Using Spatial Methods to Maximize Velocity and Optimize Spin In Pitching

Samuel Kim

Abstract

The advancement of data and technology in baseball has presented analysts and coaches with a wide variety of information that provides an unprecedented opportunity to fine-tune training regimens and improve player performance during games. While this large influx of data in recent years provides great flexibility in making adjustments that can facilitate player development, the challenge now lies in determining what changes to prioritize when developing a player. Due to the many different types of pitchers in the MLB, some may want to prioritize maximizing velocity, while others may want to focus on the movement of their pitches by finding a way to optimize the spin they put on the ball. Therefore, by using areal and point-reference spatial methods, analysis will be done on how one can maximize velocity, along with analysis on how one can optimize their spin to create the pitch movement they desire.

Introduction

Why We Want To Maximize Velocity

As we look around in the world of sports, there are clear tangibles that analysts can look at to get a good idea of how effective the player can be. For wide receivers in football; speed, along with the ability to change body movements quickly is key. In basketball, the ability to make shots, dribble, and pass separate the great point guards in the league from the rest. In swimming or running, the quickness and stamina of the athlete often defines which athletes get the best times. In baseball, a key tangible that is often looked at for pitchers is how hard the pitcher can throw the ball; specifically, how hard the pitcher can throw his fastball. Even just last year in 2022, the two relief pitcher of the year award winners — Edwin Diaz and Emmanuel Clase[1] — showed the importance of throwing hard as they both averaged over 99 MPH on their fastballs.[2] The importance of velocity in a fastball to the success of a pitcher makes sense, after all, the faster a ball reaches the hitter, the less time the hitter has to react to that ball. Take for example a baseball that's thrown at 90 MPH, which is around 3.9 MPH slower than the average Major League Baseball fastball.[3] The 90 MPH pitch reaches the batter from the pitcher's hand in approximately 400 milliseconds. Given that it takes around 150 milliseconds for a Major League Baseball player to swing a bat, this gives the hitter 250 milliseconds to identify where the ball is going, and decide whether they should swing or not. For reference, an average human blink is around 300 milliseconds, which means that a hitter has less than a blink of an eye to react to a below average fastball in terms of velocity.

In a study done by numerous medical specialists and the Orthopedic Surgery and Biostatistics department at Wake Forest, researchers used mixed-model analysis on various professional pitchers to identify how changing certain arm mechanics, particularly, how hard one decides to torque their elbow and shoulder rotation led to an increase in risk of arm injury.[4] In a separate study with the same pitchers, researchers were also able to identify an association between an increasing pitch velocity and an increased risk of arm injury.[5] Therefore, using these two studies, although it may increase injury risk, one can intuitively identify how an increase to the force of torque on an elbow or shoulder rotation can increase pitch velocity. Going off of this study, I will aim to analyze another large factor in arm mechanics: where the pitcher decides to release the ball. I was only able to find one similar study that analyzed release points, which was done by numerous members of Japanese institutions. In the study, researchers used paired t-test analysis to find a statistically significant difference in 3-D release points (x, y, and z coordinates of where the ball was released) between fastballs and sliders, along with fastballs and curveballs.[6] However, this study failed to use spatial analysis on these coordinates, and only looked at the type of pitch, not the velocity of the pitch. When purely looking at velocity, pitch types will not be needed in the analysis, as the velocity of a pitcher's other pitches are largely influenced by the speed of their fastball since they will try to throw all their pitches the exact same way to avoid giving the hitter any hints as to which pitch is coming.

Therefore, we ask the question: is there a certain release point of a fastball that will help maximize the velocity of the pitch? In other words, do pitchers that typically throw the ball fast have similar release points to other pitchers that also typically throw the ball fast?

Why We Want To Optimize Spin

In addition to the aforementioned heightened injury risk of throwing baseballs at high velocities, there are also genetic limitations that do not allow for some pitchers to throw as hard as others. Instead of only targeting pitchers who throw hard; analysts, coaches, and the pitchers themselves found creative ways to outsmart hitters, instead of simply overpowering them with high velocities. Although it only takes around 400 milliseconds for even a below average 90 MPH fastball to reach the hitter from the pitchers hand, MLB hitters are good enough to detect the spin of the ball to help identify the type of pitch, along with the likely movement that the pitch will take to anticipate where they should swing.[7] This is because when the ball is in flight, it experiences a force called Magnus force, which is perpendicular to the spin axis of the baseball and helps push the ball a certain direction.[8] For example, a typical 4-seam fastball is forced down by gravity with no movement, due to being thrown with a focus on backspin. On the other hand, a slider is thrown with a focus on side spin, allowing the baseball to move sideways. Even in the little time they have to react, MLB hitters can use the direction of the spin to discern between the fastball and slider, helping them predict where the ball will end up.

Although the impact that the direction of the spin had on the movement of the baseball was intuitively understood through physics, there was still a lack of concrete data that could lead to any large discoveries in the movement of a baseball. That is, until 2015, when Major League Baseball revolutionized the application of statistics in baseball with the implementation of Statcast. By using a combination of radar technology and high speed cameras, Statcast is able to record around seven terabytes of different player and baseball movement data each game that were previously unable to be obtained.[9] With a large influx of newfound data, physicists and statisticians in baseball were able to find another factor other than the direction of the spin that contributed to the movement of a baseball pitch: spin rate; which is how fast the ball spins through the air, measured in Revolutions Per Minute. Using spin rate as a test case, researchers were able to find that an increased fastball spin rate led to a pitch that dropped less than a pitch at the same velocity and spin direction with lower spin rate.[10] However, the rate at which the ball was spinning was not the only new discovery; physicists also discovered an additional side force on the baseball, which is a result from the asymmetric flow separation caused by the odd orientation of rough seams on the baseball. In other words, there was spin that contributed to the movement of the baseball (active spin), in addition to spin that actually changes the direction of the spin during the flight (gyro spin).[11] This phenomenon was coined “seam-shifted wake,” and it wasn’t long before analysts, coaches, and pitchers decided to experiment with different orientations of the seams at the point of release to get the ideal combination of active and gyro spin to give the pitch a sense of unpredictability to the hitter. It is important to note that there is no universal way to make a ball move a certain way due to the differing factors in a pitcher, such as their force of torque in their elbow and shoulder or their release point; therefore, finding the ideal combination of active and gyro spin for a pitcher takes a lot of trial and error using high speed cameras to record the spin accurately.

A prime example comes from Atlanta Braves pitcher Max Fried, who finished second in the National League Cy Young Award (given to the best starting pitcher) race in 2022.[12] Despite his success, Fried finished in the 56th percentile in fastball velocity, averaging 94 MPH on his fastball. Instead of relying on his non-overpowering fastball, Fried was one of the first pitchers to rely on seam-shifted wake to confuse hitters. In addition to his 4-seam fastball, Fried also throws a sinker, which is a variation of a fastball, where the direction of the spin axis is slightly tilted, creating a greater magnus force downwards and sideways than a typical 4-seam fastball, allowing for downwards and sideways movement of a fastball. Typically, pitchers will force this tilt in the spin direction of the baseball; however, Fried threw both pitches with less than 75% active spin, allowing for a change in direction in the air. As a result, Fried's fastball and sinker on average started at a nearly indistinguishable spin direction of 195 and 187.5 degrees, respectively (pure backspin is 180 degrees), but ended up at 187.5 and 217.5 degrees, respectively by the time the pitches reached the plate.[13] Due to the change in spin, Fried's 4-seam fastball had 18.5 inches of downwards movement with 1.7 inches of side movement, while his sinker had 21.6 inches of downwards movement with 11 inches of side movement. This means that these two pitches had more than a foot of difference in total movement, even though the spin was almost identical out of Fried's hand to the hitter.

The optimal amount of active spin for a certain pitch largely depends on that pitcher's initial spin direction of their pitches, along with how much they emphasize disguising each pitch as Max Fried did with his fastball and sinker as opposed to simply creating the most movement on their pitches. Therefore, instead of attempting to determine the best combination of active spin and gyro spin for each pitch, it is of interest to understand how the amount of active spin affects the movement of each pitch. This allows for more flexibility of analysts, coaches, and players, as they can mix and match different amounts of active spin for pitches to create the most optimal combination of disguising pitches and applying movement to the pitches.

