Who is the greatest? A Bayesian analysis of Test match cricketers

Pete Philipson $^1$ & Richard Boys $^2$

$^1$Northumbria University

$^2$Newcastle University

New England Symposium on Statistics in Sports
September 26th, 2015
Overview

1. Motivation
2. Data and notation
3. Initial models
4. Modelling extensions
5. Results and analysis
6. Future directions
“Dealing with comparisons in cricket is harder and more complex than in most other sports... Averages can be a guide, but are not conclusive since pitches and conditions have changed.”

- Sir Donald Bradman, taken from *Harold Larwood* by Duncan Hamilton.
The Demon and The Little Master
Earliest work on the statistical modelling of cricket scores was undertaken by Elderton (1945) and Wood (1945) who considered a geometric distribution.

Pollard et al (1977) found runs scored by teams to follow the negative binomial distribution.

Scarf et al (2011) confirmed this finding.

Kimber and Hansford (1993) also considered the geometric distribution.

1. Little evidence against this model in the upper tail but rejected its validity for low scores, mainly due to the excess of ducks in the data.

2. Independence of cricket scores for a batsman: no major evidence of autocorrelation via a point process approach.
Aim of the analysis

To rank ‘all’ Test match batsmen & bowlers who have ever played Test match cricket, noting that
- different players have different abilities and mature and decline at different rates
- taking account of
  - the opposition they are playing against
  - whether it is a home or away fixture
  - the match innings
- overlapping careers form a bridge from present to past:
  - A plays against B, who plays against C, who plays against D, ...
- information is compromised by
  - law changes, depth of competition, technological advances, game conditions, game focus
The data and some notation

The focus here is on Test match cricket and the data consists of ≈ 65K innings for batsmen and ≈ 30K for bowlers encompassing
- 1022 batsmen and 578 bowlers
- 139 years from Test 1 in 1877 to Test 2179 in August 2015
- A Test match consists of up to four innings
- Axiom of cricket that battling last is difficult
- Many more Test matches are played today than at the time of the first Test:

<table>
<thead>
<tr>
<th>Test Match #</th>
<th>Year</th>
<th>Elapsed years</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1877</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>1908</td>
<td>31</td>
</tr>
<tr>
<td>500</td>
<td>1960</td>
<td>52</td>
</tr>
<tr>
<td>1000</td>
<td>1984</td>
<td>24</td>
</tr>
<tr>
<td>2000</td>
<td>2011</td>
<td>27</td>
</tr>
</tbody>
</table>
Data and notation (continued)

- **Inclusion criteria**
  1. Batsmen > 20 innings played
  2. Bowlers > 2000 balls bowled
  3. $17 < \text{Age} < 42$ due to scarcity of data outside this range

- **Batting data are innings by innings scores**

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Runs</th>
<th>Inns</th>
<th>Opposition</th>
<th>Ground</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A Bacher</td>
<td>4</td>
<td>1</td>
<td>Eng</td>
<td>Lord’s</td>
<td>23</td>
</tr>
<tr>
<td>2</td>
<td>A Bacher</td>
<td>37</td>
<td>3</td>
<td>Eng</td>
<td>Lord’s</td>
<td>23</td>
</tr>
</tbody>
</table>

- **Bowling data are aggregated by innings**

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Balls</th>
<th>Runs</th>
<th>Wkts</th>
<th>Inns</th>
<th>Opposition</th>
<th>Ground</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A Cotter</td>
<td>84</td>
<td>44</td>
<td>0</td>
<td>1</td>
<td>Eng</td>
<td>Sydney</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>A Cotter</td>
<td>111</td>
<td>41</td>
<td>3</td>
<td>3</td>
<td>Eng</td>
<td>Sydney</td>
<td>20</td>
</tr>
</tbody>
</table>
Some example data: Runs scored by Ian Bell against Australia between 2009 and 2015
Initial models

- Model runs scored per innings ($X_{ijk}$) for batsmen and wickets taken ($W_{ijk}$) by bowlers using the Poisson distribution:
  1. $X_{ijk} \mid \lambda_{ijk} \sim \text{Poisson}(\lambda_{ijk})$
  2. $W_{ijk} \mid \mu_{ijk} \sim \text{Poisson}(\mu_{ijk})$

