Journal of the
Royal Statistical Society
SERIES A
Statistics
in Society

J. R. Statist. Soc. A (2011)
174, *Part* 2, *pp.* 471–497

On the distribution of runs scored and batting strategy in test cricket

Philip Scarf,

University of Salford, UK

Xin Shi

Manchester Metropolitan University Business School, UK

and Sohail Akhtar

University of Salford, UK

[Received February 2010. Revised September 2010]

Summary. Negative binomial distributions are fitted to partnership scores and innings scores in test cricket. For partnership scores, we use a parametric model that allows us to consider run rate as a covariate in the distribution of runs scored and hence to use run rate as a surrogate for batting strategy. Then we describe the implied influence of run rate on match outcome probabilities given the state of the match at some point during the third innings; we refer to such a point in the match as the current position. Match outcome probabilities are calculated by using a model for the outcome given the end of the third-innings position, and a model for transitions from the current position to the end of the third-innings position, with transition probabilities considered as a function of run rate. Although the run rate is not wholly in the control of the batting side, our approach at least allows a captain or team analyst to consider the match outcome probability if the team can bat towards a target at a particular run rate. This will then at least indicate whether an aggressive or defensive batting strategy is desirable.

Keywords: Cricket; Logistic regression; Negative binomial distribution; Strategy

1. Introduction

In this paper, principally, we do two things. Firstly, we model the distribution of runs scored, with runs scored considered for innings and for partnerships. Secondly, we consider quantitative decision support for batting strategy in the third innings. For innings, the distribution of runs scored is considered in an exploratory manner and for general interest. For partnerships, we consider the distribution of runs scored in more detail, to model batting strategy in the third innings.

Interest in estimating the statistical distribution of runs scored goes back to Elderton and Elderton (1909), although it was not until Elderton (1945) and Wood (1945), in back-to-back papers (in this journal), that the geometric distribution was proposed and shown to be a reasonable fit. Later, the negative binomial distribution was considered, with varying success (Reep *et al.*, 1971; Pollard *et al.*, 1977). Pollard *et al.* (1977) modelled partnership data for

Address for correspondence: Philip Scarf, Centre for Operations Management, Management Science and Statistics, Salford Business School, University of Salford, Salford, M5 4WT, UK.
E-mail: p.a.scarf@salford.ac.uk

 0964–1998/11/174471

the first time, using a negative binomial distribution, and found a good fit, although Clarke (1988a) reported that for partnerships in 82 Ashes tests the fit was less than good. Kimber and Hansford (1993) provided evidence against the geometric assumption, based on an analysis of the empirical hazard of dismissal as an innings develops. They also argued for proper consideration of not-out scores in estimation of the distribution of runs scored and the calculation of a batting average, the latter being the main focus of their work. In this study, a parametric model for the runs scored in a partnership is desirable, and we pursue this approach.

Little work has been done on predicting match outcomes in test cricket: Brooks *et al.* (2002) used an ordered probit model with batting and bowling strengths, claiming to predict correctly 71% of outcomes. Allsopp and Clarke (2004) used a similar set of covariates but with the addition of first-innings lead—i.e. the lead given that each team has batted once. Thus, the match state is used to explain match outcome probabilities, and our approach in this paper is in principle the same. Baker and Scarf (2006) considered serial effects in Ashes test matches. More has been done in the analysis of 1-day international matches. Preston and Thomas (2002) in particular looked at win probability as a function of match position. Their object was the calculation of revised targets, in rain-interrupted matches, that preserve the win probability across an interruption, and they offered their method as a competitor to the well-known Duckworth–Lewis method (Duckworth and Lewis, 1998). Earlier they proposed run rate as a control variable in batting strategy in Preston and Thomas (2000). We use this idea to consider batting strategy in the third innings, presuming that choosing a batting strategy is equivalent to choosing the run rate at which to bat. Of course, the run rate is not completely in the control of the batting team—far from it in fact—and we return to this point later.

Modelling strategy in cricket is more eclectic. Clarke (1988b) investigated optimum batting rates in 1-day cricket and recommended quicker scoring earlier in an innings. Such tactics have been adopted in 1-day international matches. Preston and Thomas (2000) refined this idea to distinguish between the first and second innings. Clarke and Norman (1999) looked at tactics for protecting weaker batsmen, and at optimal deployment of the night-watchman (Clarke and Norman, 2003). Swartz *et al.* (2006) considered optimal batting order in the 1-day game. Using a ball-by-ball simulation model that allows the distribution of outcomes of an individual ball to vary by batsman, by cumulative balls bowled and by wickets lost, they

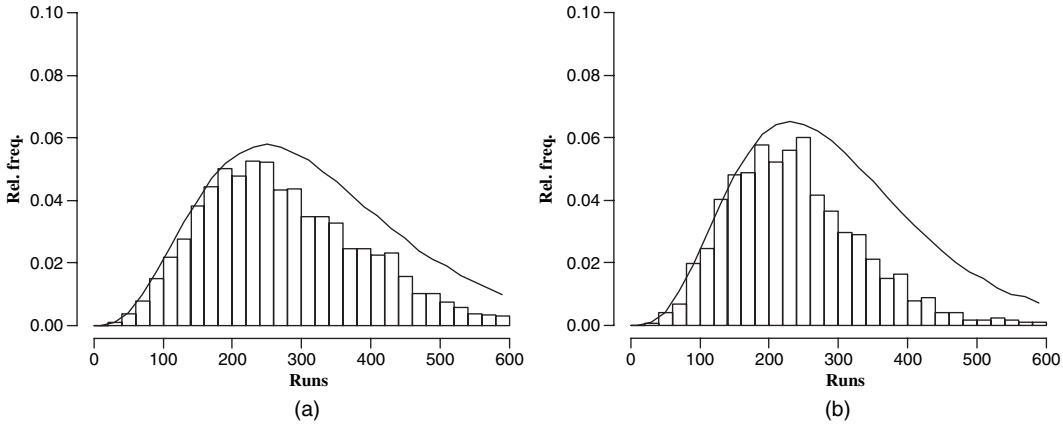

Fig. 1. Team innings scores in test cricket with the fitted negative binomial distribution (———) (1856 matches from 1877 to 2007; not-out innings have been excluded from the histograms, but included in fitting negative binomial distributions): (a) all innings, 6704 observations; (b) third innings, 1786 observations

determined an optimal batting order for the India 1-day international team. Some interesting batting orders that have not been used and that contradict the received wisdom are suggested. A purer problem, in the test match format, is to ask, given the state of a test match, at what rate should the batting team try to score? We attempt to answer this question in this paper, and we make a modest start on this problem by considering 'optimum' batting strategy during the third innings. In essence, in this particular third-innings problem, batting cautiously to ensure a large target is set for one's opponents, who bat last, is traded off against batting aggressively to ensure that sufficient time remains in the match to dismiss one's opponents in their final innings.

We use several different data sets. Quantities relating to partnerships (runs scored and balls bowled) are determined from a 'ball-by-ball' data set, the source of which is the very large archive found on the ESPNcricinfo Web site (ESPNcricinfo, 2010). The data set that we collected consists of the 341086 balls that were bowled in all 197 test matches played over the period from February 1998 to June 2004. This data set has information for each ball relating to runs scored, extras scored, extras description, wickets (0,1), innings number (in the match), over number (in the innings), ball number (in the over), batting team, bowling team, name of batsmen on strike, non-striker and bowler. When we consider match outcome modelling and are only interested in quantities that relate to overall team innings outcomes, we use an extended data set. In fact, we consider matches over the period from February 1998 to December 2007 and only those in which a final innings target was set (301 matches). We would have liked to collect ball-by-ball data over the same period so that these two data sets were effectively the same; however, we did not have the resources to do so, as the collection of the ball-by-ball data was extremely time consuming. Finally, in our exploratory analysis of innings scores, we consider all test matches played up to the end of 2007, beginning with the first ever test match (Australia *versus* England, 1877, at the Melbourne Cricket Ground).

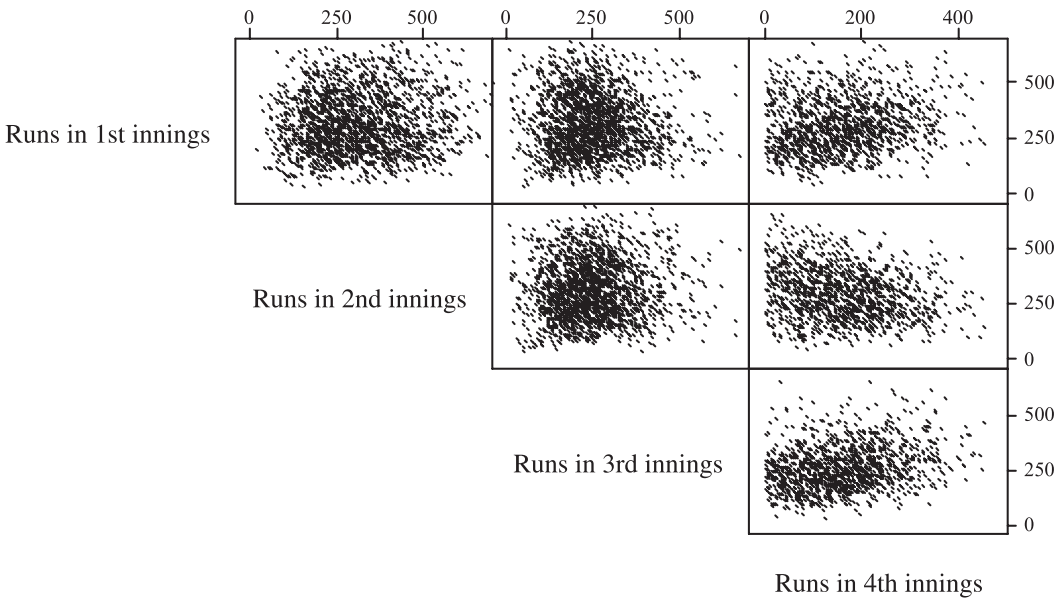

Fig. 2. Matrix scatter plot of innings scores in test cricket, excluding not-out innings (1856 matches from 1877 to 2007)

2. Distribution of runs scored

2.1. Team innings scores

The game of cricket is notorious for using the same word to mean many different things. The word 'wicket' is a case in point. This can mean

(a) the construction that is three sticks or stumps with two wooden bails on top,
(b) the strip of grass between the wickets,
(c) the area of the playing field where a wicket is cut,
(d) a batsman's turn at batting or
(e) the period during which two batsmen bat (a partnership).

