Statistical Modelling xxxx; **xx(x)**: 1–16

A truncated mean-parameterized Conway-Maxwell-Poisson model for the analysis of Test match bowlers

Peter M. Philipson[1]

[1]School of Mathematics, Statistics & Physics, Newcastle University, Newcastle upon Tyne, United Kingdom

Abstract: A truncated, mean-parameterized Conway-Maxwell-Poisson model is developed to handle under- and overdispersed count data owing to individual heterogeneity. The truncated nature of the data allows for a more direct implementation of the model than is utilized in previous work without too much computational burden. The model is applied to a large dataset of Test match cricket bowlers, where the data are in the form of small counts and range in time from 1877 to the modern day, leading to the inclusion of temporal effects to account for fundamental changes to the sport and society. Rankings of sportsmen and women based on a statistical model are often handicapped by the popularity of inappropriate traditional metrics, which are found to be flawed measures in this instance. Inferences are made using a Bayesian approach by deploying a Markov Chain Monte Carlo algorithm to obtain parameter estimates and to extract the innate ability of individual players. The model offers a good fit and indicates that there is merit in a more sophisticated measure for ranking and assessing Test match bowlers.

Key words: Conway-Maxwell-Poisson, count data, overdispersion, truncation, underdispersion

Received August 2021; **revised** November 2022; **accepted** January 2023

1 Introduction

Statistical research in cricket has been somewhat overlooked in the stampede to model football and baseball. Moreover, the research that has been done on cricket has largely focused on batsmen, whether modelling individual, partnership and team scores (Kimber and Hansford, 1993; Scarf et al., 2011; Pollard et al., 1977), or ranking Test batsmen (Brown, 2009; Rohde, 2011; Boys and Philipson, 2019; Stevenson and Brewer, 2021), or on short-format cricket via predicting match outcomes (Davis et al., 2015) or optimising the batting strategy in one day and Twenty20 international cricket (Preston and Thomas, 2000; Swartz et al., 2006; Perera et al., 2016). Indeed, to the author's knowledge, this is the first work that explicitly focuses on Test match bowlers.

Address for correspondence: Peter M. Philipson, School of Mathematics, Statistics & Physics, Newcastle University, NE1 7RU, Newcastle upon Tyne, United Kingdom.
E-mail: peter.philipson1@newcastle.ac.uk

10.1177/1471082X231178584

Test cricket is the oldest form of cricket. With a rich and storied history, it is typically held up as being the ultimate challenge of ability, nerve and concentration, hence the origin of the term 'Test' to describe the matches. For Test batsmen, their value is almost exclusively measured by how many runs they score and their career batting average, with passing mention made of the rate at which they score, if this is remarkable. Test bowlers are also primarily rated on their (bowling) average, which in order to be on the same scale as the batting average is measured as runs conceded per wicket, rather than the more natural rate of wickets per run. In this work we consider the problem of comparing Test bowlers across the entire span of Test cricket from its genesis in March 1877 to the modern day, July 2022.

Rather uniquely, the best bowling averages of all-time belong to bowlers who played more than a hundred years ago, contrast this with almost any other modern sport where records are routinely broken by current participants, with their coteries of support staff dedicated to fitness, nutrition and wellbeing along with access to detailed databases highlighting their strengths and weaknesses. The proposed model allows us to question whether the best Test bowlers are truly those who played in the late 19th and early 20th century or whether this is simply a reflection of the sport at the time. Along the way, we also deliberate whether the classic bowling average is the most suitable measure of career performance. Taking these two aspects together suggests that there are more suitable alternatives than simply ranking all players over time based on their bowling average, as seen at https://stats.espncricinfo.com/ci/content/records/283256.html.

The structure of the article is as follows. The data are described in Section 2 and contain truncated, small counts which, at the player level, are both under and overdispersed, leading to the statistical model in Section 3. Section 4 details the prior distribution alongside the computational details. Section 5 presents some of the results and the article concludes with some discussion and avenues for future work in Section 6.

2 The data

The data used in this article consists of $N = 47\,216$ individual innings bowling figures by $n = 2\,207$ Test match bowlers from the first Test played in 1877 up to Test 2 473, in July 2022. There are currently twelve Test playing countries and far more Test matches are played today than at the genesis of Test cricket (Boys and Philipson, 2019). World Series Cricket matches are not included in the dataset since these matches are not considered official Test matches by the International Cricket Council (ICC).

Bowling data for a player on a cricket scorecard comes in the form 'overs-maidens-wickets-runs'—note that in this work the first two values provide meta-information that are not used for analysis—as seen at the bottom of Figure 1. A concrete example are the bowling data for James Anderson with figures of 25.5-5-61-2. The main aspect of this to note is that the data are aggregated counts for bowlers—2 wickets were taken for 61 runs in this instance, but it is not known how many runs were conceded for each individual wicket. This aggregation is compounded across all matches to give a career bowling average for a particular player, corresponding to the average number of runs they concede per wicket taken. This measure is of a form that is understandable to fans, but counter-intuitive from the standpoint of a statistical model; this point is revisited in subsection 3.1.1.

Table 1 Frequency and percentage of Test match wickets per innings

Wickets	0	1	2	3	4	5	6	7	≥ 8
Frequency	16 468	10 935	7 599	5 406	3 262	1 964	801	281	100
Percentage	34.9	23.2	16.9	11.4	6.9	4.2	1.7	0.6	0.2

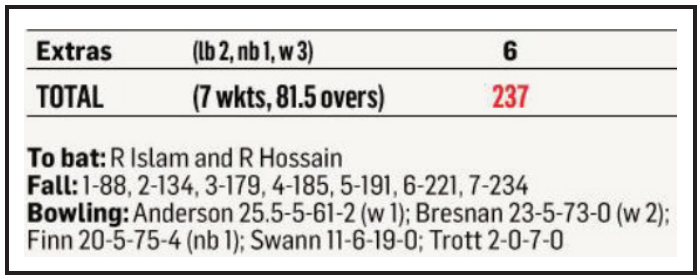

Figure 1 Bowling figures as seen on a typical cricket scorecard.

