Predicting the Winner in One Day International Cricket

Ananda Bandulasiri, Ph.D.†

Abstract

Winning a One Day International (ODI) cricket match depends on various factors related to scoring as well as the athletic strengths of the two teams. While some of these factors have been well analyzed in the literature, others have yet to be investigated. In this analysis, statistical significance for a range of variables that could explain the outcome of an ODI cricket match is explored. In particular, home field advantage, winning the toss, game plan (batting first or fielding first), match type (day or day & night), and the effect of the Duckworth-Lewis method for matches shortened due to weather interruptions will be key interests in our investigation. The comparison of the accuracy of the Duckworth-Lewis method is done by receiver operating characteristics (ROC) curves. For purposes of model-building, logistic regression is applied retrospectively to data already obtained from previously played matches. Some surprising results emerge.

Introduction

Starting around the 13th century, cricket began to emerge as a very popular worldwide game. Moreover, cricket was one of the first sports to use statistics as a tool for illustration and comparison. Although dating back to 13th century, compared to sports like baseball, there has not been much statistical modeling work done for cricket. Wood (1945) used the geometric distribution to model the total score, while Kimber and Hansford (1993) proposed a nonparametric approach based on runs scored for assessing batting performance. Chedzoy (1997) addressed the issue of the effect of umpiring errors in cricket.

The latest World Cup (March-April, 2007) enjoyed the participation of 16 nations, all qualified from a larger pool of potential qualifiers. The final participating nations included Australia, Sri Lanka, South Africa, New Zealand, West Indies, England, Pakistan, India, Zimbabwe, Bangladesh, Ireland, Bermuda, Scotland, Netherland, Canada, and Kenya. The most common and popular form of cricket is the One Day International (ODI), where over 50-*overs* (300 deliveries) per side are played. There are 11 players on a team, and the pitch is in the middle of an oval-shaped mound where most of the game action occurs. Batsmen play in pairs, but bowlers are not allowed to throw the ball; instead, they must use a "stiff-arm" action to deliver a ball. As is typical in games of sport, winning is the ultimate goal. Some studies, (De Silva, 2001), analyze the magnitude of the victory, but most consider the factors affecting winning. There are cases where the magnitude of the victory is important; and, in fact, large sums of money are routinely wagered when it comes to betting on the outcomes of ODI games (Bailey and Clarke, 2006).

Clarke (1988) used a dynamic programming model to calculate the expected score for games with rain interruptions, so that both teams have the same chance of winning the game. Popular article, Duckworth and Lewis (1998), introduced a technique for revising the target for games that are shortened due to weather interruptions. This method was well received by the cricket-playing community, and it has been using for more than 10 years. In fact, the final game between Australia and Sri Lanka in the 2007 World Cup competition used the Duckworth-Lewis (**DL**-Method) because the game was shortened to 38-*overs* each on account of rain interruption.

Factors such as winning the toss and the home team advantage affecting the results of ODI games have been studied in the literature (Clarke (2003) and De Silva (1997)). In this paper, statistical significance for a range of variables that could explain the outcome of an ODI cricket match is explored. In particular, home field advantage, winning the toss, game plan (batting first or fielding first), match type (day or day & night), and the effect of the **DL** method for matches shortened due to rain are key interests in our investigation. Logistic regression is applied to historical data for purposes of model-fitting. Our data consists of a set of the ODIs played between nations for the time period starting January 1995 to the end of the 2007 World Cup matches. Some of the matches were deleted from the analysis due to certain reasons such as abundance of bad weather or when the one team was much superior to the other (ranked teams playing non-ranked teams). Tied games were also deleted from the analysis. Therefore, we only study games having a clear decision, and the data was collected from the web page source, www.cricket.org. Despite the availability of the data beginning in the 1970s, we considered games played beginning in 1995. Due to the continuous update of cricket rules, we chose, in particular, to use this most recent data and to collect a significant amount of information during the last 12 years.

