Title: Beyond the Mound: K-Means Clustering and Principal Component Analysis for Pitching Role Evaluation

Arjun Parmar, Johnny Vecchio, Scott Smoot, Corey Grozier, & Nate Wellman

Word count: 1478/1500

Introduction

To identify pitchers who can be better off in other roles, we employed an unsupervised learning approach. Using K-means clustering to group like players together, we found that there are 22 metrics that can be used to accurately classify traditional starting and relief pitching roles. Additionally, principal component analysis was used to summarize multiple statistics and give a pitcher an overall score, allowing us to rank their performance. The model proved to be a valid analysis technique and identified a starting pitcher, Alex Cobb, and a reliever, Shelby Miller, as pitchers that should change roles to optimize the pitching rotations performance.

Methods

Data Source

All data was utilized was acquired from FanGraphs and spanned three years, 2021-2023. Furthermore, pitching statistics used were at a seasonal level for each of the three years.

Normalization

We first identified absolute variables (i.e. strikes) that may significantly differ between starting and relieving pitchers due to playing time. A Student's t-test was performed between pitching roles to confirm that these statistics were significantly different between groups. All absolute variables were significantly different ($p < 0.01$) between groups, except for catcher framing ($p = 0.55$). To normalize the identified absolute variables, we normalized by pitch count, total batters faced, and innings pitched, then compared the subsequent outcomes. Pearson product moment correlations were calculated for each variable between all three normalization techniques. There was a high level of correlation between the normalized values from pitches thrown and total batters faced ($R^2 > 0.91$), in addition to a moderate to high correlation ($0.57 < R^2 < 0.92$) between the normalized values from innings pitched and all the other methods. Due to these findings, we chose to normalize by pitches thrown, as it has the highest resolution and agreed with total batters faced. Additionally, to account for seasonal differences, pitchers were scaled to the league average and minimum of the respective year.

Pitch Arsenal

To quantify the number of pitches that a pitcher has in their arsenal, PitchingBot and Pitching+ metrics were totaled, producing two new metrics of total pitch types.

Missing Data

Of the 238 variables within our normalized data set, only 97 were complete with no missing data. To improve the dataset, we employed a random forest model to impute any incomplete data. A missingness threshold was set at 5%, thus no variables were imputed if they were missing more than 5% of their measurements, increasing the number of usable variables to 139.

Variable Selection

We next sought to improve the decision space by selecting variables that accurately classified starting and relieving pitchers in a K-means model. A logistic regression model was fit to the data and identified 97 variables of importance ($p < 0.01$). In addition, we developed a support vector machine (SVM) solution to classify pitcher role. Feature importance of the SVM was plotted and variables above the knee (inflection point) of the curve were selected, 22 variables. Both models demonstrated greater than 99% accuracy in pitching role classification.

K-Means Modeling – All Pitching Roles

To identify the best variables to categorize pitching roles, a K-means model was developed for each classification model developed in the previous step. The optimal number of clusters was determined to be three for each model. This was identified by calculating the knee within the sum of square by number of clusters plot. Variables identified by the logistic regression proved to be poor identifiers of pitching role, but the SVM-based variable K-means model demonstrated excellent accuracy in clustering pitchers of the same role together. This is seen in Figure 1, where K-means clustering using the logistic regression variables did not cluster starting and relief pitchers into homogenous groups, while the SVM-based clustering model was able to separate pitching roles. All subsequent K-means models used the SVM-based variables, as they best distinguished between pitching roles. The significant variables for classification are ordered in decreasing importance in Table 1.