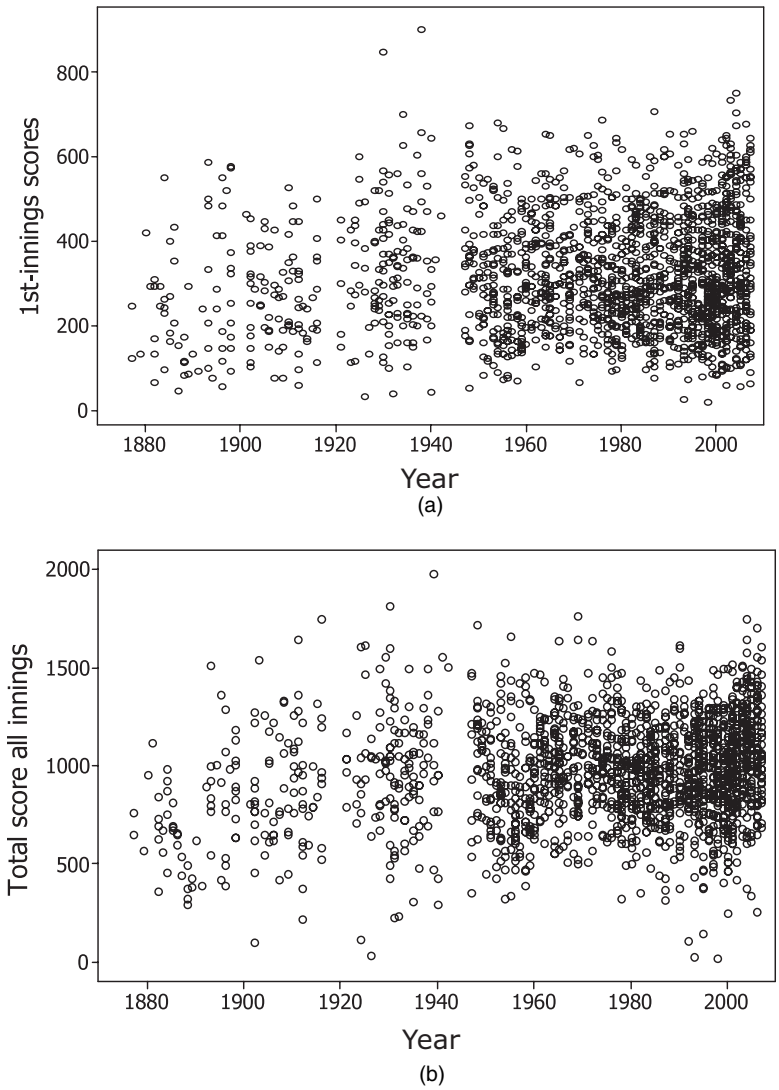

Fig. 3. (a) First-innings team scores and (b) match total scores against year (1856 matches from 1877 to 2007): not-out innings have been included

Therefore, we must be careful with our terminology. 'Innings' can mean the batting turn of a player or the entire team, and so we shall qualify the word when the meaning is ambiguous. Note that we shall use the terms 'runs' and 'scores' interchangeably.

We first look at team innings scores in an exploratory manner. There appears to be large variability in team innings scores (Fig. 1), little dependence between scores in the same match (Fig. 2) and little change in the size of scores over time (Fig. 3). Certainly, there has been an increase in the number of matches played per year over time. The negative binomial distributions in Fig. 1 model the distribution of runs scored in completed innings, and these distributions have been fitted by the method of maximum likelihood with not-out innings regarded as right censored. The histograms show completed innings only, with not-out innings excluded. The apparent 'lack of fit' in the right of the distribution can be explained by the exclusion of the not-out innings. In the histograms we have plotted the frequency relative to the total number of innings including the not-out innings. Thus, the histograms are 'missing' the not-out innings; this missing part broadly corresponds to the area to the right between the curve of the fitted negative binomial distribution and the relative frequencies, noting that higher innings scores are more likely to be for not-out innings. Approximately 14% of first innings and second innings scores are for 'not-out' innings, and this figure rises to 36% for third innings and 67% for final innings. The parameter estimates are shown in Table 1.

Table 1. Parameter estimates with standard errors for the negative binomial distribution (equation (1), with $p_0 = \theta^{\pi}$) for completed team innings scores†

Innings	π	θ
1st	4.45 (0.15)	0.0129 (0.0005)
2nd	4.55 (0.15)	0.0136 (0.0005)
3rd	4.71 (0.18)	0.0157 (0.0007)
4th	5.03 (0.31)	0.018 (0.0013)

†Not-out innings are regarded as right censored (1856 matches from 1877 to 2007).

Table 2. Descriptive statistics for the runs scored during a partnership grouped by partnership number, all innings in test matches between February 1998 and June 2004†

Partnership	*n*	*Mean*	*Standard deviation*	*Maximum*	*Median*
1	634	36.6	46.5	338	21
2	615	36.3	41.2	296	23
3	598	41.4	49.0	315	23
4	578	43.6	46.6	353	29
5	560	34.6	42.8	376	21
6	542	33.1	38.8	322	21
7	523	24.4	28.1	225	15
8	502	21.5	29.1	253	12
9	481	15.5	18.5	145	9
10	449	13.7	17.3	145	8
Overall	5482	31.0	39.6	376	17

†*n* is the number of observations.

2.2. Partnership scores

Partnership scores are calculated from the ball-by-ball data. For the 197 matches, there are 5482 partnerships. Table 2 presents descriptive statistics for all innings. Partnerships in which a batsman retired hurt present certain difficulties for partnership analysis, because where a batsman who retired hurt resumes then an 11th partnership can occur. The partnership in which a batsman retired hurt is considered as a not-out partnership (occurring in 17 out of the total of 5482). Cases where the batsman resumed have been ignored.

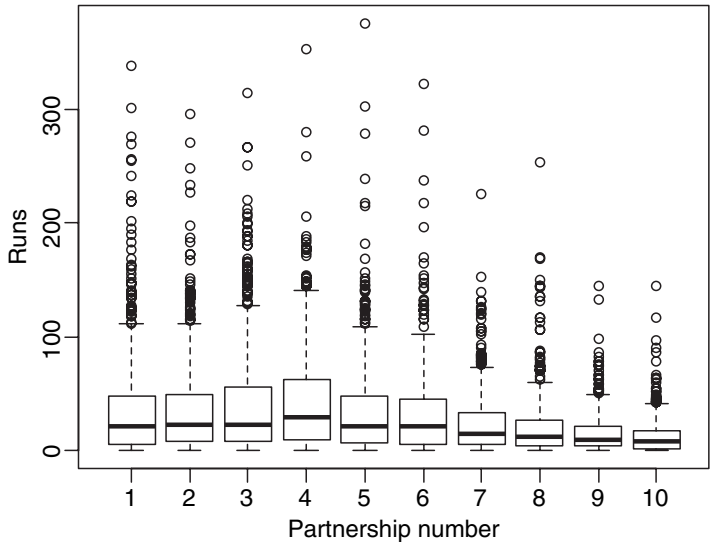

Fig. 4. Box-and-whisker plot of runs scored by partnership number: all test matches between February 1998 and June 2004

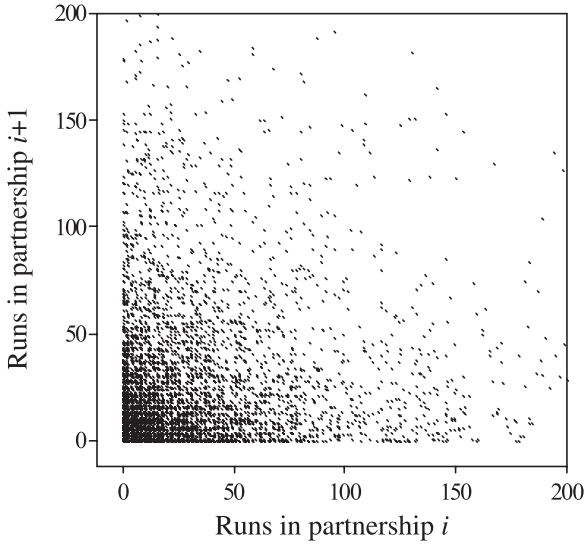

Fig. 5. Runs scored in a partnership *versus* runs scored in the previous partnership (in the same innings): all innings in all test matches between February 1998 and June 2004; correlation coefficient $\rho = 0.07$ ($p < 0.001$)

A box-and-whisker plot of partnership scores is shown in Fig. 4. The (Pearson) correlation between successive partnership scores is small (Fig. 5). Histograms of partnership scores are shown in Fig. 6(a) (all innings) and Fig. 6(b) (third innings). Fitted geometric, negative binomial and zero-inflated negative binomial distributions are drawn in Figs 6(a) and 6(b). The parameterization that we use for the zero-inflated negative binomial distribution is

$$\text{prob}(Z=z)=\begin{cases} p_0 & z=0, \\ \dfrac{(1-p_0)\,\Gamma(z+\pi)\theta^{\pi}(1-\theta)^{z}}{z!\,\Gamma(\pi)(1-\theta^{\pi})} & z=1,2,3,\ldots, \end{cases} \tag{1}$$

where the parameters of this model, p_0, θ and π, must satisfy $0<\theta$, $p_0<1$ and $0<\pi$. The implied mean and variance of this distribution are

$$E(Z)=\mu_z=\frac{(1-p_0)\pi(1-\theta)}{\theta(1-\theta^{\pi})},$$

$$\text{var}(Z)=\mu_z\left(\frac{\pi+1}{\theta}-\pi-\mu_z\right).$$

A standard negative binomial distribution is obtained by setting $p_0=\theta^{\pi}$, and this is the parameterization that we use for this distribution. When we consider batting strategy later, Z will

