Analyzing the Strike-Zone: Using K-Means Clustering to Recognize Pitching Patterns

Scott Einsidler, Finn Pounds, Ryan Eccleston-Murdock, Cris Pezoa

19 December 2019

1 Abstract

With data science on the rise over the past couple of decades, baseball organizations have become eager to utilize new quantitative analysis to guide their decisions. Of all aspects of a game of baseball, pitching is one of the most readily analyzable. Understanding what makes a pitcher effective is not only vital to any baseball organization; it can also be accomplished given the right information. While every pitcher has their own repertoire and pitching style, we would expect that we could find useful information that can further elucidate why certain pitchers are effective and why others are less so. In this paper, we take available pitching data (collected by the MLB) for various pitchers over the course of a single season, and utilize clustering algorithms (K-means) to tease out trends that support current known statistical facts about baseball, and make our own predictions as well.

When trying to interpret any set of data, an immense amount of different tools are at our disposal: clustering methods, separating hyperplanes, dimensionality reductions, etc... In this case, we chose K-means clustering and employed this on a variety of different fields of data. Rather than exploring all of these methods to understand the high dimensional data set we were given (more on what this set looks like in a later section), we want to explore the usefulness of this method of interpretation and see what information it can (and cannot) reveal to us.

2 Data Preparation and Methodology

In the MLB, two systems of data collection exist: Pitch f/x and Statcast. There is an ongoing debate within the sabermetrics community as to which one is superior. Pitch f/x was implemented in all professional stadiums in 2006. It utilizes three permanent cameras to accurately catch data for each individual pitch. In 2015, Statcast was installed in all MLB stadiums, using a combination of cameras and radars. While Statcast is able to capture more data, such as spin rate, some experts say that Pitch f/x is more accurate with certain position measurements. In our analysis, we used both systems depending on what we were looking for.

Given a large set of data, there are countless ways to pick out underlying structure. In our case, we considered which method would be most readily applicable to our given data. K-means clustering seemed to be the most natural choice. Initially, we thought this because we expect a pitcher to throw to certain parts of the strike zone more than others, and thus, clusters would emerge. Later, we realized that K-means was also appropriate because we would expect pitch types to behave similarly with respect to various other pitch attributes. An example of this is vertical acceleration: we expect a curve-ball (a pitch that is thrown with the intention of breaking significantly in the vertical direction) to have a higher vertical acceleration than a fastball (which has very little break in the vertical direction). Thus, the curveball cluster would be taller than it is wide, and the fastball cluster would be smaller and far less ellipsoidal (as opposed to circular). By characterizing the data in such a manner, we can see fundamental differences between pitch types, and also examine how effective a pitcher is at throwing certain pitches.

In an attempt to arrive at the most accurate findings we could, we took a look at data from a variety of starting pitchers, both "good" and "bad", who specialize in various kinds of pitches. It is important to note that, between hundreds of fundamentally differing pitch repertoires, no kind of pitcher is more effective than another. Every season, successful pitchers rise to the top, but their success depends more on consistency than some special recipe. The effectiveness of each kind of pitch is something that can be analyzed and quantified, as we shall now demonstrate.

3 The Ballmeans Algorithm

As we aim to glean insight on the clustering of various potantially related baseball measurements, we felt it natural to construct an algorithm to allow for easy comparison between them. Our algorithm, Ballmeans, receives a pitcher's name, a file containing their statistics, column numbers delineating the measurables, and a number k of expecting clusterings. Ballmeans then calls the MATLAB function kmeans, running a k-means algorithm on our specified data. Such an algorithm works in the following way: By iteratively updating centroids of the proposed clusters $\{\mu_j\}_{j=1}^k$ and data assignments, given clusters $(C_1, \ldots, C_k)$

1. Compute $\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$
2. Compute $y_i = \text{argmin}_{j=1,\ldots,k} \|x_i - \mu_j\|_2^2$, for $i = 1, \ldots, n$
3. Assign clusters to be $C_j = \{x_i | y_i = j\}$

The remainder of the algorithm creates a legend, referencing the seven pitch types, and plots the data along with its clusterings.