Exploratory Analysis of the Outcome of One Day International Cricket

Because the outcome of a cricket game is dichotomous (ignoring a few games with ties), we use the logistic regression model. Define,

$$Y_i = \begin{cases} 1 & : \text{winning game } i \\ 0 & : \text{losing game } i \end{cases} \qquad t_i = \begin{cases} 1 & : \text{winning coin toss in game } i \\ 0 & : \text{losing coin toss in game } i \end{cases}$$

$$b_i = \begin{cases} 1 & : \text{batting first in } i^{th} \text{ game} \\ 0 & : \text{batting second in } i^{th} \text{ game} \end{cases} \qquad d_i = \begin{cases} 1 & : i^{th} \text{ game is day and night} \\ 0 & : i^{th} \text{ game is day only} \end{cases}$$

$$h_i = \begin{cases} 1 & : i^{th} \text{ game is a home game} \\ 0 & : i^{th} \text{ game is not a home game} \end{cases}$$

Further, let **OPP** be the class variable for the opponent teams, which is coded accordingly by the SAS logistic procedure. Letting $p_i = P(Y_i = 1)$, the logistic regression model with main effects only is expressed as

$$ln\ \frac{p_i}{1\ \ p_i} = \alpha + \beta_1 OPP + \beta_2 h_i + \beta_3 d_i + \beta_4 b_i + \beta_5 t_i$$

For this part of the analysis, we carefully selected a set of matches played after 1995. The focus here concerns the two countries Sri Lanka and India. A total of 249 = 123 ("Day") + 126 ("Day & Night") matches are included for Sri Lanka, while a total of 288 = 162 ("Day") + 126 ("Day & Night") matches are included for India. Because of the similarities in geography and of the nature of the players and other conditions such as weather, there are many reasons to believe that we can fit a common model for both of these counties.

The PROC LOGISTIC (SAS 9.1) procedure was used to run the logistic analysis. Table I shows the output for the model with a second order interaction term between the variables "Coin toss, Bat first" and the "Day & Night". One can include the variable **OPP** as also part of the interaction, but we omit its consideration for model simplicity. Even though one might fit a higher order model, it is difficult to interpret the results of more complicated models.

Table I
Logistic regression model for Sri Lanka with Second Order Interaction terms between "Toss", "Bat First" and "Day & Night".

Factor	P-value	Odds Ratio	Lower 95% C.L.	Upper 95% C.L.
Intercept	0.5941	1.2519	0.5478	2.8612
Opponent	0.2465[1]			
Home Field	**<0.0001**	4.8598	2.4335	9.7054
Day & Night	0.0914	0.4392	0.8758	5.3856
Bat First	0.9971	0.9982	0.3773	2.6410
Toss	0.0271	0.3218	0.1178	0.8795
Bat First * Toss	0.4632	1.5589	0.4762	5.1039
Day & Night * Bat First	0.4318	1.6108	0.4908	5.2859
Day & Night *Toss	**0.0670**	3.0630	0.9246	10.1468

[1] This is a class variable, odds ratio is not reported.

As seen in Table I, "Home Field" advantage is highly significant with odds almost four times higher when the game played in the home country. This

result is consistent with De Silva (1997). The only interaction effect that is closer in significance is "Day & Night" verses "Coin Toss". Note from Table I that "Bat First" or type of the game ("Day & Night" or "Day" only) does not show a significant effect on the outcome of the game under a logistic model with second order interaction.

Table II shows the results for a logistic model with one interaction term, "Day & Night" verses "Toss". In addition, the existence of a "Home Field" advantage is strongly suggested by this model.

Table II
Logistic regression model for Sri Lanka with a single second order interaction term between "Toss" and "Day & Night".

Factor	P-value	Odds Ratio	Lower 95% C.L.	Upper 95% C.L.
Intercept	0.8781	0.9498	0.4921	1.8332
Opponent	0.1778[1]			
Home Field	**<0.0001**	4.8754	2.4451	9.7213
Day & Night	0.181	0.5817	0.2630	1.2866
Bat First	0.1579	1.5351	0.8468	2.7828
Toss	0.0309	0.4262	0.1964	0.9247
Day & Night *Toss	**0.0450**	3.2891	1.0267	10.5364

[1] This is a class variable; odds ratio is not reported.

Some potential reasons supporting "Home Field" advantage are effect of local fans, familiarity with the field, and expected weather conditions. Still, the interaction between factors "Day & Night" and "Toss" shows a significant effect on outcome of the game. So, "Toss" has a different influence for matches played in "Day" from those played during "Day & Night".

Tables III and IV show the winning percentages for "Day" and "Day & Night" matches. Of the 125 "Day" only matches played by Sri Lanka which we considered, that nation won 62.69% the games after having lost the coin toss, but this percentage decreases to 41.38% when the coin toss was won. In all of the proceeding tables, "W" is used for "win" and "L" is used for "loss". On one hand, this phenomenon cannot be reasonably explained; on the other hand, it seems irrational to believe that this situation is exclusively due to randomness. With the winning of the toss, a team gets to play to its peculiar strategy strengths, while the opponent is forced defend against those particular strengths. This might explain lower success percentage when winning of the coin toss for "Day" only matches.

