Arsenal/Zone Rating:  
A PitchF/X based pitcher projection system

Pei Zhe Shu

1. Introduction

Pitcher performance projection is a fundamental area in baseball analysis. Traditional projection systems, like PECOTA, ZiPS and Steamer, are based on either ERA (or RA/9), which cannot separate pitcher performance from fielder defense well, or K%, BB%, HR% and batted ball data, which are hard to derive crucial information like BABIP. These systems leave a lot to be desired.

Fortunately, PitchF/X was introduced to baseball, and by utilizing it we can analyze pitchers with much better accuracy. This paper introduces a new and improved method for pitcher projection, which we call "Arsenal/Zone rating", using PitchF/X data. The idea is that pitcher performance can be mostly judged and predicted from two aspects: arsenal rating, which corresponds to the speed and movement of the pitch, and zone rating, which is related to the location the pitch with regard to the strike zone.

Our arsenal rating and zone rating model are trained and predicted on per-pitch level data, which has a pretty decent sample size (over 2 million available pitches in 7 seasons of data) and relatively rich content (speed, movement, location, pitch count, etc.), which allows us to construct a detailed model.

Our combined projection stat, which we call arsenal-zone-combined rating, not only has a comparable result with mainstream projection systems when judged by R^2 and RMSE to actual performance data, but also excellent for predicting breakout and breakdown pitchers compared to those systems. And there's an even better result: after linearly combine our rating to basic pitching stats like K%, BB%, HR%, our system is much better than all mainstream projection systems. We'll describe the method of our system below.

2. Method

2.1. Main model

We construct a multilevel regression model to incorporate and estimate arsenal rating and zone rating. Our main model, which is on per-pitch level, and on run scale, which can be described as:

\[ r_i = a(sp_i, xmove_i, zmove_i, type_i, hand_i) + z(xloc_i, zloc_i, type_i, hand_i) + r_h + def_i + pf_{pa} \]

Here \( r_i \) is the actual result (run expectancy) of this pitch, \( a(sp_i, xmove_i, zmove_i, type_i, hand_i) \) is the arsenal rating of the pitch, with each covariate representing speed, x-movement, z-movement, pitch type, and handness (same handed or opp handed batter/pitcher) of the pitch respectively, \( z(xloc_i, zloc_i, type_i, hand_i) \) is zone rating of the pitch, with each covariate representing x-location,
z-location, pitch type and handness of the pitch, \( r_e \) is the run expectancy of the batter, \( def_i \) is the pitcher’s team defense, \( pf_{pa} \) is the park factor. We’ll explain each parameter of this model below.

2.2. Run Expectancy on pitch level
We follow the widely used fangraphs.com wOBA formula and the corresponding league average wOBA and wOBA scale[1] for run expectancy on plate appearance(PA) level. We use the following rules to determine pitch level run expectancy \( r_e \):

(1) For non-PA ending pitches, the run expectancy is the one of new pitch count minus the one of old pitch count.

(2) For PA ending pitches, the run expectancy is the one of PA result minus the one of old pitch count.

For pitch count run expectancy, we calculate the expectation of the PA result run expectancy conditioned on each pitch count. This method scatters the value of the PA result relatively equally on each pitch of PA.

2.3. Zone Rating
As both the zone rating function and pitch location distribution are relatively smooth, we use a simplified model to describe zone rating; we separate the strike zone into square grids, and use each grid as a separate covariate in our model to obtain its zone rating. A local regression is performed on the raw result to obtain a smoothed one. The graph below shows an example of the final zone rating.

To describe location of a pitch with regard to the strike zone, we need to “standardize” the strike zone, namely to adjust the strike zone on batter’s height and handness. We’ll describe this adjustment later in section 2.7.
Now that we've obtained the zone rating of each strike zone location, we need to calculate the zone rating of each pitcher. As a result, we need to know the distribution of each pitcher’s pitch location. We assume the distribution of each pitcher’s pitch location is a combination of “circles with identical size”, i.e. mixture of Gaussian with each component’s covariance matrix being identical scalar matrix. Below is a graph showing the raw data and the smoothed pitch location data.