- We assume (see Berry, Reese & Larkey, 1999)

\[
\lambda_{ijk} = \exp\{\theta_i + \delta_{y_{ijk}} + f_i(a_{ijk}) + \zeta_{h_{ijk}} + \xi_{m_{ijk}} + \omega_{o_{ijk}}\}
\]

and similarly for $\mu_{ijk}$ where

- $\theta_i$ the player-specific ability
- $\delta_{y_{ijk}}$ the year effect
- $f_i(a_{ijk})$ the player-specific ageing function
- $\zeta_{h_{ijk}}$ the playing away from home effect
- $\xi_{m_{ijk}}$ the (within-match) innings effect
- $\omega_{o_{ijk}}$ the opposition effect
Year effects

The year effects are a composite of several factors:

- **clearcut changes** such as depth of competition: more Test playing countries
- **game focus**: scoring rates are far higher in modern times and there are fewer draws
- **law changes** e.g. fewer bouncers per over allowed to make batting easier
- **other more subtle effects** e.g. technological advances and game conditions: most pitches are prepared so that they can last five days to ensure maximum profit
- There is **no regular annual fixture list** for international cricket – it operates on a broadly nine-year cycle
The ageing function for player $i$ takes the form:

$$f_i(a) = \begin{cases} 
\psi_{1i} \times \alpha_a & \text{if } a < a_M \\
\alpha_a & \text{if } a_D \leq a \leq a_M \\
\psi_{2i} \times \alpha_a & \text{if } a > a_D 
\end{cases}$$

where $a = \min(\text{age}), \ldots, \max(\text{age})$

- Maturing and declining parameters for each player are given by $\psi_{1i}$ and $\psi_{2i}$ respectively.
- The mean ageing function is piecewise constant, with $\alpha_a \equiv 0$ for the peak age.
Is equidispersion a reasonable assumption?

- The standard Poisson model imposes a strong assumption that the mean and variance of runs scored (or wickets taken) per innings are equal.

- By inspection (and intuition) the variability appears large relative to the mean.

Example (Runs scored by England batsman Graham Thorpe in 2000, aged 30):

0, 46, 40, 10, 118, 5, 79, 0, 18, 64*

where * indicates an incomplete ('not-out') innings - see later.
Modification of model for overdispersion

- Poisson model with random effects $v_{ijk}$ with $E(v_{ijk}) = 1$, i.e.
  $X_{ijk} \sim \text{Poisson}(\lambda_{ijk} \times v_{ijk})$

- Choice of distribution for the random effects
  1. Gamma distribution - i.e. $v_{ijk} \sim \text{Gamma}(\eta, \eta)$
  2. Lognormal distribution
  3. Inverse Gaussian
  4. Power transform - see Hougaard (1997)

- Allow for individual heterogeneity via $v_{ijk} \sim \text{Gamma}(\eta_i, \eta_i)$

- What to do with ‘not out’ innings for batsmen?
The ‘not out’ problem

- Affects the batting model only

Example (Runs scored by England batsman Graham Thorpe in 2000, aged 30)

0, 46, 40, 10, 118, 5, 79, 0, 18, 64*

- Occurs in approximately 11% of batted innings - see Kimber & Hansford (1993)
- Proper statistical treatment - censoring
The ‘not out’ problem

- Affects the batting model only

**Example (Runs scored by England batsman Graham Thorpe in 2000, aged 30)**

| 0, 46, 40, 10, 118, 5, 79, 0, 18, 64* |

- Occurs in approximately 11% of batted innings - see Kimber & Hansford (1993)
- Proper statistical treatment - censoring
- Standard treatment:
  
  \[
  \text{batsman’s average} = \frac{\text{total runs scored}}{\text{number of completed innings}}
  \]

- More coherent solution: treat ‘not out’ as a censored observation
Dealing with ducks: zero-inflation

- The possibility of excess zeroes (i.e. ducks) exists: Don Bradman has a modal score of zero despite averaging 99.94 runs per innings . . .

