available. For example, suppose that $\mathcal{A}$ is the set of permutations of $\{1, 2.5, 3, 3.1, 5, 6\}$. Then the solution to problem $L_{1/2}$ is 1, 3, 5, 2.5, 6, 3.1. This differs from the ordering in Theorem 2 which implies the permutation 1, 3.1, 2.5, 6, 3, 5.

References

[1] Chern-Ching Chao and Wen-Qi Liang, "Arranging n distinct numbers on a line or circle to reach extreme total variations", *Eur. J. Combin.*, (1992) **13**, 325–334.

[2] M. J. Cloud and B. C. Drachman, *Inequalities with Applications to Engineering*, Springer, New York (1998).

[3] G. L. Cohen and E. Tonkes, "Dartboard arrangements", *Electron. J. Combin.*, **8** (2001), #R4.

[4] H. A. Eiselt and G. Laporte, "A combinatorial optimization problem arising in dartboard design", *J. Oper. Res. Soc.*, **42** (1991), 113–118.

[5] P. J. Everson and A. P. Bassom, "Optimal arrangements for a dartboard", *Math. Spectrum*, **27** (1994/5), 32–34.

[6] D. McFarlan (editor), *The Guiness Book of Records 1990*, Guinness Publishing, Enfield, Middlesex (1989).

[7] K. Selkirk, "Re-designing the dartboard", *Math. Gaz.*, **60** (1976), 171–178.

DIVING INTO MATHEMATICS OR SOME MATHEMATICS OF SCUBA DIVING*

Michel de Lara (Michel Cohen de Lara)
Cermics
École Nationale des Ponts et Chaussées,
6–8 avenue Blaise Pascal,
Champs sur Marne
77455 Marne la Vallée Cedex 2,
France
mcdl@cermics.enpc.fr

Abstract

A diver is submitted to kinematic and dynamic laws for his or her movements in water: gravity, water drag force, buoyancy of Archimedes, perfect-gas law are the essential physical elements which contribute to the diver's kinematics.

The diver's safety depends on the dynamics of dissolved nitrogen in blood: Dalton's law coupled with linear dynamics provide the so-called Haldane model of compartments for dissolved nitrogen in blood. More recent and accurate models exist among which the famous Reduced Gradient Bubble Model. The diver's safety depends on the realism of such models: indeed, they are coded in tables or in wrist computers which inform the diver and prescribe diving profiles (stops and ascents) during the dive.

We present here the basic elements to build diver models which combine both types of physics. The diver state consists of position (depth), impulse, mass, air-jacket mass and dissolved nitrogen in compartments. The diver has control variables to modify his or her state: palming, jacket filling and unfilling. The state follows a controlled ordinary differential equation. Controls belong to bounded sets and linear combinations of the state must satisfy given inequalities for safety and physical reasons.

Such controlled ordinary differential equation with constraints models allow us to formulate various control problems:

- stabilisation at given depth;
- ascent at fixed speed;
- maximum sojourn time at given depth;
- minimum upward time.

At this stage, we can present the building of the model, suggest control problems and give some hints as to their resolution.

*This paper is accepted for presentation but was received too late to be submitted to the full refereeing process.