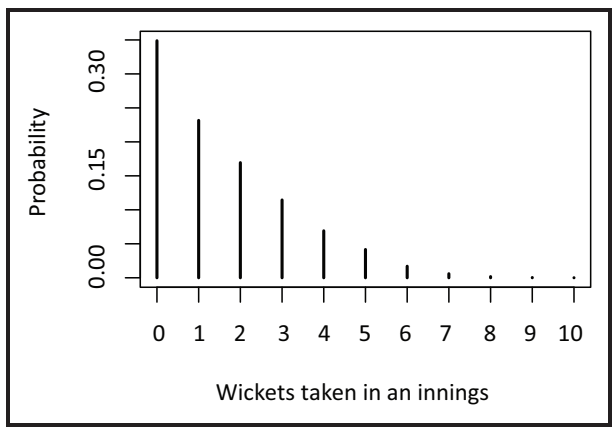

Figure 2 Probability mass function of Test match wickets taken per innings.

In each Test match innings there are a maximum of ten wickets that can be taken by the bowling team, and these wickets are typically shared amongst the team's bowlers, of which there are nominally four or five. Table 1 shows the distribution of wickets taken in an innings across all bowlers, with Figure 2 providing a visual representation. Taking six wickets or more in an innings is rare and the most common outcome is that of no wickets taken.

Alongside the wickets and runs, there are data available for the identity of the player, the opposition, the venue (home or away), the match innings, the winners of the toss and the date the match took place. These are all considered as covariates in the model.

To motivate the model introduced in section 3, the raw indices of dispersion at the player level show that around 20% of players have underdispersed counts, approximately 20% of players have equidispersed counts, with the remainder having overdispersed counts. This overlooks the effects of covariates but suggests that any candidate distribution ought to be capable of handling both under- *and* overdispersion, or *bidispersion* at the player level.

3 The model

Wickets taken in an innings are counts and a typical, natural starting point may be to consider modelling them via the Poisson or negative binomial distributions. However, in light of the bidispersion seen at the player level in the raw data, we instead turn to a distribution capable of handling such data.

The Conway-Maxwell-Poisson (CMP) distribution (Conway and Maxwell, 1962) is a generalization of the Poisson distribution, which includes an extra parameter to account for possible over- and underdispersion. Despite being introduced almost sixty years ago to tackle a queueing problem, it has little footprint in the statistical literature, although it has gained some traction in the last fifteen years or so with applications to household consumer purchasing traits (Boatwright et al., 2003), retail sales, lengths of Hungarian words (Shmueli et al., 2005), and road traffic accident data (Lord et al., 2008, 2010). Its wider applicability was demonstrated by Guikema and Goffelt (2008) and Sellers and Shmueli (2010), who recast the CMP distribution in the generalized linear modelling framework for both Bayesian and frequentist settings respectively, and through the development of an R package (Sellers et al., 2019) in the case of the latter, to help facilitate routine use.

In the context considered here, the CMP distribution is particularly appealing as it allows for both over- and underdispersion for individual players. For the cricket bowling data, we define X_{ijk} to represent the number of wickets taken by player i in his jth year during his kth bowling performance of that year. Also n_i and n_{ij} denote, respectively, the number of years in the career of player i and the number of innings bowled in during year j in the career of player i. Using this notation, the CMP distribution has probability mass function given by

$$\Pr(X_{ijk} = x_{ijk}|\lambda_{ijk}, \nu) = \frac{\lambda_{ijk}^{x_{ijk}}}{(x_{ijk}!)^{\nu}} \frac{1}{G_{\infty}(\lambda_{ijk}, \nu)}$$

with $i = 1, \ldots, 2\,207$, $j = 1, \ldots, n_i$ and $k = 1, \ldots, n_{ij}$. In this (standard) formulation, λ_{ijk} is the rate parameter and ν models the dispersion. The normalizing constant term $G_{\infty}(\lambda_{ijk}, \nu) = \sum_{r=0}^{\infty} \lambda_{ijk}^{r}/(r!)^{\nu}$ ensures that the CMP distribution is proper, but complicates analysis.

Sellers et al. (2019) implemented the CMP model using a closed-form approximation (Shmueli et al., 2005; Gillispie and Green, 2015) when λ_{ijk} is large and ν is small, otherwise truncating the infinite sum to ensure a pre-specified level of accuracy is met. Alternative methods to circumvent intractability for the standard CMP model in the Bayesian setting have been proposed by Chanialidis et al. (2018), who used rejection sampling based on a piecewise enveloping distribution and more recently Benson and Friel (2020) developed a faster method using a single, simple envelope distribution, but adapting these methods to the mean-parameterized CMP (MPCMP) distribution introduced below is a non-trivial task.

