The Relationship between Ball Speed and Anthropometrical Characteristics among Professional Baseball Pitchers: a Hybrid Evolutionary Algorithm Approach

Valandro L¹, Colombo L¹, Cao H², Recknagel F², Dun S³, Secco EL⁴

¹Department of Biology, University of Padua, Italy – ²Department of Earth & Environmental Sciences, University of Adelaide, South Australia – ³American Sports Medicine Institute, Birmingham, Alabama – ⁴Department of Computer Engineering & Systems Science, University of Pavia, Italy

Abstract

The fastest bowlers and pitchers, who are able to perform at levels that exceed those of their team-mates, should have a combination of optimal physical characteristics and techniques. The relationship between ball speed and anthropometrical characteristics of 48 professional baseball pitchers was investigated using a hybrid evolutionary algorithm (HEA). The average height and weight of the subjects were 191.4±4.7 cm and 96.2±8.2 kg, respectively. Their mean brachial index, BI, was 74.9±4.1. Performing our HEA analysis led us to discover the best rule sets between different inputs (height, weight, segment lengths) resulting in models relating ball speed to a pseudo-ectomorphic parameter. The robustness of the results was achieved by implementing a parsimonious strategy based on the Akaike criterion as the fitness function. Furthermore, the rule set resulted to be almost always weight greater than constant and after selection of the best pitchers (the lighter ones) the related distribution of speed as a function of BI was shown to be bimodal. These results could be used for talent identification and coaching technique improvement. Our finding could also have some relevance for bio-inspired robotics.

KEY WORDS: Anthropometry, Brachial Indexes, Maximum Ball Velocity, Genetic Algorithm, Akaike Criterion, Rule Discovery

Introduction

Fast bowlers and pitchers, who are able to throw the ball faster than their team-mates, should have a combination of optimal physical characteristics and techniques (Carda et al., 1994; Fleisig et al., 1996; Glazier et al., 2000; Valandro et al., 2005). Great values of height and body weight have been reported to be determinants of bowling speed in cricket (Pyne et al., 2005) and throwing sport performance in other sport disciplines (Sidhu et al., 1975; Matsuo et al., 2001).

In this work, we address the problem of predicting ball speed from anthropometric characteristics and the eligibility of rule discovering by a hybrid evolutionary algorithm (HEA). HEA uses genetic programming (GP) to generate and optimize the structure of
rule sets and a genetic algorithm (GA) to optimize the parameters of a rule set. Rules discovered by HEA have the IF-THEN-ELSE structure and allow embedding of complex functions in this work synthesized only from simple arithmetic operators (+, −, *, /). The maximum tree depth and rule set size, together with a parsimonious criterion (Akaike), control the complexity of rule sets (Cao et al., 2006).

From this preliminary work, both talent identification and coaching techniques could benefit as well as biomechanics in general, once other variables (e.g., angles) are included in the rule set discovery. Furthermore, these findings could be of some relevance for bio-inspired robotics.

**Materials and Methods**

*Subjects and Data acquisition*

Forty-eight professional baseball pitchers participated in this study. All participants had no previous history of neuro-musculo-skeletal disorders at the time of testing. Prior to the testing, each subject signed a consent form approved by the Institutional Review Board of the American Sports Medicine Institute. The humeral and radial lengths of each subject's throwing arm were measured prior to biomechanical analysis. The humeral length was measured from the acromion process to the lateral epicondyle and the radial length was measured from the humeral epicondyle to the radial styloid process. The body weight and height were self-reported by each subject (Table 1).

<table>
<thead>
<tr>
<th>Height [cm]</th>
<th>Weight [kg]</th>
<th>Arm [cm]</th>
<th>Forearm [cm]</th>
<th>Age [years]</th>
</tr>
</thead>
<tbody>
<tr>
<td>191.4 ± 4.7</td>
<td>96.2 ± 8.2</td>
<td>38.73 ± 2.03</td>
<td>28.96 ± 1.54</td>
<td>23.9 ± 3.5</td>
</tr>
</tbody>
</table>

Table 1. Average anthropometric characteristics and age of professional baseball pitchers (mean±SD).