2.4. Arsenal Rating

We use a generalized additive model (GAM) [2] to describe the arsenal rating. We assume that a pitch’s arsenal rating is sum of three components: speed-related rating, x-movement-related rating and z-movement-related rating. Each rating would then be described by a combination of multiple (usually several to fifty) thin plate splines [3].

Using our scouting knowledge, we know that different type of pitches will have different properties. For example, fastball with faster speed is highly likely to be better, while curveball’s speed has less (if any) positive impact. So we use separate arsenal rating model for different pitch types, thus pitch classification is essential to our model.

Also, even for same pitch type, identical pitches should have different “power” on different batter/pitcher handness. For example, “lateral moving” sliders are known to be better against same handed batter. So we also use separate models for same handed/opp handed batter. We also use separate zone rating model for different pitch types and same handed/opp handed batter, although this has less significant impact on our result.

We use a “soft” pitch type classification, which means each pitch has a probability ranging from [0,1] of being some pitch type. To calculate the arsenal rating of a pitcher, we average the arsenal rating of every pitch thrown by him, which is this pitch’s rating when treated as a specific pitch type, weighted by its probability of being this pitch type. We'll show how to classify pitches in section 2.5.
See the following graph for an example of arsenal rating, which shows sliders for both same handed and opp handed batters.

\[
\begin{align*}
\text{Arsenal Value for Sliders (Same Handed Batter)} \\
\text{Arsenal Value for Sliders (Opp Handed Batter)}
\end{align*}
\]

Y Axis: Arsenal Value, Top=worst, Bottom=best.
X Axis, From left to right: 1st graph-Relative Speed (Slider speed minus average Fastball speed of same pitcher, negative value, right=smaller speed difference)
2nd graph-Speed (Speed of slider, All speed used in the system are a rescaled value, right=higher speed)
3rd graph-Relative X Movement (X movement minus average Fastball X movement, right=higher lateral movement)
4th graph-Relative Z Movement (Z movement minus average Fastball Z movement, negative value, right=smaller vertical movement)

2.5. Pitch Type Classification
We first introduce the following Gaussian Mixture model (GMM) for a specific pitcher's pitches:

\[
p(s_{p_{i}}, x_{move_{i}}, z_{move_{i}}) = \sum_{t} \pi_{t} N(s_{p_{i}}, x_{move_{i}}, z_{move_{i}} | \mu_{t}, \Sigma_{t})
\]

Here \((s_{p_{i}}, x_{move_{i}}, z_{move_{i}})\) is a 3d vector contains speed, x-movement and z-movement of pitch b thrown by pitcher a, \(\pi_{t}\) is the weight of pitch cluster t, \(\mu_{t}\) and \(\Sigma_{t}\) is a 3d center vector and 3x3 covariance matrix, respectively, for pitch cluster t thrown by this specific pitcher.\(^1\)

The whole model can be described as: a pitcher's speed and movement distribution is the combination of several any shaped 3d-ellipsoid. We found this GMM model a good description of a specific pitcher's pitches, with one or several pitch clusters belonging to one actual pitch type.

\(^1\) Here we assume that the covariance matrix \(\Sigma_{t}\) can be any positive definite matrix, which means the contour of Gaussian can be an 3d-ellipsoid with any shape.
To classify pitches, we slightly alter our assumptions into a stronger one, which is the following model:

\[ p(s_{p_i}, x_{move_i}, z_{move_i}) = \sum \pi_i N(s_{p_i}, x_{move_i}, z_{move_i} | \mu_i + \text{off}_p, \Sigma_i) \]

Which means every pitcher’s pitch follows the same GMM model, now with each \( \pi_i \) representing a pitch type. The other difference is the “offset” vector \( \text{off}_p \), which is the same for any pitch throwing by a specific pitcher, regardless of pitch type.