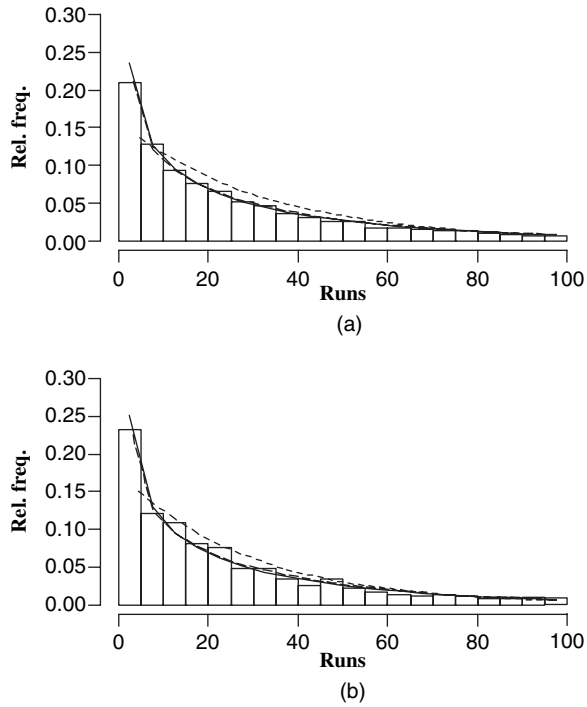

Fig. 6. Observed distribution of partnership scores (with not-out partnerships excluded), and with fitted geometric (-------), negative binomial (———) and zero-inflated negative binomial (– – –) distributions (with not-out partnerships regarded as right censored) (all matches from February 1998 to June 2004): (a) all innings partnerships, 5482 observations, Akaike information criterion values for the geometric, negative binomial and zero-inflated negative binomial distributions −23765.4, −23490.3 and −23468.3 respectively: (b) third-innings partnerships, 1412 observations, Akaike information criterion values for the geometric, negative binomial and zero-inflated negative binomial distributions −11975.3, −11916.6 and −11816.3 respectively

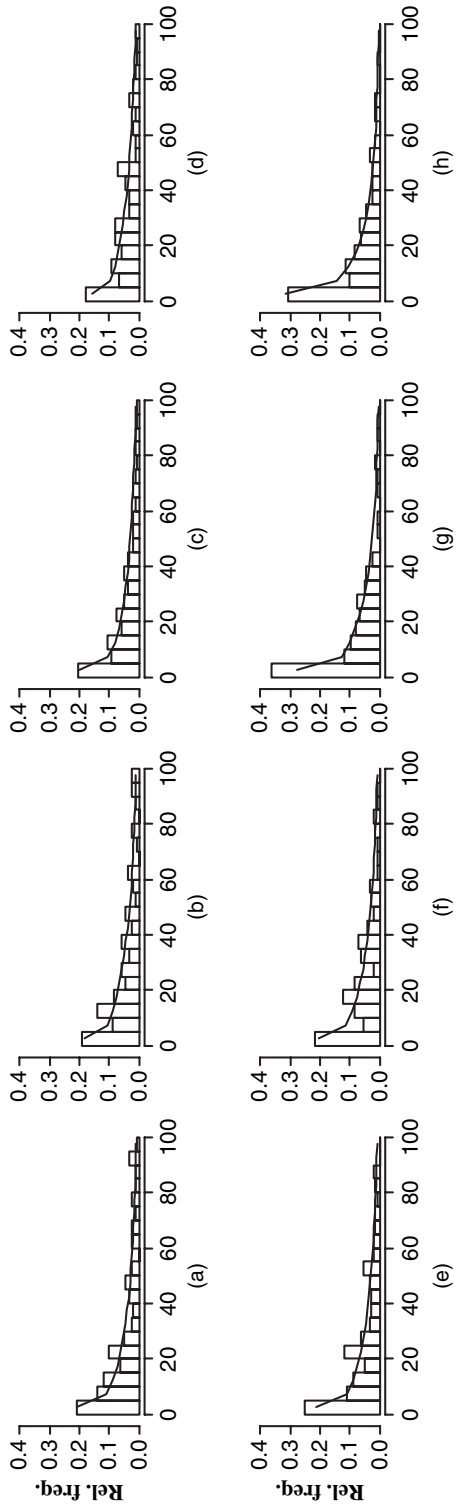

Fig. 7. Observed and fitted zero-inflated negative binomial distribution (——) of runs scored in a partnership for third innings by partnership number, with not-out partnerships regarded as right censored (all matches, from February 1998 to June 2004): (a) partnership 1; (b) partnership 2; (c) partnership 3; (d) partnership 4; (e) partnership 5; (f) partnership 6; (g) partnership 7; (h) partnership 8; (i) partnership 9; (j) partnership 10

Table 3. Maximum likelihood estimates for the zero-inflated negative binomial distributions fitted to runs scored in a partnership for all innings, by partnership number, with standard errors (all matches from February 1998 to June 2004)

Partnership	π	θ	p_0
1	0.725 (0.055)	0.019 (0.002)	0.063 (0.010)
2	0.919 (0.065)	0.024 (0.002)	0.068 (0.010)
3	0.723 (0.056)	0.017 (0.001)	0.067 (0.010)
4	0.857 (0.062)	0.019 (0.001)	0.057 (0.010)
5	0.809 (0.064)	0.022 (0.002)	0.082 (0.012)
6	0.815 (0.066)	0.023 (0.002)	0.076 (0.011)
7	0.891 (0.077)	0.033 (0.003)	0.094 (0.013)
8	0.639 (0.068)	0.029 (0.003)	0.105 (0.014)
9	0.905 (0.095)	0.05 (0.005)	0.126 (0.016)
10	0.797 (0.096)	0.051 (0.006)	0.159 (0.018)
All partnerships	0.723 (0.020)	0.022 (0.001)	0.086 (0.004)

denote runs scored in a partnership, and this parameterization allows us to model π (and θ) in terms of a covariate. Further, setting $\pi = 1$ obtains the geometric distribution.

The zero-inflated negative binomial distribution is the best model for partnership scores from among these distributions (Fig. 6). Fig. 7 shows the observed and fitted zero-inflated negative binomial distribution by partnership number. Fitted parameter values are given in Table 3. Note that when considering all innings 8.5% (467/5482) of partnerships have 0 scores. This increases to 9.8% (139/1412) for third-innings partnerships.

3. Modelling match outcome probabilities

3.1. End of third-innings position

Scarf and Shi (2005) developed a model to explain match outcome probabilities given the position at the end of the third innings. The purpose of the model is to provide decision support for setting a target at declaration (of the third innings). We briefly review their findings here. They used nominal logistic regression to model the multinomial response (win, draw or loss) as a function of match and end of third-innings covariates. An extract of the data that were used to fit this model is shown in Table 4. Fig. 8 shows a matrix scatter plot of test match outcomes with declarations and non-declarations indicated. Note that, among these matches, only two have been lost by a declaring team. In the fourth test between England and Australia at Headingley in 2001, Australia declared their second innings and set England a final innings target of 315; Australia were 3–0 ahead in the series at the time, and England won this match. In the third test between Australia and South Africa in Sydney 2006, South Africa set Australia a final innings target of 287 to win in 68 overs; Australia were 1–0 ahead in the three-match series at the time—we shall return to this match in our final example in the paper. Note that the highest final innings target that has ever been reached to win by a side batting last is 418 by the West Indies against Australia in 2003.

As there is no points system in test cricket, match outcome categories do not form a natural order. Also, for example, for the team batting third (hereafter, team A) the difference between winning and drawing is likely to be more dependent on the number of overs remaining than on the target faced; the difference between losing and drawing, in contrast, is likely to depend on

Table 4. Test match data (extract of 301 test matches, from February 1998 to December 2007)†

Date	*Home team*	*Away team*	*Venue*	*1st-innings score*	*2nd-innings score*	*Team batting 3rd (team A)*	*Follow-on enforced*	*Target set*	*Decla-ration*	*Overs remaining*	*Overs used*	*Result*
January 7th, 1998	Sri Lanka	Zimbabwe	Kandy	469	140	Zimbabwe	Yes	10	No	77	2	−1
January 14th, 1998	Sri Lanka	Zimbabwe	Colombo	251	225	Zimbabwe	No	326	No	154	113	−1
January 30th, 1998	Australia	South Africa	Adelaide	517	350	South Africa	No	361	Yes	109	108	0
February 5th, 1998	West Indies	England	Port of Spain, Trinidad	214	191	England	No	282	No	155	99	−1
February 13th, 1998	West Indies	England	Port of Spain, Trinidad	159	145	West Indies	No	225	No	207	108	−1
February 17th, 1998	South Africa	Pakistan	Bridgetown, Barbados	259	231	Pakistan	No	255	No	170	88	1
February 19th, 1998	New Zealand	Zimbabwe	Wellington	180	411	Zimbabwe	No	20	No	110	4	−1
February 27th, 1998	West Indies	England	Georgetown, Guyana	352	170	West Indies	No	380	No	152	62	1
March 6th, 1998	South Africa	Pakistan	Port Elizabeth	293	106	South Africa	No	394	Yes	142	60	1
March 12th, 1998	West Indies	England	Bridgetown, Barbados	403	262	England	No	375	Yes	109	37	0
March 14th, 1998	Zimbabwe	Pakistan	Bulawayo	321	256	Zimbabwe	No	368	Yes	105	98	0
March 21st, 1998	Zimbabwe	Pakistan	Harare	277	354	Zimbabwe	No	192	No	104	54	−1

†Variables included: match date, teams, venue, runs scored in the first two innings, team batting third and setting the target (team A), target established by team A, third-innings declaration indicator, follow-on enforced and estimated overs remaining at the start of the fourth innings (calculated on the basis of 90 overs per day). Match results are recorded as 1 (win), 0 (draw) or −1 (loss) from the point of view of team A. Other variables not shown: overs bowled in the first, second and third innings, the state of the series, weather interruptions in the first three innings and whether the toss was won by the target setting side.

both the number of overs remaining and the target faced. In this way, the target and number of overs remaining influence the match outcome categories in a non-cumulative way. Therefore, it makes sense to regard match outcome categories as nominal. Focus on a multinomial response is justified on the basis that there is little interest in the 'score' at the end of the match, and teams are not concerned with the size of a win or loss. The model is then

$$\left.\begin{aligned} Y &\sim \mathrm{MN}(p_1, p_0, p_{-1}; \textstyle\sum p_i = 1), \\ p_1 &= \exp(\alpha_1 + \beta_1^{\mathrm{T}} X)/\{1 + \exp(\alpha_1 + \beta_1^{\mathrm{T}} X) + \exp(\alpha_{-1} + \beta_{-1}^{\mathrm{T}} X)\}, \\ p_0 &= 1/\{1 + \exp(\alpha_1 + \beta_1^{\mathrm{T}} X) + \exp(\alpha_{-1} + \beta_{-1}^{\mathrm{T}} X)\}, \\ p_{-1} &= \exp(\alpha_{-1} + \beta_{-1}^{\mathrm{T}} X)/\{1 + \exp(\alpha_1 + \beta_1^{\mathrm{T}} X) + \exp(\alpha_{-1} + \beta_{-1}^{\mathrm{T}} X)\}, \end{aligned}\right\} \tag{2}$$

where the match outcome (win, draw or loss) is denoted by Y and takes values $(1, 0, -1)$, X is a vector of covariates, and p_1, p_0 and p_{-1} are the probability of a win, a draw and a loss respectively.

