The Canadian Journal of Statistics
Vol. 37, No. 2, 2009, Pages 143–160
La revue canadienne de statistique

Modelling and simulation for one-day cricket

Tim B. SWARTZ[1]*, Paramjit S. GILL[2] and Saman MUTHUKUMARANA[1]

[1]*Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada V5A 1S6*
[2]*Mathematics, Statistics and Physics Unit, University of British Columbia Okanagan, Kelowna, British Columbia, Canada V1V 1V7*

Key words and phrases: Bayesian latent variable model; cricket; Markov chain methods; Monte Carlo simulation; sports statistics; WinBUGS.

MSC 2000: Primary 62P99; secondary 62F15.

Abstract: This article is concerned with the simulation of one-day cricket matches. Given that only a finite number of outcomes can occur on each ball that is bowled, a discrete generator on a finite set is developed where the outcome probabilities are estimated from historical data involving one-day international cricket matches. The probabilities depend on the batsman, the bowler, the number of wickets lost, the number of balls bowled and the innings. The proposed simulator appears to do a reasonable job at producing realistic results. The simulator allows investigators to address complex questions involving one-day cricket matches. *The Canadian Journal of Statistics* 37: 143–160; 2009

Résumé: Cet article porte sur la simulation de matchs de cricket d'une seule journée. Étant donné qu'il y a un nombre fini d'événements possibles à chaque lancer de balle, un générateur discret. Sur un ensemble fini est développé où les probabilités de chacun des événements sont estimées à partir de données historiques provenant de matchs de cricket international d'une seule journée. Les probabilités dépendent du batteur, du lanceur, du nombre de guichets perdus, du nombre de balles lancées et des manches. Le simulateur proposé semble faire un travail raisonnable en produisant des résultats réalistes. Il permet aux chercheurs d'étudier des questions complexes concernant les matchs de cricket d'une seule journée. *La revue canadienne de statistique* 37: 143–160; 2009

1. INTRODUCTION

Simulation is a practical and powerful tool that has been used in a wide range of disciplines to investigate complex systems. When a simulation model is available, it is typically straightforward to address questions concerning the occurrence of various phenomena. One simply carries out repeated simulations and observes the frequency with which the phenomena occur.

In one-day international (ODI) cricket, there are an endless number of questions that are not amenable to experimentation or direct analysis but could be easily addressed via simulation. For example, on average, would England benefit from increasing the number of runs scored by changing the batting order of their third and sixth batsmen? As another example, what percentage of time would India be expected to score more than 350 runs versus Australia in the first innings?

To provide reliable answers to questions such as these, a good simulator for one-day cricket matches is required. Surprisingly, the development of simulators for one-day cricket is a topic that has not been vigorously pursued by academics. In the pre-computer days, Elderton (1945) and Wood (1945) fit the geometric distribution to individual runs scored based on results from test cricket. Kimber & Hansford (1993) argue against the geometric distribution and obtain

* *Author to whom correspondence may be addressed.*
E-mail: tim@stat.sfu.ca

probabilities for selected ranges of individual scores in test cricket using product-limit estimators. More recently, Dyte (1998) simulates batting outcomes between a specified test batsman and bowler using career batting and bowling averages as the key inputs without regard to the state of the match (e.g., the score, the number of wickets lost, the number of overs completed). Bailey & Clarke (2004, 2006) investigate the impact of various factors on the outcome of ODI cricket matches. Some of the more prominent factors include home ground advantage, team quality (class) and current form. Their analysis is based on the modelling of runs using the normal distribution.

In a non-academic setting, there are currently more than 100 cricket games that have been developed for use on personal computers and gaming stations where some of the games rely on simulation techniques (see www.cricketgames.com/games/index.htm for a comprehensive survey of games and reviews). However, in all of the games that we have inspected, the details of the simulation procedures have not been revealed. Moreover, the games that we have inspected suffer in various ways from a lack of realism.

One-day cricket was introduced in the 1960s as an alternative to traditional forms of cricket that may take up to 5 days to complete. With more aggressive batting, colourful uniforms and fewer matches ending in draws, one-day cricket has become extremely popular. The ultimate event in ODI cricket takes place every 4 years where the World Cup of Cricket is contested.

One-day cricket has some similarities with the sport of baseball. The game is played between two teams of 11 players where a batting team attempts to score runs against a fielding team until the first *innings* terminates. At this stage, the fielding team goes "to bat" and attempts to score more runs before the second innings terminates. A bowler on the fielding team "bowls" balls in groups of six referred to as *overs*. Bowling takes place only from one end of the *pitch* during an over, and the bowling end changes on the subsequent over. A batsman on the batting team faces a bowled ball with a bat and attempts to score *runs*. Runs are scored by running between two *stumps* located at opposite ends of the pitch. The running aspect involves two batsmen known as a *partnership* where each partner is running to the opposite stump. They may score $0, \ldots, 6$ runs according to the number of traversals between the stumps. Therefore, the batsman in the partnership who is located at the batting end faces the next ball. An innings terminates when either 50 overs are completed or 10 *wickets* are lost. A loss of a wicket can occur in various ways with the three most common being (i) a ball is caught in midair after having been batted, (ii) a bowled ball hits the stump located behind the batsman, and (iii) a batsman is *run-out* before reaching the nearest stump. When a wicket is lost, a new batsman is introduced in the batting order. This is a very brief introduction to the rules of one-day cricket, and more information is provided in the article as required. More detail on the rules (laws) of cricket is available from the Lord's website (www.lords.org).

In this article, we develop a simulator for one-day cricket matches. The approach extends the work of Swartz et al. (2006) who investigate the problem of optimal batting orders in one-day cricket for the first innings. Whereas the model used in Swartz et al. (2006) ignores the effect of bowlers, the model used in this article is more realistic in that specific batsman/bowler combinations are considered. In addition, we now provide a method of generating runs in the second innings. Given that only a finite number of outcomes can occur on each ball that is bowled, a discrete generator on a finite set is developed where the outcome probabilities are estimated from historical data involving ODI cricket matches. The probabilities depend on the batsman, the bowler, the number of wickets lost, the number of balls bowled and the current score of the match. The probabilities are obtained using a Bayesian latent variable model which is fitted using WinBUGS software (Spiegelhalter et al., 2004).

In Section 2, we develop a simulator for ODI cricket which is based upon a Bayesian latent variable model. Particular attention is given to second innings batting where the state of the match (e.g., score, wickets, overs) affects the aggressiveness of batsmen. In Section 3, we consider the

adequacy of the approach by comparing simulated results against actual data. The simulator is constructed using data from recent ODI matches. In Section 4, we demonstrate how the simulator can be used to address questions of interest. We conclude with a short discussion in Section 5.

2. SIMULATION

We consider the simulation of runs in the first innings for predetermined batting and bowling orders. We initially investigate the first innings runs since second innings strategies are affected by the number of runs scored in the first innings. By a predetermined batting and bowling order, we mean that a set of rules has been put in place which dictates the batsman and bowler at any given point in the match. These rules could be simple such as maintaining a fixed batting and bowling order. The rules for determining batting and bowling orders could also be very complex. For example, the rules could be Markovian in nature where a specified bowler may be substituted at a state in the match dependent upon the number of wickets lost, the number of overs, the number of runs and the current batsmen. The key point is that they need to be specified in advance for the purpose of simulation.