Table III
Classification of the results of the coin toss for "Day Only" matches played by Sri Lanka (Winning percentages are given in parentheses).

		Result		
		L	W	
toss	L	25 (37.31)	42 (62.69)	67
	W	34 (58.62)	24 (41.38)	58
		59	66	125

Table IV
Classification of the results of the coin toss for "Day & Night" matches played by Sri Lanka (Winning percentages are given in parentheses).

		Result		
		L	W	
toss	L	34 (53.13)	30 (46.87)	64
	W	27 (43.55)	35 (56.45)	62
		61	65	126

Of the 126 "Day & Night" matches played by Sri Lanka that we considered, that nation won 46.87% of the games after losing the coin toss, yet this percentage increases to 56.45% when winning the toss. Usually, "Day & Night" matches start around 2:00 PM local time and finish close to midnight. Field and weather conditions could change dramatically with playing time changes from day to night. Winning the toss helps a team to plan the game strategy according to the field and weather conditions, and this might be the reason behind the advantage of "Toss" for "Day & Night" matches.

Table V
Logistic regression model for India with the second order interaction terms between "Toss", "Bat First" and "Day & Night".

Factor	P-value	Odds Ratio	Lower 95% C.L.	Upper 95% C.L.
Intercept	0.4856	1.2875	0.6328	2.6196
Home Field	**0.0277**	1.8338	1.0689	3.1462
Opponent	0.2493[1]			
Day & Night	0.0029	0.2584	0.1060	0.6299
Bat First	0.1762	0.5454	0.2265	1.3131
Toss	0.0667	0.4512	0.1928	1.0562
Bat First * Toss	0.7749	1.1739	0.3914	3.5208
Day & Night * Bat First	0.7430	1.2059	0.3938	3.6927
Day & Night *Toss	**0.0005**	7.5587	2.4280	23.5311

[1] This is a class variable, odds ratio is not reported.

Results shown in Tables V, VI, and VII are consistent with the conclusions drawn for Sri Lanka.

Table VI
Classification of the result for the coin toss for "Day Only" matches played by India (Winning percentages are given in parentheses).

		Result		
		L	W	
toss	L	33 (45.21)	40 (54.79)	73
	W	55 (61.80)	34 (38.20)	89
		88	74	162

Table VII
Classification of the result for the coin toss for "Day & Night" matches played by India (Winning percentages are given in parentheses)

		Result		
		L	W	
toss	L	44 (74.58)	15 (25.42)	59
	W	32 (47.76)	35 (52.24)	67
		76	50	126

Effectiveness of the Duckworth-Lewis Method for Interrupted Matches

As mentioned earlier, for One Day International cricket matches each team gets only 50-*overs* (300 legal deliveries). If the weather (mainly rain) interrupts the game during play, at least one team does not get the intended number of *overs*. In this situation, the game is shortened and a revised target score should be given to decide the winner. For example, if the innings of the team which bats second are interrupted by rain, a target score for the second team should be reduced to compensate for the lost *overs* due to rain interruption.

Duckworth and Lewis (1998) introduced a way of revising targets for games which are shortened due to inclement weather. This method was well received by the cricket playing community and has been using for more than 10 years. While there are other methods, such as the Average Run Rate, the Most Productive Overs, the Parabola Method, and the Clark Curves technique, the Duckworth Lewis (**DL**) method was able to overcome most of the shortcomings of these other procedures (See Duckworth Lewis, 1998 for a brief review of these methods). The **DL** method is based on the idea that the batting team has two resources:
(i) it has certain number of *overs* to face, and (ii) it has a limited number of *wickets* in hand. When $\boldsymbol{w}$ *wickets* have been lost, but $\boldsymbol{u}$ *overs* are still left, the revised relationship follows an exponential decay model given by

$$Z(\boldsymbol{u},\boldsymbol{w}) = Z_0(\boldsymbol{w})[1-\exp\{-b(\boldsymbol{w})\boldsymbol{u}\}],$$

where, $Z_0(\boldsymbol{w})$ is the asymptotic average total from the last $10-\boldsymbol{w}$ *wickets* in unlimited *overs* and $b(w)$ is the exponential decay constant. For more model details, refer to Duckworth and Lewis (1998). While this technique is the best current method for the revised target, we investigate the effectiveness of the **DL** predictive win/loss percentages.