3.1 Truncated mean-parameterized CMP distribution

The standard CMP model is not parameterized through its mean, however, restricting its wider applicability in regression settings since this renders effects hard to quantify, other than as a general increase or decrease. To counter this, two alternative parameterizations via the mean have been developed (Huang, 2017; Ribeiro Jr et al., 2018; Huang and Kim, 2019), each with associated R

packages (Fung et al., 2019; Elias Ribeiro Junior, 2021). The mean of the standard CMP distribution can be found as

$$\mu_{ijk} = \sum_{r=0}^{\infty} \frac{r\lambda_{ijk}^{r}}{(r!)^{\nu} G_{\infty}(\lambda_{ijk}, \nu)},$$

which, upon rearranging, leads to

$$\sum_{r=0}^{\infty}(r - \mu_{ijk})\frac{\lambda_{ijk}^{r}}{(r!)^{\nu}} = 0. \tag{3.1}$$

Hence, the CMP distribution can be mean-parameterized to allow a more conventional count regression interpretation, where λ_{ijk} is a nonlinear function of μ_{ijk} and ν under this reparameterization. Huang (2017) suggested a hybrid bisection and Newton-Raphson approach to find λ_{ijk} and applied this in small sample Bayesian settings (Huang and Kim, 2019), whereas Ribeiro Jr et al. (2018) used an asymptotic approximation of $G_{\infty}(\lambda_{ijk}, \nu)$ to obtain a closed form estimate for λ_{ijk}. The appeal of the former is its more exact nature, but this comes at considerable computational cost in the scenario considered here as the iterative approach would be required at each MCMC iteration, and, in this case, for a large number of (conditional) mean values. The approximation used by the latter is conceptually appealing due to its simplicity and computational efficiency, but is likely to be inaccurate for some of the combinations of μ_{ijk}, ν encountered here, and the level of accuracy will also vary across these combinations.

Irrespective of mean parameterization and method, the above model formulation has two obvious flaws: the counts lie on a restricted range, that is, $0, \ldots, 10$, and there is no account taken of how many runs the bowler conceded in order to take their wickets. For the first issue, the model can be easily modified using truncation, which, in this case, leads to a simplified form for the CMP distribution, which is exploited below:

$$\begin{aligned}\Pr(X_{ijk} = x_{ijk}|\lambda_{ijk}, \nu) &= \frac{\lambda_{ijk}^{x_{ijk}}}{(x_{ijk}!)^{\nu}} \frac{1}{G_{\infty}(\lambda_{ijk}, \nu)} \frac{1}{\Pr(X_{ijk} \leq 10|\lambda_{ijk}, \nu)} \\ &= \frac{\lambda_{ijk}^{x_{ijk}}}{(x_{ijk}!)^{\nu}} \frac{1}{G_{10}(\lambda_{ijk}, \nu),}\end{aligned}$$

where $G_{10} = \sum_{r=0}^{10} \lambda_{ijk}^{r}/(r!)^{\nu}$ is used to denote the finite sum. This yields the truncated mean-parameterized CMP distribution (MPCMP_{10}), where the subscript denotes the value at which the truncation occurs. Furthermore, the infinite sum in (3.1) is replaced by the finite sum

$$\sum_{r=0}^{10}(r - \mu_{ijk})\frac{\lambda_{ijk}^{r}}{(r!)^{\nu}} = 0. \tag{3.2}$$

Since λ_{ijk} is positive, there is a single sign change in (3.2) when $\mu_{ijk} > r$, which, by Descartes' rule of signs, informs us that there is a solitary positive real root. Hence, a solution for λ_{ijk} can be found

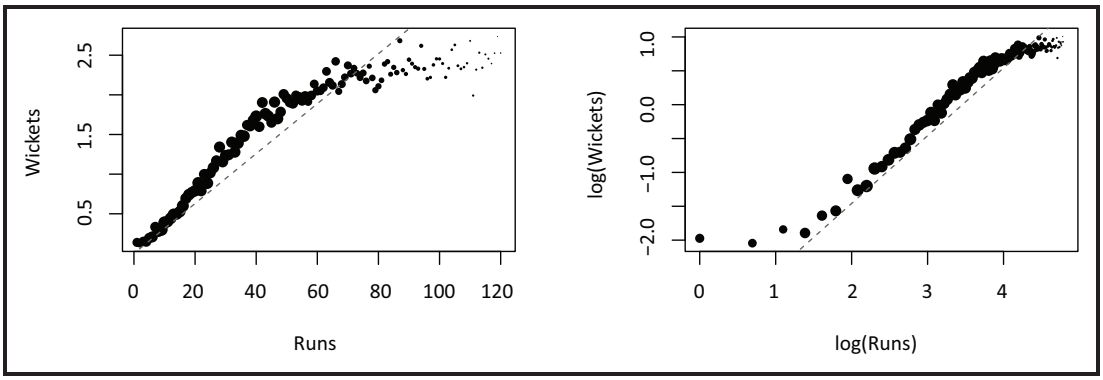

Figure 3 Average wickets taken for values of runs on the linear (left) and log (right) scales. The size of the data points reflects the amount of data for each value of runs; the dashed line represents the relationship under the traditional cricket bowling average.

without recourse to approximations or less scaleable iterative methods in this case and we can directly solve the tenth order polynomial using the `R` function `polyroot`, which makes use of the Jenkins-Traub algorithm. The substantive question of the relationship between wickets and runs is deliberated in the next subsection.

3.1.1 Functional form of runs

As noted earlier, data are only available in an aggregated form. That is, the total number of wickets is recorded alongside the total number of runs conceded. Historically, this has been converted to a cricket bowling average by taking the rate of runs conceded to wickets taken, chiefly to map this on to a similar scale as to the classic batting average. However, in the usual (and statistical) view of a rate this is more naturally expressed as wickets per run, rather than runs per wicket, and this rate formulation is adopted henceforth. In either event, the number of runs conceded conveys important information since taking three wickets at the cost of thirty runs is very different to taking the same number of wickets for, say, ninety runs.

By looking at the mean number of wickets for each value of runs the relationship between wickets and runs can be assessed, see Figure 3; note that there is very little data for values of runs exceeding 120 so the plot is truncated at this point. The nonlinear relationship rules out instinctive choices such as an offset or additive relationship—this makes sense as the number of wickets taken by a bowler cannot exceed ten in a (within-match) innings, which suggests that the effect of runs on wickets is unlikely to be wholly multiplicative (or additive on the log-scale).