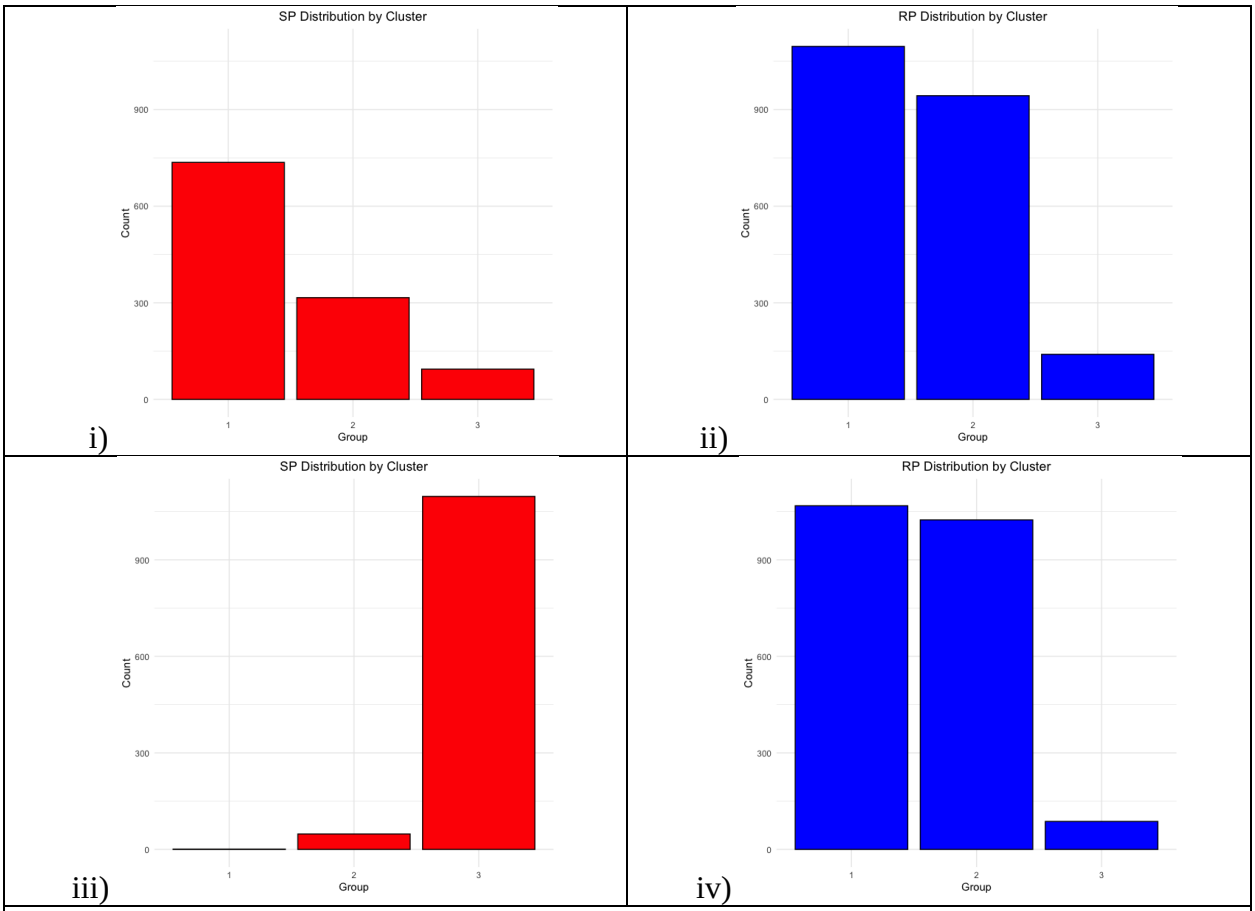
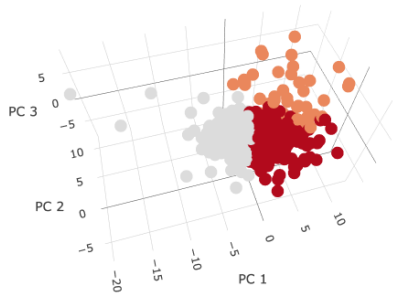


Figure 1. K-means decision space optimization.
i) Logistic regression-based cluster distribution of starters. ii) Logistic regression-based cluster distribution of relievers. iii) SVM-based cluster distribution of starters. iv) SVM-based cluster distribution of relievers.

K-Means Modeling – Starting and Relief Pitchers

A K-means model was developed to cluster starters that are the most alike using the SVM model variables. Using the same technique to identify the ideal number of clusters, three clusters were determined to be optimal. Similarly, a K-means model was developed using the same variables for relief pitchers and three clusters were optimal. Tables 2 and 3 compare the variables of importance between each of the clusters for starters and relievers, respectively. Each table displays which group contains the highest and lowest statistics, on average, within a pitching role. To specify, if a variable was determined to be at its highest or lowest within Group 1, then it cannot appear in another group.

Starting Pitchers



Relief Pitchers

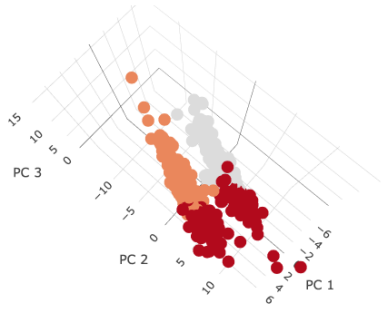


Figure 2. Three-Dimensional Plots of Starting and Relief Pitcher Clusters
Pitchers projected onto the first three components of the SVM-based PCA and colored by cluster.