Three different brachial indexes, expressed in terms of length ratios, were compared:

1) The brachial index is defined as:

\[
\text{BI} = \frac{\text{forearm}}{\text{arm}} \cdot 100
\]

2) Another index, derived from *De architectura* by Marcus Vitruvius Pollio (27/23 B.C.) and from Leonardo Da Vinci’s *Homo Vitruvianus* (1490), is defined as:

\[
\text{BI}_{LV} = \frac{\text{arm} + \text{forearm} + \text{hand}}{\text{forearm} + \text{hand}}
\]

3) An efficacious brachial index is defined as:
BI* = \frac{(\text{forearm} + \text{hand}/2)}{\text{arm}}

Hand lengths were derived from allometry.

Before the testing, each pitcher was given ample time for his normal warm-up routine. When the subject was ready, data were collected with the subject pitching off an ATEC (Athletic Training Equipment Company, 18 Santa Cruz, AZ) indoor pitching mound towards a strike zone target located above home plate, 18.4 meters away from the pitching rubber. Each subject was instructed to pitch ten fastballs with maximal effort. The subject was given 30-60 seconds of rest time before each pitch.

Ball speed was measured in an indoor laboratory with a Jugs radar gun (Jugs Pitching Machine Company, Tualatin, OR) from behind the home plate.

*Hybrid Evolutionary Algorithm*

*HEA* has been designed in order to discover predictive rule sets. It firstly evolves the structure of the rule sets by using GP, and secondly optimizes the random parameters in the rule set by using a GA. Rules discovered by *HEA* have the IF-THEN-ELSE structure and allow embedding of complex functions synthesized from various predefined arithmetic operators. The main framework of *HEA* for the rule discovery in the studied data is represented in Figure 1. Cao et al. (2006) provided a detailed description of *HEA* including the design of the genetic operators in GP and GA.

![Figure 1. Conceptual diagram of the Hybrid Evolutionary Algorithm for the discovery of predictive rule sets in professional baseball pitchers.](image-url)
In our analysis, we used the first 24 data points of the measured anthropometric data for training and the remaining 24 data points for testing the generalization behavior of the resulting rule sets. Several analyses of 300 runs were conducted independently. The maximal rule set size was set to be four. All the experiments were performed on a Hydra supercomputer (IBM eServer 1350 Linux) with a peak speed of 1.2 TFlops by using C++ programming language.

The HEA was implemented by using anthropometric data as inputs and the measured or observed ball speed (ball speed$_{obs}$) and the predicted ball speed (ball speed$_{pre}$) as outputs (Table 2a-b). We started with simple and easily measurable anthropometric variables (Martin et al., 1988).

<table>
<thead>
<tr>
<th>Height [m]</th>
<th>Weight [kg]</th>
<th>Forearm length [m]</th>
<th>Arm length [m]</th>
<th>Ball Speed$_{obs}$ [m/s]</th>
<th>Ball Speed$_{pre}$ [m/s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_1$</td>
<td>$W_1$</td>
<td>$F_1$</td>
<td>$A_1$</td>
<td>$s_1$</td>
<td>$S_1$</td>
</tr>
<tr>
<td>$H_2$</td>
<td>$W_2$</td>
<td>$F_2$</td>
<td>$A_2$</td>
<td>$s_2$</td>
<td>$S_2$</td>
</tr>
<tr>
<td>$H_3$</td>
<td>$W_3$</td>
<td>$F_3$</td>
<td>$A_3$</td>
<td>$s_3$</td>
<td>$S_3$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>$H_N$</td>
<td>$W_N$</td>
<td>$F_N$</td>
<td>$a_N$</td>
<td>$s_N$</td>
<td>$S_N$</td>
</tr>
</tbody>
</table>