4 Limitations of K-Means

The drawback of this algorithm is that it is heavily reliant on K-means which as two pitfalls: Locating a minimum may only yield a local minimum as the function to be minimized $F(C_1, \ldots, C_k) = \sum_{j=1}^k (\sum_{x_- \in C_j} \|x_i - \mu_j\|_2^2)$ is not convex. In addition, the effectiveness of this method is contingent on the initial guess of $\mu_1, \ldots, \mu_k$ and the consequent clusters $(C_1, \ldots, C_k)$. For this reason, running the algorithm multiple times may result differently, and drastically different if the data is not clearly clusterable.

5 Assumptions

During the formulation of our hypothesis, we made a critical assumption: a lower batting average against (BAA) is correlated to a better pitcher. BAA is defined to be the batting average of opponent batters against a particular pitcher. Batting average against is calculated by dividing the number of hits against a pitcher over the total number of at bats against him [2]. This assumption is valid because inherently a pitcher that allows less hits will give his own team a better chance to win. Thus, it is clear to see why a lower BAA would correlate to a more skilled pitcher.

6 Analysis: Spin Rate and its Effectiveness on Fastballs and Breaking Balls

One result we found through our analysis was that the ball's angular velocity often seems to determine what makes certain pitchers more effective at not letting up runs. Physically, this

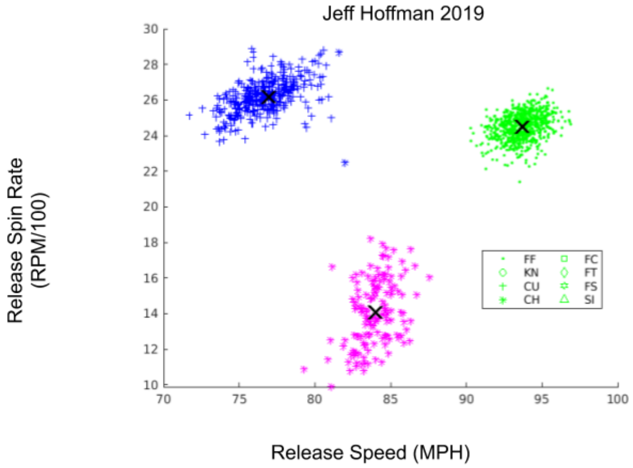
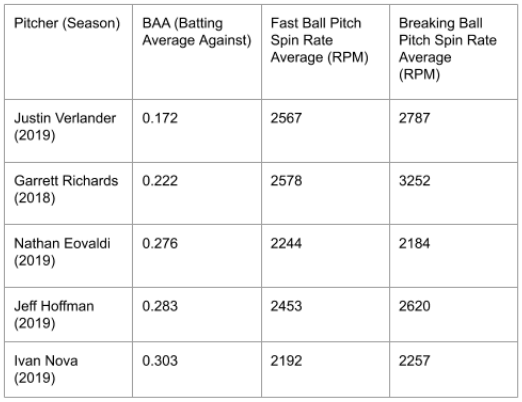


Figure 1: Spin Rate vs. Release Speed of Jeff Hoffman's 2019 season

can be explained by a physical phenomenon called the Magnus effect. The Magnus effect occurs when a spinning ball approaches terminal velocity through a fluid. The surface drag of the ball creates an airflow imbalance, which leads to a pressure difference in the air around the ball. Thus, by conservation of momentum, the ball experiences a constant force towards the side that is spinning with the perceived wind. In the case of a four-seam fastball, the backspin imparted by the pitcher forces the ball to lift as it approaches the plate, keeping the ball higher away from the ground (than a pitch with no backspin would be). A similar effect happens with curveballs, where the ball dips down as a result of topspin.

While physics can explain the path a baseball takes toward the plate, it alone cannot predict whether this effects how important spin rate is to a pitcher's success. Our hypothesis before we did this analysis was that pitchers that threw curveballs, sliders, and fastballs with high spin rates would be more effective pitchers than those with lower spin rates because the amount of variation in the paths of each pitch would be greater, making it harder for a batter to predict where each pitch is going to go.