Smoothing splines in the form of cubic B-splines are adopted to capture the nonlinear relationship, with knots chosen at the quintiles of runs (on the log scale). This corresponds to internal knots at 18, 35, 53 and 77 along with the boundary knots at 1 and 298 on the runs scale. Various nonlinear models were also considered but did not capture the relationship as well as the proposed spline.

3.1.2 Opposition effects

As the data span 146 years it is perhaps unreasonable to assume that some effects are constant over time. In particular, teams are likely to have had periods of strength and/or weakness whereas

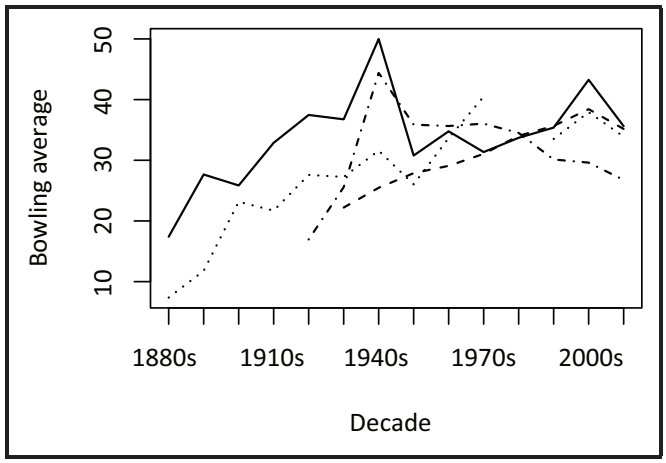

Figure 4 Mean bowling averages across decades for a selection of opposing countries: Australia (solid line), India (dashed), South Africa (dotted) and West Indies (dot-dash).

conditions, game focus, advances in equipment and technology may have drastically altered playing conditions for all teams at various points in the Test cricket timeline. A plot of the mean bowling average across decades for a selection of opposition countries is given in Figure 4. Here we use the conventional bowling average on the y-axis for ease of interpreation, but the story is similar when using the rate. Clearly, the averages vary substantially over time as countries go through periods of strength and weakness and game conditions and rules evolve. As such, treating them as time invariant effects does not seem appropriate. Note also that some countries started playing Test cricket much later than 1877 (see the lines for West Indies and India in Figure 4) and teams appear to take several years to adapt and improve.

This motivates the inclusion of dynamic opposition effects, where we again adopt smoothing splines, this time with knots chosen at the midpoints of decades (of which there are fourteen); this is a natural timespan, both in a general and sporting sense; cricket followers will often discuss the West Indies of the 1980s or Australia of the 1990s for instance. For identifiability, the opposition effect of Australia in 2022 is chosen as the reference value.

3.2 Log-linear model for the mean rate

As well as runs conceded by a bowler and the strength of opposition, there are several other factors that can affect performance, some of which are considered formally in the model. Introducing some notation for available information: r_{ijk} is the number of runs conceded by bowler i during the kth innings in the jth year of his career, y_{ijk} represents the year of this same event, o_{ijk} indicates the opposition (there are twelve Test playing countries in these data), h_{ijk} indicates whether the innings took place in the bowler's home country (1 = home, 2 = away), m_{ijk} is the within-match innings index and t_{ijk} represents winning or losing the toss (1 = won, 2 = lost).

These remaining terms in the model for the wicket-taking rate are assumed to be additive and, hence, the mean log-rate is modelled as

$$\log \mu_{ijk} = \sum_{p=1}^{8} \beta_p B_p \left\{\log r_{ijk}\right\} + \sum_{q=1}^{q_o} \omega_{q,o_{ijk}} B_{q,o_{ijk}}(y_{ijk}) I(\text{Opp} = o_{ijk}) + \theta_i + \zeta_{h_{ijk}} + \xi_{m_{ijk}} + \gamma\, I(t_{ijk} = 1) I(m_{ijk} = 1) \tag{3.3}$$

where θ_i represents the ability of player i and the next three terms in the model are game-specific, allowing home advantage, pitch degradation (via the match innings effects), and whether the toss was won or lost, respectively, to be taken into account. Home advantage is ubiquitous in sport (Pollard and Pollard, 2005), and it is widely believed that it is easier to bowl as a match progresses to the third and fourth innings due to pitch degradation, and winning the toss allows a team to have 'best' use of the playing/weather conditions. The effect of losing the toss is ameliorated as the match progresses so we anticipate that this only effects the first innings of the match, and this effect is captured by γ.

The design matrices B_p and $B_{q,o_{ijk}}$ are the spline bases for (log) runs and each opposition detailed in the previous two subsections, with associated parameters β_p and $\omega_{q,o_{ijk}}$. Note that the summation index for the opposition spline varies across countries owing to their different spans of data—as seen in subsection 3.1.2—with the total number of parameters for each opposition denoted denoted by $q_o, o = 1, \ldots, 12$.

For identifiability purposes we set $\zeta_1 = \xi_1 = 0$, measuring the impact of playing away via ζ_2 and the innings effects relative to the first innings through ξ_2, ξ_3 and ξ_4. We impose a sum-to-zero constraint on the player abilities, thus in this model $\exp(\theta_i)$ is the rate of wickets per innings taken by player i bowling at home in the first innings of a Test match against Australia in 2022.

CMP regression also allows a model for the dispersion. Here, recognizing that we may have both under and overdispersion at play, we opt for a player-specific dispersion term $\nu_i, i = 1, \ldots 2\,207$. Naturally, this could be extended to include covariates, particularly runs, but this is not pursued here since the model is already heavily parameterized (for the mean).