Data

For both analyses, much of the data was collected through back end web scraping of Baseball Savant,[14] which is where all publically available data recorded by Statcast exists. The first dataset contained a row for every pitch thrown in the 2022 MLB season, the last full season that took place at the time of this paper. For each pitch — of which there were 709,580 — there were 99 variables that included information such as the outcome of the pitch, the velocity of the pitch, movement data of the pitch, the release point of the pitch, and batted ball metrics of the pitch if the hitter was able to hit the pitch into play.

Fastball Velocity Data

Using this initial dataset, a new dataset was created to help answer whether the velocity of a fastball can be maximized using the release point of the pitch. Over the course a baseball season, the amount of fastballs thrown by a pitcher will vary largely on the pitcher himself; therefore, in the aim of avoiding any issues with the count of fastballs a pitcher threw; the average fastball velocity, along with the average x and y release point were recorded for all pitchers who threw over 100 fastballs in 2022. Furthermore, I will only be analyzing right handed pitchers to account for the naturally different release points that a left-handed pitcher will have, due to throwing the ball from the other side of his body. It is important to note that pitchers can alter the direction of spin of the baseball to create three variations of fastballs: the 4-seam fastball, sinker, and cutter. The sinker will typically move downwards towards the arm-side of the pitcher (right for right-handed pitchers), the cutter will typically move downwards towards the glove-side of the pitcher (left for right-handed pitchers), and the 4-seam fastball will typically move straight relative to the other two pitches; however, it will not have as much of a drop downwards as cutters or sinkers. There are some pitchers who actually throw multiple variations of the fastball; however, to account for any differences in velocity and release points among variations of the fastball, each fastball variation will be treated as their own pitch. For example, if a pitcher threw a 4-seam fastball over 100 times, while also throwing a sinker over 100 times in the 2022 MLB season, the average velocities and average release points for both the pitcher's sinker and fastball will be placed in the dataset as separate rows. After accounting for all of the aforementioned caveats, there were 658 rows in our dataset, with each row representing a fastball that was thrown by a right-handed pitcher over 100 times throughout the entire 2022 MLB season.

Although the average x and y release points can be treated as a spatial object for our response variable, the average velocity of the fastball is not a metric that can be analyzed using spatial methods. Therefore, we will transform the dataset into an areal dataset by using "pseudo" spatial variables. The average release points for each fastball will be laid out on the data space, with 17 different grids evenly also laid out among the data space. Each average release point will then therefore be placed into a grid that have other similar average release points, creating areal units that contain the count of average release points in each unit, along with the average velocity in each unit. The number of different grids is entirely arbitrary; however, after extensive analysis on how changing the amount of grids would affect how many average release points were placed in each grid, along with making sure that average release points in the same grid were not too dissimilar, I believed that 17 grids brought the most balanced areal dataset. Of these 17 grids, only 13 of the grids contained release points; therefore, these 4 grids will be taken out of the dataset as there is no provided velocity information at these units. The amount of grids is a very large caveat in our analysis, and further discussion will take place in the discussion section below. Table 1 below displays the summary statistics of this newly created areal dataset. It is important to note that the grid number is numbered based on the original 17 grids.

Table 1: Summary Statistics of Spatial Grids

Grid #	Min X Release (Feet)	Mean X Release (Feet)	Max X Release (Feet)	Min Y Release (Feet)	Mean Y Release (Feet)	Max Y Release (Feet)	Min Velocity (MPH)	Mean Velocity (MPH)	Max Velocity (MPH)	Count In Grid
1	-3.45	-3.43	-3.40	6.07	6.09	6.11	92.42	92.52	92.62	2
2	-3.03	-2.42	-2.16	5.86	6.10	6.50	86.49	93.06	98.91	56
3	-2.13	-1.54	-0.91	5.85	6.17	7.00	84.29	93.48	100.81	209
4	-0.89	-0.57	0.16	5.89	6.33	7.04	86.36	94.00	99.65	39
5	-3.66	-3.54	-3.42	4.65	5.06	5.47	93.38	94.51	95.65	2
6	-3.35	-2.55	-2.14	4.63	5.42	5.85	85.34	93.30	99.08	133
7	-2.13	-1.66	-0.90	4.84	5.53	5.85	85.53	93.66	100.20	183
8	-0.88	-0.69	-0.12	5.47	5.82	5.69	91.40	94.00	95.40	11
9	-4.37	-3.82	-3.38	3.50	3.99	4.50	89.00	90.92	93.09	9
10	-3.34	-2.76	-2.15	3.52	4.16	4.60	85.96	90.83	95.14	9
13	-4.11	-3.81	-3.50	2.13	2.74	3.35	86.44	88.81	91.18	2
14	-2.47	-2.47	-2.47	2.70	2.70	2.70	83.62	83.62	83.62	1
17	-4.12	-4.10	-4.09	1.17	1.63	2.09	83.23	84.86	86.49	2

Pitch Movement Data

Unfortunately, although the x and y direction of a pitch's movement was recorded at a pitch level, the amount of active or gyro spin the pitch had was not a metric in the dataset that contained every pitch thrown in 2022. However, Baseball Savant was able to provide a separate dataset of the percentage of active spin a pitcher was able to average in each of his pitches over the course of the 2022 season. This dataset contained active spin information about 7 different pitches; 4-seam fastballs, sinkers, cutters, splitters, changeups, sliders, and curveballs. Since the type of pitches each pitcher throws is different, we aim to estimate how the amount of active spin in a pitch will affect the movement of each of these pitches, as there is no pitch that is “more important” than the others for movement. It is also important to note that the only publicly available pitch movement data at a pitch level considers the movement of the pitch if gravity were to be removed. However, even though the vertical movement would be higher for all pitchers without gravity (since a ball will not drop as fast); the difference in movements between pitchers does not change, allowing us to still analyze how different active spin values affect movement based on direction.

As mentioned in the introduction, previous studies have shown that spin rate also contributes to a pitch’s movement; therefore, the percentage of active spin a pitcher was able to average on a pitch will be applied to the average spin rate of that pitch, creating a new variable that represents the amount of active spin on a pitch. Since spin rate is calculated on a very large RPM scale, this new variable will be logged to avoid issues with skewed data. I could not find an efficient and accurate way to also incorporate spin direction in this variable; therefore, further discussion on not deciding to incorporate spin direction will be discussed in the discussion section. This leads to the creation of 7 different datasets, one for each pitch, with each dataset containing the average x and y movement of the pitch, along with the amount of active spin a pitcher was able to average for that pitch. Therefore, the spatial variable of the average x and y movements of the pitch will be analyzed, with each average x and y point of movement being marked by an average active spin value. Furthermore, in order to avoid any issues of sample size and the naturally different movements between a left-handed and a right-handed pitcher, only the pitch movements that were thrown by a right-handed pitcher over 100 times were recorded. For example, for a pitcher to have his splitter movement analyzed, that pitcher would have had to throw his splitter 100 times, and be right-handed. After filtering for right-handed pitchers that threw a specific pitch over 100 times, I discovered that there were less than 150 pitchers in the splitter, cutter, and curveball datasets; so we will exclude these pitches from our analysis, citing a lack of sample size. Table 2 below displays the summary statistics for each of our datasets, where HM represents horizontal movement and VM represents vertical movement.