- Ducks account for $\approx 10\%$ of observations (excluding censored zeroes)

- Are all players equally vulnerable at the start of their innings?

- A zero-inflated model is adopted here

$$Pr(X_{ijk} = 0 | \cdot) = \pi_i + (1 - \pi_i) Pr(X_{ijk} = 0 | \lambda_{ijk}, \nu_{ijk})$$
Offset in the bowling model?

- The standard way of presenting a bowling average is on the same scale as that of a batting average, namely runs per wicket.

- Earlier model looks at wicket rate only - this is flawed. Need to factor in the number of runs conceded too.

- An offset model is adopted here

\[
\mu_{ijk} = \text{Runs}_{ijk} \times \exp\left\{\theta_i + \delta_{y_{ijk}} + f_i(a_{ijk}) + \zeta_{h_{ijk}} + \xi_{m_{ijk}} + \omega_{o_{ijk}}\right\}
\]

- We can then take \(e^{-\theta_i}\) as being on the same scale as the standard cricket bowling average

- The right modelling choice?
Identifiability issues

- Peak age taken as 30, i.e. $\alpha_{30} = 0$ with $a_M = 28$ and $a_D = 32$
- Most recent year taken as reference year $\delta_{139} = 0$
- Home effect $\zeta_1 = 0$, away effect $\zeta_2$
- Innings effect $\xi_1 = 0$: remaining innings effects $\xi_2, \xi_3, \xi_4$
- Opposition effect: order countries alphabetically $\omega_1 = 0$ – reference country is Australia
Identifiability issues

- Peak age taken as 30, i.e. $\alpha_{30} = 0$ with $a_M = 28$ and $a_D = 32$
- Most recent year taken as reference year $\delta_{139} = 0$
- Home effect $\zeta_1 = 0$, away effect $\zeta_2$
- Innings effect $\xi_1 = 0$: remaining innings effects $\xi_2, \xi_3, \xi_4$
- Opposition effect: order countries alphabetically $\omega_1 = 0$ – reference country is Australia

Thus $\exp(\theta_i)$ is the expected number of runs scored by player $i$ in the batting model and $\exp(-\theta_i)$ is the runs per wicket taken by player $i$

- when playing at his ‘peak’ age
- at home against Australia
- in the first innings of a Test match
- taking place in 2015
The following specifications for the prior distributions were used:

- Use a simple (backwards) random walk to smooth year effects (with $\delta_{139} = 0$)
  
  $\delta_\ell = \delta_{\ell+1} + N(0, \sigma^2_\delta), \quad \ell = 1, \ldots, 138$

  $\sigma^2_\delta \sim IGa(2, 0.01)$

- Second order random walks to smooth ageing effects
  
  $\alpha^*_a = 2\alpha^*_a + 1 - \alpha^*_{a+2} + N(0, \sigma^2_\alpha), \quad a = 17, \ldots, 42$

  $\sigma^2_\alpha \sim IGa(2, 0.01)$
Computational algorithms

- Code written in R
- Metropolis-within-Gibbs sampling scheme
- Discarded 5K iterations as burn-in
- Ran chain for further 250K iterations, thinning by 50, to get 5K (almost) independent draws
### Summary table: Batsmen