The factors which impact on outcomes in cricket are extensive, e.g. home advantage, state of the series, the teams' strengths, the umpires and pitch conditions, and some of these effects have been estimated (Allsopp and Clarke, 2004; Brooks *et al.*, 2002; Ringrose, 2006). We are concerned principally with match state covariates, and outline model statistics for various fitted models are shown in Table 5. These are based on a larger data set than that considered by Scarf and Shi (2005). Table 6 presents the maximum likelihood estimates for the chosen 'best' model from among the set of models that was considered. Table 6 indicates that the loss–draw probability ratio depends strongly on both current lead and the number of overs remaining. However, the win–draw probability ratio depends strongly on the number of overs remaining only. An ordinal logistic regression model could not capture this non-cumulative dependence on the covariates.

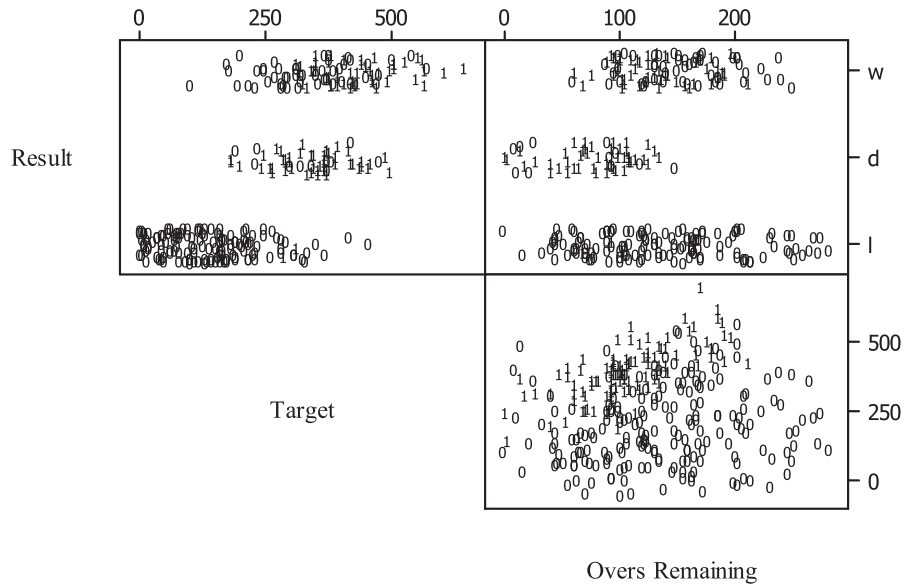

Fig. 8. Matrix scatter plot for test matches between February 1998 and December 2007 in which a final innings target was set: outcome (win w, draw d or loss l) plotted against the target and overs remaining at the start of the fourth innings (the data points have been perturbed slightly to improve legibility): 1, declaration of the third innings; 0, otherwise

Table 5. Results of fitting the multinomial logistic regression model to 301 test match outcomes (from February 1998 to December 2007) for various sets of predictors†

Model	*Number of parameters*	*Likelihood*	*AIC*	R^2 *(%)*
$T + \mathrm{OR} + \mathrm{RR}_{12} + \mathrm{W\%D} + D$	12	−125.38	274.76	81.57
$T + \mathrm{OR} + \mathrm{RR}_2 + \mathrm{W\%D}$	10	−130.39	280.77	80.46
$T + \mathrm{OR} + \mathrm{RR}_1 + \mathrm{W\%D}$	10	−134.41	288.81	79.55
$T + \mathrm{OR} + \mathrm{RR}_{12} + \mathrm{W\%D}$	10	−131.31	282.62	80.26
$T + \mathrm{OR} + \mathrm{RR}_2$	8	−137.03	290.07	78.95
$T + \mathrm{OR} + \mathrm{RR}_1$	8	−141.42	298.84	77.90
$T + \mathrm{OR} + \mathrm{RR}_{12}$	8	−138.87	293.73	78.51
$T + \mathrm{OR} + \mathrm{W\%D}$	8	−137.37	290.74	78.87
$T + \mathrm{OR}$	6	−142.13	296.27	78.24
$T + \mathrm{OR}$ (ordinal)	4	−198.08	404.16	61.32
$T + \mathrm{OR} + T^2 + \mathrm{RR}_{12} + \mathrm{W\%D}$	12	−130.14	284.28	80.52
$T + \mathrm{OR} + \mathrm{OR}^2 + \mathrm{RR}_{12} + \mathrm{W\%D}$	12	−129.61	283.22	80.64
$T + \mathrm{OR} + T{*}\mathrm{OR} + \mathrm{RR}_{12} + \mathrm{W\%D}$	12	−129.62	283.24	80.64

†The Akaike information criterion AIC and Nagalkerke's (1991) R^2 are shown, with covariates target set (T), number of overs remaining (OR), the run rate in the first two innings (RR_{12}), the win percentage difference (W%D) and the declaration indicator (D).

Table 6. Fitted parameter estimates for the chosen model (fourth row in Table 5) with covariates target set (T), number of overs remaining (OR), the run rate in the first two innings (RR_{12}) and the win percentage difference (W%D), with standard errors and *p*-values†

	Parameter	*Coefficient*	*Standard error*	*p-value*
Win or draw (1–0)	Intercept	−4.8841	1.6633	0.003
	Number of overs remaining OR	0.0518	0.0087	0.000
	Target T	−0.0069	0.0033	0.037
	Run rate in first two innings RR_{12}	0.7260	0.4947	0.142
	Win percentage difference W%D	0.0167	0.0122	0.170
Loss or draw (−1–0)	Intercept	−2.0650	1.9776	0.296
	Number of overs remaining OR	0.0538	0.0093	0.000
	Target T	−0.0290	0.0038	0.000
	Run rate in first two innings RR_{12}	1.6975	0.6258	0.007
	Win percentage difference W%D	−0.0276	0.0156	0.077

†301 test matches (from February 1998 to December 2007).

Although, the run rate in the second innings, RR_2, appears to be a better predictor than the run rate in the first two innings, RR_{12}, interpretation of this former covariate is not straightforward. This is because, on occasions, RR_2 is the run rate of the team batting third but in their first innings—they may have followed on—this occurred in 12 of the 301 matches. We could consider a new but similar covariate: if team B are batting last and chasing the target, then we can define RR as the run rate of team A in their first innings. This may also include a time effect, however. Therefore, we instead use the run rate in the first two innings, RR_{12}, in the final model. The size of the two first-innings totals, S, might also be included, but arguably S, the run rate in the first two innings, RR_{12}, and the number of overs remaining, OR, will be collinear.

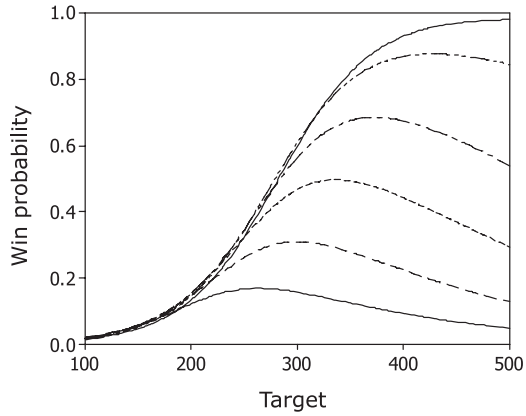

Fig. 9. Win probability for the team batting third as a function of target established and number of overs remaining ($RR_{12} = 3.12$; W%D = 0): ———, 60 overs remaining; – – –, 80 overs remaining; -------, 100 overs remaining; -—-—, 120 overs remaining; —··—, 150 overs remaining; ···—···, 200 overs remaining

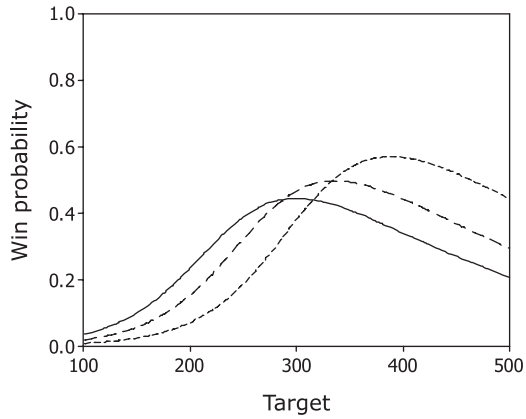

Fig. 10. Win probability for the team batting third against target set, with 100 overs remaining (W%D = 0): ———, $RR_{12} = 2.5$; – – –, $RR_{12} = 3.12$; -------, $RR_{12} = 4$

Team strengths are considered simply by calculating the difference in win percentage in the previous 20 matches between the reference team and their opponents. This we label W%D. A similar covariate based on the International Cricket Council ratings (International Cricket Council, 2010) might be used but these ratings were not available to us for matches before 2003. Rather than using the winning records of teams, team effects could be considered in various other ways: as a fixed effect for each team; as a random effect in a generalized linear mixed model; as a fixed effect for the home team (and perhaps a random effect for the away team). In the last of these, the decision support model would then be designed to consider the 'optimal' declaration for any team setting a target when the match is played in a particular country. Similarly, we might consider a fixed ground effect—batting last in Lahore might be a very different prospect from batting last at Edgbaston for any team. With data on more matches, fixed country effects and even ground effects might be estimated. A model with countries (of the reference team and opponent), or grounds or both as random effects may indeed be estimable with the data that are available to us. However, such models would be less useful for prediction. Therefore,

we compromise and consider only the strength difference covariate, win percentage difference (W%D), corresponding to an additional two parameters in the model.