In one-day cricket, there are a finite number of outcomes arising from each ball bowled. Suppose that the first innings terminate on the mth ball bowled where $m \leq 300$. Ignoring certain rare events (such as scoring 5 runs), and temporarily ignoring wide-balls and no-balls, let X_b denote the outcome of the bth ball bowled, $b = 1, \ldots, m$ where

$$X_b = \begin{cases} 1 & \text{if a wicket is taken} \\ 2 & \text{if the batsman scores 0 runs} \\ 3 & \text{if the batsman scores 1 run} \\ 4 & \text{if the batsman scores 2 runs} \\ 5 & \text{if the batsman scores 3 runs} \\ 6 & \text{if the batsman scores 4 runs} \\ 7 & \text{if the batsman scores 6 runs} \end{cases} \tag{1}$$

and set $X_{m+1} = \cdots = X_{300} = 0$. Note that the coding in (1) includes the possibility of scoring due to byes and and leg byes. Byes and leg byes occur when the batsman has not hit the ball with his bat but decides to run.

Using square brackets to generically denote probability mass functions, the joint distribution of $X_1, \ldots, X_{300}$ can be written as

$$[X_1, \ldots, X_{300}] = [X_{300} \mid X_0, \ldots, X_{299}]\,[X_{299} \mid X_0, \ldots, X_{298}] \cdots [X_2 \mid X_0, X_1][X_1 \mid X_0] \tag{2}$$

where we define $X_0 = 0$ for notational convenience. Note that the conditional probabilities in (2) suggest that the scoring distribution for a given ball depends on the scoring up to that point in time in the first innings. The proposed simulation algorithm is facilitated by the structure in (2) whereby the first outcome X_1 is generated for the first ball bowled, then the second outcome X_2 is generated for the second ball bowled conditional on X_1. The first innings continue until either the overs are completed ($b = 300$) or all wickets are lost (wickets $= 10$). Let v denote the probability of a wide-ball or a no-ball. The simulation algorithm generates a uniform(0,1) random variable u, and if $u < v$, this signals a wide-ball or no-ball condition. In this case, a single run is added to the batting team but the ball is not counted. In addition, further runs may be scored on the no-ball or wide-ball according to the probability ϕ_k for the kth outcome, $k = 1, \ldots, 7$. The following simulation algorithm generates the number of runs R scored in the first innings:

wickets $= 0$
$R = 0$
for $b = 1, \ldots, 300$
  if wickets $= 10$
  then
    $X_b = 0$
  else
    generate $u \sim$ uniform$(0, 1)$ $\star$
    if $u < v$
    then
      generate $Y \sim$ multinomial$(1, \phi_1, \ldots, \phi_7)$
      $R \leftarrow R + 1 + I(Y = 3) + 2I(Y = 4) + 3I(Y = 5) + 4I(Y = 6) + 6I(Y = 7)$
      go to step $\star$
    else
      generate $X_b \sim [X_b \mid X_0, \ldots, X_{b-1}]$
      $R \leftarrow R + I(X_b = 3) + 2I(X_b = 4) + 3I(X_b = 5) + 4I(X_b = 6) + 6I(X_b = 7)$
      wickets $\leftarrow$ wickets $+ I(X_b = 1)$

For the sake of simplicity, the above algorithm does not distinguish a run-out from other forms of wickets. Runs may still be accumulated on a ball prior to a run-out, and the dismissed batsman may be either of the two batsmen in the partnership. We remark that we have estimated the probability of run-outs and have accounted for run-outs in our Fortran implementation of the above algorithm. The proposed simulation algorithm is simple and requires only that we be able to generate from the multinomial$(1, \phi_1, \ldots, \phi_7)$ distribution and the conditional finite discrete distributions given by $[X_b \mid X_0, \ldots, X_{b-1}]$. In the following subsection, we describe a method to compute the conditional distributions by modelling ball by ball outcomes in one-day cricket matches.

2.1. Modelling

The conditional distributions $[X_b \mid X_0, \ldots, X_{b-1}]$ depend on many factors including

- the batsman;
- the bowler;
- the number of wickets lost;
- the number of balls bowled;
- the current score of the match;
- the opposing team;
- the location of the match;
- the coach's advice;
- the condition of the pitch, etc.

For the first innings, we consider the first four factors and define p_{ijwbk} as the probability corresponding to outcome $k = 1, \ldots, 7$ as described in (1), where the the ith batsman, $i = 1, \ldots, I$ faces the jth bowler, $j = 1, \ldots, J$ when $w = 0, \ldots, 9$ wickets have been lost and the bth ball is about to be bowled, $b = 1, \ldots, 300$. Since wide-balls and no-balls have been excluded as possible outcomes of X_b, we have $\sum_k p_{ijwbk} = 1$ for all i, j, w, b. With estimates $\hat{p}_{ijwbk}$, the simulation

 DOI: 10.1002/cjs

algorithm generates outcomes according to

$$\text{Prob}(X_b = k \mid X_0, \ldots, X_{b-1}) = \hat{p}_{ijwbk}. \tag{3}$$

Our data are based on 472 ODI matches from January 2001 until July 2006 amongst the 10 full member nations of the International Cricket Council (ICC). These matches are those for which ball by ball commentary is available on the Cricinfo website (www.cricinfo.com) and include almost all matches amongst the 10 nations during the specified time period. We note that a Powerplay rule was introduced for a 10-month trial period beginning August 2005 and the Supersub rule was introduced for a portion of the trial period. Although we believe that these temporary rules had some effect on scoring, we do not account for the presence and absence of the temporary rules in our modelling. In the 472 matches, 257,922 balls were bowled involving $I = 435$ batsman and $J = 360$ bowlers. In the first innings, 138,439 balls were bowled. In Table 1, we record the number of matches for which data were collected on the ICC teams. Except Bangladesh, we observe that there is reasonable balance in the number of matches played.

Over these matches, we calculate $\hat{v} = 8289/257{,}922 = 0.032$ as the total number of wide-balls and no-balls divided by the total number of balls bowled. The conditional probabilities ϕ_k are similarly estimated by frequencies.

Before describing the model used in the estimation of p_{ijwbk}, we offer some preliminary comments based on our experience in fitting numerous models over several years. Most importantly, our goal is to obtain a model which fits well and provides a realistic simulator. With regard to estimation and prediction, there are many situations for which data do not exist. For example, a given batsman i may never have faced a given bowler j in the third over when two wickets are lost. To predict what may happen in these situations it is necessary to borrow information from similar situations. For example, it is relevant in the above problem to consider how other batsman/bowler combinations fared in the third over when two wickets are lost, and it is also relevant to consider how batsman i fared against bowler j in other stages of a match. With nearly 1,000 batsmen and bowlers, our experience suggests that it is not a good idea to have multiple parameters for each batsman and bowler since this leads to excessive parameterization. When the number of parameters is excessive, computation may be overwhelming and unreliable estimates may arise. We strive for parsimonious model building, assigning only a single parameter to each batsman and assigning only a single parameter to each bowler.