Comparisons of the Accuracy of the Duckworth-Lewis Prediction

A sample of 105 uninterrupted (fully played) recent ODI cricket matches was carefully selected. To keep the effect of team superiority differences at a minimum, only Australia, England, South Africa, New Zealand, West Indies, Sri Lanka, India and Pakistan were included in the 105 matches.

Even though the **DL** method can be applied in more complicated situations involving multiple interruptions, here we analyze only the one simple situation in which the first team finishes the game without any interruptions, but the second team has to stop batting before the end of 50-*overs*. For example, Table 8 shows data for interruption occurring at 30-*overs*, while Table 9 shows data for interruption occurring at 40-*overs*. Our question is, what would have been the outcome if we must apply the **DL** method at 30-*overs* and at 40-*overs*? When we mention winning, our focus is on the team who bats first and the game outcome, in terms of winning, for them. Tables 8 and 9 show the actual outcome as well as the **DL** predicted outcome for an imaginary interruption. The question is how accurate is the **DL** method. Analyzing the data in Table 8 (interruption at 30-*overs*), a highly significant relationship is found between the **DL** method and reality, having a Chi Square value less than 0.0001. This is exactly the situation one expects. As seen in Table 8, when the **DL** method declares that the team who bats first loses the game due to an interruption at 30-*overs*, it is accurate only 56.90% of the time if we had let the game to continue until the end. Moreover, **DL** agrees with reality 80.85% of the time when the prediction is for the first batting team to win the game.

Table IX shows that the **DL** method does a better job with the interruption at 40-*overs*, having a highly significant Chi Square value still less than 0.0001. Here, prediction of a loss for the first team batting is 62.90% when compared to the actual outcome, but this percentage rises to 93.02% for the prediction of a win for the first team batting. Apparently, the **DL** method does a very good job in predicting first batting team's victory in this situation, but does a relatively poor job predicting the loss. This is consistent with one of the main criticisms that the **DL** method produces a decision in favor of the team suffering from the interruption. As comparison percentages of the prediction for the Run Rate method is given in Tables X and XI. Run rate is calculated as the ratio between the total score and the total number of overs. In the case that a team loses all 10 wickets before the 50 overs, we still have to use the 50 as the denominator of the ratio. At 30-overs prediction for the loss of team 1 is accurate 62.86% of the time, while the prediction of the victory of team 1 is

correct up to 71.43%. At 40-overs the two numbers are 64.71% and 71.83% respectively. Unlike in the **DL**-method, accuracy of the prediction of Run Rate method is not getting better with the number of overs that the team 2 has faced before the interruption.

Table VIII
Comparison of the actual outcome with the DL method outcome at 30 overs

		Actual Outcome: L	Actual Outcome: W	
DL Out-come	L	33 (56.90)	25 (43.10)	58
	W	9 (19.15)	38 (80.85)	47
		42	63	105

Table IX
Comparison of the actual outcome with the DL method outcome at 40 overs

		Actual Outcome: L	Actual Outcome: W	
DL Out-come	L	39 (62.90)	23 (37.10)	62
	W	3 (6.98)	40 (93.02)	43
		42	63	105

Cohen's Kappa, $k = \frac{f_0 \quad f_c}{N \quad f_c}$ is one of the commonly used statistics used to measure the agreement of two raters, where f_0 is the number of concordant pairs for observed data and f_c is the number of concordant pairs produced just by chance alone. N denotes the total number of observations. Here, it is found that at the 30-*over* interruption, $k = 0.3657$ while for 40-*over* interruption, $k = 0.5220$. This means that for the 30-*over* interruption the **DL** method is in agreement with the actual result 36.57% of the time, while for the 40-*over* interruption agreement increases to 52.20%.

Table X
Comparison of the actual outcome with the Run Rate method outcome at 30 overs

		Actual Outcome: L	Actual Outcome: W	
DL Out-come	L	22 (62.86)	13 (37.14)	35
	W	20 (28.57)	50 (71.43)	70
		42	63	105

Table X1
Comparison of the actual outcome with the Run Rate method outcome at 40 overs

		Actual Outcome: L	Actual Outcome: W	
DL Out-come	L	22 (64.71)	12 (35.29)	34
	W	20 (28.17)	51 (71.83)	71
		42	63	105

For the Run Rate method, it is found that at the 30-*over* interruption, $k = 0.3265$ while for 40-*over* interruption, $k = 0.3443$. This means that for the 30-*over* interruption the Run Rate methods and the actual result are in agreement 32.65% of the time, while for the 40-*over* interruption agreement increases to 34.43%, which is not much of an improvement.