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>E(Runs)</th>
<th>SD(Runs)</th>
<th>E(π)</th>
<th>Median rank (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DG Bradman</td>
<td>98.72</td>
<td>16.49</td>
<td>0.07</td>
<td>1 (1-5)</td>
</tr>
<tr>
<td>2</td>
<td>RG Pollock</td>
<td>70.62</td>
<td>14.57</td>
<td>0.03</td>
<td>8 (1-135)</td>
</tr>
<tr>
<td>3</td>
<td>JB Hobbs</td>
<td>69.39</td>
<td>10.83</td>
<td>0.05</td>
<td>8 (2-84)</td>
</tr>
<tr>
<td>4</td>
<td>GS Sobers</td>
<td>66.80</td>
<td>8.16</td>
<td>0.06</td>
<td>11 (2-62)</td>
</tr>
<tr>
<td>5</td>
<td>GA Headley</td>
<td>64.67</td>
<td>14.07</td>
<td>0.04</td>
<td>16 (1-211)</td>
</tr>
<tr>
<td>6</td>
<td>KF Barrington</td>
<td>62.92</td>
<td>7.90</td>
<td>0.02</td>
<td>19 (3-93)</td>
</tr>
<tr>
<td>7</td>
<td>SR Tendulkar</td>
<td>61.69</td>
<td>5.33</td>
<td>0.01</td>
<td>22 (5-73)</td>
</tr>
<tr>
<td>8</td>
<td>WR Hammond</td>
<td>61.56</td>
<td>8.72</td>
<td>0.02</td>
<td>23 (3-133)</td>
</tr>
<tr>
<td>9</td>
<td>SM Gavaskar</td>
<td>61.11</td>
<td>6.51</td>
<td>0.02</td>
<td>24 (4-90)</td>
</tr>
<tr>
<td>10</td>
<td>GS Chappell</td>
<td>60.61</td>
<td>7.25</td>
<td>0.06</td>
<td>26 (4-108)</td>
</tr>
<tr>
<td>11</td>
<td>H Sutcliffe</td>
<td>60.24</td>
<td>8.79</td>
<td>0.03</td>
<td>27.5 (3-165)</td>
</tr>
<tr>
<td>12</td>
<td>JE Root</td>
<td>59.37</td>
<td>10.19</td>
<td>0.02</td>
<td>34 (2-219)</td>
</tr>
<tr>
<td>13</td>
<td>JH Kallis</td>
<td>59.31</td>
<td>5.29</td>
<td>0.03</td>
<td>30 (6-100)</td>
</tr>
<tr>
<td>14</td>
<td>SR Waugh</td>
<td>58.91</td>
<td>5.98</td>
<td>0.04</td>
<td>33 (5-109)</td>
</tr>
<tr>
<td>15</td>
<td>IVA Richards</td>
<td>58.89</td>
<td>6.52</td>
<td>0.02</td>
<td>33 (5-116)</td>
</tr>
<tr>
<td>16</td>
<td>Javed Miandad</td>
<td>58.54</td>
<td>6.09</td>
<td>0.01</td>
<td>34 (6-114)</td>
</tr>
<tr>
<td>17</td>
<td>BC Lara</td>
<td>58.34</td>
<td>5.56</td>
<td>0.03</td>
<td>35 (6-113)</td>
</tr>
<tr>
<td>18</td>
<td>L Hutton</td>
<td>58.27</td>
<td>7.53</td>
<td>0.02</td>
<td>36 (5-148)</td>
</tr>
<tr>
<td>19</td>
<td>ED Weekes</td>
<td>58.12</td>
<td>8.65</td>
<td>0.05</td>
<td>39 (4-181)</td>
</tr>
<tr>
<td>20</td>
<td>AR Border</td>
<td>57.94</td>
<td>5.63</td>
<td>0.02</td>
<td>37 (8-113)</td>
</tr>
</tbody>
</table>
Results - Batting boxplots and ranking confidence intervals
Results - Game specific effects for batting

Motivation
Data and notation
Initial models
Modelling extensions
Results and analysis
Future directions

Pete Philipson & Richard Boys
NESSIS: Who is the greatest?
Results - overdispersion and zero-inflation

- Posterior mean for $\eta = 0.88$ with IQR (0.76, 1.03)
- Posterior mean for $\pi = 0.06$ and IQR (0.04, 0.08)
- Evidence that overdispersion and zero-inflation are present - Geometric distribution not appropriate for majority of players
- Comparison with other rankings
  1. Highest career batting average
  2. International Cricket Council rankings
Results - Batting ageing function
Results - Individual player profiles

- SR Waugh
- SR Tendulkar
- Javed Miandad
- KF Barrington
- RG Pollock
- JB Hobbs