TABLE 1: The number of ODI matches for which data were collected on the ICC teams.

ICC nation	Number of matches
Australia	106
Bangladesh	48
England	89
India	125
New Zealand	94
Pakistan	108
South Africa	101
Sri Lanka	118
West Indies	77
Zimbabwe	78

To obtain the estimates $\hat{p}_{ijwbk}$ in (3), we develop a Bayesian latent variable model which is related to the classical cumulative-logit models for ordinal responses (Agresti, 2002). We initially assume that batting is taking place in the first innings and we temporarily ignore the state of the first innings (e.g., the number of overs completed and the number of wickets lost). We imagine that there is a latent continuous variable U which describes the "quality" of the batting outcome. For example, one might imagine that a ball batted near a fielder has a lower rating than a similar ball batted slightly further away from the fielder. Although U is unobserved (latent), there is an observed data variable X which is related to U as follows:

$$\begin{aligned}
X = 1 \text{ (dismissal)} &\leftrightarrow a_0 < U \leq a_1 \\
X = 2 \text{ (0 runs)} &\leftrightarrow a_1 < U \leq a_2 \\
X = 3 \text{ (1 runs)} &\leftrightarrow a_2 < U \leq a_3 \\
X = 4 \text{ (2 runs)} &\leftrightarrow a_3 < U \leq a_4 \\
X = 5 \text{ (3 runs)} &\leftrightarrow a_4 < U \leq a_5 \\
X = 6 \text{ (4 runs)} &\leftrightarrow a_5 < U \leq a_6 \\
X = 7 \text{ (6 runs)} &\leftrightarrow a_6 < U \leq a_7
\end{aligned}$$

where $-\infty = a_0 < a_1 < \cdots < a_6 < a_7 = \infty$. Note that increasing values of X correspond to higher quality of batting outcomes and that $a_1, \ldots, a_6$ are unknown parameters.

We now define a single batsman characteristic $\mu_i^{(1)}$ for the ith batsman and a single bowler characteristic $\mu_j^{(2)}$ for the jth bowler. We assume that the quality of the batting outcome U is expressed as

$$U = \mu_i^{(1)} - \mu_j^{(2)} + \epsilon \tag{4}$$

where $\mu^{(1)} = 0$ represents an average batsmen, $\mu^{(2)} = 0$ represents an average bowler and ϵ is a random variable whose distribution determines the quality of the batting outcome for an average batsman and an average bowler. From (4), we observe that a good batsman ($\mu_i^{(1)} > 0$) increases the quality of the batting outcome. Similarly, a good bowler ($\mu_j^{(2)} > 0$) decreases the quality of the batting outcome. Letting F denote the distribution function of ϵ, we write

$$\begin{aligned}
\text{Prob}(X \leq k) &= \text{Prob}(U \leq a_k) \\
&= \text{Prob}\left(\mu_i^{(1)} - \mu_j^{(2)} + \epsilon \leq a_k\right) \\
&= \text{Prob}\left(\epsilon \leq a_k - \mu_i^{(1)} + \mu_j^{(2)}\right) \\
&= F\left(a_k - \mu_i^{(1)} + \mu_j^{(2)}\right)
\end{aligned} \tag{5}$$

for the batting outcome $k = 1, \ldots, 7$.

To gain a better appreciation of the model given by (5), refer to Figure 1 where the density function of the logistic distribution (i.e., $F(\epsilon) = 1/(1 + \exp(-\epsilon))$) has been chosen using typical values of $a_1, \ldots, a_6$. The logistic distribution is symmetric about 0 and has longer tails than the normal distribution. We observe that the area under the density function between a_{k-1} and a_k corresponds to the probability that $X = k$. The effect of better batsmen ($\mu_i^{(1)} > 0$) and weaker bowlers ($\mu_j^{(2)} < 0$) causes a simultaneous leftward shift in the vertical lines and hence decreases the probability of dismissal and increases the probability of scoring 6 runs.

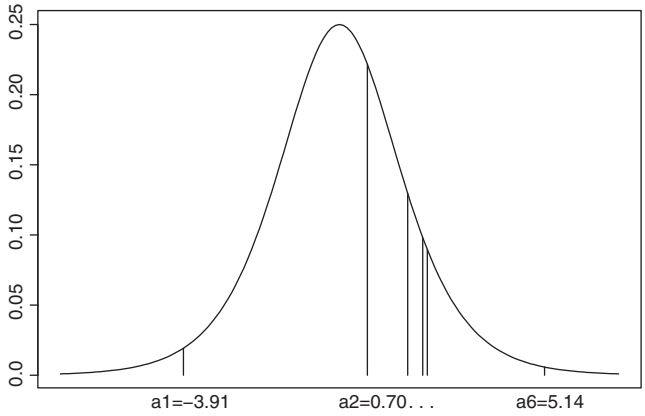

FIGURE 1: Logistic density function and typical parameters for the model in (5).

A difficulty with the model given by (5) is that it does not account for the variability in aggressiveness which batsmen display during the first innings. For example, it is well known that batsmen become more aggressive in the final overs if few wickets have been lost. Increasing aggressiveness corresponds to an increase in the probability of dismissal ($k = 1$), a decrease in the probability of scoring 0 runs ($k = 2$) and an increase in the probability of scoring 6 runs ($k = 7$). The effect of increasing aggressiveness on the remaining four outcomes $k = 3, 4, 5, 6$ is not as obvious and is situation dependent. Note that aggressiveness is a characteristic that affects the parameters a_k and is different from the quality of batsmen $\mu_i^{(1)}$ and the quality of bowlers $\mu_j^{(2)}$. In the spirit of parsimonious model building, in Table 2 we propose nine situations where aggressiveness is assumed constant. Situations 1, 2, and 3 are motivated by the fielding restriction which is in place during the first 15 overs. We view overs 16–35 as the middle period of the first innings where any change in aggressiveness is due to wickets lost. The final period of the first innings (overs 36–50) can lead to very aggressive batting if few wickets are lost. We note that situations 4, 7, and 8 are particularly rare and the parameters corresponding to these situations may not be well estimated. However, this is not a great concern as simulation rarely takes place during these periods.

TABLE 2: The nine first innings situations where aggressiveness is assumed constant.

Situation	Over	Wickets lost	Percentage of dataset (%)
1	1–15	0–3	30.3
2	16–35	0–3	25.3
3	36–50	0–3	5.0
4	1–15	4–6	1.2
5	16–35	4–6	14.0
6	36–50	4–6	15.2
7	1–15	7–9	0.0
8	16–35	7–9	1.7
9	36–50	7–9	7.3

The fourth column provides the percentage of balls in the dataset that correspond to the given situation.