Table 2a. Input variables used in this study were height, weight, forearm length and arm length (4 inputs). The prediction of ball speed was performed by means of HEA in order to discover rules between variables. Ball speed observed was used to calculate the fitness function.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Units</th>
<th>Mean±SD</th>
<th>Min/Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>[m]</td>
<td>1.914 ± .047</td>
<td>1.80 – 2.01</td>
</tr>
<tr>
<td>Height</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_2$</td>
<td>[kg]</td>
<td>96.2 ± 8.2</td>
<td>72.0 – 112.5</td>
</tr>
<tr>
<td>Weight</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_3$</td>
<td>[pure number]</td>
<td>74.9±4.1</td>
<td>65.0 – 86.1</td>
</tr>
<tr>
<td>BI</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2b. Input and Output variables (3 inputs) used in the evolutionary modeling of rule sets.

The inputs are randomly combined by using the following simple operators $+$, $-$, $\times$, $/$ and the best fitting models are selected on the base of number of variables, parameters and error (see Results). Figure 2 outlines the flowchart of the detailed algorithm for the rule discovery.
Figure 2 - Flowchart of the Hybrid Evolutionary Algorithm.

**Fitness Function**

**HEA** was implemented in C++ by using anthropometric data as inputs and ball speed_{obs.} and ball speed_{pre.} as outputs.

Furthermore, the following corrected Akaike Information Criterion (**AIC_c**) was introduced in order to reduce overall model complexity (de Leeuw, 1992):

\[
AIC_c = AIC + \frac{2k \cdot (k + 1)}{(n - k - 1)}
\]

where **AIC** contains both the error and the number of parameters **k**, while **n** is the sample size.

This allowed us to compare pairs of models (THEN and ELSE) derived from each supercomputer run.

The following fitness function was used:
\[ AIC = AIC_{\text{THEN}} + AIC_{\text{ELSE}} \]

The AIC values used in the fitness function were then compared with other measures of goodness of fit, such as the coefficient of correlation (R) and the Root Mean Square Error (RMSE).

**Results**

The analysis with HEA allowed us to identify the most relevant variables to predict both average and maximum ball speeds. Weight was the most selected variable.

Here, we do not present any model including arm length and arm proportions because the solutions of HEA were not robust. The comparison of the three BI after analysis with HEA showed that the efficacious brachial index was much more frequent than the classical BI and superior to the BI \( L-V \) (see Materials and Methods).

Fitting of average and maximum ball speeds with the THEN and ELSE models produced a trivial solution (4 inputs; best AIC):

\[
\begin{align*}
\text{IF} & \quad \text{weight} < \theta_M \\
\text{THEN} & \quad S_1, S_3, \ldots, S_N = K_f \\
\text{ELSE} & \quad S_2, S_6, \ldots, S_{N-1} = K_2 \\
K_2 & < K_1
\end{align*}
\]

Where \( \theta_M \) is a weight threshold, \( S_i \) is the pitcher’s ball speed and \( K_1, K_2 \) are constants.

After performing 300 more runs (using 3 inputs), three pseudo-ectomorphic models were obtained (Figure 3). Being most subjects fitted by only one of the models (THEN and ELSE), the rule discovery tool becomes an equation discovery one (Table 3).
Figure 3. \( R^2 \)-values for linear correlation between ball speed observed and predicted by model with best AIC (model a), 8th best AIC (model b) and 3rd best AIC (model c) (training data, N=23).

![Graph showing observed vs predicted ball speed with linear models and \( R^2 \) values]

Table 3. Best ectomorphic models fitting the data. \( F(x) \) is ball speed (mph), \( x1 \) is height (cm) and \( x2 \) represents weight (kg).