The explanatory power of the declaration indicator variable is good and not surprising since it may incorporate many factors, possibly unmeasurable, which lead to a captain declaring or otherwise. However, it would not make sense to use it as a covariate in a model to support decision-making regarding declaration.

Fig. 9 shows the win probability from the fitted model as a function of target set and number of overs remaining. Note that the win probability increases to a peak and then decreases—if a very large target is set, the team batting last will not attempt to play for a win and a draw becomes more likely. Fig. 10 shows the effect of RR_{12} on the win probability. It appears that the effect of a high value of RR_{12} is to make a draw less likely and so a win more likely when the target is large, and a loss more likely when the target is small. It appears therefore that this covariate represents playing conditions in a manner that we would expect. The range of values of RR_{12} that was considered is not unreasonable (RR_{12} has mean 3.12 and standard deviation 0.47), although the size of the effect on the outcome of a match is larger than expected.

4. Decision support for batting strategy

4.1. Setting a target at declaration

A batting team is not required to complete its innings. The batting captain may declare, at any point in the innings of his team, the 'innings over' and ask the opposing team to bat. The purpose of a third-innings declaration is to provide sufficient time to dismiss the opponents in their final innings. Test cricket is time limited and, if a team is to win, all innings must be completed within 5 days. Consequently, in timing a declaration, a batting captain is essentially trading off the lead, and consequently the probability of losing, against the time remaining in the match, and consequently the probability of drawing. In the first two innings in a game, the teams aim to establish their position. The third innings can then be played more strategically. The second innings may be declared, but often a decision about a second-innings declaration is less finely balanced because it takes place earlier in the match. First-innings declarations are rare. So the question arises: is there an optimum time to declare a third innings closed? The match outcome probabilities that are associated with various end of third-innings positions that are shown in Table 7 could provide decision support. These probabilities are calculated by using the model that was described in the previous section (equation (2)), and show win, draw and loss probabilities conditional on the target established (or lead + 1) and the number of overs remaining at the end of the third innings. Of course, the batting team is not guaranteed to reach a certain end-of-innings position given the current position, and the probabilities are therefore to be interpreted as 'if one does indeed reach a particular position at the end of the third innings, then all else being equal these would be the win, draw and loss probabilities when in that position'. Our ultimate goal is to determine win, draw and loss probabilities given a particular position during the third innings. These probabilities are considered in the next section.

A captain would of course take account of other factors such as the state of the series, the state of the pitch and possibly the weather. Since test matches are always played as part of a series, typically comprising three or five matches between the same two teams, the attitude of the side batting third to risk will depend very much on the state of the series. It is an overall win in the series that is most important. Generally, declaring captains act conservatively.

Third-innings declarations occur in the order of 31% of test matches played (in terms of our current database), and so the timing of a declaration is important although not a universal problem in the game. More often during the third innings the batting team is in a less commanding

Table 7. Example of match outcome probabilities given the target and number of overs remaining†

Target established	*Probabilities for the following projected run rates:*															
	3				*4*				*5*				*6*			
	Overs remaining	*% win*	*% draw*	*% loss*	*Overs remaining*	*% win*	*% draw*	*% loss*	*Overs remaining*	*% win*	*% draw*	*% loss*	*Overs remaining*	*% win*	*% draw*	*% loss*
310	136	56	4	39	137	56	4	39	137	56	4	39	138	56	4	40
330	129	64	8	28	132	64	7	29	133	65	7	29	134	65	6	29
350	123	68	13	19	127	69	11	20	129	70	10	20	131	71	9	20
370	116	66	22	12	122	70	17	13	125	72	15	13	128	74	13	13
390	109	60	33	7	117	68	24	8	121	71	21	8	124	73	18	9
410	103	52	44	4	112	62	33	5	117	67	28	5	121	71	24	5
430	96	41	57	2	107	54	43	3	113	61	36	3	118	67	30	3
450	89	30	69	1	102	45	53	1	109	54	44	2	114	60	38	2
470	83	22	78	0	97	36	63	1	105	46	53	1	111	54	45	1
490	76	14	86	0	92	28	72	0	101	38	62	0	108	46	53	1

†The current lead is 299, the current number of overs remaining is 142, $RR_{12} = 3.5$ and $W\%D = 0$, assuming progression towards the target at differing (projected) run rates in the third innings from the current position, and assuming that two overs are lost for a change of innings. The probabilities are conditional probabilities of match outcome (win, draw or loss) given the target and number of overs remaining at the beginning of the fourth innings.

position and is merely aiming to set as large a target as possible or may be attempting to save the game having conceded a large lead on the first innings. Target setting in 1-day matches differs from that in test matches. This is because in 1-day matches there is no notion of playing out the time remaining for a draw.

Broadly speaking, the approach that is described here cannot present an 'optimal' solution, because the probability of winning will not be maximum when the probability of losing is minimum; the decision problem is a multiple-criteria problem. In English county cricket, in contrast, a points system is used and so it would be possible to consider an objective function, such as the expected number of points achieved in the match or more interestingly the probability of winning the championship. Using the latter objective, a team would then act differently with regard to declarations depending on whether the opposition was a close competitor for the championship or otherwise. Thus, if team A is considering a declaration, then we would expect a cautious target if both team A and team B are contenders for the championship title. If team A are contenders but team B are not, then a much less cautious target would be optimal.

4.2. Third-innings batting strategy

Consider now the problem of determining the optimum batting strategy during the third innings. The team batting third, the reference team, must decide whether to bat defensively or aggressively as it plays its innings. Suppose that the reference team aims to set a target for the team batting last. Call this target aimed for T. Further, suppose that the reference team aims to bat towards this target at run rate x. Thus we suppose that T and x are the decision variables in this formulation. The probability of a win for the reference team will depend on T, x and the current position. Broadly speaking, if the reference team is in a strong position, then the probability of a draw will increase as T increases and x decreases. Conversely, the draw probability will decrease, and both the win and the loss probabilities will increase, as T decreases and x increases. Thus batting aggressively (large x) is more risky.

To determine the match outcome (win, draw or loss) probabilities given the current position, we fix T and x and then condition on reaching a particular end of third-innings position. Using the match outcome model that was considered in Section 3.1, and relaxing the conditioning (by considering all possible end of third-innings positions given the target aimed for, T, and the run rate in the remainder of the third innings, x), we can find the probability of a win, draw or loss given the current position. Of course, factors other than just the current position and batting strategy in the remainder of the third innings will influence the outcome of the match. We do not attempt to quantify these factors. Thus the model that is developed can only guide captains as they make decisions about a declaration strategy. We would expect them to modify model outcomes in the light of their experience regarding local conditions.

So, proceeding with the detail about how to calculate outcome probabilities given the current position, denote the current position (from the point of view of the team batting third—the reference team) by $P = (r, V_r, w)$, where r is the current lead, V_r the current number of overs remaining and w the current third-innings wickets lost. Let t be the actual target that has been set—this will be at most the target aimed for, and less if the reference team are all out beforehand. Thus $t \leqslant T$.

Let Y denote the outcome of the match. Then, using $\text{prob}(Y = y \mid P, x, T)$ to denote the match outcome probabilities given the current position P and the choice of the decision variables, it follows that

$$\text{prob}(Y = y \mid P, x, T) = \sum_{t=r+1}^{T} \text{prob}(Y = y \mid t, V_t)\, \text{prob}(t \mid P, x, T), \tag{3}$$

where $\text{prob}(t|P,x,T)$ is the probability distribution of the target set given the current position P and the choice of the decision variables, and $\text{prob}(Y=y|t,V_t)$ is the probability of outcome Y (win, draw or loss) given the target set t and number of overs remaining at the end of the third innings, V_t. Note that V_t is determined by V_r, t and x: $V_t = V_r - \{(t-r-1)/x\} - 2$ (assuming that two overs are lost for a change of innings).

To proceed with the probability calculation in equation (3), we seek a suitable model for $\text{prob}(t|P,x,T)$. Let Z be the total number of further runs added by the reference team in their third innings from the current position if they complete each remaining partnership, so that $t = \min(Z+r+1, T)$, i.e. $Z+r$ would be the lead if the reference team batted until all 10 wickets had been lost. At the current position, there are w wickets down, and so

$$Z = Z'_{w+1} + \sum_{k=w+2}^{10} Z_k \qquad w \leqslant 8, \tag{4}$$

where Z'_{w+1} is the additional number of runs added in the current partnership and Z_k is the number of runs scored in the kth partnership, $k = w+2, \ldots, 10$. We next assume that $Z_k \sim \text{NB}\{\pi_k(x), \theta_k\}$, with parameters θ_k a function of partnership number, and π_k, and hence the mean number of runs scored, a function of the run rate x. Thus, given knowledge about the distribution of runs, Z, that the reference team could add from the current position given the chosen run rate x, we can determine the probability distribution of the actual target set t given the chosen target aimed for, T. The number of overs remaining in the match at the end of the third innings is a deterministic function of the number of overs remaining at the current position and t and x. Thus $\text{prob}(t|P,x,T)$ in equation (3) can be calculated; $\text{prob}(Y=y|t,V_t)$ can be calculated by using equation (2).