Table 2: Summary Statistics of Each Pitch Type Dataset

Pitch Type	Amount of Pitchers	Min HM (Feet)	Mean HM (Feet)	Max HM (Feet)	Min VM (Feet)	Mean VM (Feet)	Max VM (Feet)	Min Log Active Spin (RPM)	Mean Log Active Spin (RPM)	Max Log Active Spin (RPM)
4-Seam Fastball	346	-1.44	-0.61	0.25	-1.09	1.33	1.74	6.89	7.61	7.80
Slider	276	-0.86	0.50	1.69	-0.60	0.16	1.18	5.21	6.69	7.80
Sinker	213	-1.65	-1.26	-0.38	-0.75	0.70	1.61	7.19	7.53	7.75
Changeup	160	-1.61	-1.20	-0.44	-0.71	0.50	1.19	6.68	7.36	7.77

Methods

How We Will Maximize Velocity

Now that we have an areal dataset represented by our created grids, we will analyze the Moran’s I — a coefficient of spatial autocorrelation (Paez, Chapter 23.8) — on the residuals of a linear model that aims to predict the log count of pitch release points in each grid based on the mean fastball velocity in each grid. By running a Monte Carlo simulation – a permutation bootstrap test, in which the observed values are randomly assigned to our grids, and Moran’s I is computed a set number of times (50,000 in our case) – we are able to receive a p-value for our Moran’s I statistic, helping us determine whether our newly created areal data is spatially autocorrelated (Bivand, 289). When running our Monte Carlo Simulation, we will test out various neighborhood matrices to create spatial weighting in our areal units to ensure that no spatial patterning that remains in the residuals.

There are 3 different types of neighborhood matrices that I will apply: row-standardized, binary, and k-nearest neighbors. A binary neighborhood matrix is fairly straightforward; if a grid is adjacent to another grid, then the relationship between those two grids are given a 1, and a 0 otherwise. However, it may be a case that not all adjacent neighbors have the same impact on a unit. For example, let us imagine that unit 1 is adjacent to unit 2 and unit 3. Unit 2 has three other adjacent neighbors, and unit 3 has no other adjacent neighbors. Since unit 2 has to consider three other neighbors, and unit 3 only has to account for unit 1, then it may not be fair to say that unit 1 impacts each of its adjacent units the same. A row-standardized matrix accounts for any other neighbors a unit may have by making the weight of a relationship between two units proportional to the total number of neighbors each unit has. Looking back on our example then, the impact of neighbors for unit 1 would be 1/2 for each of its two adjacent neighbors, while the spatial weight of neighbors for unit 2 would be 1/4 for each of its four adjacent neighbors. The spatial influence of the lone neighbor for unit 3 would stay the same as a binary neighborhood matrix, as it would still be 1 (Bivand, 270). While row-standardized matrices standardize the weights between adjacent neighbors, k-nearest neighbor matrices regulate the number of neighbors a specific unit has, based on the value of k (Bivand, 268). Since there is little prior knowledge on the spatial process of our areal units, we will apply $k = 1$, $k = 2$, $k = 3$, and $k = 4$ nearest neighbors matrices, along with the binary and row-standardized matrices for extensive analysis on our data. Among these matrices, the neighborhood matrix that generates the lowest p-value for our Moran's I statistic in its Monte Carlo simulation will be used for final analysis.

If the simulation with the lowest p-value reveals that the counts of pitch release points in the grids suggest spatial autocorrelation, then this would suggest that based on the neighborhood matrix that was used, the spatial distribution of the residuals of the grids show clustering. If the Moran's I statistic is positive, it would indicate that grids with a high count of pitch release points tend to be surrounded by grids with a high count of pitch release points; however, if the Moran's I statistic is negative, it would indicate that grids with a high count of pitch release points tend to be surrounded by grids with a low count of pitch release points (Paez, Chapter 23.7). To account for this spatial structure among the residuals, there is motivation to use spatial regression instead of linear regression, due to the fact that spatial regression models will take into account the spatial autocorrelation of the residuals; in turn, avoiding any biased coefficient estimates and inefficient standard errors that may come from a model that assumes that the grids are independent of each other.

There are two types of spatial regression models; a simultaneous autoregressive model (SAR), along with a conditional autoregressive model (CAR). In our case, not only will a SAR model run regression on how average velocity in a grid would impact the log count of pitch release points for a given grid, but it would also use a regression on how the log counts of pitch release points in the neighboring grids would impact the log count of pitch release points for a given grid (Bivand, 293). A CAR model is similar; however, instead of the additional regression analysis being on the neighboring grid counts, a CAR model will run a regression on how the residuals of pitch release points in the neighboring grids would impact the log count of pitch release points in a given grid (Bivand, 298). Each of these models will be run to further confirm whether the counts of neighboring grids have an effect on the count of a grid, by running a likelihood ratio test on the λ value (Bivand, 294). If we see statistically discernible evidence that $\lambda \neq 0$, then that indicates that the counts of neighboring grids have an effect on the count of a grid. To compare the performance of each model, we will use the model's Akaike's Information Criterion (AIC), which is the weighted sum of the log-likelihood of the model, along with the number of fitted coefficients (Bivand, 298). Essentially, AIC is able to assess how well a model fits the data, while accounting for how much information is given to the model, with a lower value meaning a better fit model.

Once the "best" model is picked, the coefficients of the mean pitch velocity will be analyzed to determine what their effect on the counts of the grids is. This way, we will be able to analyze the relationship between the counts of the grids, along with the relationship between mean velocity against the count at each grid. If we find that both of these relationships are statistically discernible, we can then intuitively analyze the relationship between the grids' mean velocities.

How We Will Optimize Spin

In order to analyze the active spin at each pitch movement, I will run point-reference (or geostatistical) analysis on each pitch type dataset. In order to estimate spatial correlation in each dataset, stationary assumptions must first be made. In particular, I will assume that the process that generated the samples in our data was composed of a mean and residual, creating the function

$$Z(s) = X\beta + e(s)$$

where $Z(s)$ represents the random process, $X\beta$ represents the mean function, and $e(s)$ represents the residual at location $$s$$ (Bivand, 218). Under this assumption, if we also assume that the mean is constant, we would expect that the spatial correlation in Z does not depend on the location s; but rather, on separation distance h. Therefore, this will allow us to then apply a variogram, given by

$$\gamma(h) = \frac{1}{2}E(Z(s) - Z(s+h))^2$$

where $\gamma(h)$ represents the semivariance of $Z(s)$ and h represents the separation distance. If we hold the assumption that any spatial correlation in Z is only dependent on the separation distance h, then we can then use the variogram to measure whether pitch movements that have a lower separation distance would have similar active spin metrics. Under spatial autocorrelation in our data, we would expect an increase in semivariance $\gamma(h)$ to also be associated with an increase in our separation distance h.

In order to verify whether an increase in the data's semivariance and separation distance is not due to chance, we first must plot a variogram cloud, which is the plot of all possible squared differences of observation pairs $(Z(s_i) - Z(s_j))^2$ against their separation distance h. From here, we can then take the averages of variogram cloud values over intervals created by the distance h (Bivand 220). We then plot this sample variogram over a set number of additional variogram computations (200 in our case), of which each variogram is computed from the same data, but after randomly re-assigning active spin values to different pitch movement locations. If we analyze that our sample variogram differs from these 200 random variograms, then we can assume that our data is distinguishable from complete spatial randomness, allowing for the application of spatial prediction, which is a method to predict $Z(s)$ at unknown locations using the spatial dependence between points (Bivand, 232).

The least biased predictor of unknown locations is given by the kriging estimate

$$\hat{Z}(s_0) = x(s_0)\hat{\beta} + v'V^{-1}(Z(s) - X\hat{\beta})$$

where $x(s_0)$ represents the spatial predictors at prediction locations, X represents the spatial predictors as measured locations, $x(s_0)\hat{\beta}$ is the estimated mean at unknown location s_0, and $v'V^{-1}$ are the simple kriging weights (Bivand, 232). To produce a kriging estimate, the sample variogram once again has to be analyzed to fit a specific variogram model to our dataset. Each variogram model component has a model type; such as a linear, power, gaussian, exponential, or various other statistical distributions. The partial sill is then analyzed, which is the vertical extent of the model component; followed by the range, which represents the horizontal extent of the model component. Finally, the nugget variance of our model component may be specified if there is a need to account for any measurement errors or micro-variability in our variogram model (or in some cases, both) (Bivand, 224-225). For each pitch type dataset, three different model types will be analyzed to see how well it fits that particular dataset. For each of these model types; the partial sill, range, and nugget variance will be optimized through the trial-and-error of trying different parameter values to see which combination of parameters fit each dataset the best.