Having proposed the nine situations in Table 2, we modify the model given by (5) to take into account the varying levels of aggressiveness. We introduce a new subscript $l = 1, \ldots, 9$ to denote the nine situations and we note that l is really a function of w (wickets lost) and b (ball currently facing) as indicated in (3). Consider then a batting outcome where batsman i faces bowler j in the lth situation and outcome k is recorded. The likelihood contribution due to the event is

$$F\left(a_{lk} - \mu_i^{(1)} - \Delta_l + \mu_j^{(2)}\right) - F\left(a_{l,k-1} - \mu_i^{(1)} - \Delta_l + \mu_j^{(2)}\right) \tag{6}$$

where $\Delta_l = 0$ for $l = 1, 2, 3$, $\Delta_l = \Delta^{(1)}$ for $l = 4, 5, 6$, and $\Delta_l = \Delta^{(1)} + \Delta^{(2)}$ for $l = 7, 8, 9$. The complete likelihood is therefore the product of terms of the form (6) over all balls bowled in the dataset. The additional parameters $\Delta^{(1)}$ and $\Delta^{(2)}$ provide a link to situations where batsmen do not typically bat. For example, batsmen who bat in positions 7, 8, and 9 in the batting order are usually bowlers and are usually not very good batsmen. If the model given by (6) were fit without the Δ terms, then the $\mu_i^{(1)}$ values for these batsmen would be relative only to batsmen who batted in situations 7, 8, and 9. We therefore adjust situational skill levels using the Δs. Recall that our intention is to develop a simulator that allows experimentation whereby batsmen may bat in atypical positions in the batting order. The estimation of $\Delta^{(1)}$ and $\Delta^{(2)}$ is feasible due to data corresponding to batsmen who crossover. For example, $\Delta^{(1)}$ is estimable since there are some batsmen who have data in at least one of situations 1, 2, or 3 and who have data in at least one of situations 4, 5, or 6. We remark that our approach to handling situational skill levels and aggressiveness in the first innings might be better handled in some sort of continuous fashion rather than imposing nine states where homogeneity is assumed.

From (6) the primary parameters of interest are $\Delta^{(1)}$, $\Delta^{(2)}$, the a_{lk}s, the $\mu_i^{(1)}$s and the $\mu_j^{(2)}$s. This corresponds to $1 + 1 + 9(6) + 435 + 360 = 851$ unknown primary parameters. In a Bayesian formulation, parameters have prior distributions and we assign the following prior distributions

$$\begin{aligned}
\Delta^{(1)} &\sim \text{uniform}(0, 1) \\
\Delta^{(2)} &\sim \text{uniform}(0, 1) \\
a_{lk} &\sim \text{normal}(0, \sigma^2) \quad \text{where } \sigma^{-2} \sim \text{gamma}(1.0, 1.0) \\
\mu_i^{(1)} &\sim \text{normal}(0, \tau^2) \\
\mu_j^{(2)} &\sim \text{normal}(0, \tau^2) \quad \text{where } \tau^{-2} \sim \text{gamma}(1.0, 1.0)
\end{aligned}$$

for batting outcomes $k = 1, \ldots, 6$, for situations $l = 1, \ldots, 9$, for batsmen $i = 1, \ldots, I = 435$ and for bowlers $j = 1, \ldots, J = 360$. The prior distributions are assumed independent except for the order restriction $a_{l1} < a_{l2} < \cdots < a_{l6}$ for $l = 1, \ldots, 9$. The notation $Y \sim \text{gamma}(a, b)$ implies $\text{E}(Y) = a/b$ and $\text{Var}(Y) = a/b^2$. Note that although the prior distributions are somewhat diffuse, prior knowledge is used in the prior specification. For example, it is known that batsmen who bat in positions 1, 2, and 3 in the batting order are generally better than batsmen who bat in positions 4, 5, and 6 in the batting order who are in turn generally better than batsmen who bat in positions 7, 8, and 9 in the batting order. This knowledge implies that $\Delta^{(1)} > 0$ and $\Delta^{(2)} > 0$. The prior distributions for σ^{-2} and τ^{-2} make use of the knowledge that the bulk of the probability for the logistic distribution F lies in the interval $(-5, 5)$. Also, the modelling of common prior means for the $\mu_i^{(1)}$ and the modelling of common prior means for the $\mu_j^{(2)}$ is sensible in that it produces average characteristics for batsmen and bowlers for whom little data has been collected. Finally, note that our prior specification introduces only two extra hyperparameters (σ and τ).

Our model can be specified within the WinBUGS platform, and upon providing the data, WinBUGS constructs a Markov chain whose output consists of generated parameters. Since the Markov chain converges to the posterior distribution, the generated parameters can be averaged to

obtain approximations of the posterior means. When the parameters are estimated, the probabilities $\hat{p}_{ijwbk}$ for the simulator (3) are obtained by substituting the relevant primary parameter estimates into (6). One of the features of the model is that it permits inference on various batsman/bowler combinations at different stages of a match even when they have not faced one another in actual competition.

In our WinBUGS implementation, we used a burn-in of 10,000 iterations and a further 10,000 iterations for parameter estimation. This required approximately 1 day of computation. Standard diagnostics such as trace plots, autocorrelation plots and varied starting values were used to assess convergence. We experimented with changes to the vague prior distributions (particularly the gammas) and found that our results were not sensitive to the prior specification.

There is an important remaining point that needs to be made concerning the fitting of the Bayesian latent variable model. In order to fit the model, we require ball by ball data to specify the terms in (6). To obtain the data, the ball by ball commentary log from the Cricinfo website was parsed into a convenient format. For this, we used a Java script that extracted the relevant details on a ball by ball basis, and stored the data in a tabular form. For example, codes were created to index batsmen and bowlers, and outcomes were categorized according to (1).

2.2. Generating Runs in the Second Innings

Up until now, we have considered only the generation of first innings runs. It is evident that the conditional distributions $[X_b \mid X_0, \ldots, X_{b-1}]$ for the second innings also depend on the current score of the match. For example, it is well known that when the team batting first scores an unusually high number of runs, the team batting second becomes more aggressive in its batting style.

One idea is to modify the model given by (6) so that batting outcome probabilities also depend on the score of the match. One could imagine introducing additional subscripts to the a_{lk} terms to denote the score of the match. The problem with such an approach is that there would be many more parameters and very few replicate observations for the purposes of estimation.

Our approach is to leave the model given by (6) as it stands, and view the p_{ijwbk} terms in (3) as the first innings probabilities which share some characteristics with the second innings probabilities. We account for the current score in the second innings by modifying the conditional distributions $[X_b \mid X_0, \ldots, X_{b-1}]$ used in the algorithm for simulation. Consider then the stage of the second innings where batsman i faces bowler j, w wickets have been lost and the bth ball is about to be bowled. For notational convenience, we suppress the subscripted notation $ijwb$. Then, referring to (1) and ignoring wide-balls and no-balls, the expected number of runs that the batsman scores on the current ball is

$$\mathrm{E}_1(p) = p_3 + 2p_4 + 3p_5 + 4p_6 + 6p_7$$

and the expected proportion of resources that the batsman consumes on the current ball is

$$\begin{aligned}\mathrm{E}_2(p) &= (x+y)p_1 + xp_2 + xp_3 + xp_4 + xp_5 + xp_6 + xp_7 \\ &= x + yp_1\end{aligned}$$

where the proportion of resources lost x due to the current ball and the proportion of resources lost y due to a wicket are known quantities that are available from the Duckworth/Lewis resource table (Duckworth & Lewis, 1998, 2004). For the batsman to become more aggressive during the match, this implies a change in the probabilities $p = (p_1, \ldots, p_7)$. We make the assumption that the batsman modifies his overall batting behaviour from p to $p' = (p'_1, \ldots, p'_7)$ according to the current score in the match.