To test our hypothesis, we took single season data for pitchers that threw curveballs, sliders, and fastballs, and used our algorithm to associate spin rate with effective speed. In the pitchers we analyzed, a fast speed pitch cluster usually took form, consisting mostly of four seam fastballs for most pitchers. A breaking ball pitch cluster also formed, consisting of curveballs and sliders. Figure 1 is a graph of these clusters for Jeff Hoffman's 2019 season. It is important to note that the unit for spin rate is $\frac{RPM}{100}$. This scaling was in order to mitigate the influence spin rate would have on the clusters. If we did not scale the data, spin rate would have a much higher influence over the clusters than speed would because the units of RPM have a much larger range (from 1000 to 3000, potentially) than speed does (from 70 to 100). This greater effect on the euclidean distance unnecessarily complicates the calculations K-means depends on. Though dividing spin rate by 100 is not a perfect solution, it was enough mitigate scale skew in the clusters of pitchers we examined enough



Pitcher (Season)	BAA (Batting Average Against)	Fast Ball Pitch Spin Rate Average (RPM)	Breaking Ball Pitch Spin Rate Average (RPM)
Justin Verlander (2019)	0.172	2567	2787
Garrett Richards (2018)	0.222	2578	3252
Nathan Eovaldi (2019)	0.276	2244	2184
Jeff Hoffman (2019)	0.283	2453	2620
Ivan Nova (2019)	0.303	2192	2257

Figure 2: Table of six pitchers comparing centroid positions and BAA

for our analysis.

After running this on several pitchers, we noticed that ones with lower Batting Average Against (BAA) - a baseball statistic determined by the number of hits batters get against a pitcher over the total number of at bats - tended to have centroids in the breaking ball and fast speed pitch clusters. Figure 2 is a table showing this result for several pitchers.

For all besides Garrett Richards, we can see by this table that this trend holds true. While this relationship between higher spin rates and batting average against is not a perfect correlation and can be only applied to pitchers with a specific (and ideally narrow) repertoire of pitches, we can see that it occurs a surprising amount of the time.

7 Further Research

Regarding the baseball data deluge, many different angles of study point towards potentially worthy results. Even just using K-means, we could apply the model we used to analyze the sequence of pitches in each game to recognize pitch patterns the batter might rely on, or subconsciously tend towards. Within the clusters themselves, a Principal Component Analysis (PCA) technique could reveal the principal direction in which pitches vary. This could be a critical tool for batters to align their swings with the variance patters of specific pitchers, or to supplement the training tools available to a professional pitcher.

Currently, the MLB does not disclose the research they are doing as to not give teams an advantage over their opponents. This might suggest that, given these contemporary data clustering tools, a team's data is nearly as valuable as a pitcher's list of secret signals, or some similarly compromising piece of information. Some teams, however, do not dedicate resources to this mathematical analysis of the data, and instead rely more on traditional techniques to make their prediction-dependent decisions. Through our exploration of clustering algorithms' insights into pitching data, one could attempt to use this information to determine whether players are overvalued or undervalued, which would help determine how to assemble a team in a cost-effective manner.

Teams could very well attempt to get more measurements on each pitch. As increasingly accurate pitch information becomes available, more patterns and trends can be analyzed to further understand the statistics behind what makes a pitcher skilled. At some point, we might have data that fully encapsulates the path that each pitch takes, as precisely as possible, from release to the plate. Having this available would unveil myriad aspects of the mechanics of pitching, which could lead to a greater understanding as to why certain pitchers outperform others.

8 Conclusion

With the aid of our baseball-leveraged data clustering algorithm, Ballmeans, we were able to produce rather fruitful results. Our hypothesis proved to hold relatively true: pitchers releasing the ball with a higher spin rate struck out more batters. This conclusion, though, only points to the many possible applications of newer data science methods to the sport. Our clustering function takes multiple columns of a pitching data sheet as an input, yielding a wealth of potentially interesting combinations. So, beyond our prediction that attributes the success of these various pitchers to the physical phenomenon the Magnus effect, we also predict significant changes in the field of baseball statistics due to these emerging tools and the relationships they are capable of revealing.

9 References

1. https://baseballsavant.mlb.com/statcast_search
2. http://m.mlb.com/glossary/standard-stats/batting-average