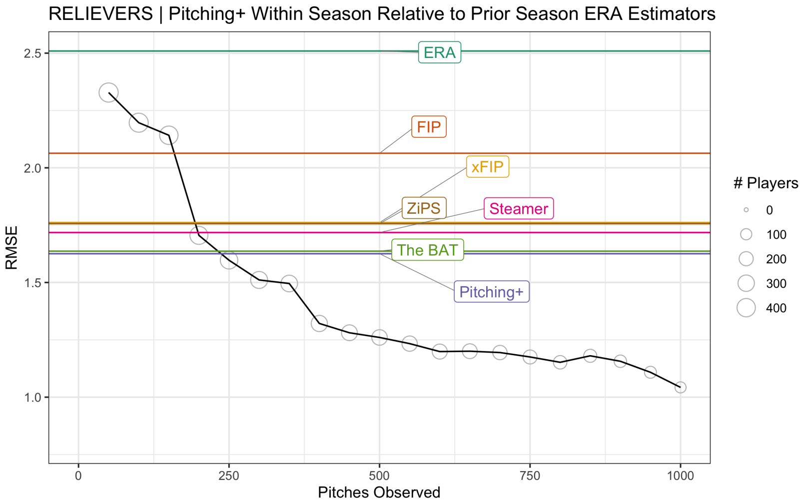
Pitch trajectory density estimation for predicting future outcomes

Scott Powers and Vicente Iglesias

Saberseminar 2023

RICE SOCIAL SCIENCES
Department of Sport Management

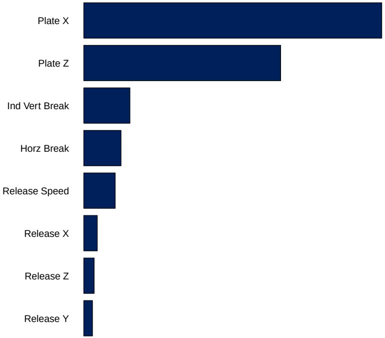
Pitch Modeling



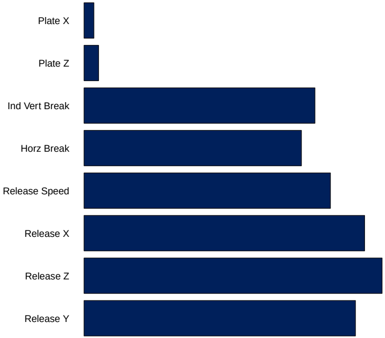
https://library.fangraphs.com/pitching/stuff-location-and-pitching-primer/

The Conundrum

Variable Importance[1]



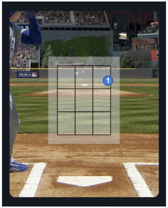
Variable Reliability[2]



[1] fractional contribution of each feature's splits to gradient boosting pitch model
[2] (between-pitcher variance) / (total variance); varies by pitch type (here: RHB FB)

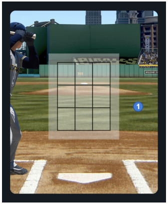
An Example

Pitch A



- Fastball on 0-0 count
- 91 mph w/ 15 inches rise
- located on the edge of the zone
- 68% called strike, 11% foul, 8% ball in play, 8% called ball, 5% swinging strike (-0.04 runs)

Pitch B



- Fastball on 0-0 count
- 98 mph w/ 20 inches rise
- located a foot off of the plate
- 99.6% called ball ($+0.04$ runs)