Consider now the situation where f runs have been scored in the first innings, s runs have been scored in the second innings and the proportion of resources remaining in the second innings is r. To win, the team batting second needs to score $f - s + 1$ runs in the remainder of the match relative to the r resources that are available. This suggests that the run to resource ratio for the batsman should be at least $(f - s + 1)/r$. If $(f - s + 1)/r > \mathrm{E}_1(p)/\mathrm{E}_2(p)$, then the team batting second is on the verge of losing/tying, and we assume that the batsman becomes more aggressive in his batting style. We therefore propose that the batsman modifies his style from p to p' where

$$p'_2 = cp_2, \quad c \in (0, 1). \tag{7}$$

The idea is that a more aggressive batsman swings the bat more often and is less likely to score 0 runs (i.e., $p'_2 < p_2$). When $c = 1$, the batsman is behaving in a neutral fashion (i.e., not extra aggressive), and when $c = 0$, the batsman has reached his limit of aggressive behaviour where scoring 0 runs is impossible. Accordingly, when $(f - s + 1)/r > \mathrm{E}_1(p)/\mathrm{E}_2(p)$, we set

$$c = \frac{r\mathrm{E}_1(p)}{(f - s + 1)\mathrm{E}_2(p)}. \tag{8}$$

We propose that when $c \in (0, 1)$ as in (8), the decrease in probability from p_2 to p'_2 results in an increase in the probability of dismissal

$$p'_1 = p_1 + \delta p_2(1 - c), \quad \delta \in [0, 1] \tag{9}$$

and a proportional increase in the run-scoring probabilities

$$p'_i = \left(\frac{1 - p_1 - (c + \delta(1 - c))p_2}{1 - p_1 - p_2}\right) p_i, \quad i = 3, \ldots, 7. \tag{10}$$

It is easy to establish that $p'_1, \ldots, p'_7$ form a simplex, and hence constitute a probability distribution. Observe that the free parameter $\delta \in [0, 1]$ determines the degree to which the aggressive behaviour affects dismissals (9) and run-scoring (10). When $\delta = 1$, all of the aggressive behaviour increases the dismissal probability, and when $\delta = 0$, all of the aggressive behaviour increases the run-scoring probabilities.

When a batsman is aggressive in the second innings (i.e., $c \in (0, 1)$), it is easy to establish that

$$\mathrm{E}_1(p') = \left(\frac{1 - p_1 - (c + \delta(1 - c))p_2}{1 - p_1 - p_2}\right) E_1(p) \geq \mathrm{E}_1(p) \tag{11}$$

and

$$\mathrm{E}_2(p') = E_2(p) + \delta p_2(1 - c)y \geq \mathrm{E}_2(p). \tag{12}$$

In modifying his behaviour from p to p', the inequalities (11) and (12) imply that the batsman simultaneously increases the expected number of runs scored and the expected number of resources consumed on the current ball. Moreover, it is straightforward to show that both $\mathrm{E}_1(p')$ and $\mathrm{E}_2(p')$ are decreasing functions of $c \in (0, 1)$. These consequences correspond to our intuition of more aggressive batting.

The remaining detail in modelling aggressive batting in the second innings is the determination of the parameter $\delta \in [0, 1]$. Although an aggressive batsman is attempting to increase his run production $\mathrm{E}_1(p')$, the quantity which really determines the quality of batting is $\mathrm{E}_1(p')/\mathrm{E}_2(p')$. We argue that a batsman is unable to modify his batting style from p to p' so as to make

$E_1(p')/E_2(p') > E_1(p)/E_2(p)$; for if he were able to do this, he would do it all the time, in both the first and second innings. However, when a team is on the verge of losing (i.e., $(f - s + 1)/r > E_1(p)/E_2(p)$), we suggest that a batsman may be willing to sacrifice $E_1(p')/E_2(p')$ with the benefit of increased run production $E_1(p')$. In other words, we require $E_1(p')/E_2(p') \leq E_1(p)/E_2(p)$ and that $E_1(p')/E_2(p')$ be an increasing function of $c \in (0, 1)$. Using the expressions in (11) and (12), it is possible to show that $E_1(p')/E_2(p')$ is an increasing function of $c \in (0, 1)$ provided $\delta > E_2(p)/(E_2(p) + y(1 - p_1 - p_2)) \in (0, 1)$. Since a batsman would naturally desire $E_1(p')/E_2(p')$ to be as large as possible, we therefore set

$$\delta = \frac{E_2(p)}{E_2(p) + y(1 - p_1 - p_2)}. \tag{13}$$

If $(f - s + 1)/r < E_1(p)/E_2(p)$, the team batting second is on the verge of winning, and although it may not be optimal, batsmen become more cautious. The tendency to become more cautious when protecting a lead is widely acknowledged in many sports including American football, basketball, and ice hockey. Following the above development for aggressive batting, a similar modification in a batsman's style from p to p' can be obtained. The idea is that a more cautious batsman provides greater protection of the wicket and is more likely to score 0 runs. Specifically, we set c and δ according to (8) and (13), respectively, and we determine the probabilities $p'_1, \ldots, p'_7$ according to (7), (9), and (10). In this case, increasing c corresponds to increasing cautiousness. We note that $p'_2 > p_2$ and $p'_i < p_i$ for the remaining probabilities $i = 1, 3, 4, 5, 6, 7$. We keep in mind that it may be necessary to reduce c to an upper bound to ensure that the probability vector p' forms a simplex. We note that the only way that the limit of cautious behaviour is reached (i.e., $c = 1/p_2$) is when $(f - s + 1)/r = 0$ which means that the batting side has already won the match and batting has terminated.

There is a final modification in our approach to second innings batting which we now consider. In many sporting activities it is an advantage to have the final offensive opportunity in a game. For example, this is the case in baseball where it is widely viewed as advantageous to bat in the bottom innings since strategy varies according to the number of runs required. Similarly, it is generally beneficial in golf to play in the last foursome of a tournament since the score to beat is known. It therefore seems reasonable that a second innings batting advantage should also exist in one-day cricket. An advantage in second innings batting seems to go hand in hand with the quotation from Sir Francis Bacon that "knowledge is power". However, upon looking at empirical data, de Silva & Swartz (1997) found no such advantage in second innings batting for ODI cricket. A possible explanation is that the strategic advantage in second innings batting is offset by the deterioration of the pitch during the second innings. As a match progresses, the pitch often becomes worn down, and batting becomes more difficult as bowling is subject to more erratic bounces. Although our procedure for generating runs in the second innings takes strategy into account through modified aggression levels in batting, our procedure fails to account for the deterioration in the pitch. We therefore introduce the "pitch variable" η in (6) which is activated in second innings batting but not in first innings batting. We define η as an offset to the parameters a_{l1}, modifying a_{l1} to $a_{l1} + \eta$. This has the effect of increasing the probability of dismissal in second innings batting. Since our estimation procedure is based on first innings data only, we do not treat the pitch variable η as a standard parameter which is estimated but rather we treat η as a tuning parameter. We have set $\eta = 0.15$ (a very small adjustment relative to the size of a_{l1}) to coincide with the findings of de Silva & Swartz (1997). The treatment of η as a tuning parameter may be more appropriate than viewing η as a pitch variable. As pointed out by a referee, it is not always the case that batting conditions deteriorate during the course of a match.