We labeled a training set for the classification task. From the training set and the model above we get a Naïve Bayes classification result using this model, and we add in the raw speed and movement data, to form a multinomial logistic regression model, to obtain the final result. See the graph below for an example of our classification result.

![Pitch classification result for Cliff Lee's pitches in 2013(left) and Jon Lester's pitchers in 2013(right).](image)

We follow the pitch type system of brooksbaseball.net. As a result, our label set classify pitches into 6 types: FF(fastball), SI(sinker), CH(changeup, including splitter), FC(cutter), SL(slider), CU(curve, including faster ones and slower ones).

We found the model above has good accuracy for classifying pitches: We obtained 88.3% accuracy on our test set. While this might not seem to be too high, we found nearly all of the classification errors “borderline pitches”, namely FF/SL, FC/SL or SL/CU. Those pitch types are naturally hard to classify, and our label set sometimes essentially provides an arbitrary label.

This multinomial logistic regression model also naturally gives probabilistic classification result, which we use to calculate arsenal rating, as we discussed in section 2.4. This probabilistic result provides a nice alternative solution to the borderline pitches.

**2.6. Pitch Movement Calculation and Pitch Calibration**

We follow A. Nathan’s method[4] to calculate pitch movement. We remove drag effect from the data, and remove gravity effect from data as well. As a result, we only consider the Magnus effect of the pitch.

Also, Pitch F/X raw data has a known issue: it is poorly calibrated[5]. Precisely speaking, in some ballparks and/or some games, it will systematically report a slightly higher or lower number of
some attribute (including velocity and movement) of every pitch thrown. This is what we want to correct, as we can see below, the raw data can be pretty weird.

We use the model in section 2.5 for clustering pitches to solve this problem, also with an “offset” vector. This time we use an offset for each game. See the following graph for an example.

![Graph showing pitch clustering](image)

*Billingsley's pitches in 2008. Left graph=original data, Right graph=data after calibration, both with clusters. The original data contains a right-sided “tail” (obvious poor calibration), which is corrected after calibration.*

### 2.7. Strike Zone

We use a simple definition of the strike zone. A closed curve which satisfies the following condition: a pitch locates on the curve has 50% probability of being called a strike. Following this definition, we use a logistic regression on raw called-strike/ball data to determine the actual strike zone.

As we're actually more concerned about the impact of height and handness to the strike zone rather than its shape, we can just approximate the strike zone with a rectangle with same center of mass and same area. Add in the assumption that the center of mass of the z-side of the rectangle has a linear relationship with batter height, and the x-side parameters of the rectangle as well as length of the z-side are irrelevant with batter height, we arrive at the following strike zone formula:

- **Right-handed batter:**
  \[ -0.991 \leq x \leq 0.946 \text{ and } 0.906 + 0.1281 \times height \leq z \leq 2.591 + 0.1281 \times height \]

- **Left-handed batter:**
  \[ -1.139 \leq x \leq 0.755 \text{ and } 0.617 + 0.1729 \times height \leq z \leq 2.306 + 0.1729 \times height \]

As we can see, this is pretty similar with Mike Fast's strike zone formula[6]. We use this result to adjust a pitch's strike zone location according to batter height and handness.

### 2.8. Other parameters

We use a simple linear regression for the projection of both team defense and park factor: the projected value is the linear combination of previous several years of observed value. We use the average of UZR and DRS score for team defense, and observed wOBA-based park factor as raw value. For park factor, 5 years of previous data are used, while for team defense, 3 years are used, as team defense has a much higher year-to-year variance, especially when using raw models like UZR and DRS.
We assume a Gaussian prior distribution for run expectancy of batters\(^2\). This is to eliminate the effect of each pitcher’s context, or different (combined) ability of the batters he pitched to.