Age

Runs

NESSIS: Who is the greatest?
### Summary table: Bowlers

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Player ability</th>
<th>Median rank (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GA Lohmann</td>
<td>16.38</td>
<td>1 (1-30)</td>
</tr>
<tr>
<td>2</td>
<td>SF Barnes</td>
<td>19.98</td>
<td>7 (1-83)</td>
</tr>
<tr>
<td>3</td>
<td>MD Marshall</td>
<td>20.50</td>
<td>10 (2-38)</td>
</tr>
<tr>
<td>4</td>
<td>CEL Ambrose</td>
<td>21.15</td>
<td>13 (3-48)</td>
</tr>
<tr>
<td>5</td>
<td>FH Tyson</td>
<td>21.18</td>
<td>12 (1-101)</td>
</tr>
<tr>
<td>6</td>
<td>Sir RJ Hadlee</td>
<td>21.64</td>
<td>16 (3-56)</td>
</tr>
<tr>
<td>7</td>
<td>J Garner</td>
<td>21.98</td>
<td>18 (3-63)</td>
</tr>
<tr>
<td>8</td>
<td>R Peel</td>
<td>22.14</td>
<td>15 (2-229)</td>
</tr>
<tr>
<td>9</td>
<td>JJ Ferris</td>
<td>22.14</td>
<td>15 (1-248)</td>
</tr>
<tr>
<td>10</td>
<td>ERH Toshack</td>
<td>22.20</td>
<td>18 (1-206)</td>
</tr>
<tr>
<td>11</td>
<td>J Briggs</td>
<td>22.42</td>
<td>19 (2-196)</td>
</tr>
<tr>
<td>12</td>
<td>Imran Khan</td>
<td>22.53</td>
<td>22 (5-71)</td>
</tr>
<tr>
<td>13</td>
<td>W Barnes</td>
<td>23.14</td>
<td>22 (1-317)</td>
</tr>
<tr>
<td>14</td>
<td>GD McGrath</td>
<td>23.20</td>
<td>29 (7-79)</td>
</tr>
<tr>
<td>15</td>
<td>MA Holding</td>
<td>23.28</td>
<td>30 (6-95)</td>
</tr>
<tr>
<td>16</td>
<td>AA Donald</td>
<td>23.33</td>
<td>30 (6-88)</td>
</tr>
<tr>
<td>17</td>
<td>M Muralitharan</td>
<td>23.43</td>
<td>31 (9-80)</td>
</tr>
<tr>
<td>18</td>
<td>AK Davidson</td>
<td>23.53</td>
<td>30 (5-120)</td>
</tr>
<tr>
<td>19</td>
<td>DW Steyn</td>
<td>23.56</td>
<td>32 (7-102)</td>
</tr>
<tr>
<td>20</td>
<td>J Cowie</td>
<td>23.79</td>
<td>31 (1-273)</td>
</tr>
</tbody>
</table>
Results - Bowling boxplots and ranking confidence intervals
Results - Game specific effects for bowling

Motivation
Data and notation
Initial models
Modelling extensions
Results and analysis
Future directions

Pete Philipson & Richard Boys
NESSIS: Who is the greatest?
Posterior mean for $\eta = 4.21$ with 95% interval (3.97, 4.47)

No zero-inflation in bowling model

Evidence of overdispersion - much less extreme than in batting model $\rightarrow$ bowling data are aggregated

Comparison with other rankings

1. Lowest career bowling average
2. International Cricket Council rankings
Results - Bowling year effects

![Graph showing year effects over time]

Pete Philipson & Richard Boys  NESSIS: Who is the greatest?
Results - Bowling ageing function
Model checking: Zero-inflation

Number of ducks

Probability

7000 7200 7400 7600

0.000 0.001 0.002 0.003 0.004
Future directions

- Covariate rather than offset model for bowling
- Consider alternative ageing functions
- Comparing alternative models and goodness-of-fit
- Incorporate other data, i.e. ground, batting position, handedness?
- Application to one day international (ODI) data
- Twenty20 cricket?
References

Berry, S., Reese, S. and Larkey, P.
Bridging Different Eras in Sports
JASA, 1999.

Kimber, A. C. and Hansford, A. R.
A Statistical Analysis of Batting in Cricket
JRSS Series A, 1993

Hougaard, P., Ting Lee, M-L. and Whitmore, A.
Analysis of Overdispersed Count Data by Mixtures of Poisson Variables and Poisson Processes