Two types of kriging estimates will be applied to each of our pitch type datasets: simple and ordinary kriging. Simple kriging assumes that β is known, which means that the estimate for $\hat{\beta}$ will not be used in the formula. When running simple kriging on our datasets, each pitch type dataset will use the average log active spin for their β value. An ordinary kriging will only apply the intercept for $x(s_0)$ and X, with no other covariate information at each point. Unfortunately, due to the nature of the dataset, it was difficult to apply any spatial covariates at each location in Z; therefore, universal kriging will not be applied to our data.

In order to validate which of ordinary or simple kriging should be the final kriging estimate applied to each of our datasets, we will run leave-one-out cross validation on each model for each pitch type dataset. Leave-one-out cross validation is the process of removing one observation from a dataset, then using the remaining data to create a model, then using the newly created model to predict the excluded observation. We can denote $Z(s_i)$ as the observed data point of the omitted observation s_i, and $Z(\hat{s}_{-1})$ as the prediction at s_i when this observation is removed. Ideally, the mean error $\frac{1}{N}\sum(Z(s_i) - Z(\hat{s}_{-1}))$ and the mean-squared error $\frac{1}{N}\sum(Z(s_i) - Z(\hat{s}_{-1}))^2$ will be as small as possible, since kriging is supposed to be unbiased. Although both metrics will be assessed, the mean-squared error will be prioritized in picking the "best" kriging estimate.

Results

Maximizing Velocity

To begin our areal analysis of maximizing velocity, we first run a Monte Carlo simulation of Moran's I on the residuals of the linear model below:

$$LogCount = \beta_0 + \beta_1 AVGVelocity + \epsilon$$

Table 3 below displays the Moran's I statistic, along with the corresponding p-value based on our Monte Carlo simulations of row-standardized and binary neighborhood matrices, as well as k = 1, k = 2, and k = 3 nearest neighbors matrices:

Table 3: Moran's I Statistic and Corresponding P-Value For Monte Carlo Simulations of Various Neighborhood Matrices

Neighborhood Matrix	Moran's I Statistic	P-Value
Row-Standardized	0.061	0.17
Binary	0.098	0.11
K = 1 Nearest Neighbor	0.86	0.0013
K = 2 Nearest Neighbor	0.12	0.19

K = 3 Nearest Neighbor	0.15	0.11
K = 4 Nearest Neighbor	0.049	0.18

The results show that the Monte Carlo simulation with the lowest p-value used a k = 1 nearest neighbor neighborhood matrix, which produced a Moran's I statistic of 0.86, along with a p-value of 0.0013. Based on these two statistics, there is indication that there is statistically discernible evidence of clustering in the spatial distribution of the residuals in our linear model. Furthermore, since the Moran's I statistic is positive, this also indicates that high counts in grids tend to be around grids that also have a high count of release points. All the other Monte Carlo simulations that used different neighborhood matrices did not reveal statistically discernible evidence of clustering in the spatial distribution of the residuals in our linear mode; therefore, we should take the results of any further analysis with the k =1 nearest neighbor neighborhood matrix with caution.

Now that we have our desired neighborhood matrix, we can apply it to a CAR and SAR model for AIC comparison. The SAR and CAR models test the following hypothesis test:

$H_0 : \lambda = 0$ vs. $H_A : \lambda \neq 0$

and table 4 displays the λ values, p-values, and AIC values for each of our spatial regression models, along with the AIC value of a non-spatial linear model.

Table 4: Summary Statistics For Our Linear and Spatial Regression Models

Model	λ	P-Value	AIC
Linear	N/A	N/A	52.6
SAR	0.63	0.0013	44.28
CAR	0.60	0.036	50.19

With a lower AIC value than the other two models, the SAR model will be analyzed for final analysis. For the SAR model, a λ value of 0.63 and a p-value of 0.0013 indicates that there is statistically discernible evidence that high counts of neighboring grids have a positive effect on the count of a grid. Table 5 shows the estimated coefficients of the following final model:

$$LogCount = \beta_0 + \beta_1 AVGVelocity + \lambda W(LogCount - \beta_1 AVGVelocity) + e$$

where W is the k = 1 nearest neighbor neighborhood matrix and e represents the error terms in our units.

Table 5: Summary Of $$\beta$$ Coefficients In Our Final Model

	Coefficient	Standard Error	P-Value
$\beta_0(Intercept)$	-22.86	7.65	0.0028
$\beta_1(AVGVelocity)$	0.28	0.085	0.0011

After fitting a SAR model, the coefficient for average fastball velocity is 0.28, with a p-value of 0.0011, which means that average fastball velocity is statistically discernible to the log count of pitch release points in our created grids, even after considering the spatial clustering of the residuals in our linear model. Since the coefficient is positive, for each grid, as the log count of pitch release points goes up, the average fastball velocity in that grid also increases. Due to the nature of baseball being a billion dollar industry, teams are not willing to reveal any information publicly; often asking front-office and coaching staff employees to sign a non-disclosure agreement that states that they cannot reveal any new information discovered at work. As more technology and data becomes available, analysts and coaches likely push pitchers to change their mechanics — such as their release points — to replicate success that other pitchers in the MLB have experienced with certain mechanical changes. Therefore, although there is not a lot (if any at all) literature on the relationship between a release point of a fastball and its velocity; with baseball team employees dedicating their lives to find even the smallest advantages, it would intuitively make sense that there would be more pitchers that throw from certain release points that allow them to throw their fastball harder.

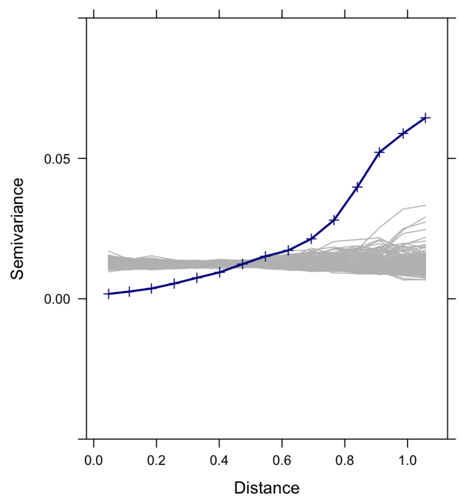
Optimizing Spin

For each pitch type dataset, similar steps were taken to analyze how the amount of active spin in a pitch impacts how the pitch moves, as outlined in our methods section. Interpretation of complete spatial randomness and the choice of the variogram model applied to the kriging estimates were the same among all pitches; therefore, extensive analysis on testing against complete spatial randomness and choosing the variogram model will be written on the 4-seam fastball dataset, but removed in the individual pitch subsections to avoid redundancy.

Complete Spatial Randomness and Variogram Model Choice

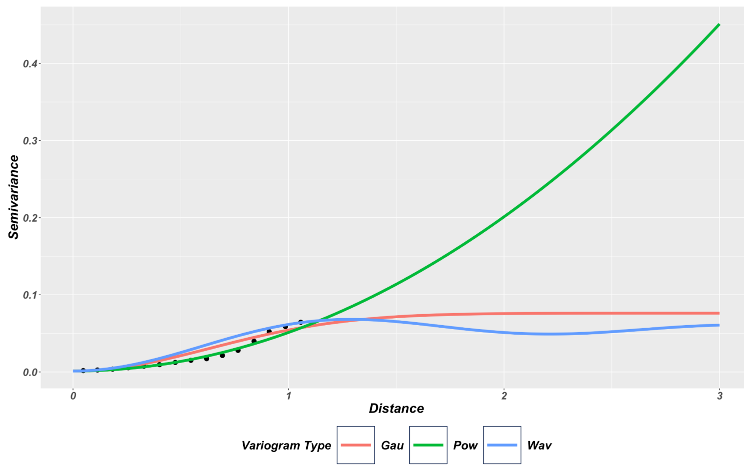
To begin our analysis of how the amount of active spin in a 4-seam fastball impacts how it moves, we analyze how the sample variogram of our 4-seam fastball dataset differs from 200 random variograms that were computed by randomly shuffling active spin values to different pitch movements in our 4-seam fastball dataset. Figure 1 below shows a graphical representation of this comparison:

Figure 1: Graphical Representation Of The Comparison Between The Sample Variogram Of Our 4-Seam Fastball Dataset and 200 Random Variograms



For the figure above, the blue line represents the sample variogram for the pitch movements and active spin values in our 4-seam fastball dataset, while the gray shadow represents the “envelopes” of our simulation of the 200 random variograms. These envelopes essentially account for the variability of our simulation, and allows for clear comparison between the sample and random variograms. Since the sample variogram does not consistently fall within the simulation envelopes, there is evidence that the active spin values in our 4-seam fastball data are distinguishable from complete spatial randomness, allowing us to continue with the spatial prediction of pitch movements and their corresponding active spin values in our 4-seam fastball dataset. The three three model types that initially fit the sample variogram of our 4-seam fastball dataset the best were wave, gaussian, and power. Figure 2 compares each of these fits to the sample variogram.