3. TESTING MODEL ADEQUACY

As a type of cross-validation procedure, we fit the Bayesian latent variable model using only first innings data. Although this reduces the size of the dataset (by roughly 50%), it permits us to compare simulated results for the second innings with actual second innings results that were not used in determining the parameter estimates.

The model was fit using WinBUGS software where posterior means were calculated for the 853 model parameters. Although the WinBUGS program requires two hours of computation, once the parameter estimates are obtained, they can be used over and over again as inputs to the simulation program.

Another advantage of the Bayesian formulation concerns the use of parameter estimates in the simulation program. It is a widely held belief that the performances of batsmen and bowlers are not constant. For example, batsmen have good days and bad days, and this can be related to their health or any number of reasons. In a Bayesian formulation, we need not use the same parameter estimates $\mu^{(1)}$ and $\mu^{(2)}$ (posterior means) for batsmen and bowlers over all matches. Alternatively, at the beginning of a match, the $\mu^{(1)}$ and $\mu^{(2)}$ values can be generated from their respective posterior distributions to reflect match by match variation in performance.

Now, there are countless ways that one might test the adequacy of the model. In Table 3, we provide the estimated probabilities p_{ijwbk} for some batsmen/bowler combinations at different states of a match. We have also included the expected number of runs per over for each combination. We have presented batting outcome probabilities when Alistair Cook of England is batting against Glenn McGrath of Australia, and against Nazmul Hossain of Bangladesh. At the beginning of a match (i.e., ball 1, 0 wickets), we observe that with probabilities 0.681 and 0.078, Cook scores 0 runs and 4 runs respectively against McGrath. At the beginning of a match, these probabilities change to 0.626 and 0.100 respectively when Hossain is bowling. These changes are consistent with the general belief that McGrath is a better bowler than Hossain. We then investigate a situation where batsmen ought to become more aggressive (ball 271 when 2 wickets are lost). Indeed, the probability that Cook scores 0 runs decreases substantially to 0.338 and 0.285 depending on whether McGrath or Hossain is the bowler. We also note a curious result concerning the probability of scoring 4 runs. Even though batsmen are more aggressive on ball 271 with 2 wickets than at the beginning of a match (ball 1 with 0 wickets), the fielding restriction that is in place at the beginning of a match enables batsmen to score 4's at a higher rate. This batting behaviour is observed in Table 3 and has been verified by looking at empirical data. In Table 3, we also investigate the case of ball 271 when 4 wickets are lost which according to common knowledge should be a less aggressive batting situation than ball 271 when 2 wickets

TABLE 3: Batting probabilities p for various states and the expected number of runs per over $E(R)$ where CM denotes the Cook/McGrath matchup and CH denotes the Cook/Hossain matchup.

State of the match	Dismissal	Zero	One	Two	Three	Four	Six	$E(R)$
CM (ball 1, 0 wickets)	0.024	0.681	0.165	0.038	0.010	0.078	0.004	3.7
CM (ball 271, 2 wickets)	0.039	0.338	0.452	0.077	0.006	0.069	0.018	6.1
CM (ball 271, 4 wickets)	0.033	0.352	0.435	0.085	0.007	0.071	0.017	6.1
CH (ball 1, 0 wickets)	0.018	0.626	0.191	0.047	0.012	0.100	0.006	4.5
CH (ball 271, 2 wickets)	0.030	0.285	0.472	0.094	0.008	0.088	0.024	7.1
CH (balls 271, 4 wickets)	0.025	0.297	0.454	0.103	0.008	0.091	0.022	7.1

are lost. Accordingly, we observe that the probability of 0s increase and the probability of 1s decrease in the less aggressive situation.

In Table 4, we investigate the adjustment from p to p' in the second innings. Consider again the matchup between the batsman Cook and the bowler Hossain. Suppose that Bangladesh has scored $f = 250$ runs in the first innings. Suppose further that ball $b = 183$ is about to be bowled and $w = 3$ wickets have been lost in the second innings (situation 2). From the Duckworth/Lewis table, we therefore have the proportion of resources lost $x = 0.0027$ due to the current ball and the proportion of resources lost $y = 0.044$ due to a wicket. This might be considered as a "middle point" of the second innings since the proportion of resources used is $R(w, b) = 0.4993$, and therefore, the proportion of resources remaining is $r = 0.5007$. In this case, the estimated parameters give $E_1(p) = 0.9723$ and $E_2(p) = 0.00393$. We now investigate the outcome probabilities p'_{ijwbk} when England has scored $s = 127, 90, 60$ runs. When $s = 127$, then $(f - s + 1)/r = 247.7 \approx E_1(p)/E_2(p) = 247.5$ and England is on pace to draw the match, and $p' = p$ (i.e., no adjustment). When $s = 90$, Cook should become more aggressive ($c = 0.768$), and when $s = 60$, Cook should become even more aggressive ($c = 0.648$). The entries in Table 4 appear reasonable and support these tendencies.

We now compare actual runs versus simulated runs. For this, we consider the 23 matches between Sri Lanka and India from November 1998 through March 2007 in which Sri Lanka batted first. The 23 matches consist of 15 matches from the original dataset (2001–2006) used for model fitting and 8 matches outside of the training period. We simulate 1,000 first innings results for Sri Lanka based on representative batting and bowling orders employed during the time period. The resultant QQ plot comparing the actual runs and the simulated runs is given in Figure 2. We observe excellent agreement and we remark that satisfactory plots are also observed for other pairs of teams that we investigated. In comparing wickets taken, the actual results also compare favourably with the simulated results.

We also investigate the effect of the second innings adjustment. The difficulty in this exercise is that given the number of first innings runs between two teams, replicate observations tend not to occur. Therefore, to address goodness-of-fit, we provide some evidence that the second innings adjustment p' is an improvement over having the second innings team bat in a neutral fashion (i.e., $p' = p$). Consider then simulated matches between Australia and the other nine ICC teams where Australia is batting in the second innings and where the batting and bowling lineups resemble those used in the 2007 World Cup matches. We generate first innings runs for the other teams, and then second innings runs for Australia with the proposed batting adjustment p' based on the target scores. The simulation is repeated for 1,000 hypothetical matches for each of the 9 teams. We observe that Australia uses their full 50 overs in 8.5% of the simulated matches. The small percentage seems sensible since Australia rarely uses all 50 overs in matches that they win. In matches that Australia loses, at some point when they are falling behind, they become

TABLE 4: Second innings batting probabilities p' and the expected number of runs per over for the Cook/Hossain matchup when Bangladesh has scored $f = 250$ runs in the first innings.