Group 1		Group 2		Group 3	
Max Stats	**Min Stats**	**Max Stats**	**Min Stats**	**Max Stats**	**Min Stats**
EV	wFB_per_c	ZSwing_pct_sc	RE24	wFB_per_c	EV
ZContact_pct_sc	Pace			Pace	ZContact_pct_sc
LA	LOB_pct_plus			RE24	LA
Contact_pct	LOB_pct			LOB_pct_plus	Contact_pct
Contact_pct_sc	Location_plus			LOB_pct	Contact_pct_sc
kwERA	Stuff_plus			Location_plus	kwERA
ZContact_pct	Pitching_plus			Stuff_plus	ZContact_pct
E_minus_F	SwStr_pct			Pitching_plus	E_minus_F
tERA	ZSwing_pct_sc			SwStr_pct	tERA
xFIP	CSW_pct			CSW_pct	xFIP
xFIP_minus					xFIP_minus

Table 2. Starter Clusters

Group 1		Group 2		Group 3	
Max Stats	**Min Stats**	**Max Stats**	**Max Stats**	**Min Stats**	**Max Stats**
wFB_per_c	EV	EV	wFB_per_c	ZContact_pct_sc	Pace
Pace	ZContact_pct_sc	LOB_pct_plus	Contact_pct	LA	RE24
RE24	LA	LOB_pct	Contact_pct_sc	Contact_pct	SwStr_pct
Location_plus	LOB_pct_plus	SwStr_pct	kwERA	Contact_pct_sc	CSW_pct
Stuff_plus	LOB_pct	CSW_pct	ZContact_pct	kwERA	
Pitching_plus	E_minus_F		Location_plus	ZContact_pct	
			Stuff_plus	E_minus_F	
			tERA	tERA	
			xFIP	xFIP	
			Pitching_plus	xFIP_minus	
			xFIP_minus	ZSwing_pct_sc	
			ZSwing_pct_sc		

Table 3. Relief Clusters

Identify Clusters with Role Change Potential

To identify starting and relief pitchers that can change roles, we first conducted a Student's t-test to identify stats that are significantly different between each cluster and to ensure the validity of the clustering. All variables were significantly different ($p < 0.01$) between the three starting and three relief groups, demonstrating validity of the clustering model. Next, the average minimum distance was calculated between each starting group and all relief pitchers. Starting Group 3 was found to be the closest to all relief pitchers. Furthermore, when calculating the average minimum distance between relief groups and all starters, it was found that Relief Group 1 was closest to starting pitchers.

Principal Component Analysis Player Ranking

Once potential players for role changes were identified, we conducted a principal component analysis (PCA) to rank the athletes. First, we used both pitching roles to perform the PCA on the variables identified by the SVM model. Next, PCA scores for each component were compared to ERA, WAR, Dollars, FIP, and xFIP to identify which component's scores best reflected pitcher performance. Components 1 and 3 showed significantly higher correlation to these variables than other components (Figure 3) when a Pearson product moment correlation was performed. We chose Component 1 to score players since it had a high positive correlation with WAR and Dollars. Finally, Relief Group 1 and Starting Group 3 pitcher statistics were projected onto Component 1 (from the overall PCA model) to obtain PCA scores. Athletes with the top ten scores through all three years are presented in Table 4. (N.B. only pitchers with more than 25 innings pitched were included).