Figure 2: Comparison of Wave, Gaussian, and Power Model Types to 4-Seam Fastball Sample Variogram



Although a wave or power model type appears to fit our sample variogram well, there is an issue with applying a variogram model other than a gaussian in the context of each of our datasets. With pitch movement based on physics and spin, there is a limit to how much spin can contribute to a pitch's movement. As mentioned before, baseball analysts, coaches, and pitchers dedicate their lives to find every little advantage they can find in a baseball game. With the importance of spin universally known in the baseball industry, along with the fact that MLB pitchers are the best pitchers in the world, I will hold the assumption that the reason there are no higher active spin values or pitches that move more is due to the fact that it is likely physically impossible to produce more active spin or movement in a pitch. Even if it was physically possible to produce more movement in a pitch based on the amount of active spin, I will also hold the assumption that it has been tested and proven ineffective; because if it was effective, it would likely show up in our data.

Therefore, since I am assuming that effective pitchers will not have higher active spin values or movements than in our data, it is unfair to assume that increasing active spin values even further will produce the same distribution of distances between pitch movements. This immediately eliminates a power model, as it assumes that as you increase the distance between two pitch movements even further than what our sample variogram provides, the semivariance of the active spin will also increase. A wave model is likely the second best fit; however, under the assumptions that I mentioned, a distribution that levels off after a certain distance, such as the gaussian distribution is determined to be the best fit for each of the pitch type datasets. Table 6 below displays the model component parameters for the variogram model that is applied to the kriging estimated for each of our pitch type datasets.

Table 6: Model Component Parameters For Our Final Variogram Models

Pitch Type	Model Type	Partial Sill	Range	Nugget Variance
4-Seam Fastball	Gaussian	0.070	0.90	0.0012
Slider	Gaussian	0.290	0.60	0.033
Changeup	Gaussian	0.045	0.45	0.0064

Sinker	Gaussian	0.014	0.40	0.0032

The parameters for each of these model components will be applied to their respective dataset for kriging estimation.

4-Seam Fastball

Based on the average logged value of active spin in our 4-seam fastball dataset, a value of approximately 7.61 was applied to our simple kriging estimate as the β value, while the $\hat{\beta}$ value was estimated based only the intercept for $x(s_0)$ and X in our ordinary kriging estimate. Table 7 shows the mean error, along with the mean-squared error of our kriging estimates after leave-one-out cross validation.

Table 7: Mean Error and Mean-Squared Error Of Kriging Estimates For 4-Seam Fastballs After Leave-One-Out Cross Validation

Kriging Estimate	Mean Error	Mean-Squared Error
Simple Kriging	-0.00044	0.0026
Ordinary Kriging	0.00013	0.0025

With a mean error and mean-squared error closer to 0, the ordinary kriging estimate will be the final kriging estimate applied to our 4-seam fastball dataset. Figure 3 displays the predicted log active spin based on the horizontal and vertical movement of a 4-seam fastball, while figure 4 shows the variance in the predictions of the log active spin based on the horizontal and vertical movement of a 4-seam fastball.

Figure 3: Predicted Log Active Spin

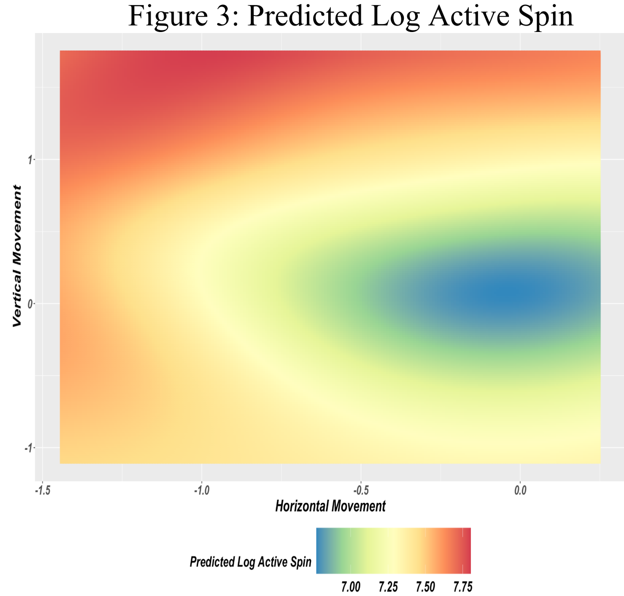
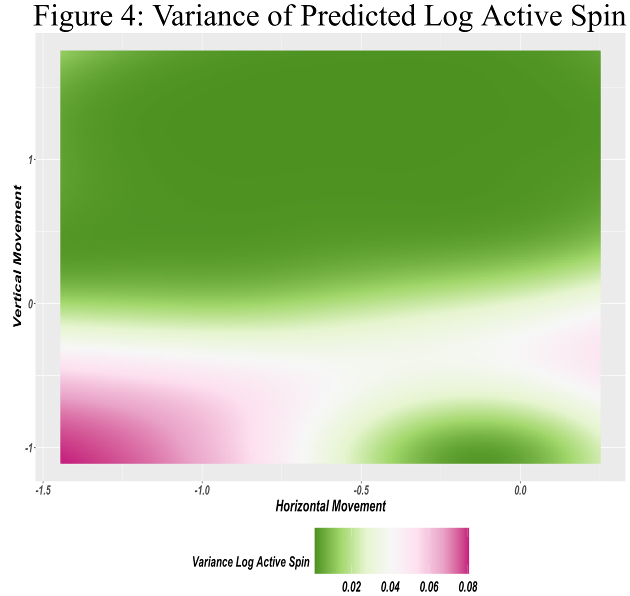


Figure 4: Variance of Predicted Log Active Spin



According to figure 3, there is a high predicted log active spin for fastballs that have high vertical movement, while there also appears to be a relatively high predicted log active spin value for fastballs that move to the left. As both the horizontal and vertical movement get closer to 0, there is a low predicted log active spin. Surprisingly, pitches that had negative vertical movements without gravity still had a relatively higher predicted log active spin than pitches that had no vertical movements without gravity. However, while also considering figure 4, there appears to be a higher variance of predicted log active spin values for negative vertical movements when compared to positive vertical movements; therefore, we should be cautious with predicted values at negative vertical movements. Among these negative vertical movements, there is a clear disparity in variance between pitches that move a foot or more to the left compared to any pitches that move to the right of that, so we should be extra cautious with pitches that have a negative vertical movement without gravity and move to the left.

4-seam fastballs are thrown with backspin, which is why intuitively, the high predicted log active spin for fastballs that have high vertical movements makes a lot of sense, since the more active spin that is in a 4-seam fastball, the more spin that is contributing to the lack of drop that comes when throwing a ball with backspin. The low predicted log active spin values when the vertical and horizontal movements are closer to 0 also intuitively make sense, as we would expect a ball with a lack of spin contributing to the movement to move as little as possible. At first, I was surprised that 4-seam fastballs with negative vertical movement had a higher predicted log active spin than 4-seam fastballs with zero vertical movement because I expected a 4-seam fastball to drop lower as the log active spin decreased. However, what has to be considered is that active spin is any spin that contributes to movement, not just spin that contributes to the desired movement of the pitch. Therefore, an explanation as to why 4-seam fastballs with relatively high predicted log active spin values had negative vertical movement is due to them not being thrown with pure backspin. It can be very difficult to perfect a 4-seam fastball with perfect backspin; so I would speculate that the 4-seam fastballs thrown with negative vertical movement were due to being thrown with more sidespin than desired, as pitches thrown with more sidespin drop faster than pitches thrown with more backspin due to physics.