England runs (s)	Dismissal	Zero	One	Two	Three	Four	Six	Runs/over
127	0.016	0.452	0.397	0.058	0.008	0.062	0.008	5.0
90	0.029	0.347	0.466	0.068	0.010	0.072	0.009	5.8
60	0.036	0.293	0.501	0.073	0.010	0.078	0.010	6.3

In the second innings, $w = 3$ wickets have been lost, ball $b = 183$ is about to be bowled and England has scored s runs.

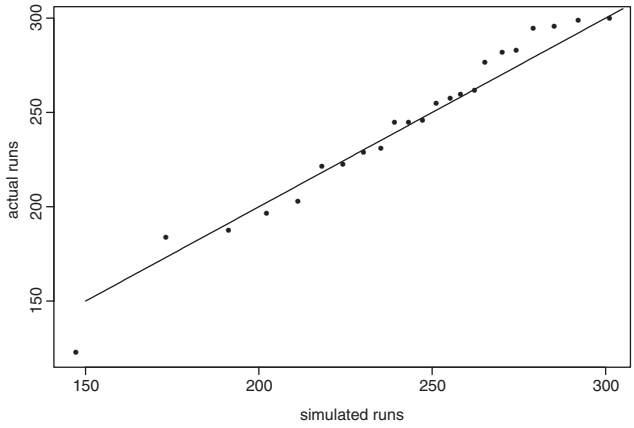

FIGURE 2: QQ plot corresponding to first innings runs for Sri Lanka batting against India.

desperate (aggressive), and typically consume all of their wickets before using the allotted 50 overs. When we repeat the simulation with neutral batting in the second innings (i.e., Australia behave as they would in the first innings), Australia uses all 50 overs 13% of the time. To get a sense of the percentages using actual data, we look at all 83 matches from 2000 to 2006 where Australia batted second, and observe that Australia used the full 50 overs only 7% of the time. This suggests that there is merit in our modification of aggressiveness in second innings batting.

4. ADDRESSING QUESTIONS VIA SIMULATION

Having developed a simulator for ODI cricket matches, there is no limit to the number and type of questions that may be posed. The greatest utility of the simulator occurs for circumstances in which there is limited empirical data. In these cases, without a simulator, the best that one can do is to rely on hunches with respect to the questions of interest. In this section, we give a flavour for the types of questions that might be posed. We see these types of applications as being of value not only to cricket devotees but also to selection committees and team strategists. We note that each of the simulations described below requires less than 1 min of computation.

4.1. Question 1

Adam Gilchrist is often an opening batsman for Australia, and Australia has not played the West Indies often in recent history. We are interested in the probability of Gilchrist hitting a century as an opening batsman against the West Indies when Australia is batting in the first innings and the West Indies are using a bowling lineup from the 2007 World Cup. Based on 1,000 first innings simulations, Gilchrist reaches a century 5.1% of the time. The result appears consistent with Gilchrist's actual batting performances where Gilchrist made a century 8 times as an opening batsman in 138 first innings ODI matches (5.8%) throughout his career (1996–2008).

4.2. Question 2

England has occasionally sent Alastair Cook and Matt Prior as opening batsmen. In other matches, they used Ian Bell and Michael Vaughan as opening batsmen. We are interested in the performance in the untested opening partnership of Alastair Cook and Ian Bell. More specifically, we consider the length of the partnership (i.e., the number of overs prior to losing the first wicket) for Cook

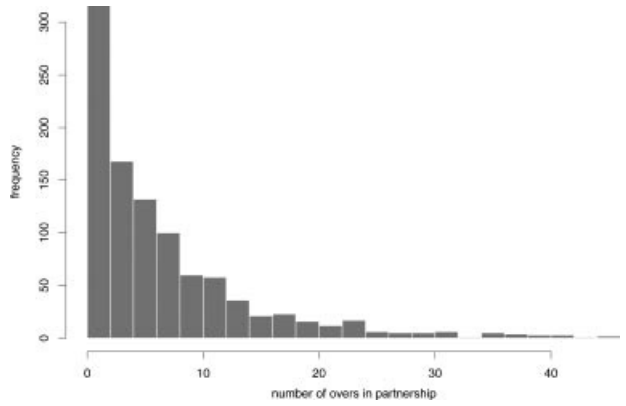

FIGURE 3: Histogram of the length of the partnership (in overs) of the untested opening partnership of Alastair Cook and Ian Bell.

and Bell when they are batting in the first innings against Pakistan where Pakistan uses a bowling lineup comparable to the lineup used in their December 15, 2005 match against England. In Figure 3, we provide a histogram of the number of overs in the length of their partnership based on 1,000 simulations. We observe that the median and the mean length of the partnership is 5 overs and 7.1 overs respectively. It appears very unlikely for Cook and Bell to have a partnership exceeding 20 overs.

4.3. Question 3

Consider a match between New Zealand and Sri Lanka where Sri Lanka is batting in the second innings and New Zealand has scored an impressive 300 runs in the first innings. We are interested in the probability that Sri Lanka can overcome the barrier and win the match. Based on 1,000 simulations, and taking batting/bowling lineups used in the 2007 World Cup matches, we observe that Sri Lanka wins 10.0% of the simulated matches. The result is consistent with Sri Lanka's first innings performance in the current decade where Sri Lanka has scored over 300 runs in 9 out of 121 matches (i.e., 7.4% of the time).

4.4. Question 4

Muttiah Muralitharan of Sri Lanka is a spin bowler and is widely regarded as one of the best bowlers in cricket. The question arises as to his value to the Sri Lankan team. Consider a match between Sri Lanka and India where India is batting in the first innings and India's batting lineup is based on their 2007 World Cup team. We use a bowling lineup comparable to Sri Lanka's bowling lineup in the 2007 World Cup where Muralitharan was a prominent bowler. Based on 1,000 simulations, we observe that India scores 247 runs on average. When we make Muralitharan the only Sri Lankan bowler (which is of course against the rules), then India scores only 185 runs on average. Clearly, if every bowler on Sri Lanka were as good as Muralitharan, Sri Lanka would have a much better team. When we instead replace Muralitharan with Upul Chandana (a more typical bowler), then India scores 267 runs on average.

4.5. Question 5

Here is a crazy question that surely few people have contemplated. What would happen if we reverse a team's batting order? We consider a batting order used by Australia in the 2007 World Cup

matches. Based on 1,000 simulations against each of the other nine teams using their 2007 World Cup bowling lineups, Australia produces on average 272 runs, losing 6.3 wickets in 48.8 overs during the first innings. This compares favourably with empirical data over the data collection period where Australia produces on average 273 runs, losing 6.8 wickets in 49.8 overs during the first innings. When we reverse the Australian batting lineup, and simulate 1,000 matches against each of the other 9 teams, Australia produces on average 234 runs, losing 7.3 wickets in 48.4 overs during the first innings. A simple explanation for the difference in expected runs is that higher-scoring batsmen tend to be placed at the beginning of the lineup. When the batting order is reversed, they often do not get an opportunity to bat or they bat for shorter periods of time.