4 Computational details and choice of priors

R (R Core Team, 2022) was used for all model fitting, analysis and plotting, with the `splines` package used to generate the basis splines for runs and the temporal opposition effects. The Conway-Maxwell-Poisson distribution is not included in standard Bayesian software such as `rstan` (Stan Development Team, 2020) and `rjags` (Plummer, 2004) so bespoke code was written in R to implement the model. The code and data are available on GitHub at https://github.com/petephilipson/MPCMP_Test_bowlers.

For analysis, four MCMC chains were run in parallel with 1 000 warm-up iterations followed by 5 000 further iterations. Model fits took approximately eight hours on a standard MacBook Pro. A Metropolis-within-Gibbs algorithm is used in the MCMC scheme, with component-wise updates for all parameters except for those involved in the spline for runs, $\boldsymbol{\beta}$. For these parameters a block update was used to circumvent the poor mixing seen when deploying one-at-a-time updates, with the proposal covariance matrix based on the estimated parameter covariance matrix from a frequentist

fit using the `mpcmp` package. In order to simulate from the CMP distribution, to enable posterior predictive checking, the `COMPoissonReg` (Sellers et al., 2019) package was used, albeit with some modifications to ensure the counts are truncated. Plots were generated using `ggplot2` (Wickham, 2016) and highest posterior density intervals were calculated using `coda` (Plummer et al., 2006).

4.1 Prior distribution

For the innings, playing away and winning the toss effects we use zero mean normal distributions with standard deviation 0.25, reflecting a belief that these effects are likely to be quite small on the multiplicative wicket-taking scale, with effects larger than 50% increases or decreases deemed unlikely (with a 5% chance a priori). The same prior distribution is used for the player ability terms, $\boldsymbol{\theta}$, reflecting that while heterogeneity is expected we do not expect players to be, say, five and ten times better/worse in terms of rate. The coefficients for the splines in both the runs and opposition components are given standard normal priors. We recognize that smoothing priors could be adopted here, but as we have already considered the choice of knots we do not consider such an approach here.

For the dispersion parameters we work on the log-scale, introducing $\eta_i = \log(\nu_i)$ for $i = 1, \ldots, 2\,207$. We adopt a prior distribution that assumes equidispersion, under which the MPCMP model is equivalent to a Poisson distribution. Due to the counts being small we do not expect the dispersion in either direction to be that extreme, allowing a 5% chance for η_i to be a three-fold change from the a priori mean of equidispersion. Hence, the prior distribution for the log-dispersion is $\eta_i \sim N(0, 0.5 \log 3)$ for $i = 1, \ldots, 2\,207$.

5 Results

5.1 Functional form for runs

A plot of wickets against runs using the posterior means for β is given in Figure 5. This clearly shows the non-linear relationship between wickets and runs and suggests that using the standard bowling average, which operates in a linear fashion, overlooks the true nature of how the number of wickets taken varies with the number of runs conceded. An important ramification of this is that the standard bowling average overestimates the number of wickets taken as the number of runs grows large, whereas the true relationship suggests that the rate starts to flatten out for values of runs larger than 50. Returning to the figure, we clearly see much more uncertainty for larger values of runs, where, as seen earlier, the data are considerably more sparse.

5.2 Opposition effects

A plot of the fitted posterior mean profiles for each opposition is given in Figure 6. The largest values of all occur for South Africa when they first played (against the more experienced England and Australia exclusively); most teams struggle when they first play Test cricket, as shown by the largest posterior means at the left-hand side of each individual plot (since we are modelling the rate, which is the inverse of the traditional average). Overall, this led to the low Test bowling averages of the

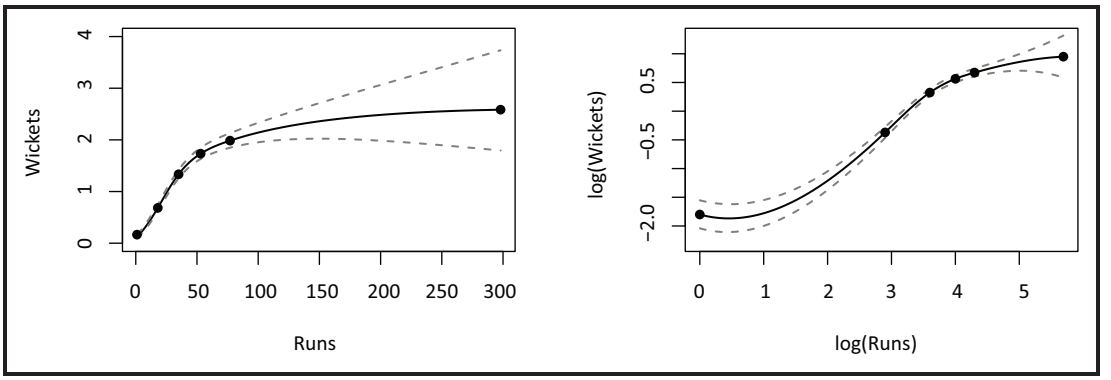

Figure 5 Mean wickets taken against runs on the linear (left) and log (right) scales with 95% HDI interval uncertainty bands; solid circles represent the knots for the spline.