It now remains to model $Z_k \sim \text{NB}\{\pi_k(x), \theta_k\}$. We use the parameterization that is implied by equation (1) in Section 2.2 (with $p_0 = \theta^\pi$), and we consider various forms for $\pi_k(x)$ in Table 8. Since $E(Z) = \pi(1/\theta - 1)$ for this parameterization of the negative binomial distribution, it is simpler and more natural to allow π_k to vary with x rather than θ_k. Fig. 11, and the notion that there is a chosen run rate x at which the mean score is maximum, suggests a gamma function

Table 8. Log-likelihood LL, number of parameters n and Akaike information criterion value AIC for various models of the distribution of runs scored in partnership k, $Z_k \sim \text{NB}\{\pi_k(x), \theta_k\}$ $(k = 1, \ldots, 10)$†

Model	*LL*	*Number of parameters*	*AIC*
$\pi_k = \alpha_k$, $\theta_k = \theta$ constant	−6034.0	11	12090.1
$\pi_k = \alpha$ constant, θ_k	−6005.6	11	12033.3
$\pi_k = \alpha_k$, θ_k	−6002.8	20	12045.6
$\pi_k = \alpha_k$, θ_k, $p_{0,k}$ (zero-inflated negative binomial)	−5988.3	30	12036.6
$\pi_k = \alpha_k x^\beta \exp(-\gamma x)$, $\theta_k = \theta$ constant	−5462.7	13	10951.3
$\pi_k = \alpha_k x^\beta \exp(-\gamma x)$, $\theta_k = \theta$, $p_{0,k} = p_0$ constant (zero-inflated negative binomial)	−5803.5	14	11635.0
$\pi_k = \alpha_k x^\beta \exp(-\gamma x)$, θ_k	−5411.1	22	10866.1
$\pi_k = \alpha x^\beta \exp(-\gamma x)$, θ_k	−5418.0	13	10862.1
$\pi_k = \alpha x^{\beta_k} \exp(-\gamma x)$, θ_k	−5411.4	22	10866.9
$\pi_k = \alpha x^\beta \exp(-\gamma_k x)$, θ_k	−5409.5	22	10863.0
$\pi_k = \alpha_k x^{\beta_k} \exp(-\gamma x)$, θ_k	−5407.0	31	10876.0
$\pi_k = \alpha x^\beta \exp\{-(\gamma x + \psi x^2)\}$, θ_k	−5416.8	14	10861.7

†The data comprise all partnerships in the third innings ($n = 1412$) (all matches from February 1998 to June 2004).

for $\pi_k(x)$. Note that a run rate that is greater than 6 is rare (157 partnerships among 5482). It might appear that the zero-inflated negative binomial distribution with $\pi_k(x)$ a function of x is a candidate model here. However, when $x=0$ we require that $\text{prob}(Z=0)=1$, because in reality the run rate is 0 if and only if the partnership score is 0. This property does not hold for the model $Z_k \sim \text{ZINB}\{\pi_k(x), \theta_k, p_{0,k}\}$. The run rate is 0 (no runs scored) in approximately 8% of partnerships.

We use the model in the eighth row of Table 8 for the distribution of partnership scores as a function of run rate; parameter values are in Table 9. Although the number of not-out innings

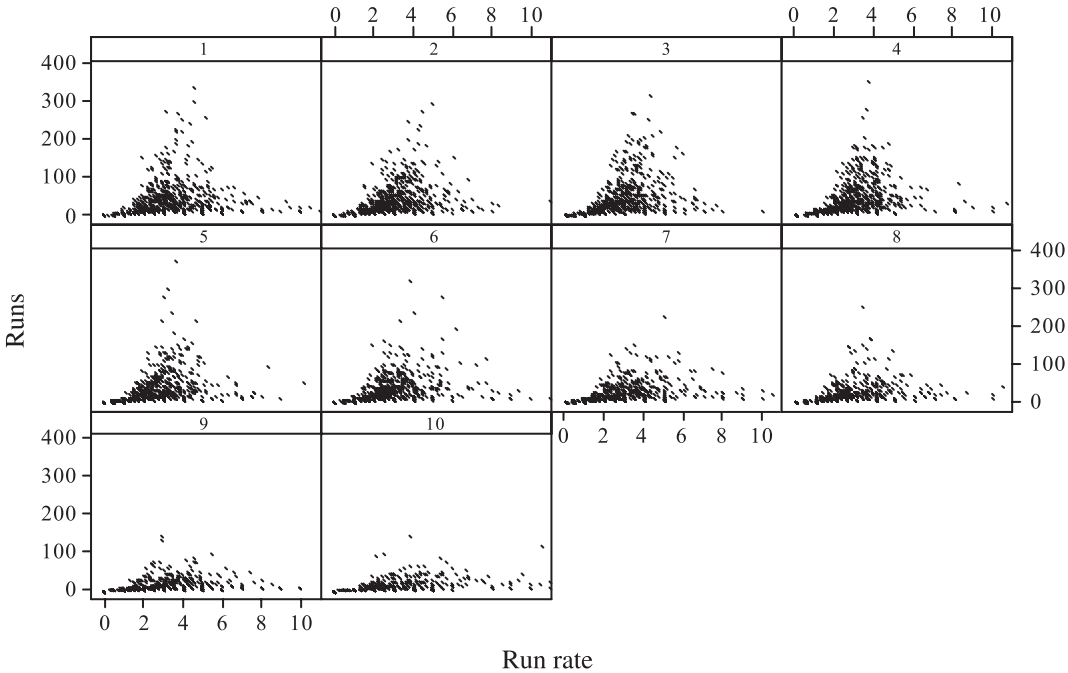

Fig. 11. Scatter plots of runs scored in a partnership against run rate by partnership number (all innings; all matches from February 1998 to June 2004)

Table 9. Maximum likelihood estimates (with standard errors) for the model in the eighth row of Table 8

Parameter	*Estimate (standard error)*	*Parameter*	*Estimate standard error*
θ_1	0.045 (0.003)	α	0.837 (0.045)
θ_2	0.037 (0.003)	β	1.440 (0.107)
θ_3	0.033 (0.003)	γ	0.295 (0.033)
θ_4	0.034 (0.003)		
θ_5	0.042 (0.003)		
θ_6	0.039 (0.003)		
θ_7	0.061 (0.005)		
θ_8	0.062 (0.005)		
θ_9	0.089 (0.008)		
θ_{10}	0.099 (0.009)		

is small, 148 (2.70%) and 35 (2.48%) partnerships for all innings and third innings respectively, we suppose that the not-out partnerships provide right-censored observations. The final model in Table 8 has minimum Akaike information criterion AIC among those shown; however, in the interest of simplicity and parsimony, we use the model in the eighth row. Fig. 12 shows the scores (third innings only), along with the fitted and observed means for the chosen 'best' model among those tried. The lack of fit when the run rate is small is not so important for the batting strategy problem since we are particularly interested in the score for moderate values of the run rate. The observed means are calculated by grouping the data.

We can now attempt to find the distribution of Z, the further number of runs added in the innings if the innings is completed. If the current position is at the fall of a wicket, then $Z'_{w+1} = Z_{w+1}$. If at the current position a partnership is some way through, there may be some justification in assuming a lack-of-memory property, so that the runs scored and the further runs scored follow the same distribution. Furthermore, the distribution of Z (equation (4)) will not be straightforward to calculate. For simplicity, we shall approximate the distribution of Z, and suppose that $Z \sim \mathrm{NB}(\pi_Z, \theta_Z)$, with π_Z and θ_Z obtained by equating moments. Thus, setting $E(Z) = \mu_Z = \Sigma_{k=w+1}^{10} \mu_k$ and $\mathrm{var}(Z) = \sigma_Z^2 = \Sigma_{k=w+1}^{10} \sigma_k^2$, and noting that

$$E(Z) = \mu_Z = \pi_Z(1-\theta_Z)/\theta_Z,$$

$$\mathrm{var}(Z) = \mu_Z\left(\frac{\pi_Z+1}{\theta_Z} - \pi_Z - \mu_Z\right),$$

we can determine (π_Z, θ_Z). Consequently, our match outcome probability calculations are approximations. We would expect that exact calculation of the distribution of Z will make

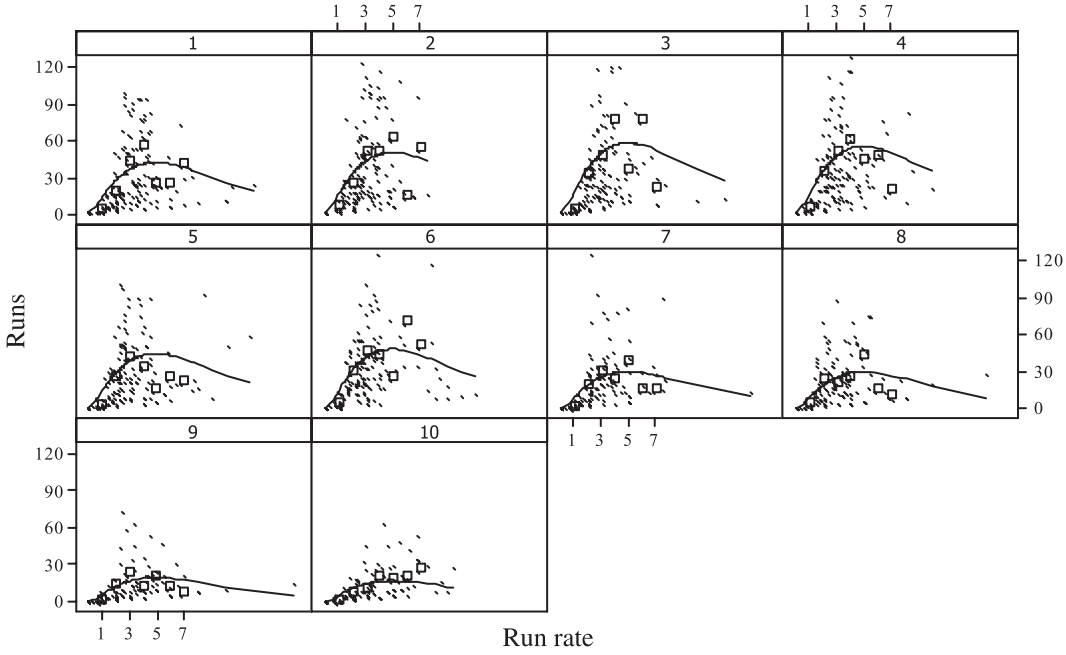

Fig. 12. Scatter plots of runs scored in a partnership against run rate by partnership number (third innings): fitted mean numbers of runs (———) and observed means (□) calculated in each interval (0.5,1.5], (1.5, 2.5], ... (all matches from February 1998 to June 2004)

Table 10. Match outcome probabilities given the current position as a function of target aimed for, T, and (projected) run rate x†