4.6. Question 6

We now determine the probability that one team defeats another team. Using typical batting and bowling lineups taken from matches during the 2007 World Cup of Cricket, Table 5 provides estimated probabilities based on 10,000 simulations between each pair of ODI teams. Accordingly, Australia is clearly the best team, and Bangladesh and Zimbabwe are the weakest teams. The probabilities for the other teams roughly agree with the authors' beliefs although we note that the probabilities are sensitive to the choice of the batting and bowling lineups. Referring to the row and column averages, we observe that the probability of winning is nearly the same for first and second innings batting. This corresponds to the observations of de Silva & Swartz (1997) and provides a justification for the tuning parameter η introduced at the end of Section 2.

5. DISCUSSION

In this article, a simulator for ODI cricket is developed. One of the virtues of the approach is that the characteristics of individual batsmen and bowlers are used to generate ball by ball outcomes. As time progresses and more matches are played, the database may be updated to reflect changes in player characteristics.

TABLE 5: Estimated probabilities of the row team defeating the column team where the row team corresponds to the team batting first.

Batting first	Batting second										
	Australia	Bangladesh	England	India	New Zealand	Pakistan	South Africa	Sri Lanka	West Indies	Zimbabwe	Average
Australia		0.88	0.69	0.65	0.72	0.81	0.58	0.64	0.74	0.96	0.74
Bangladesh	0.14		0.29	0.23	0.34	0.42	0.19	0.25	0.35	0.77	0.33
England	0.29	0.75		0.43	0.54	0.61	0.36	0.43	0.54	0.89	0.54
India	0.33	0.75	0.54		0.57	0.65	0.40	0.46	0.57	0.89	0.57
New Zealand	0.30	0.81	0.50	0.44		0.64	0.40	0.46	0.59	0.92	0.56
Pakistan	0.16	0.58	0.31	0.27	0.37		0.22	0.28	0.37	0.78	0.37
South Africa	0.44	0.89	0.63	0.57	0.71	0.77		0.60	0.72	0.96	0.70
Sri Lanka	0.37	0.81	0.58	0.52	0.61	0.70	0.46		0.63	0.92	0.62
West Indies	0.23	0.67	0.41	0.34	0.47	0.55	0.30	0.36		0.85	0.46
Zimbabwe	0.05	0.33	0.12	0.10	0.15	0.21	0.08	0.12	0.16		0.15
Average	0.74	0.28	0.55	0.61	0.50	0.40	0.67	0.60	0.48	0.12	

The final column are the row averages and correspond to the average probabilities of winning when batting in the first innings. The final row are the average probabilities of winning when batting in the second innings.

Beyond obvious uses in betting, the simulator may be used to help teams determine optimal strategies. In ODI cricket, it is not so easy to test ideas as a team may only play 20 matches per year and a team does not typically play all ICC nations. For example, it is not clear how changes in the batting and bowling orders affect a match. The simulator allows a team to easily investigate the results of making changes to the batting and bowling orders.

Our simulator was developed with an attempt to realistically model one-day cricket, and it appears to do a reasonable job of reflecting major tendencies. Nevertheless, there are improvements that might be made in future implementations of the simulator. For example, rather than treat wide-balls and no-balls as aggregate characteristics, it may be possible to incorporate wide-balls and no-balls as individual characteristics of bowlers. It is also possible to include a home-field advantage term in the Bayesian latent variable model. Home field advantage has been estimated to be worth roughly 16 runs for the home team (de Silva, Pond & Swartz, 2001).

There are other modelling issues that we have not considered yet may have an impact on scoring. For example, in many sports there tends to be an aging effect where younger players improve, reach a plateau and then experience a decline in performance. Modelling the aging effect is a challenge (Berry, Reese & Larkey, 1999) and appears to be sport specific. We hope that by discarding old data and regularly updating our database, we might mitigate the aging effect by retaining data reflective of current performance. We might also consider the possibility that batsmen are more vulnerable to dismissal when they first come to the crease. Another modelling challenge involves the recognition that some batsmen struggle with certain types of bowlers (e.g., fast bowlers, spin bowlers).

Immediately prior to the World Cup of Cricket held in the West Indies in March 2007, a significant rule change was made with respect to fielding restrictions. This rule is known as the Powerplay rule which differs slightly from the temporary Powerplay rule considered during the 10-month period beginning August 2005. The new Powerplay rule has the potential of affecting the periods of constant aggressiveness as assumed in Table 2. In a few years time, when more data have been collected subject to the new rule, we plan on altering the definition of the situations in Table 2 and refitting our model. Of course, our simulations are only as good as the data which have been collected. Therefore it is advisable to regularly update our database and eliminate data that is deemed to have occurred too far in the past.

ACKNOWLEDGEMENTS

The authors thank the Editor, the Associate Editor and five referees for helpful comments that lead to an improvement in the article. Swartz and Gill were partially supported by Discovery grants through the Natural Sciences and Engineering Research Council of Canada.

BIBLIOGRAPHY

A. Agresti (2002). *Categorical Data Analysis*, 2nd edition, Wiley, New York.

M. J. Bailey & S. R. Clarke (2004). Market inefficiencies in player head to head betting on the 2003 cricket world cup. In *Economics, Management and Optimization in Sport*, S. Butenko, J. Gil-Lafuente & P. M. Pardalos, editors, Springer-Verlag, Heidelberg, pp. 185–202.

M. J. Bailey & S. R. Clarke (2006). Predicting the match outcome in one day international cricket matches, while the match is in progress. *Journal of Science and Sports Medicine*, 5, 480–487.

S. M. Berry, C. S. Reese & P. D. Larkey (1999). Bridging different eras in sports. *Journal of the American Statistical Association*, 94, 661–676.

B. M. de Silva & T. B. Swartz (1997). Winning the coin toss and the home team advantage in one-day international cricket matches. *The New Zealand Statistician*, 32, 16–22.

B. M. de Silva, G. R. Pond & T. B. Swartz (2001). Estimation of the magnitude of victory in one-day cricket. *The Australian and New Zealand Journal of Statistics*, 43, 259–268.

F. C. Duckworth & A. J. Lewis (1998). A fair method for resetting targets in one-day cricket matches. *Journal of the Operational Research Society*, 49, 220–227.

F. C. Duckworth & A. J. Lewis (2004). A successful operational research intervention in one-day cricket. *Journal of the Operational Research Society*, 55, 749–759.

D. Dyte (1998). Constructing a plausible test cricket simulation using available real world data. In *Mathematics and Computers in Sport*, N. de Mestre & K. Kumar, editors, Bond University, Queensland, Australia, pp. 153–159.

W. E. Elderton (1945). Cricket scores and some skew correlation distributions. *Journal of the Royal Statistical Society, Series A*, 108, 1–11.

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

D. Spiegelhalter, A. Thomas, N. Best & D. Lunn (2004). *WinBUGS User Manual Version 1.4.1*, Medical Research Council Biostatistics Unit, Cambridge.

T. B. Swartz, P. S. Gill, D. Beaudoin & B. M. de Silva (2006). Optimal batting orders in one-day cricket. *Computers & Operations Research*, 33, 1939–1950.

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

Received 14 January 2008
Accepted 2 March 2009