Two Sources of Noise

1. Random variation in the outcome given the pitch trajectory
    - This is addressed by Pitching+, PitchingBot, etc.

2. Random variation in the pitch trajectory itself
    - This is NOT addressed by Pitching+, PitchingBot, etc.

The Approach

1. Fit a model to predict pitch outcome given its trajectory
   - We use gradient boosting, not the focus today

2. Estimate the probability distribution over pitch trajectories
   - Depends on pitcher, batter side, count, etc.

3. Apply the model 1. to the distribution 2.
   - As opposed to applying the model to the observed pitches

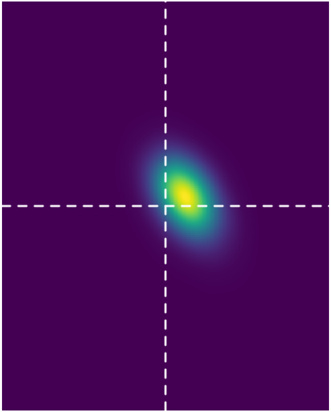
Bayesian Hierarchical Model

Within each pitch type:

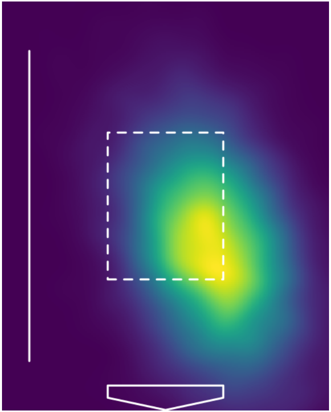
- We model each pitch as multivariate normal in 9 dimensions
  - x/y/z release point, x/y/z release velocity, x/y/z acceleration
- Each pitcher has 81 parameters:
  - $9 \times 4 = 36$ parameters for **mean**
    - Main effect plus interactions w/ balls, strikes, batter side
  - $9 \times 1 = 9$ parameters for **variance**
  - $\binom{9}{2} = 36$ parameters for **correlation** between dimensions
- Each (ball, strike, batter side) combo has 18 parameters:
  - 9 parameters for mean, 9 parameters for variance
- We find the maximum *a posteriori* (MAP) model fit using the `optimize` function (automatic differentiation) from cmdstanr