Receiver operating characteristic (ROC) curves

A Receiver Operating Characteristic (ROC) curve visually demonstrates the tradeoff between sensitivity and specificity as a function of varying a classification threshold. It is a common practice to use ROC curves to measure the accuracy of predictions by different methods. Although this method has been used primarily in medical and engineering fields, it could be used effectively in sports as well. More precisely, an ROC plots the sensitivity versus (1- specificity), and the area under the curve gives a measure of the prediction. So, the ideal best prediction should have one square unit of area under the ROC, where it achieves both 100% sensitivity and 100% specificity (which, in reality, rarely happens). Consequently, when we compare two methods, the one with the greater area under its ROC is judged best. Table XII shows the ranking for **DL** method based on the magnitude of the victory predicted. This is based on the 105 matches we considered earlier where we know the actual status. As in the previous section "win" means the victory of the team who bats first. Rankings were created in such a manner so as to represent the magnitude of the victory. For example if the difference between the **DL** revised target for team 2 and the actual score of team 2 (Revised **DL** target – Actual score of team 2) is less than -10, a ranking of 1 was assigned, which is a strongly unfavorable situation for team 1. The other ranks were similarly assigned. Table XIII shows the sensitivity and specificity values for several different threshold values.

Table XI1
Ranked predictions by DL method at 40 overs

		1	2	3	4	5	Total
True Status	Loss	27	10	2	2	1	42
	Win	11	11	1	4	36	63
	Total	38	21	3	6	37	105

The ranking shown above was done using the following scale:

IF (Revised **DL** target – Actual score of team 2) < **-10** then Rank = **1 (strongly negative*)**
IF **-10** ≤ (Revised **DL** target – Actual score of team 2) < **-2** then Rank = **2 (negative)**

IF **-2** ≤ (Revised **DL** target – Actual score of team 2) ≤ **2** then Rank = **3 (not clear)**

IF **2** < (Revised **DL** target – Actual score of team 2) ≤ **10** then Rank = **4 (positive)**

IF (Revised **DL** target – Actual score of team 2) > **10** then Rank = **5 (strongly positive)**

* Positive means the victory of team 1

Table XIII

Sensitivity and Specificity for different thresholds for the DL method at 40 overs.

≥1

	W	L
+	63	42
-	0	0
Total	63	42

Sensitivity = 63/63 = 1
Specificity = 0/42 = 0

≥2

	W	L
+	52	15
-	11	27
Total	63	42

Sensitivity = 52/63 = 0.825
Specificity = 27/42 = 0.642

≥3

	W	L
+	22	5
-	41	37
Total	63	42

Sensitivity = 41/63 = .651
Specificity = 37/42 = .881

≥4

	W	L
+	40	3
-	23	39
Total	63	42

Sensitivity = 40/63 = .635
Specificity = 39/42 = .929

≥5

	W	L
+	36	1
-	27	41
Total	63	42

Sensitivity = 36/63 = .571
Specificity = 41/42 = .976

>5

	W	L
+	0	0
-	63	42
Total	63	42

Sensitivity = 0/63 = 0
Specificity = 42/42 = 1

Figure I shows the ROC curve for the **DL** method at the 40 over interruption. A reference curve, which is for random guessing, (example: Toss a coin and decide the winner) is also shown. It is very clear that the **DL** prediction does an outstanding job when compared to random guessing. The area under the curve is considerably larger for the **DL** ROC. Figure II shows the curves for the

interruptions at 30 and 40 overs. Here, the area under the ROC curve for 40 overs has the largest area. This implies that when closer to the end of the game, the accuracy of **DL** method gets better.

Figure I
ROC curves for the DL method at 40 overs.