Target aimed for	*Probabilities for the following projected run rates:*																	
	2			*3*			*4*			*5*			*6*			*8*		
	% win	*% draw*	*% loss*	*% win*	*% draw*	*% loss*	*% win*	*% draw*	*% loss*	*% win*	*% draw*	*% loss*	*% win*	*% draw*	*% loss*	*% win*	*% draw*	*% loss*
260	39	1	60	39	0	60	39	0	60	39	0	61	39	0	61	39	0	61
280	50	2	48	50	1	49	50	1	49	50	1	49	50	1	50	50	1	50
300	60	4	36	61	2	37	61	1	38	61	1	38	61	1	38	61	1	39
320	66	8	26	69	4	27	70	2	28	70	2	28	70	2	28	70	1	29
340	67	16	17	75	6	19	77	4	19	77	3	20	78	2	20	78	2	20
360	62	26	12	77	11	12	81	6	13	82	4	13	83	4	14	83	3	14
380	54	38	8	75	17	8	82	9	9	85	6	9	86	5	9	86	4	10
400	45	48	7	70	24	5	81	14	6	85	9	6	87	7	6	88	5	7
420	39	55	6	63	33	4	77	19	4	84	12	4	87	9	4	88	7	5
440	35	59	6	55	42	3	72	25	3	81	16	3	85	12	3	87	9	4
460	33	61	6	48	50	2	66	32	2	77	21	2	83	15	2	86	11	3
480				42	55	2	60	38	2	73	26	2	80	18	2	84	13	3
500				39	59	2	54	44	1	68	31	1	77	21	2	82	15	3

†England *versus* West Indies, 2007, first test of four. England first innings 553 runs in 142 overs; West Indies first innings 437 runs in 117 overs ($RR_{12} = 3.82$; $W\%D = 15$); current position (at tea on the fourth day), England second innings 105 runs for 2 wickets. Reference team, England; lead, 221; number of overs remaining, 127; third-innings wickets down, 2. England declared having added 179 runs in 35 overs, setting the West Indies a target of 401 runs in 92 overs; owing to rain, only 20 overs were bowled on the final day and the match was drawn.

Table 11. Match outcome probabilities given the current position as a function of target aimed for, T, and (projected) run rate x†

Target aimed for	*Probabilities for the following projected run rates:*																	
	2			*3*			*4*			*5*			*6*			*8*		
	% win	*% draw*	*% loss*	*% win*	*% draw*	*% loss*	*% win*	*% draw*	*% loss*	*% win*	*% draw*	*% loss*	*% win*	*% draw*	*% loss*	*% win*	*% draw*	*% loss*
200	7	7	86	7	6	87	7	6	87	7	6	87	7	6	87	7	6	87
220	9	18	72	10	14	76	10	12	78	10	12	79	10	11	79	10	10	79
240	10	39	51	12	28	60	13	24	63	13	21	66	13	20	67	13	19	68
260	9	63	28	13	48	39	14	40	46	15	35	49	16	32	52	16	30	53
280	6	81	13	11	67	22	14	58	28	16	52	32	17	47	35	18	44	37
300	4	89	7	8	81	11	12	73	15	15	66	19	16	62	22	18	58	24
320	3	92	5	6	89	5	9	83	8	12	78	10	14	73	13	16	69	15
340	3	93	5	4	93	3	7	89	4	10	85	6	12	81	7	14	77	9
360	2	93	5	3	94	2	5	92	3	8	89	3	10	86	5	12	82	6
380	2	93	5	3	95	2	4	94	2	6	92	2	8	88	3	10	84	5
400	2	93	5	2	96	2	3	95	2	5	93	2	7	90	3	9	86	5
420	2	93	5	2	96	2	3	95	2	4	94	2	6	91	3	9	87	4

†Australia *versus* South Africa, 2006, third test of three, Australia leading the series 1–0; South Africa first innings 451 runs in 155 overs; Australia first innings 359 runs in 95 overs ($RR_{12} = 3.24$; $W\%D = -18$); current position (the start of the final day), South Africa second innings 94 runs for 3 wickets. Reference team, South Africa; lead, 186 runs; number of overs remaining, 90; third-innings wickets down, 3. South Africa added 100 runs in 20 overs, setting Australia a target of 287 runs in 68 overs, Australia reached 288 for 2 wickets in 61 overs to win by 8 wickets.

Table 12. Match outcome probabilities given the current position as a function of target aimed for, T, and (projected) run rate x†

Target aimed for	*Probabilities for the following projected run rates:*																	
	2			*3*			*4*			*5*			*6*			*8*		
	% win	*% draw*	*% loss*	*% win*	*% draw*	*% loss*	*% win*	*% draw*	*% loss*	*% win*	*% draw*	*% loss*	*% win*	*% draw*	*% loss*	*% win*	*% draw*	*% loss*
180	11	2	87	11	2	87	11	2	87	11	2	87	11	2	87	11	2	87
200	16	6	78	16	5	79	16	4	80	16	4	80	16	4	80	16	4	80
220	21	15	65	21	11	68	22	9	69	22	8	70	22	8	71	22	7	71
240	23	32	45	26	21	53	27	17	56	28	15	57	28	13	58	29	12	59
260	20	54	25	28	36	36	31	28	40	33	24	43	34	22	44	35	20	46
280	15	74	12	26	54	21	31	43	26	35	36	29	37	32	31	38	29	32
300	9	86	5	20	69	10	28	57	15	33	49	18	37	44	20	39	40	21
320	5	92	2	15	81	5	23	69	8	29	61	10	34	55	12	37	50	13
340	4	95	1	10	88	2	18	78	4	24	71	5	29	64	6	33	59	7
360	3	96	1	7	92	1	14	84	2	19	78	3	24	72	3	28	67	4
380	2	97	1	5	95	1	11	88	2	15	84	1	20	79	2	24	74	2
400	2	97	1	3	96	0	9	90	1	11	88	1	16	83	1	20	79	1
420				2	97	0	7	91	1	8	91	0	12	87	1	17	82	1
440				2	98	0	7	92	1	7	93	0	10	89	0	14	85	1
460				2	98	0	6	93	1	5	94	0	8	91	0	12	87	1

†New Zealand *versus* England, 2008, first test of three. New Zealand first innings 470 runs in 139 overs; England first innings 348 runs in 173 overs ($RR_{12} = 2.62$; $W\%D = -10$); current position (at tea on the fourth day), New Zealand second innings 55 runs for 1 wicket. Reference team New Zealand; lead, 177 runs; number of overs remaining, 120; third-innings wickets down, 1. New Zealand added 122 runs in 36 overs, setting England a target of 300 runs in 81 overs; England were 110 runs all out in 55 overs.

only a very small difference. It would have been convenient to use the model in the fifth row of Table 8 with θ not varying by partnership number. Then the distribution of Z would be exactly negative binomial, since, if $Z_1 \sim \mathrm{NB}(\pi_1, \theta)$ and $Z_2 \sim \mathrm{NB}(\pi_2, \theta)$ independent, then $Z = Z_1 + Z_2 \sim \mathrm{NB}(\pi_1 + \pi_2, \theta)$.

Thus we have the components of equation (3) for the calculation of the match outcome probability given a third-innings position and a target aimed for and a chosen run rate. Tables 10–12 show three examples. Negative binomial probabilities were calculated by using Stirling's approximation (Johnson *et al.*, 1993). From Table 10, for example, from the current position, the team batting third have a 0.70 probability of winning if they aim for a target of 400 and bat at a run rate of 3 runs per over. If the team bats more aggressively at a run rate of 6 runs per over, say, then the win probability increases to 0.87, whereas the loss probability increases only marginally from 0.05 to 0.06. Here then the batting strategy decision is relatively straightforward. In contrast Table 11 illustrates that the batting strategy decision problem is more complex and that generally speaking as the win probability increases (as a result of changing batting 'strategy') the loss probability also increases; if South Africa (the reference team) aim for 280 at 3 runs per over, their win and loss probabilities are 0.11 and 0.22 respectively; if they aim for 280 at 6 runs per over these probabilities become 0.17 and 0.35 respectively. The relative rates of increase in these probabilities, although appearing constant in both these examples, generally depend on the current position. If the team batting third have a strong lead at the current position, then increasing the run rate will increase both the win and the loss probabilities but the loss probability will increase relatively more slowly. Thus the optimum strategy will depend very much on the captain's attitude to risk (of a loss).

Table 10 clearly illustrates a limitation of our approach; the calculations are based on the assumption of no loss of overs from the current position due to an interruption, caused by bad weather for example. Weather interruptions are outside the scope of the modelling, but nonetheless they are a very important consideration when setting a final innings target. A further limitation is that the run rate is only a surrogate for batting strategy. The run rate will not be in the complete control of the batting side. Therefore, strictly, the run rate in Tables 10–12 should be interpreted in the sense 'if the batting side can and do bat at x runs per over then the match outcome probabilities are...' rather than in the sense that 'if we choose to bat at x runs per over then...'.

The entries in Tables 10–12 have been implemented on a spreadsheet that allows for the updating of the calculations as the current position changes. Thus, it is implied that the support about the decision is provided continuously; this allows for 'over-by-over' and 'run-by-run' updating. The spreadsheet implementation has the potential for practical use in test matches.

4.3. Latent variable model for strategy

A further limitation in the model for observed scores that was developed above is that the observed run rate may only be a reasonable indicator of strategy in longer innings. For shorter innings, broadly speaking, the observed run rate must provide less information about the batting strategy. Observed outcomes in short innings will be more variable for various possible reasons; for example

(a) a batsman may intend to bat very aggressively, but a low run rate may result if the batsman is quickly dismissed or
(b) very high run rates in a test match are difficult to sustain and so are likely to be associated with shorter innings.

Consequently, one might attempt to model the outcome of a partnership in a more general way. For example, the runs scored and the run rate in a partnership might be considered as a bivariate pair (Z, U) that are conditionally independent given an underlying unobserved strategy S. Here, Z and U are manifest variables and S a latent variable in the sense of Bartholomew (2002). The distribution of Z given $S = s$ we might denote as $f_{Z|S}(z|s, \phi_1)$ where ϕ_1 is a vector of unknown parameters and was specified above as a negative binomial distribution: $Z|S = s \sim \mathrm{NB}\{\pi(s), \theta\}$. The distribution of U given $S = s$ denoted as $f_{U|S}(u|s, \phi_2)$ (with parameters ϕ_2) might be such that $E(U) = s$ and $\mathrm{var}(U) \sim 1/b$ where b is the number of balls (or overs) bowled. Over the notional population of test matches, we may further suppose that S itself is a random variable with density $f_S(s, \phi_3)$ (with parameters ϕ_3). The joint distribution for Z and U is then

$$\int f_{Z|S}(z|s, \phi_1)\, f_{U|S}(u|s, \phi_2)\, f_S(s, \phi_3)\, \mathrm{d}s.$$

The likelihood for observed scores and run rates based on this joint density can then be constructed. There may be identifiability issues with the model, although, if the parameterization of U and S is fixed with respect to partnership number, the effective increase in the number of model parameters may be relatively small. If we use a gamma density for f_S and further if we assume that $\mathrm{var}(U) = \sigma^2/b$, then three additional parameters are sufficient to specify the model. Note that the model chosen in Table 8 has 13 parameters.