<table>
<thead>
<tr>
<th>Model</th>
<th>Equation</th>
<th>AIC Rank</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>( y = 0.491x + 19.326 )</td>
<td>( R^2 = 0.516 )</td>
<td>1 constant</td>
</tr>
<tr>
<td>b</td>
<td>( y = 0.599x + 15.323 )</td>
<td>( R^2 = 0.512 )</td>
<td>1 constant</td>
</tr>
<tr>
<td>c</td>
<td>( y = 0.568x + 16.427 )</td>
<td>( R^2 = 0.520 )</td>
<td>1 constant</td>
</tr>
</tbody>
</table>

The testing has been extended to the remaining unseen data (data not seen from the algorithm) (N=17) for all three models (Figure 4a). Seven subjects with extreme body mass index (BMI) values, were excluded as outliers.
### Table

<table>
<thead>
<tr>
<th>Model</th>
<th>Observed Speed (ms(^{-1}))</th>
<th>Predicted Speed (ms(^{-1}))</th>
<th>Equation</th>
<th>R(^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>34, 38, 42</td>
<td>34, 38, 42</td>
<td>y = 0.371x + 23.165</td>
<td>0.165</td>
</tr>
<tr>
<td>b</td>
<td></td>
<td></td>
<td>y = 0.445x + 20.249</td>
<td>0.169</td>
</tr>
<tr>
<td>c</td>
<td></td>
<td></td>
<td>y = 0.437x + 20.609</td>
<td>0.160</td>
</tr>
</tbody>
</table>

### Figure 4a

**R\(^2\)-values for linear correlation between ball speed observed and predicted for the three pseudo-ectomorphic models (unseen testing data, N = 17).**

### Figure 4b

**R\(^2\)-values for linear correlation between ball speed observed and predicted for the three pseudo-ectomorphic models (unseen testing data, N = 7).**
A further testing of the generalization behavior of the three models has been performed on 10 professional baseball pitchers whose data were collected with the same protocol (Figure 4b). Again 30% of the athletes were excluded from the analysis following the same criteria.

The maximum ball speed predicted by the three models (weight 72 kg; height 183 cm) is always less than 100 mph, the model c being the closest (97.32 mph = 43.51 m/s = 156.64 km/h). Model b improves the coefficient of correlation in both the testing procedures, both in the remaining athletes (N=17) as well as in the post-hoc collected ones (N=7).

### Table 4a. Coefficient of correlation, R, for observed versus predicted ball speed (data not fitted by most models in previous runs and for which the difference between ball speed observed and predicted was greater than 2.5 m/s were considered outliers).

<table>
<thead>
<tr>
<th>Model</th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>0.718</td>
<td>0.721</td>
<td>0.715</td>
</tr>
</tbody>
</table>

### Table 4b. Coefficient of correlation, R, for observed versus predicted ball speed (data for which the difference between ball speed observed and predicted was greater than 2.5 m/s were considered outliers).

<table>
<thead>
<tr>
<th>Model</th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>0.401</td>
<td>0.411</td>
<td>0.400</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>a</th>
<th>b</th>
<th>c</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>0.684</td>
<td>0.699</td>
<td>0.666</td>
</tr>
</tbody>
</table>
Bimodality of Ball Speed as a Function of BI

The correlation of BI with ball speed (Figure 5) was not significant (p = 0.89), while after selection of lighter subjects throwing faster, a bimodal distribution of speed appeared, but the linear correlation was still obviously not significant (p = 0.98) (Figure 6).
Discussion

Increases in height and weight have been observed and associated in elite baseball athletes to improved performance (Silver, 2005: see http://www.baseballprospectus.com/article.php?articleid=3979).

In this study, high values of height and inversely small values of body weight were reported to be determinants of throwing sport performance contrarily to Pyne et al (2005) for cricket. His finding could be reconciled with our data considering a parabolic function of ball speed as a function of body weight.

Since the ectomorphic model derived from baseball pitchers was also used for predicting speed in fast-medium bowlers using anthropometric data from Glazier et al. (2000) (in preparation), we propose that, although Pyne and co-authors’ work has relevant practical implications, it could be misleading when indicating weight as a main variable in the senior-junior comparison. Matsuo et al. (2001) found that a faster throwing group of pitchers was significantly taller than the slower throwing group. In addition, subjects in the taller group have longer arms, as discussed by Garda et al. (1994). According to their somatotype analysis, future work should investigate not only the relevance of weight, but also the distribution of masses for speed prediction.