Slider

As we did with the 4-seam fastballs, we applied the average logged value of active spin in our slider dataset — a value of approximately 6.69 — to our simple kriging estimate as the β value, while the $\hat{\beta}$ value was estimated based only the intercept for $x(s_0)$ and X in our ordinary kriging estimate. Table 8 displays both the mean error and the mean-squared error of our kriging estimates after leave-one-out cross validation.

Table 8: Mean Error and Mean-Squared Error Of Kriging Estimates For Sliders After Leave-One-Out Cross Validation

Kriging Estimate	Mean Error	Mean-Squared Error
Simple Kriging	0.0068	0.038
Ordinary Kriging	0.0035	0.036

Once again, the ordinary kriging will be used as the final kriging estimate for our slider data, as it has a mean error and mean-squared error that is closer to 0. Figure 5 showcases the predicted log active spin based on the horizontal and vertical movement of a slider, while figure 6 shows the variance in the predictions of the log active spin based on the horizontal and vertical movement of a slider.

Figure 5: Predicted Log Active Spin

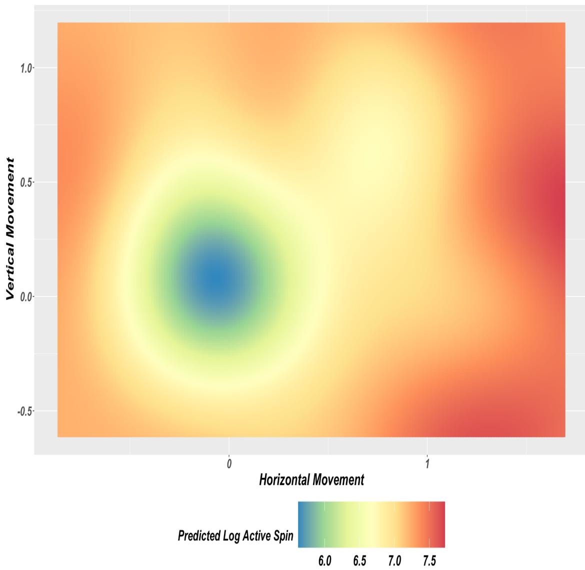
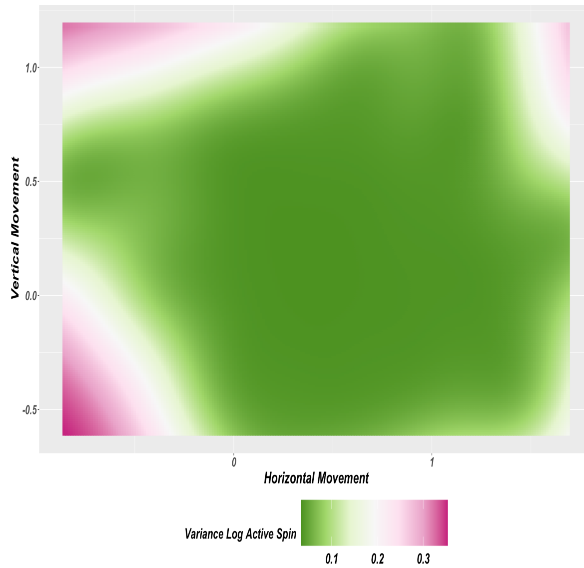


Figure 6: Variance of Predicted Log Active Spin



Referencing figure 5, there is a high predicted log active spin value as the horizontal movement for sliders moves to the right. Similar to the 4-seam fastballs, there is a low predicted log active spin for horizontal and vertical movements close to 0. Other than this however, there does not appear to be much of a pattern for vertical movements; however there appears to be a consistent high predicted log active spin for sliders that vertically move negatively, and move more than 1 foot to the right. The predicted log active spin tends to become lower as the horizontal movement moves to the left; however, there is a sudden increase in predicted log active spin for sliders with a negative horizontal movement. After looking at figure 6, this sudden increase should be taken with caution, as there is an increase in variance for sliders with negative horizontal movement, especially in sliders with negative vertical movement, or vertical movement that is higher than 1.

Instead of throwing the baseball with backspin like you would when throwing a 4-seam fastball, sliders are thrown with a focus in sidespin, which is why it intuitively makes sense that if a right-handed pitcher throws a slider with high active spin, it would move across his body more, creating more positive horizontal movement. The rest of the intuition is very similar to the intuition as to how a 4-seam fastball moves, as any pitch that has a lack of spin contributing to the movement will likely move as little as possible. Furthermore, the pitches thrown with negative horizontal movement are likely from sliders that are thrown with more topspin than side spin, making it move less horizontally. This is why sliders with negative horizontal movement still have a relatively higher predicted log active spin than sliders with no movement, as this top spin would still contribute to the movement of the pitch, even though the top spin is canceling out what a slider is “supposed to do”.

Sinker

For analysis on sinkers, we once again applied the average logged value of active spin to our simple kriging estimate as the β value. This value ended up being 7.53 for sinkers, while the $\hat{\beta}$ value was once again estimated based only on the intercept for $x(s_0)$ and X in our ordinary kriging estimate. The mean error and mean-squared error of our kriging estimates leave-one-out cross validation is showcased in table 9.

Table 9: Mean Error and Mean-Squared Error Of Kriging Estimates For Sinkers After Leave-One-Out Cross Validation

Kriging Estimate	Mean Error	Mean-Squared Error
Simple Kriging	-0.00071	0.00540
Ordinary Kriging	-0.00036	0.00544

With a mean error and mean-squared error closer to 0, the sinker dataset will also use an ordinary kriging as the final kriging estimate. Based on the horizontal and vertical movement of a sinker, figure 7 will display the predicted log active spin, and figure 8 will display the variance of these predictions.

Figure 7: Predicted Log Active Spin

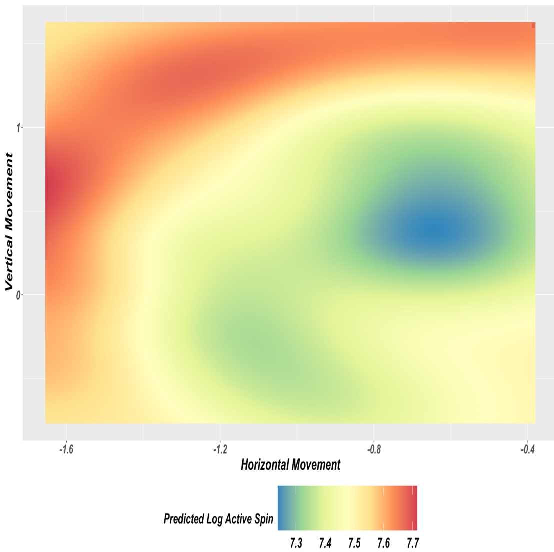
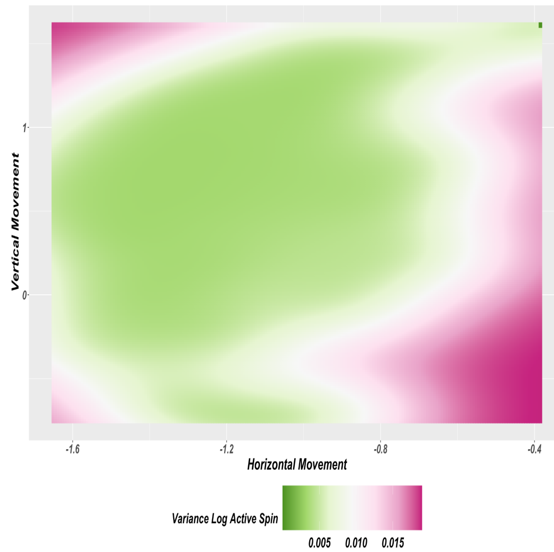


Figure 8: Variance of Predicted Log Active Spin



In figure 7, there is high predicted log active spin when the vertical movement is high; however, as movement moves more to the left, the areas with high predicted log active spin slightly drop, creating a spiral like pattern. Interestingly, there was no one in our dataset who threw a sinker with 0 horizontal movement, and the area with the lowest predicted log active spin is not when the vertical movement is at 0. Using figure 8, we can see that the highest variance is at the right edge, along with the bottom right, bottom left, and top left corners; these areas tend to correspond to areas where the predicted log active spin is neither high or low. The area with the lowest predicted log active spin as well as the areas with high predicted log active spin seems to have a relatively low variance.