### 2.9. Adjustment to our main model

The output of our model till now is a pitcher’s zone rating and arsenal rating based on his actual (past) data of a specific period. Combining them together, and add in the projected team defense and park factor, we obtain a raw combined rating. While this is already a decent estimate of pitcher’s future performance, we feel the need to do some relatively minor but still important adjustment, which can be described using the following combined formula:

\[
ar_p = (\sum a + \sum z + \text{def}_i + p_{f_{pa}}) \cdot \text{pitch}_p + l_{\_\text{avg}} + \text{hand}_p + pc_{\_\text{value}}_p
\]

Here \(ar_p\) is the combined result of our model, which we will call “arsenal-zone-combined rating” from now on. \(\sum a\) is the sum of the arsenal rating, discussed in section 2.4, \(\sum z\) is the sum of the zone rating, discussed in section 2.3. \(\text{def}_i\) is the team defense, \(p_{f_{pa}}\) is the park factor, both of which are discussed in section 2.8. \(\text{pitch}_p\) is the estimate value of pitches thrown by this pitcher per 9 innings. \(l_{\_\text{avg}}\) is the league average RA/9. \(\text{hand}_p\) is an adjustment based on pitcher handness. \(pc_{\_\text{value}}_p\) is an adjustment based on pitcher’s past pitch-count distribution. We’ll explain each of those parameters below.

**\(\text{pitch}_p\):** This is used to scale the rating from run-per-pitch scale to RA/9 scale. To predict the estimate value of pitches thrown by this pitcher per 9 innings, we use the observation that this value is in fact the multiplication of two factors: pitches per PA, and PA per 9 innings (roughly same as 9 times WHIP). We just use the previous season’s value for pitches per PA, and use a linear model for PA per 9 innings, with arsenal rating and zone rating as parameters.

**\(l_{\_\text{avg}}\):** The zone rating and arsenal rating are trained on multi-year data, and the baseline is “league average”. So when we scale these rating back to RA/9, we should use a different league average RA/9 for each year and for AL/NL. We use a linear model (of previous several years’ RA/9) to predict next year’s RA/9, to prevent overfitting.

**\(\text{hand}_p\):** We use different models for same-handed/opp-handed batter, but due to lack of data on lefty pitchers, we combined righty and lefty pitchers in the same arsenal rating model, which is not very satisfactory. Our adjustment on this is to give all lefties a “bonus”. We use out-of-sample data (we use 2008-11 data to predict 2012-14) to obtain the bonus value, to prevent overfitting.

**\(pc_{\_\text{value}}_p\):** We sum the zone and arsenal rating of each pitch (and thus, each pitch type) of a pitcher to obtain his final rating. But according to our scouting knowledge, combination of several types of “strong” pitch probably would lead to even higher level of performance than summing the pitches’ arsenal and zone ratings alone. We have several ways to compensate for this, and after

---

\(^2\) This method is often called “mixed model” or “hierarchical model”.

---

2016 Research Papers Competition
Presented by:

MIT Sloan Sports Analytics Conference
March 11 - 12, 2016 Boston Convention and Exhibition Center
testing it on data, we found a pitch-count-distribution based adjustment is the best one. That is, pitchers run into “good” pitch count more often will lead to even better combination result.

2.10. Adjustment for prediction purpose
For prediction purposes, past data is not exactly equal to future data. Although a specific pitcher’s arsenal and zone rating are largely stable through years (we’ll discuss this in section 3.3), a pitcher’s arsenal rating is probably decreasing every year (at least after around age 24-25), due to advancing age and accumulated pitch “miles”. We use a direct age-based adjustment, based on kernel regression result of 2008-2011 data.

Regression to the mean is also a widely adopted method in projection systems. We also utilize this idea, and regress each pitcher to league average RA/9 according to his innings pitched in last year (the year we calculate his arsenal and zone rating). The regression coefficient is picked by cross-validation on between-season arsenal-zone-combined rating.

