# Clustering Batted Balls to Project Next-Year wOBA

Tad Berkery ‘24, Justin Nam ‘24

Project Mentors: Mr. Sig Mejdal, Professor Anton Dahbura

Johns Hopkins University | Whiting School of Engineering | Baltimore, MD

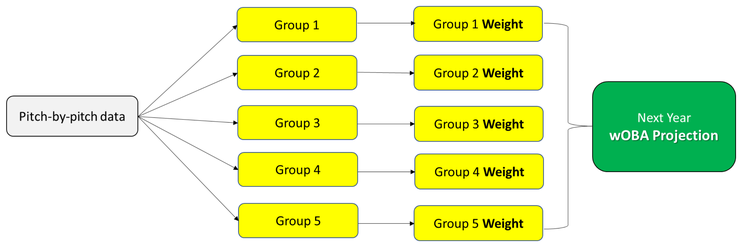
Design Day 2022

## Introduction

Baseball has a well-established accounting system for batted balls in singles, doubles, triples, and homeruns that is wonderfully easy to see and intuitive. However, while reasonable at the box score level and casual fan level, this accounting system doesn’t maximize predictability in a year-to-year context. In the context of the immense value of accurately projecting future performance, is it possible to design an accounting system that groups batted balls in a predictive manner? In an analytical sense, can we group batted balls into clusters and assign a weight to the frequency of batted balls in each category as a means for projecting a given batter’s future weighted on-base average (wOBA)?

## Objectives, Materials, & Methodology

Our core objective is to define predictive clusters for batted balls and identify weights to apply to each cluster to form a projection for next-year wOBA. This was accomplished by utilizing K-means clustering to identify similar groups of batted balls and then using linear regression to properly weight each group in forming a wOBA projection for the following year. This process was repeated for numerous combinations of batted ball metrics, and we kept track of which combinations and associated weights formed the strongest projections. Our data set was created using data from *Baseball Savant.*



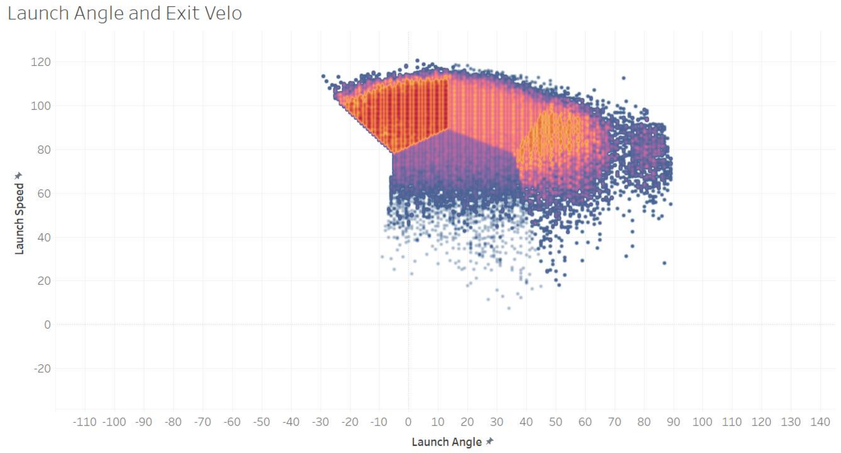

## Results

### Overall Model Results and Effectiveness

For the 2019 season, our model was able to explain 63% of the variation in weighted on-base average when adjustments for number of plate appearances and sample size were implemented. This is a strong result given that using 2018 wOBA on its own would only explain 38% of the variation in 2019 wOBA.

## Results

The optimal cluster was shaped by four variables: horizontal spin of the pitch, vertical spin of the pitch, launch angle, and exit velocity. The impact of launch angle and exit velocity on the cluster into which a batted ball is clustered is evident, with colors differentiating clusters.




The weights for our five clusters are illustrated as follows:

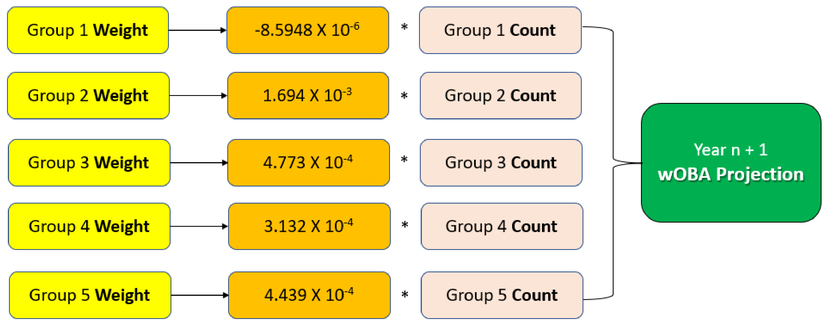



**Table 1—Most Accurate Projections**

| Player | Model Projection | 2019 PA | Real 2019 xwOBA | 2020 PA | Real 2020 xwOBA | Model vs. Actual |
|---|---|---|---|---|---|---|
| Josh VanMeter | 0.289 | 260 | 0.319 | 79 | 0.289 | 0.000 |
| José Iglesias | 0.384 | 530 | 0.273 | 150 | 0.384 | 0.000 |
| Robbie Grossman | 0.326 | 482 | 0.324 | 192 | 0.325 | 0.001 |
| Josh Bell | 0.296 | 613 | 0.379 | 223 | 0.297 | 0.001 |
| Adalberto Mondesi | 0.265 | 443 | 0.278 | 233 | 0.266 | 0.001 |
| Nick Senzel | 0.292 | 414 | 0.294 | 78 | 0.291 | 0.001 |
| Adam Eaton | 0.298 | 656 | 0.321 | 176 | 0.299 | 0.001 |
| Tommy La Stella | 0.347 | 321 | 0.347 | 228 | 0.346 | 0.001 |
| Nick Ahmed | 0.304 | 625 | 0.320 | 217 | 0.302 | 0.002 |
| Paul DeJong | 0.312 | 664 | 0.324 | 174 | 0.315 | 0.003 |
| Rio Ruiz | 0.280 | 413 | 0.291 | 204 | 0.276 | 0.004 |
| J.D. Martinez | 0.312 | 657 | 0.389 | 237 | 0.316 | 0.004 |
| Mike Zunino | 0.258 | 289 | 0.273 | 84 | 0.263 | 0.005 |
| Nick Markakis | 0.303 | 469 | 0.349 | 141 | 0.308 | 0.005 |
| Freddy Galvis | 0.304 | 589 | 0.276 | 159 | 0.299 | 0.005 |

**Table 2—Least Accurate Projections**

| Player | Model Projection | 2019 PA | Actual 2019 xwOBA | 2020 PA | Actual 2020 xwOBA | Model vs. Actual |
|---|---|---|---|---|---|---|
| Juan Soto | 0.259 | 659 | 0.407 | 196 | 0.474 | 0.215 |
| Bryce Harper | 0.277 | 682 | 0.383 | 244 | 0.453 | 0.176 |
| Fernando Tatis Jr. | 0.248 | 372 | 0.344 | 257 | 0.419 | 0.171 |
| Hanser Alberto | 0.411 | 550 | 0.292 | 231 | 0.259 | -0.152 |
| Teoscar Hernández | 0.252 | 464 | 0.311 | 207 | 0.396 | 0.144 |
| Marcell Ozuna | 0.300 | 549 | 0.368 | 267 | 0.435 | 0.135 |
| Luke Voit | 0.264 | 510 | 0.360 | 234 | 0.388 | 0.124 |
| Brandon Lowe | 0.259 | 327 | 0.338 | 224 | 0.382 | 0.123 |
| Ronald Acuña Jr. | 0.298 | 715 | 0.390 | 202 | 0.421 | 0.123 |
| George Springer | 0.277 | 556 | 0.390 | 222 | 0.397 | 0.120 |
| Daniel Murphy | 0.387 | 478 | 0.289 | 132 | 0.269 | -0.118 |
| Corey Seager | 0.311 | 541 | 0.327 | 232 | 0.427 | 0.116 |
| Mike Trout | 0.308 | 600 | 0.451 | 241 | 0.421 | 0.113 |
| Rhys Hoskins | 0.267 | 705 | 0.337 | 185 | 0.378 | 0.111 |
| Paul Goldschmidt | 0.287 | 682 | 0.354 | 231 | 0.394 | 0.107 |

**Table 3—Best Projections of Significant Performance Changes**

| Player | Model Projection | 2019 PA | Actual 2019 xwOBA | 2020 PA | Real 2020 xwOBA | Model vs. Actual | Real Performance Change |
|---|---|---|---|---|---|---|---|
| Josh Bell | 0.296 | 613 | 0.379 | 223 | 0.297 | 0.001 | -0.082 |
| Victor Caratini | 0.293 | 279 | 0.336 | 132 | 0.284 | 0.009 | -0.052 |
| Javier Báez | 0.283 | 561 | 0.327 | 235 | 0.274 | 0.009 | -0.053 |
| Bryan Reynolds | 0.319 | 546 | 0.353 | 208 | 0.310 | 0.009 | -0.043 |
| Danny Santana | 0.292 | 511 | 0.342 | 63 | 0.281 | 0.011 | -0.061 |
| José Martínez | 0.300 | 373 | 0.346 | 98 | 0.288 | 0.012 | -0.058 |
| Mitch Garver | 0.251 | 359 | 0.373 | 81 | 0.239 | 0.012 | -0.134 |
| Tim Anderson | 0.342 | 518 | 0.319 | 221 | 0.357 | 0.015 | 0.038 |
| Edwin Encarnación | 0.291 | 486 | 0.363 | 181 | 0.275 | 0.016 | -0.088 |
| Brian Dozier | 0.280 | 482 | 0.329 | 16 | 0.263 | 0.017 | -0.066 |
| Yoán Moncada | 0.271 | 559 | 0.354 | 231 | 0.288 | 0.017 | -0.066 |
| Justin Smoak | 0.278 | 500 | 0.366 | 132 | 0.260 | 0.018 | -0.106 |
| Nomar Mazara | 0.308 | 469 | 0.332 | 149 | 0.290 | 0.018 | -0.042 |
| Carson Kelly | 0.285 | 365 | 0.336 | 129 | 0.267 | 0.018 | -0.069 |
| Dexter Fowler | 0.311 | 574 | 0.336 | 101 | 0.292 | 0.019 | -0.044 |

## Conclusion and Further Research

Clustering batted balls is a strong method for predicting performance. Our clustering-based model is more effective than relying on previous-year wOBA at predicting subsequent-year wOBA, projecting up to 63% of the variance depending on season. Like any model, it successfully projects numerous players and breakouts while missing on others. Of note, our model’s biggest misses are disproportionately star players who are among the league leaders in walks. Our model makes projections solely using batted balls (excludes walks). Thus, as a product of our approach, it makes sense it would underestimate the players who walk the most. Updating our projections to include walks is likely to improve its performance and is a focus for future research.