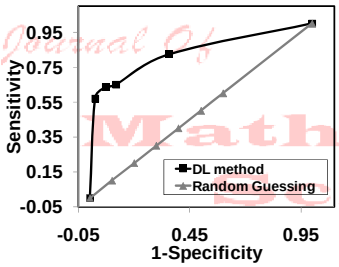
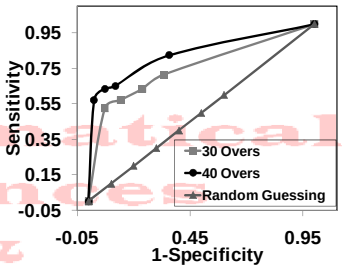


Figure II
ROC curves for the interruptions at 30 and 40 overs for the DL method.



ROC curves were developed for the **Run Rate** method using a similar ranking method:

IF (Run Rate for team 1 - Run Rate for team 2)< **-0.25** then Rank = **1 (strongly negative)**

IF **-0.25** ≤ (Run Rate for team 1 - Run Rate for team 2)< **-0.05** then Rank = **2 (negative)**

IF **-0.05** ≤ (Run Rate for team 1 - Run Rate for team 2) ≤ **0.05** then Rank = **3 (not clear)**

IF **0.05** < (Run Rate for team 1 - Run Rate for team 2) ≤ **0.25** then Rank = **4 (strongly positive)**

IF (Run Rate for team 1 - Run Rate for team 2) > **0.25** then Rank = **5 (strongly positive)**

Figures III and IV show comparisons between the **Run Rate** method and the **DL** method. Clearly, the **DL** method does a better job than the **Run Rate** method, and the difference is larger for the 40-over interruption.

Figure III
ROC curves for the interruptions at 30 overs.

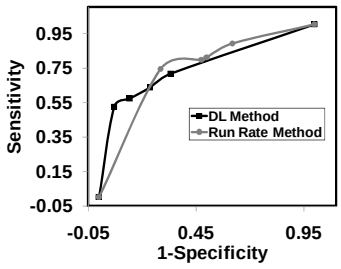
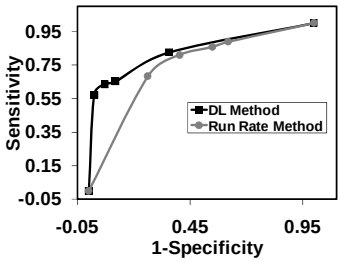


Figure IV
ROC curves for the interruptions at 40 overs.



Conclusions

In this article, we have statistically studied the importance of "home field" advantage on One Day International cricket. In addition, the strange result of the disadvantage of winning the coin toss for day time matches has also been observed. However, we have seen that winning the coin toss gives competitive advantage for "Day & Night" matches.

The Duckworth Lewis method, which is the best current method for revising targets for interrupted cricket matches, seems to statistically render a favorable decision to the team most impacted by the interruption. It could be argued that this is not a fair comparison, as the **DL** method does not have available the same amount of information to decide the true winner. However, the comparisons given here clearly indicate that the **DL** method can and should be improved.

† *Ananda Bandulasiri, Ph.D.*, Sam Houston State University, Texas, USA

References

Baily M. and Clarke S.R. (2006) Predicting the match outcome in one day international cricket matches, while the game is in progress, *Journal of sports Science and Medicine*, 5, 480-487.

Bennet J., (1998) "Statistics in Sports", Oxford University Press Inc., pp 83-101.

Chedzoy O.B., (1997) The effect of umpiring errors in cricket, *The Statistician*, 46, 4, 529-540.

Clarke, S. R. (1998) Dynamic programming in one-day cricket-optimal scoring rates, *Journal the Operational Research Society*, 39, 331-337.

Crowe, S. M. and Middeldorp (1996) A comparison of leg before wicket rates between Australian and their visiting teams for test cricket series played in Australia, 1977-94. *The Statistician*, 45, 255-262.

De Silva, B. M., and Swartz, T.B. (2001) Estimation of the magnitude of the victory in one-day cricket. *Austr. New Zeal. J. Sta*tist., 43, 1369-1373.

De Silva, B. M., and Swartz, T.B. (1997) Winning the coin toss and the home team advantage in one-day international cricket. *Austr. New Zeal. J. Sta*tist., 32, 16-22.

Duckworth, F. and Lewis, T. (1998) A fair method for resetting the target in interrupted one-day cricket matches. *J. Oper. Res. Soc*. , 49, 22-28.

Kimber, A.C. and Hansford A.R. (1993) A statistics analysis of batting in cricket, *Journal of the Royal Statistical Society Series A*, 156, 443-455.

Wood, G.H. (1945), Cricket scores and geometrical progression, *Journal of the Royal Statistical Society Series A*, 108, 12-22. www.cricket.org.