3. Result and Discussion

3.1. Predictive power test
Using our method described above, we’re able to produce the result of the arsenal rating and zone rating models, and each pitcher’s arsenal-zone-combined rating. This result is on RA/9 scale, and can be used to predict pitcher’s future performance.

We use PitchF/X from MLB.com and play-by-play data from retrosheet.org from season 2008 to season 2014 to train our arsenal rating and zone rating models. To classify pitches, we additionally require the pitcher involved to have at least 1500 pitches (roughly 95 innings) pitched during a single season, which restrict our analysis to starters only. The full 7-season dataset contains around 4.9 million pitches, and with the additional 1500-pitch requirement, the dataset contains around 2.4 million pitches.

To test the predictive power of our combined RA/9 scale result, we use 2012-2014 data, with the additional requirement of pitcher pitching at least 50 inning in the respective season, and pitch for only one team during both the current season and previous season. We have a total of 181 qualified seasons of pitchers, all of them starters. We predict pitcher’s season performance using the previous season’s arsenal-zone-combined rating only, with the only adjustment being the age adjustment and regression to the mean described in section 2.10.

We use four projection systems as our comparison baseline: Marcel, PECOTA, ZiPS, and context-based-FIP (cFIP). The Marcel projection system is a pure replacement-level “baseline” system.

---

3 We run a partial least squares (PLS) regression of 2008-2011 data with a lot of potential features and choose the main component (which almost only contains pitch-count-distribution).
4 We use cross-validation to maximize the log likelihood of the data after regression-to-the-mean. The result is that the prior counts as around 20 innings.
5 This not-changing-team-for-two-seasons requirement is mainly because we use park factor and team defense to combine our final rating, and we know the exact team each pitcher pitched for in the second season, but we’re not sure how each projection system handles this issue, so to avoid unfair advantage for our rating, we just remove those pitchers who changed team.
but is designed to have a good weighted Root Mean Standard Error(RMSE). The PECOTA projection system and ZiPS system are two mainstream systems, both of which utilize “comparable pitchers” to predict pitcher’s future performance. ZiPS consistently rank highly in pitcher projection comparison[8]. Those three systems above are on ERA scale, so we linearly scale it to RA/9 scale. cFIP is a system that adjust K%, BB% and HR% according to opponent, park, catcher and umpire, and use those result to combine an FIP-like rating. This rating is originally on a 100-point scale, but it can also be linearly scaled to RA/9 scale. Also, this rating is not a complete projection system, as it is calculated only on one season of data. We also use previous season's cFIP only to project next season, with an age adjustment and regression to the mean, both with similar method described in section 2.10.

<table>
<thead>
<tr>
<th>Result of projection system</th>
<th>Marcel</th>
<th>PECOTA</th>
<th>ZiPS</th>
<th>cFIP</th>
<th>arsenal-zone-combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$ with next season's RA/9</td>
<td>0.1866</td>
<td>0.1975</td>
<td>0.2043</td>
<td>0.2183</td>
<td>0.2017</td>
</tr>
<tr>
<td>RMSE with next season's RA/9</td>
<td>0.860</td>
<td>0.881</td>
<td>0.874</td>
<td>0.867</td>
<td>0.864</td>
</tr>
</tbody>
</table>

The table above shows the correlation($R^2$, higher means better) and RMSE(lower means better) with real RA/9 data for our system and four baseline systems. We can see that our system is 3rd in $R^2$ and 2nd in RMSE, and not too much behind the leaders. These comparison are on those 181 qualified 2012-2014 starting pitchers, and the same set of data are used for all $R^2$ and RMSE comparisons below.