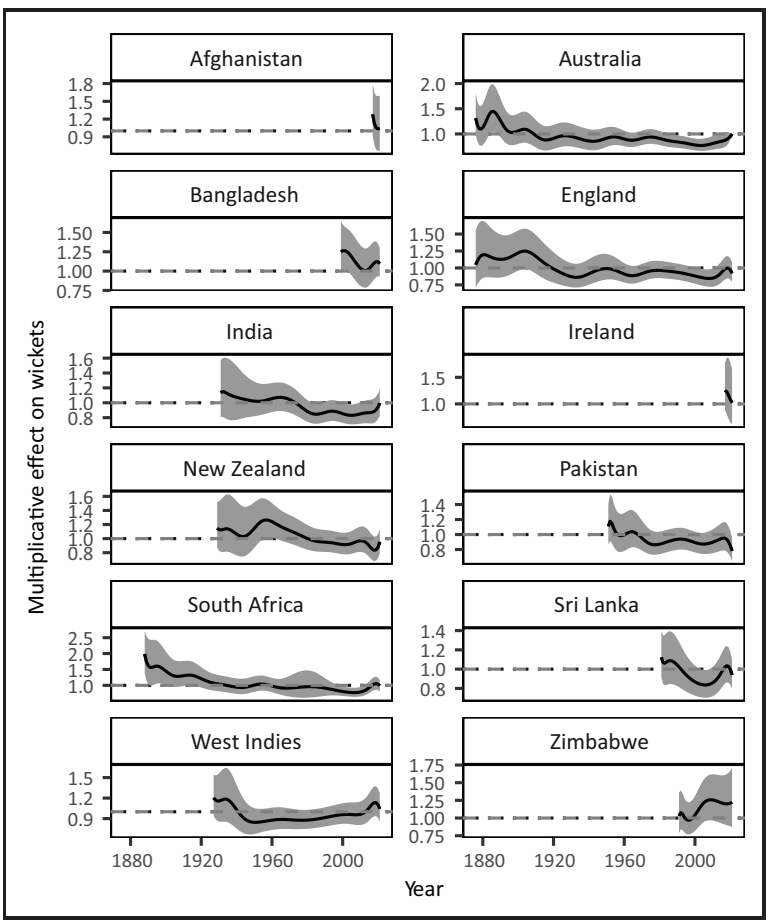

Figure 6 Posterior mean and 95% HDI bands for each opposition.

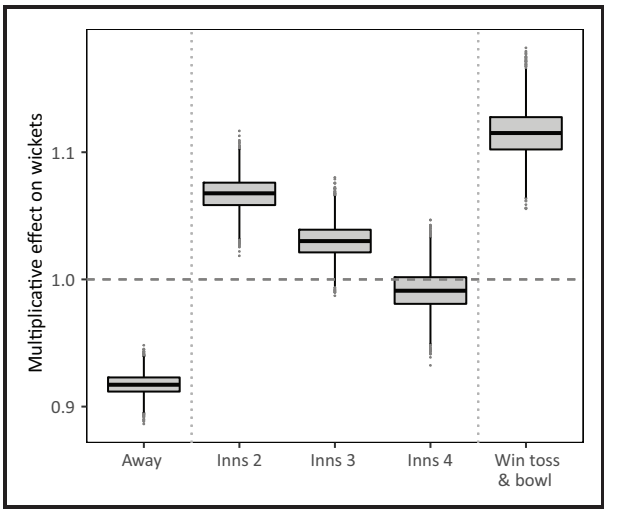

Figure 7 Boxplots of the posterior distributions for playing away (ζ_2), match innings 2–4 (ξ_2, ξ_3, ξ_4) and winning the toss and bowling (γ)

1880s-1900s that still stand today as the lowest of all-time and adjusting for this seems fundamental to a fairer comparison and/or ranking of players.

5.3 Game-specific effects

The posterior means and 95% highest density intervals (HDIs) for the innings effects—on the multiplicative scale—are 1.07 (1.04–1.09), 1.03 (1.01–1.06) and 0.99 (0.96–1.02) for the second, third and fourth innings respectively, when compared to the first innings. Similarly, the effect of playing away is to reduce the mean number of wickets taken by around 10%; posterior mean and 95% HDI are 0.92 (0.90–0.93). The impact of bowling after winning the toss is the strongest of the effects we consider, with a posterior mean and 95% HDI interval of 1.12 (1.08–1.15). Boxplots summarizing the game-specific effects are shown in Figure 7.

5.4 Player rankings

The top thirty bowlers, as ranked by their posterior mean ability, θ_i, are given in Table 2. We also include the posterior dispersion parameter for each player along with information on the era in which they played (via the date of their debut) and the amount of available data as measured through the total number of innings they bowled in (n_i).

The rankings under the proposed model differ substantially from a ranking based on bowling average alone. As a case in point, the lowest, that is, best, bowling average of all time belongs to GA Lohmann, who is ranked 9th in our model. Similarly, Muttiah Muralidaran, who tops our list, is 50th on the list of best averages at the time of writing. The main ramifications of using the proposed model is that players from the early years are ranked lower, and spinners are generally

Table 2 Top 30 Test match bowlers ranked by posterior mean multiplicative rate

Rank	Name	Debut	Innings	Player ability $E(e^{\vartheta})$	Player ability $SD(e^{\vartheta})$	Dispersion $E(\nu)$
1	M Muralidaran	1992	230	2.02	0.08	0.83
2	SF Barnes	1901	50	1.91	0.18	0.62
3	WJ O'Reilly	1932	48	1.81	0.18	0.71
4	Sir RJ Hadlee	1973	150	1.78	0.10	0.58
5	JJ Ferris	1887	16	1.78	0.24	1.02
6	CV Grimmett	1925	67	1.77	0.14	0.79
7	AA Donald	1992	129	1.73	0.10	1.04
8	MJ Procter	1967	14	1.72	0.25	1.12
9	GA Lohmann	1886	36	1.71	0.22	0.35
10	MD Marshall	1978	151	1.71	0.10	0.76
11	T Richardson	1893	23	1.70	0.21	0.93
12	J Cowie	1937	13	1.69	0.25	0.88
13	SK Warne	1992	271	1.69	0.08	0.64
14	CEL Ambrose	1988	179	1.68	0.10	0.69
15	DW Steyn	2004	171	1.68	0.09	0.71
16	GD McGrath	1993	241	1.68	0.09	0.65
17	R Ashwin	2011	162	1.68	0.09	0.75
18	JC Laker	1948	86	1.67	0.14	0.63
19	CTB Turner	1887	30	1.66	0.20	0.66
20	Mohammad Asif	2005	44	1.65	0.17	0.74
21	DK Lillee	1971	132	1.65	0.10	0.84
22	K Rabada	2015	95	1.63	0.10	1.23
23	FH Tyson	1954	29	1.62	0.20	0.64
24	H Ironmonger	1928	27	1.62	0.21	0.55
25	Imran Khan	1971	160	1.62	0.10	0.70
26	SE Bond	2001	32	1.62	0.16	1.24
27	Waqar Younis	1989	154	1.61	0.10	0.76
28	FS Trueman	1952	126	1.61	0.10	0.86
29	PJ Cummins	2011	81	1.61	0.12	0.88
30	AK Davidson	1953	82	1.6	0.14	0.62