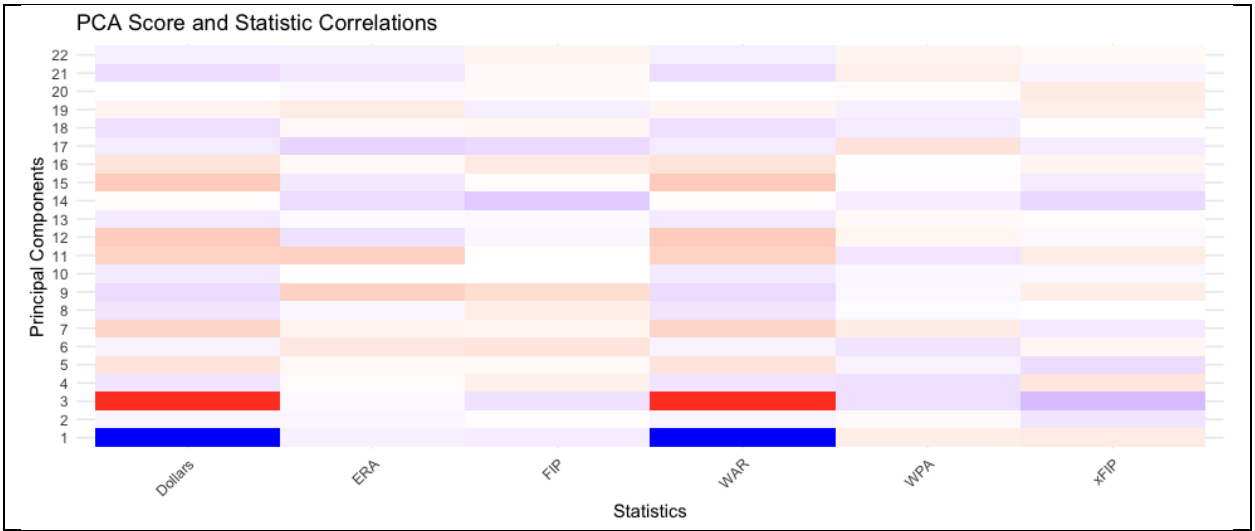


Figure 3. PCA Score and Player Statistic Association

	Starter to Reliever			Reliever to Starter		
	Name	**Season**	**Innings Pitched**	**Name**	**Season**	**Innings Pitched**
1	Walker Buehler	2021	207.2	Daniel Bard	2021	65.2
2	Lucas Giolito	2022	161.2	Colin Holderman	2022	28.1
3	Johnny Cueto	2022	153.1	Seth Lugo	2021	46.1
4	Alek Manoah	2021	111.2	Evan Marshall	2021	27.1
5	Jaime Barria	2023	27.1	Mark Melancon	2022	56.0
6	Trevor Bauer	2021	107.2	Kevin Ginkel	2022	29.1
7	Chris Bassitt	2021	157.1	Tyler Rogers	2022	75.2
8	Luis Castillo	2021	187.2	Miguel Castro	2021	68.1
9	Trevor Cahill	2021	35.2	Ian Kennedy	2021	56.1
10	Drew Rasmussen	2021	42.0	Emmanuel Clase	2022	72.2

Table 4. Top Ten Prospects

Results

Group Analysis

Comparing the descriptive comparisons from Tables 2 and 3, we can identify what statistics are predictive of potential role changes. Interestingly, compared to their peers in the same roles, Relief Group 1 and Starting Group 3 had very similar maximum and minimum statistics. This indicates that while they have a high potential to change roles completely, they could also be dominant swingmen.

2024 Season Pitching Role Change Recommendations

Ranking all starting pitchers in specifically 2023, the top 2 prospects to become relief pitchers were Jaime Barria and Chris Flexen. Notably, both athletes are relief pitchers, who played the role of starting pitcher for a select portion of the season, demonstrating that our model can accurately find pitchers that could potentially change roles and be successful. For the 2023 pitchers that could benefit in a role change, the model identified a true starter that should be a relief pitcher, Alex Cobb of the San Francisco Giants. Additionally, the model found a relief pitcher that should be a starting pitcher, Shelby Miller of the Detroit Tigers. The role predicting statistics for both players are shown in Table 5.

	Alex Cobb	Shelby Miller
Current Role	Starting	Relief
New Role	Relief	Starting
wFB_per_c	0.14	3.41
Pace	18.17	19.48
EV	89.77	89.55
RE24	6.81	14.71
ZContact_pct_sc	0.91	0.85
LA	1.30	14.57
Contact_pct	0.81	0.77
Contact_pct_sc	0.81	0.77
ZContact_pct	0.91	0.84
kwERA	4.47	4.56
LOB_pct_plus	106.37	124.08
LOB_pct	0.76	0.89
Location_plus	101.85	98.98
E_minus_F	-0.14	-1.99
Stuff_plus	90.98	110.93
tERA	4.16	3.36
xFIP	3.51	4.53
Pitching_plus	100.45	101.41
SwStr_pct	0.09	0.10
xFIP_minus	80.24	103.35
ZSwing_pct_sc	0.60	0.62
CSW_pct	0.28	0.29

Table 5. Role Change Pitchers' Statistics

The leading predicting statistic was weighted fastball runs per 100 pitches (wFB_per_c), which measures the impact of a pitcher's fastball on opponent runs. Cobb prevented 0.14 runs and Miller prevented 3.41 runs per 100 pitches, demonstrating their fastball could consistently prevent runs. Another predictive statistic is stuff (from Pitching+, Stuff_plus), which considers the physical characteristics of a pitch, like velocity, spin rate, movement, and release point. According to FanGraphs[1], a reliever transitioning to a starting role should see his stuff drop 5.5 points. From the data provided, the average stuff rating for pitchers was 94.71. If Cobb and Miller were to transition roles, their stuff ratings would both be above the average, thus making great relief and starting pitcher candidates.

Conclusion

Our current model demonstrates that using data provided by FanGraphs, we can accurately and validly identify players that may be more successful in transitioning roles from

[1] Appelman, David. "Pitch Type Linear Weights." FanGraphs Baseball, May 20, 2009. https://blogs.fangraphs.com/pitch-type-linear-weights/.

reliever to starter and *vice versa*. Additionally, this model can be modified to cluster and rank players for various cases, such as role, performance, or health. Future models should utilize other normalization techniques and should examine a model and develop a more robust PCA scoring system with the intention to capture more than WAR and Dollars as metrics. Overall, our model demonstrated a promising approach to identifying pitchers who could benefit from changing roles and could be used for further analysis to identify other key metrics from the professional pitching population.