| Breakout-Breakdown prediction result using arsenal-zone-combined rating |
|---|---|---|---|---|---|---|---|
| Baseline System | Percentage | Top 1/3 | Middle | Bottom 1/3 | Bottom 1/3 | Middle | Top 1/3 |
| PECOTA | 20% | 58% | 25% | 17% | 44% | 50% | 6% |
| | 33% | 58% | 27% | 15% | 45% | 43% | 12% |
| ZiPS | 20% | 56% | 17% | 28% | 44% | 42% | 14% |
| | 33% | 57% | 17% | 27% | 40% | 43% | 17% |
| cFIP | 20% | 50% | 28% | 22% | 56% | 28% | 17% |
| | 33% | 47% | 25% | 28% | 50% | 33% | 17% |

The table above shows the “breakout-breakdown” prediction of our rating. This prediction works as follow(using ZiPS baseline as example): sort all 2012-2014 qualified starting pitchers according to “arsenal-zone-combined rating minus ZiPS projection”. Check the pitchers with the highest(indicating breakthrough, or lowest, indicating breakout) 20%(or 33%) result, to see where their “real RA/9 minus ZiPS projection” falls within all pitchers: top third, middle, or bottom third.

The result shows our system can statistically significantly pick out the breakout and breakthrough (or underestimated and overestimated by each baseline projection system, respectively) pitchers with

---

6 Marcel system’s good RMSE property can also be found in [8].
7 We allow projection systems to have an “offset” when computing the RMSE.
regard to every baseline projection system, as in every breakout prediction, the “top 1/3” percentage (red) is way bigger than “bottom 1/3” (green), and vice versa in breakdown prediction.

This breakout-breakdown prediction shows our system should have some distinct advantage over the baseline projection systems. We use the following test to confirm this observation.

### 3.2. Combination with K%, BB% and HR%
As our arsenal-zone-rating system haven’t utilize the most stable component stats of pitchers (K%, BB%, HR% adjusted by park or estimated HR% based on FB%), and every projection system have already used those, we naturally arrive at the following guess: if we combine the arsenal-zone-rating with these stable pitching components, we should obtain a better combined model than combine the baseline systems with these components.

#### Result of combined system with xFIP

<table>
<thead>
<tr>
<th>Index(weight)</th>
<th>Marcel+xFIP</th>
<th>PECOTA+xFIP</th>
<th>ZiPS+xFIP</th>
<th>cFIP+xFIP</th>
<th>arsenal-zone-combined+xFIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>R^2 (50/50)</td>
<td>0.2263</td>
<td>0.2315</td>
<td>0.2293</td>
<td>0.2124</td>
<td>0.2528</td>
</tr>
<tr>
<td>RMSE (50/50)</td>
<td>0.834</td>
<td>0.834</td>
<td>0.836</td>
<td>0.846</td>
<td>0.819</td>
</tr>
<tr>
<td>R^2 (optimal)</td>
<td>0.2268</td>
<td>0.2315</td>
<td>0.2298</td>
<td>0.2209</td>
<td>0.2529</td>
</tr>
<tr>
<td>RMSE (optimal)</td>
<td>0.833</td>
<td>0.834</td>
<td>0.836</td>
<td>0.845</td>
<td>0.819</td>
</tr>
</tbody>
</table>

The table above shows this guess is true. After linearly combining⁸ every model with xFIP (can be seen as a linear model of K%, BB% and estimated HR% based on FB%) in previous season, our model comes out on top, with a pretty big advantage (R^2 advantage is similar to ZiPS vs Marcel).

Of course, cFIP has a natural disadvantage in our combination test above, as it’s basically an adjusted FIP with almost nothing else. But cFIP can also been viewed as a linear model of stable pitching component (in fact, its components are more stable than raw K%, BB%, and HR%), so we can combine cFIP with our arsenal-zone-combined-rating, and with every baseline projection system (other than cFIP itself). This is shown in the table below, and our model comes out on top again, with a similar big advantage.