One can think of a sinker as a fastball, but with a little bit more intended side spin than a traditional 4-seam fastball. The intended side spin is the opposite side spin of a slider, which is why there are only negative horizontal movements, compared to the slider that had positive horizontal movements. The slightly added side spin also makes the ball drop slightly more than a 4-seam fastball, which would make the spiral like pattern in figure 7 intuitively make sense. For a well thrown sinker, the more spin that is contributing to the movement will allow the ball to dip and move sideways more than a 4-seam fastball. The high predicted log active spin at high vertical movements and horizontal movements closer to 0 are likely due to pitchers throwing their sinker with less side spin than intended; in turn, throwing a pitch that is more similar to a 4-seam fastball than a pure sinker. It is hard to intuitively explain why there are no horizontal movements of 0, or why the lowest predicted log active spin is not at 0 vertical movement like the other two pitches. It may be a case that pitchers who throw sinkers are very efficient at throwing the pitch, with even the sinkers that move the least still having slight movement. If this is the case, there may simply not be enough data on sinkers that have no movement, which may explain why the variance increases as the horizontal movement moves closer to 0.

Changeup

As we have done with all other pitches, the average logged value of active spin in our changeup dataset (7.36) will be used as the β value in our simple kriging, while the ordinary kriging will be estimated as we have done before. Once again, leave-one-out cross validation was done on these two kriging estimates, with table 10 showing the mean error and the mean-squared error of these estimates.

Table 10: Mean Error and Mean-Squared Error Of Kriging Estimates For Changeups After Leave-One-Out Cross Validation

Kriging Estimate	Mean Error	Mean-Squared Error
Simple Kriging	0.00042	0.0103
Ordinary Kriging	0.00096	0.0104

Unlike the other three pitches that we have analyzed, the changeup dataset will use a simple kriging as the final kriging estimate. Figure 9 and figure 10 displays the predicted log active spin and the variance for these predictions, respectively, based on the horizontal and vertical movement of a changeup.

Figure 9: Predicted Log Active Spin

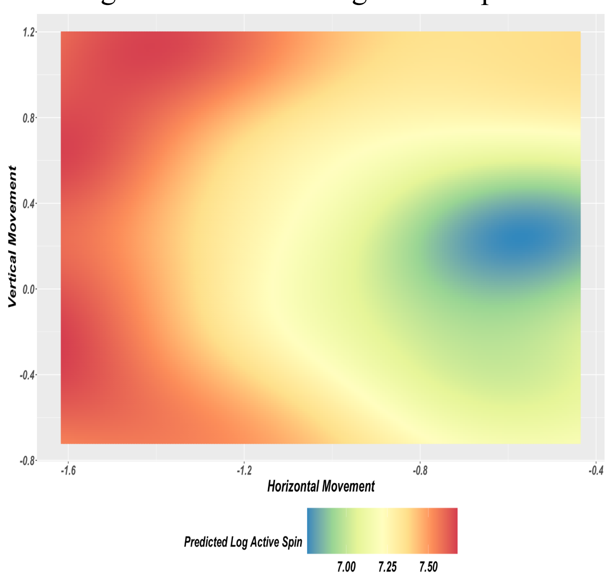
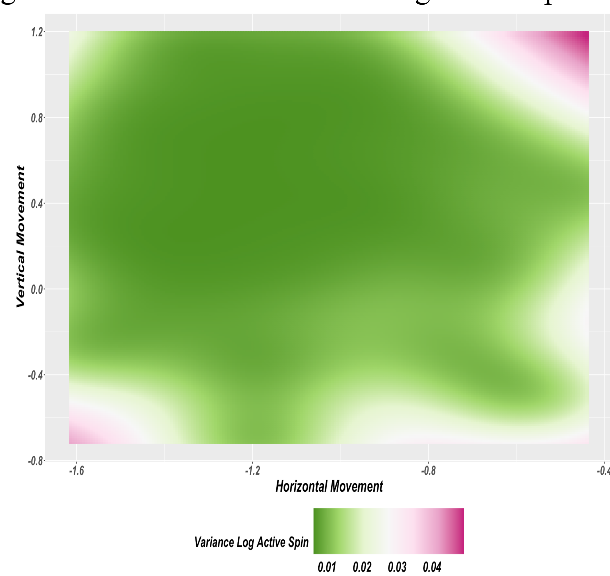


Figure 10: Variance of Predicted Log Active Spin



In figure 9, there is a high predicted log active spin as the movement of the changeup goes more left. Similar to the sinker, there appears to be a spiral-like pattern in the predicted log active spin of changeups; as the horizontal movement moves more right from the very left, vertical movements around 0 tend to have less predicted log active spin. Also similar to the sinker, no one threw a changeup with 0 horizontal movement, and the lowest predicted log active spin appears to be for changeups slightly above 0 vertical movement. Figure 10 reveals a plot with low variance all around, other than the corners. Still there does not seem to be a clear pattern in the variance in relation to the predicted values to warrant any caution.

Changeups are thrown with a similar spin as a sinker, but with more focus on downward movement, as changeups tend to be much slower than sinkers. When comparing figure 9 to the sinker kriging estimate in figure 7, we can see that the spiral appears to have shifted, with a higher predicted log active spin for changeups with low vertical movement, as opposed to a higher predicted log active spin for sinkers with high vertical movement. Once again, the no horizontal movements of 0, along with the fact that the lowest predicted log active spin is not exactly at 0 is likely due to pitchers who throw changeups being very efficient with the intended movement of the pitch. For changeups that move to the left, the high predicted log active spin values for both high and low vertical movements can intuitively be explained by most of the active spin being applied to the side spin, with a very small difference in the amount of back spin affecting how the ball moves vertically. It appears that a small change in the amount of backspin in a changeup affects the vertical movement of the pitch much more than a 4-seam fastball or sinker would.

Discussion

Maximizing Velocity

In an attempt to maximize fastball velocity through pitch release point, data was collected on every fastball that was thrown over 100 times by a right-handed pitcher, with each row containing the average x and y release point of the fastball, along with the average velocity of the fastball. This data was then split into 17 evenly distributed grids that were placed on top of all release points in our data, with each grid containing the count of pitch release points in that grid, along with the average velocity of fastballs in that grid. Using a Monte Carlo simulation of Moran's I statistic on the residuals of a linear model that used the average velocity in a grid to predict the log count in the grid, we find statistically discernible evidence that that grids with a high count of pitch release points tend to be surrounded by grids with a high count of pitch release points. Then, by using a simultaneous autoregressive model, we also found statistically discernible evidence that the increase in the average velocity in a grid was associated with an increase in the log count of release points in that grid. Intuitively then, this would mean that pitchers that throw high velocities tend to have similar release points, while pitchers that throw their fastballs at low velocities are likely to have similar release points. Since the conclusions in this data are only drawn from right-handed fastballs in the MLB in 2022, we can not extrapolate our findings to fastballs from any other year or baseball league, along with left-handed pitchers. Furthermore, we also can not extend our findings to pitches other than fastballs, as the velocities of other pitches may not have the same relationship with release points. Also, these are observational data, so no causal inferences can be made.

There are a couple of caveats that must be mentioned, as the results of this study may change based on these caveats. The first is the amount of grids that were chosen to divide our release points into areal units. I ended up choosing 17; however, with this choice, there were 4 grids with 0 observations that we had to remove from our data completely, along with multiple grids with low counts of release points. While it was possible to have less grids (in turn, having less grids with a low count of release points), there were concerns with pitchers in the same grid having non similar release points (some of them being a foot apart at both the x and y release point). To take extra caution with these low counts, I decided to log the count in each grid, but there is still a possibility that the low counts in grids affected our results greatly. The other caveat is associating sinkers and cutters as fastballs. Although cutters and sinkers are the closest pitches to 4-seam fastballs, there are still small differences in the pitch that may lead one to believe that these pitches should be treated separately. Still, I believe that cutters and sinkers are still thrown with the same intention as 4-seam fastballs in terms of release point, just with different starting spin at the point of release; therefore, this is why I decided to include all three pitches as fastballs.