ranked higher. The latter is an artefact of the model, in that bowling long spells and taking wickets is now appropriately captured through the nonlinear effect for runs.

We also see from Table 2 that five players—all seam bowlers interestingly—have underdispersed data judging from the posterior means of ν_i, verifying that a model capable of handling both underdispersed and overdispersed counts is required for these data.

5.5 Model fitting

The performance of the model is evaluated using the posterior predictive distribution. Namely, due to the small number of observed counts, it is viable to compare the model-based posterior predictive probability for each observed value of 0, 1, …, 10 and compare that to the observed probability in each case. A summary of this information is given in Figure 8, where we see excellent agreement between the observed and expected proportions at each value of wickets.

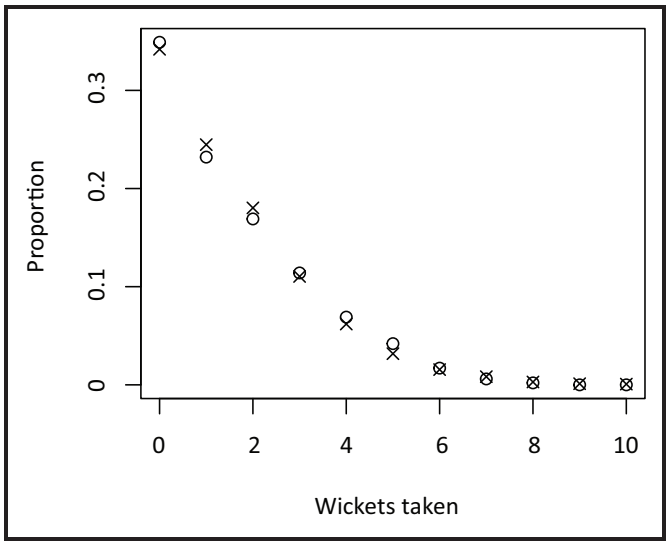

Figure 8 Observed (grey circles) and model-based (crosses) probabilities for each value of wickets.

6 Discussion

The outcome of interest considered here was the number of wickets taken by a bowler in an innings, which was modelled using a truncated mean-parameterized Conway-Maxwell-Poisson distribution. The model allows for comparisons between players from different eras through the inclusion of dynamic opposition effects alongside axiomatic game-specific effects of home advantage, pitch conditions and the best use thereof. However, there are other factors that were not considered here, principally due to the aggregated nature of the data, which precludes looking at the ability of bowlers to break partnerships, take top-order wickets or to perform at their best in the most important matches, in their most crucial stages.

Other classic count models were considered, but it was found that Poisson and negative binomial regression models failed to adequately describe the data. Alternative count models capable of handling both under- and overdispersion may perform equally well, such as those based on the gamma count, generalized Poisson, double Poisson or Poisson-Tweedie distributions, although each of these has some limitations. Looking further afield, an exponentially tilted multinomial model (Rathouz and Gao, 2009) may offer increased flexibility over the aforementioned count data models. Alternative models are not pursued further here, owing to the good performance of the chosen model and the nontrivial implementation of such non-standard models (that would require bespoke coding) in this big data setting.

Additional metrics not formally modelled here are bowling strike rate and bowling economy rate, which concern the number of balls bowled (rather than the number of runs conceded) per wicket and runs conceded per over respectively. Whilst both informative measures, they are typically viewed as secondary and tertiary respectively to bowling average in Test cricket, where time is less constrained than in shorter form cricket. Hence, future work looking at one day international or

Twenty20 cricket could consider the triple of average, strike rate and economy rate for bowlers, alongside average and strike rate for batsmen, thereby requiring a multivariate count data model.

One limitation of this work is that we have ignored possible dependence between players, the taking of wickets can be thought of as a competing resource problem and the success (or lack of) for one player may have an impact on other players on the same team. Indeed, the problem could potentially be recast as one of competing risks. However, this argument is rather circular in that a bowler would not concede many runs if a teammate is taking lots of wickets and this is taken into account through the modelling of the relationship between wickets and runs.

The MPCMP model may have broader use in other sports where there may be bidispersed data, for example modelling goals scored in football matches—stronger and weaker teams are likely to exhibit underdispersion; hockey, ice-hockey and baseball all have small (albeit not truncated) counts as outcomes of interest with bidispersion likely at the team or player level, or both. Moving away from sports to other fields, the MPCMP model could be used to model parity, which is known to vary widely across countries with heavy underdispersion (Barakat, 2016), or scores in the popular web-based word game, Wordle. It could also be appropriate for longitudinal counts with volatile (overdispersed) and stable (underdispersed) profiles at the patient level, where the level of variability may be related to a clinical outcome of interest in a joint modelling setting. Indeed, it will have broader use in any field where counts are subject to bidispersion.

This work has shown that truncated counts subject to bidispersion at some hierarchical level can be handled in a mean-parameterized CMP model, based on a large dataset, without too much computational overhead. Further methodological work is needed to implement the model in the case of non-truncated counts and to seek faster computational methods.