#### Result of combined system with cFIP

<table>
<thead>
<tr>
<th>Index(weight)</th>
<th>Marcel+cFIP</th>
<th>PECOTA+cFIP</th>
<th>ZiPS+cFIP</th>
<th>arsenal-zone-combined+cFIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>R^2 (50/50)</td>
<td>0.2499</td>
<td>0.2484</td>
<td>0.2486</td>
<td>0.2693</td>
</tr>
<tr>
<td>RMSE (50/50)</td>
<td>0.822</td>
<td>0.829</td>
<td>0.829</td>
<td>0.812</td>
</tr>
<tr>
<td>R^2 (optimal)</td>
<td>0.2499</td>
<td>0.2487</td>
<td>0.2486</td>
<td>0.2693</td>
</tr>
<tr>
<td>RMSE (optimal)</td>
<td>0.822</td>
<td>0.8287</td>
<td>0.829</td>
<td>0.812</td>
</tr>
</tbody>
</table>

---

⁸ We use a 50/50 weight for every model combination for the following reason: 50/50 is extremely close to MAP estimation of this prior weight (linear combination of model can be seen as a Bayesian averaging of models) of every pair of models. We also use an optimal prior weight (based on cross-validation on R^2), and found the optimal weight for most model combination pairs are close to 50/50 in most cases (every pair is in 45/55-55/45 range, except cFIP+xFIP) as well. We show the result for both.
3.3. Year-to-year correlation
We are also interested in the year-to-year correlation of same pitcher’s arsenal-zone-combined rating, with or without the regression to the mean adjustment. We found even without the regression to the mean, it has a much higher $R^2$ than that of cFIP (which is in turn much higher than every basic pitching stat and linear models based on it, such as xFIP and SIERA, as shown in [7]), and has a similar $R^2$ with that of ZiPS, only lower than PECOTA, as shown in the table below, during seasons 2012-2014. So it’s a pretty stable rating across years.

<table>
<thead>
<tr>
<th>Year-to-year R$^2$</th>
<th>system</th>
<th>arsenal-zone-combined (w/o regression to the mean)</th>
<th>arsenal-zone-combined (w/o regression to the mean)</th>
<th>cFIP (w/o regression to the mean)</th>
<th>cFIP</th>
<th>PECOTA</th>
<th>ZiPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>system</td>
<td>0.6628</td>
<td>0.6787</td>
<td>0.4106</td>
<td>0.4475</td>
<td>0.8327</td>
<td>0.6703</td>
</tr>
</tbody>
</table>

3.4. Discussion
All our result above shows a good projection system for pitchers with enough previous season’s data (around 95 innings). This limits our result to qualified starting pitchers. However, as the arsenal-zone-combined rating is stable enough across years, and is based on per-pitch level data which has a bigger sample size than per-PA data, not to mention arsenal rating (the main part of the model) itself uses features (speed and movement) that have way less variance than other per-pitch stats like plate discipline stats, it should theoretically generalize well to starting pitchers with less previous data, as well as relief pitchers, although some adjustment must be made. The main problem probably lies on pitch classification, as our system requires GMM and Naïve-Bayes classification result, which fit well for large data set but can be problematic with smaller ones. Luckily there’re already manually-labeled pitch classification data (for example, the one on Brooksbaseball.net), so it’s possible to overcome this problem.

4. Conclusion
We present a PitchF/X based model: Arsenal/Zone rating, to describe pitcher’s true pitching ability, and to predict pitcher’s future performance.

The final model, which we call arsenal-zone-combined rating, is a good pitching performance predictor by itself. It has comparable result with mainstream projection systems, and it has distinct advantage in picking out the breakout and breakdown pitchers than those mainstream systems.

When linearly combined with stable pitching stats (K%, BB% and HR% adjusted by park or estimated HR% based on FB%), the arsenal-zone-combined rating is much better than the mainstream projection systems, and can be seen as one of the best pitching projection system currently available.
References