Analyzing an exiguous sample of front-on fast-medium bowlers, subjects displaying more ectomorphic characteristics were faster (Valandro et al, 2005). In this work, the HEA discovered simple models with good prediction power. Although Matsuo et al. (2001) found a non-significant difference between the weights of faster and slower
In our work weight was the most commonly selected parameter to derive the two sets of points to be fitted with the THEN and ELSE models. HEA offered a rough simplistic solution, although not highlighting any relevant variables, where the fitting of pitchers’ speed was possible with merely a constant $K_1 = 83.79$ for the heavier group and a constant $K_2 = 87.39$ for the lighter group. It nicely explains how the lighter athletes were able to throw faster (THEN MODEL) than those athletes weighing above a certain threshold ($\theta_M = 92$ kg) (ELSE MODEL).

The model (a) with the best AIC was reported here because it could fit the data with only one constant. However the most reliable model is the pseudo-ectomorphic one (b), which allows the estimation of the maximum ball speed theoretically possible and improves the error after generalization to all professional pitchers (8th best AIC). The maximum ball speed predicted using a formula containing only height and weight variables was $94.49 \text{ mph} = 42.25 \text{ m/s} = 152.09 \text{ km/h}$, in any case less than 100 mph even when using the lightest subject (72 kg; 183 cm).

The three models presented were not able to predict all points, leaving space to biomechanical interpretation (e.g., Fleisig et al., 1996). It should be noted, however, that the subjects not fitted by the three models were consistently the same, indicating the robustness of the method. Indeed, the introduction of the corrected $AIC_c$, that takes into account the small sample size, led us to select the most parsimonious models increasing the robustness of HEA (Akaike, 1973; Sugiura, 1978; see also de Leeuw, 1992 for a review). AIC trends were almost always comparable to those of coefficient of correlation and other error functions, meaning that each run with good AIC was associated with small error functions.

The HEA approach is an adaptive method, which mimics processes of biological evolution, genetic variation and natural selection and searches for suitable representations of a problem solution by means of genetic operators and the principle of survival of the fittest (see Holland, 1975; Goldberg, 1989; Valandro et al., 2000; Cao et al, 2006). It seems very promising and will soon be applied in the discovery of a more refined and general model and in the evaluation and guide of coaching techniques.

Here, we did not consider models containing arm length and body proportions because HEA results were not robust. However, the analysis with HEA showed that the efficacious brachial index was much more frequent than the classical BI and superior even to the $BI_{LV}$ index.

The possibility of a bimodal distribution of speed as a function of BI after selecting a sub-sample as indicated by the HEA (Rule set, weight<$w_2$) is of great interest and could be explained by the minimum jerk theory (Secco et al., 2005). The greater usability, i.e. the amount of all the admissible reachable positions (Caimmi et al, 2005) of the subjects with $BI^* = 1$, could imply a less efficient movement control by the brain which in turn would yield smaller ball velocity with respect to a slightly lower usability but with more efficient movement control and then higher ball velocity. This bimodal finding was also observed in the frequency of brachial indexes of Olympic rowing athletes, both males and females, indicating that this is possibly quite a general phenomenon (Valandro et al., unpublished).
Future work will try to interpret such kinanthropometric result (bimodality) by means of biomechanical analysis (Bartlett et al., 1996; Escamilla et al., 2001).

Acknowledgements. We would like to acknowledge Prof. Giovanni Magenes and Benedetto Salvato for discussion and for supervising this work. Thanks to Francesca De Conti for carefully reading through the manuscript. Thanks to Dr. Roberto Caimmi for helpful discussion and careful reading through the manuscript. Last, but not least we would like to acknowledge Dr. Glenn Fleisig, the Director of the American Sport Medicine Institute for providing the data and for helpful discussion.
References