Acknowledgements

Thanks to the late Professor Richard Boys, this work is dedicated in memory of him. The genesis of this work was a Masters project under his kind supervision farther back in time than I would care to remember.

Further thanks to the reviewer and editor whose comments helped improve the article and provided some thought-provoking content to help shape future research.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author received no financial support for the research, authorship and/or publication of this article.

References

Barakat BF (2016) *Generalised Poisson distributions for modelling parity*. Technical report, Vienna Institute of Demography Working Papers.

Benson A and Friel N (2020) *Bayesian inference, model selection and likelihood estimation using fast rejection sampling: the Conway-Maxwell-Poisson distribution*. Bayesian Analysis.

Boatwright P, Borle S and Kadane JB (2003) A model of the joint distribution of purchase quantity and timing. *Journal of the American Statistical Association*, **98**, 564–572.

Boys RJ and Philipson PM (2019) On the ranking of test match batsmen. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **68**, 161–179.

Brown HSI (2009) Comparing batsmen across different eras: The ends of the distribution justifying the means. *Economic Analysis & Policy*, **39**, 443–454.

Chanialidis C, Evers L, Neocleous T and Nobile A (2018) Efficient Bayesian inference for COM-Poisson regression models. *Statistics and Computing*, **28**, 595–608.

Conway RW and Maxwell WL (1962) A queuing model with state dependent service rates. *Journal of Industrial Engineering*, **12**, 132–136.

Davis J, Perera H and Swartz TB (2015) A simulator for Twenty20 cricket. *Australian & New Zealand Journal of Statistics*, **57**, 55–71.

Elias Ribeiro Junior E (2021) cmpreg: Reparametrized COM-Poisson Regression Models. R package version 0.0.1.

Fung T, Alwan A, Wishart J and Huang A (2019) The mpcmp Package for Mean-Parameterised Conway-Maxwell-Poisson Regression.

Gillispie SB and Green CG (2015) Approximating the Conway-Maxwell-Poisson distribution normalization constant. *Statistics*, **49**, 1062–1073.

Guikema SD and Goffelt JP (2008) A flexible count data regression model for risk analysis. *Risk Analysis: An International Journal*, **28**, 213–223.

Huang A and Kim A (2019) Bayesian Conway-Maxwell-Poisson regression models for overdispersed and underdispersed counts. *Communications in Statistics—Theory and Methods*, **50**.

Huang A (2017) Mean-parametrized Conway-Maxwell-Poisson regression models for dispersed counts. *Statistical Modelling*, **17**, 359–380.

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

Lord D, Guikema SD and Geedipally SR (2008) Application of the Conway-Maxwell-Poisson generalized linear model for analyzing motor vehicle crashes. *Accident Analysis & Prevention*, **40**, 1123–1134.

Lord D, Geedipally SR and Guikema SD (2010) Extension of the application of Conway-Maxwell-Poisson models: Analyzing traffic crash data exhibiting underdispersion. *Risk Analysis: An International Journal*, **30**, 1268–1276.

Perera H, Davis J and Swartz TB (2016) Optimal lineups in Twenty20 cricket. *Journal of Statistical Computation and Simulation*, **86**, 2888–2900.

Plummer M, Best N, Cowles K and Vines K (2006) CODA: Convergence diagnosis and output analysis for MCMC. *R News*, **6**, 7–11. URL http://CRAN.R-project.org/doc/Rnews/.

Plummer M (2004). JAGS: Just another Gibbs sampler.

Pollard R, Benjamin B and Reep C (1977) Sport and the negative binomial distribution. In: *Optimal strategies in sports*, edited by Ladany SP and Machol RE, pages 188–195. North Holland: Amsterdam.

Pollard R and Pollard G (2005) Long-term trends in home advantage in professional team sports in North America and England (1876–2003). *Journal of Sports Sciences*, **23**, 337–350.

Preston I and Thomas J (2000). Batting strategy in limited overs cricket. *Journal of the Royal*

Statistical Society: Series D (The Statistician), **49**, 95–106.

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rathouz PJ and Gao L (2009) Generalized linear models with unspecified reference distribution. *Biostatistics*, **10**, 205–218.

Ribeiro Jr EE, Zeviani WM, Bonat WH, Demétrio CG and Hinde J (2018) Reparametrization of COM–Poisson regression models with applications in the analysis of experimental data. *Statistical Modelling*, 1471082X19838651.

Rohde N (2011) An economical ranking of batters in Test cricket. *Econ. Papers*, **30**, 455–465.

Scarf P, Shi X and Akhtar S (2011) On the distribution of runs scored and batting strategy in test cricket. *Journal of the Royal Statistical Society, Series A*, **174**, 471–497.

Sellers K, Lotze T and Raim A (2019) *COMPoissonReg: Conway-Maxwell Poisson (COM-Poisson) Regression*. URL https://CRAN.R-project.org/package=COMPoissonReg. R package version 0.7.0.

Sellers KF and Shmueli G (2010) A flexible regression model for count data. *The Annals of Applied Statistics*, **4**, 943–961.

Shmueli G, Minka TP, Kadane JB, Borle S and Boatwright P (2005) A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**, 127–142.

Stan Development Team (2020). *RStan: the R interface to Stan*. URL http://mc-stan.org/. R package version 2.19.3.

Stevenson OG and Brewer BJ (2021) Finding your feet: A Gaussian process model for estimating the abilities of batsmen in test cricket. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **70**, 481–506.

Swartz TB, Gill PS, Beaudoin D and DeSilva BM (2006) Optimal batting orders in one-day cricket. *Computers & Operations Research*, **33**, 1939–1950.

Wickham H (2016) *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. URL https://ggplot2.tidyverse.org.