Dylan Cease's Slider vs RHB in 0-0 Counts

Predicted Break Chart

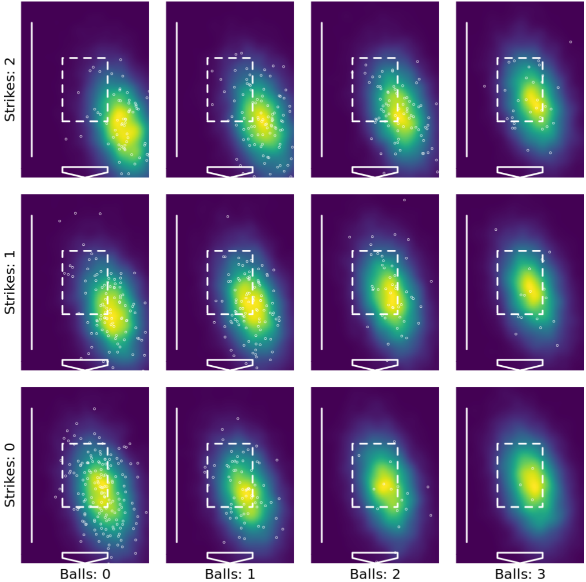


Predicted Plate Location

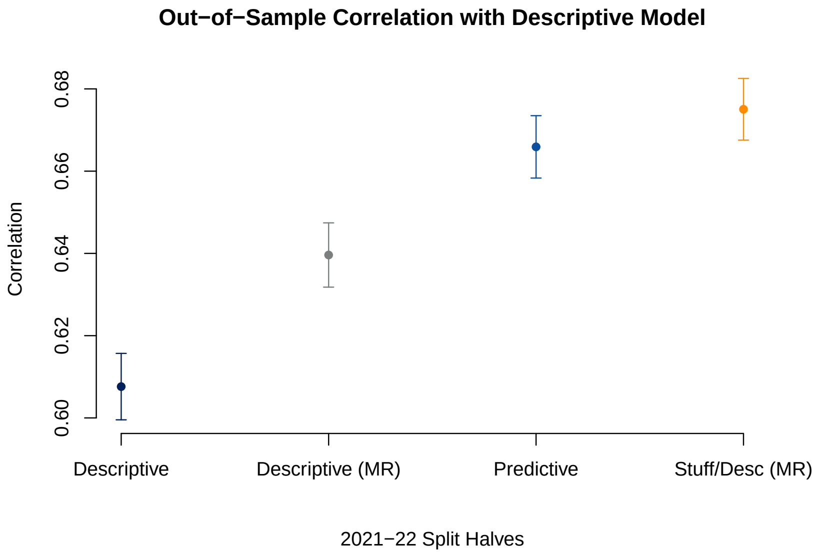


saberpowers.shinyapps.io/predictive-pitch-score

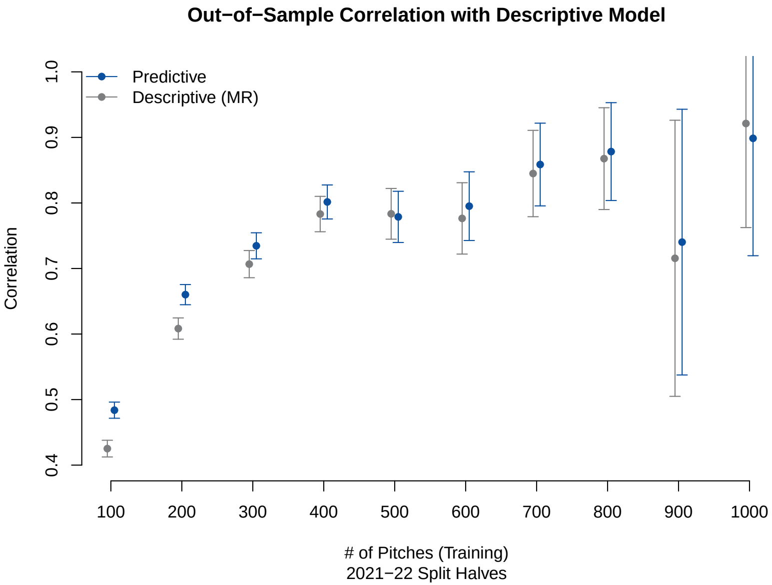
Dylan Cease's Slider vs RHB in All Counts



Does It Work?


Out−of−Sample Correlation with Descriptive Model
Correlation
0.60
0.62
0.64
0.66
0.68
Descriptive
Descriptive (MR)
Predictive
Stuff/Desc (MR)
2021−22 Split Halves

Does It Work?


Out−of−Sample Correlation with Descriptive Model
Predictive
Descriptive (MR)
Correlation
0.4
0.5
0.6
0.7
0.8
0.9
1.0
100
200
300
400
500
600
700
800
900
1000
# of Pitches (Training)
2021−22 Split Halves

Leaderboard

Pitcher	PT	#	Stuff	Desc Score	Pred Score
Tyler Glasnow	all	1087	-8	-15	-18
Zack Wheeler	all	1912	-5	-17	-16
Spencer Strider	all	2164	-11	-14	-16
Sandy Alcantara	all	2094	-7	-12	-14
Pablo López	all	1662	1	-10	-13
Logan Webb	all	2183	-8	-12	-13
Shane McClanahan	all	1820	-2	-11	-13
Bobby Miller	all	1067	-3	-10	-13
Hunter Greene	all	1322	-8	-10	-12
Bryce Miller	all	1179	-3	-12	-11

Showing 1 to 10 of 140 entries

Previous 1 2 3 4 5 … 14 Next

saberpowers.shinyapps.io/predictive-pitch-score

Conclusions

Takeaways:

1. Think more about the second source of noise (random variation in the pitch trajectory itself)
2. Pitch modeling predictions don't capture *all* of the predictive information in a pitch

What's coming up next:

- Better (simpler?) parameterization for distribution model
- Relax Gaussian assumption (unimodal with specific tails)

Where to Find Us

saberpowers.shinyapps.io/predictive-pitch-score
github.com/saberpowers/predictive-pitch-score

linkedin.com/in/saberpowers
twitter.com/saberpowers

twitter.com/fiftycente1