Alternatively, it would be interesting to model jointly the number of runs scored and the number of overs or balls bowled, B, conditional on the latent strategy S. Then the target set and the time at which it was set (the number of overs remaining) would follow directly from these quantities (when specified for each of the remaining partnerships). Then equation (3) would require modification. Determination of the joint distribution of Z and B could be developed directly and would be an interesting topic for further study. One might even consider modelling the joint distribution, conditional on S, of U and a binary random variable W that indicates whether an individual ball results in a wicket or not. This approach might be considered the most natural given that the latent batting strategy may be most directly related to the scoring rate and the probability of a fall of a wicket.

In another refinement, we might develop the model for U given $S = s$ further by using information relating to whether the innings was declared closed or not. If an innings is declared closed, we might suppose than the batsmen were confident and that therefore $E(U) = s$. If the innings were not declared closed (and the team batting third were all out before they reached their desired target), we might suppose that $E(U) < s$. Parametrically, we might have

$$E(U) = \begin{cases} s & \text{if } D = 1, \\ s - \delta & \text{if } D = 0, \end{cases}$$

for some additional parameter $\delta \geqslant 0$. Such a model would then use the information in D to some extent. Previously in this paper we had discounted the use of information regarding declaration, but this related to the model of match outcome given the end of third-innings position. One might legitimately use declaration information in inference for the outcome of the third innings as the declaration (or otherwise) is itself a part of the third-innings outcome.

We expect that the latent variable approach would tend to reduce the apparent effect of decision variables on the outcome probabilities. In other words, the model that we use in Section 4.2 tends to underestimate variability in outcomes. Clear strategies would on the whole be less apparent than in Tables 10–12. Using declaration information might conversely strengthen the effect. Returning to the question of high variability in the run rate when the number of balls

bowled is small, a simpler approach would be to left-censor observed scores arbitrarily, to refit the simple model and to consider sensitivity to the left censoring point.

5. Discussion

The aim of this paper is to model quantitatively optimum batting strategy in the third innings in test cricket. We would like to be able to model strategy given any match position. However, looking at the third innings has two benefits. Firstly, some progress can be made with the statistical problem. Secondly, batting is perhaps more strategic during this innings than in others—in the second and first typically teams will merely attempt to score as much as possible, and in the final innings a team will be either trying to win or save a game. We approach the statistical problem by supposing that the third-innings run rate and the target that the side batting third aims to set its opponent are decision variables, i.e. we suppose that these are within the control of the batting side, and the batting side will, given the current match state, choose a run rate and a target that are most desirable, be it to maximize the probability of a win or to minimize the probability of a loss, or some combination of the two. Of course, the run rate is not strictly in the control of the batting side, and therefore we think of the run rate as a surrogate for batting strategy. The run rate is merely a random variable that depends to some (unknown) extent on the batting strategy. Therefore, the output from the decision support model that we propose should be used by a team batting third to consider how match outcome probabilities vary with run rate in the remainder of the third innings and target aimed for, rather than as indicating how the side should bat for the remainder of its third innings.

To model the runs that are scored in the third innings, we look at the runs that are scored in each partnership and model these with negative binomial distributions. Partnership scores are assumed to be independent. Although there is evidence to contradict this latter assumption, the actual correlation is very small. Furthermore, there is a small difference between the distribution of team innings scores and that implied by independent partnership scores distributions. This difference is partly explained by the fact that the matches that are included in the data set which was used to model the partnership scores are only a subset of those used to model the team innings scores. A more refined approach for modelling the added runs in the third innings given the current position might consider the dependence structure between partnership scores and use a multivariate negative binomial distribution of the type that was discussed in McHale and Scarf (2007).

The problem that we address is a special case of the more general problem of determining playing strategy given the state of the match. To date, the most general approach to this problem has been described by Preston and Thomas (2000, 2002), although they looked at 1-day international cricket. The problem can be stated generally as follows: if $X(t)$ is the state of the match at time t, and Y is the outcome of the match, what is prob$\{X(t_1)|X(t_0), S\}$ $(t_0 < t_1)$, and so what playing strategy S should be adopted in the period (t_0, t_1)? In this paper, we use the run rate as a surrogate for S. In other sports, e.g. football, it is more difficult to measure the playing strategy. One might attempt to use the positions of players on the pitch, and modern data collection systems are sufficient to calculate the 'centre of gravity' of a team over time (Di Salvo *et al.*, 2006). Alternatively, one could allow the decision maker to explore different $X(t_1)$ scenarios (which are plausible given $X(t_0)$ and S) by considering prob$\{Y|X(t_1)\}$ and the decision maker's own subjective probability about the transition from $X(t_0)$ to $X(t_1)$ if he adopts strategy S in the period (t_0, t_1). This approach could be implemented in test cricket by discretizing time by session, or by lap in track cycling. The fact that the opponent will also make strategic choices is a complication. Modelling matches as dynamic games would be an interesting way forward.

'In-the-running' betting odds (e.g. the spread for the third-innings total) might be used to rescale the predicted probabilities, to take account of unmeasured factors. Conversely, the model that was developed here might be used to exploit inefficiencies in betting markets. The state of the pitch and deterioration in the pitch might be measured by using time-related run rate and strike rate, adjusted for the strengths of the batting and bowling attacks.

One wonders whether a quantitative approach like that described here is useful for the experienced coach and captain, and therefore whether it can provide a competitive edge. Perhaps decision makers already have an intuition about match outcomes that is more than sufficient for their purpose. Perhaps those aspects of a match that we do not quantify, such as the state of the pitch, and weather conditions, are so influential that they render our analysis too simple to be helpful. This said, the analysis in this paper might provide a tool that allows a decision maker to explore various options rapidly, while subjectively adjusting the model outputs to accommodate local conditions.

Acknowledgements

The authors are grateful to the referees for their comments. These not only suggested improvements for this paper but also suggested some interesting directions for modelling developments.

References

Allsopp, P. E. and Clarke, S. R. (2004) Rating teams and analysing outcomes in one-day and test cricket. *J. R. Statist. Soc.* A, **167**, 657–667.

Baker, R. and Scarf, P. (2006) Predicting the outcomes of annual sporting contests. *Appl. Statist.*, **55**, 225–239.

Bartholomew, D. J. (2002) Old and new approaches to latent variable modelling. In *Latent Variable and Latent Structure Models* (eds G. A. Marcoulides and I. Moustaki), pp. 1–14. New York: Erlbaum.

Brooks, R. D., Faff, R. W. and Sokulsky, D. (2002) An ordered response model of test cricket performance. *Appl. Econ.*, **34**, 2353–2365.

Clarke, S. R. (1988a) Test statistics. In *Statistics in Sport* (ed. J. Bennett), pp. 83–103. London: Arnold.

Clarke, S. R. (1988b) Dynamic programming in one-day cricket—optimal scoring rates. *J. Oper. Res. Soc.*, **39**, 331–337.

Clarke, S. R. and Norman, J. M. (1999) To run or not?: some dynamic programming models in cricket. *J. Oper. Res. Soc.*, **50**, 536–545.

Clarke, S. R. and Norman, J. M. (2003) Dynamic programming in cricket: choosing a night watchman. *J. Oper. Res. Soc.*, **54**, 838–845.

Di Salvo, V., Collins, A., McNeill, B. and Cardinale, M. (2006) Validation of Prozone®: a new video-based performance analysis system. *Int. J. Perform. Anal. Sport*, **6**, 108–119.

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

Elderton, W. (1945) Cricket scores and some skew correlation distributions (an arithmetical study). *J. R. Statist. Soc.* A, **108**, 1–11.

Elderton, W. P. and Elderton, E. M. (1909) *Primer of Statistics*. London: Black.

ESPNcricinfo (2010) The home of cricket. ESPNcricinfo, Bristol. (Available from `http://www.cricinfo.com/`.)

International Cricket Council (2010) International Cricket Council team rankings. International Cricket Council, Dubai. (Available from `http://www.icc-cricket.yahoo.net/match_zone/team_ranking.php`.)

Johnson, N. L., Kotz, S. and Kemp, A. W. (1993) *Univariate Discrete Distributions*, 2nd edn. New York: Wiley.

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

McHale, I. and Scarf, P. A. (2007) Modelling soccer matches using bivariate discrete distributions with general dependence structure. *Statist. Neerland.*, **61**, 432–445.

Nagalkerke, N. J. D. (1991) A note on a general definition of the coefficient of determination. *Biometrika*, **78**, 691–692.

Pollard, R., Benjamin, B. and Reep, C. (1977) Sport and the negative binomial distribution. In *Optimal Strategies in Sports* (eds S. P. Ladany and R. E. Machol), pp. 118–195. New York: North-Holland.

Preston, I. and Thomas, J. (2000) Batting strategy in limited overs cricket. *Statistician*, **49**, 95–106.

Preston, I. and Thomas, J. (2002) Rain rules for limited overs cricket and probabilities of victory. *Statistician*, **51**, 189–202.
Reep, C., Pollard, R. and Benjamin, B. (1971) Skill and chance in ball games. *J. R. Statist. Soc.* A, **134**, 623–629.
Ringrose, T. J. (2006) Neutral umpires and leg before wicket decisions in test cricket. *J. R. Statist. Soc.* A, **169**, 903–911.
Scarf, P. A. and Shi, X. (2005) Modelling match outcomes and decision support for setting a final innings target in test cricket. *IMA J. Mangmnt Math.*, **16**, 161–178.
Swartz, T. B., Gill, P. S., Beaudoin, B. and de Silva, B. M. (2006) Optimal batting orders in one-day cricket. *Comput. Ops Res.*, **33**, 1939–1950.
Wood, G. H. (1945) Cricket scores and geometrical progression. *J. R. Statist. Soc.* A, **108**, 12–22.