In terms of future studies, there are a couple of different directions that can be taken with this topic. If the caveat of choosing the number of grids is not an issue, one can add more information to the spatial regression models, such as the average spin rate of a fastball, or the spin direction of a fastball to see if there are any further relationships between the count of release points in a grid and different pitching metrics. Furthermore, it is also possible to explore the relationship between the count of release points in a grid and the metrics of other pitches, such as a slider or curveball. Of course, with any analysis with spatial regression models, complete spatial randomness will have to be tested on the residuals of the linear model, as we did with the Monte Carlo simulations of Moran's I on our linear model. Another option a future study can take is using the pitch release points as point-reference data, and having the velocity of that point as a mark; very similar to the analysis we did for optimizing spin in different pitch movements.

Optimizing Spin

For optimizing the amount of active spin in relation to the movement of baseball pitches, data was collected on every pitcher that threw a certain pitch over 100 times. In each pitch type dataset, each row contained the average horizontal and vertical movement without gravity for that pitcher, along with the logged average amount of active spin that pitcher was able to generate for that pitch. Unfortunately, pitchers who threw splitters, cutters, and curveballs had to be taken out of our final analysis with concerns of sample size. However, for 4-seam fastballs, sliders, changeups, and sinkers; we were able to find statistically discernible evidence that the active spin values were distinguishable from complete spatial randomness, allowing us to continue with the spatial prediction of pitch movements and their corresponding active spin values in all of these pitch type datasets. Gaussian variograms with varying parameter values were deemed to be the best fit for our datasets, and the kriging estimate that produced the mean error and mean-squared error closest to 0 was used as our final kriging estimate. Based on the plots of our final kriging estimates, there was no issues interpreting the results through intuitiveness; however, there were some concerns with a lack of data points at unexpected movements of a particular pitch (such as a 4-seam fastball that has a lot of negative vertical movement, or a sinker that has little horizontal movement). Once again, since the conclusions are only drawn from right-handed pitchers in the MLB in 2022, we can not extrapolate our findings to pitchers from any other year or baseball league, along with left-handed pitchers. No causal inferences can be made with this data, since these are observational data.

In terms of caveats with our results, one big issue may be that we did not incorporate spin direction in our analysis. Although the movement of a pitch has shown to be mainly based on its active spin, the direction of the spin still matters due to the nature of physics, especially for pitches that have a high active spin, since the direction of the spin is not changing. For baseball pitches, there is no clear definition of what a spin direction has to be; therefore, there may be a lot of differences in the spin direction of the same type of pitches. Many of the interpretations were under the assumption that differences in movement when the predicted log active spin was the same were caused by differing spin directions (more sidespin or topspin than intended), but these interpretations should be taken with caution in case there were no differences in the same direction. Another caveat is that universal kriging was not considered, since there was no way to get a metric for a baseball pitch at every point in our spatial space. As more data is collected, this may one day be possible by using data over multiple years; however, due to the evolution of pitching, there may be concerns of treating data over multiple years as similar data. For example, in 2012, when there was a lack of technology in baseball, pitchers and analysts were at a much bigger disadvantage than pitchers and analysts today. These days, it is possible to obtain exact measurements, leading to more accurate data, and better adjustments that were simply not possible previously. I would expect this technology to continue developing, and there is a possibility that methods used today are one day seen as outdated.

In terms of future studies, although there are concerns of using data over multiple years, I would still be interested in seeing how a universal kriging model estimate would perform if there was one day enough data to incorporate it. This could also allow someone to incorporate spin direction; if there is one day enough data to have a metric for spin direction at every point in our spatial space, then I believe that this covariate could have a major impact on our final kriging estimate. Another option for incorporating spin direction would be finding a way to possibly predict active spin based on the spin direction using a non-spatial model, then using the predictions of active spin derived from this model as the mark for our pitch movements.

References

1. MLB. "Reliever of the Year Award Winners: History." *MLB.Com*, www.mlb.com/awards/reliever-of-the-year. Accessed 5 May 2023.
2. Savant, Baseball. "Statcast Percentile Rankings." *Baseballsavant.Com*, baseballsavant.mlb.com/leaderboard/percentile-rankings?type=pitcher&year=2022&team=&sort=14&sortDir=desc. Accessed 29 May 2023.
3. Bhatta, Deepak. "Average Fastball Velocity by Year." *Average Fastball Velocity By Year*, 11 May 2023, mlbrun.com/average-fastball-velocity#:~:text=Average%20fastball%20velocity%20by%20year%20is%20recorded%20from%2093.1%20mph,to%2093.9%20mph%20in%202022.
4. Anz AW, Bushnell BD, Griffin LP, Noonan TJ, Torry MR, Hawkins RJ. Correlation of torque and elbow injury in professional baseball pitchers. Am J Sports Med. 2010 Jul;38(7):1368-74. doi: 10.1177/0363546510363402. Epub 2010 Apr 16. PMID: 20400752.
5. Bushnell BD, Anz AW, Noonan TJ, Torry MR, Hawkins RJ. Association of maximum pitch velocity and elbow injury in professional baseball pitchers. Am J Sports Med. 2010 Apr;38(4):728-32. doi: 10.1177/0363546509350067. Epub 2010 Jan 21. PMID: 20093420.
6. Hashimoto Y, Nagami T, Yoshitake S, Nakata H. The relationship between pitching parameters and release points of different pitch types in major league baseball players. Front Sports Act Living. 2023 Apr 24;5:1113069. doi: 10.3389/fspor.2023.1113069. PMID: 37168522; PMCID: PMC10164925.
7. Sarris, Eno. "What Can Hitters Actually See out of a Pitcher's Hand?" *FanGraphs Baseball*, 15 Sept. 2016, blogs.fangraphs.com/what-can-hitters-actually-see-out-of-a-pitchers-hand/.
8. Smith AW, Smith BL. Using baseball seams to alter a pitch direction: The seam shifted wake. Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology. 2021;235(1):21-28. doi:10.1177/1754337120961609
9. M. Lage, J. P. Ono, D. Cervone, J. Chiang, C. Dietrich and C. T. Silva, "StatCast Dashboard: Exploration of Spatiotemporal Baseball Data," in IEEE Computer Graphics and Applications, vol. 36, no. 5, pp. 28-37, Sept.-Oct. 2016, doi: 10.1109/MCG.2016.101.
10. Cinque ME, LaPrade CM, Abrams GD, Sherman SL, Safran MR, Freehill MT. Ulnar Collateral Ligament Reconstruction Does Not Decrease Spin Rate or Performance in Major League Pitchers. Am J Sports Med. 2022 Jul;50(8):2190-2197. doi: 10.1177/03635465221097421. Epub 2022 May 26. PMID: 35616521.
11. Sarris, Eno. "What Is 'seam-Shifted Wake' and Which Pitchers Benefit Most from It?" *The Athletic*, 21 Jan. 2021, theathletic.com/2337324/2021/01/22/seam-shifted-wake-sleeper-pitchers/.
12. Person. "Here Are the Cy Young Award Vote Totals." *MLB.Com*, 17 Nov. 2022, www.mlb.com/news/2022-mlb-cy-young-award-voting-results.
13. "Statcast Spin Axis All Pitches." *Baseballsavant.Com*, baseballsavant.mlb.com/leaderboard/spin-direction-pitches?year=2022&min=q&sort=7&sortDir=asc&pitch_type=ALL&throws=&playerName=FRIED&team=&pov=Bat. Accessed 1 June 2023.
14. Bill Petti, "How to Build a Statcast Database from Baseballsavant, v3.0," Root Mean Squared Musings (B, April 2, 2021), https://billpetti.github.io/2021-04-02-build-statcast-database-rstats-version-3.0/.
15. Bivand, Roger S., et al. *Applied Spatial Data Analysis with R*. Springer, 2013.
16. Paez, Antonio. *An Introduction to Spatial Data Analysis and Statistics: A Course in R*.