Investigating if MLB pitchers are self-selecting for high or low drag balls

by

Jessica Sonner

Submitted to the Department of Mechanical Engineering
in partial fulfillment of the requirements for the degree of

Bachelor of Science in Mechanical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Author .....
Department of Mechanical Engineering
May 8, 2022

Certified by .....
Anette (Peko) Hosoi
Neil and Jane Pappalardo Professor of Mechanical Engineering
Thesis Supervisor

Accepted by .....
Kennith Karmin
Associate Professor of Mechanical Engineering
Undergraduate Officer

Investigating if MLB pitchers are self-selecting for high or low drag balls

by

Jessica Sonner

Submitted to the Department of Mechanical Engineering
on May 8, 2022, in partial fulfillment of the
requirements for the degree of
Bachelor of Science in Mechanical Engineering

Abstract

The drag coefficient of 4928 batted balls in the MLB in 2019 were measured and matched with 121 pitchers who threw the ball to examine if they are statistically more or less likely to select a high or low drag ball to throw. It was hypothesized that (1) pitchers are routinely selecting for either high or low drag balls, and (2) the regular selection of high drag balls are associated with "better performing" pitchers, or pitchers who have less home runs scored against them, as high drag balls travel less far through the air than low drag balls. However, it was found that pitchers in this data set do not appear to be selecting for higher or lower drag balls, and the average drag coefficient of the balls they select is not correlated with that of higher or lower performing pitching statistics.

Thesis Supervisor: Anette (Peko) Hosoi
Title: Neil and Jane Pappalardo Professor of Mechanical Engineering

Acknowledgments

I would like to first and foremost thank my advisor Peko for her continued support, leanings, and guidance throughout the entire experience. Thank you for your kindness and generosity in your teachings, for your ability to take an immensely complex idea and distill it down into an understandable way. Thank your for your excitement and enthusiasm towards my own learnings and education, and it is with you I am inspired to want to make a deep impact into research and sport.

I would also like to thank Sarah Fay and the rest of Peko's lab for their support and continued guidance, and always being an encouraging and accepting resource.

Contents

List of Figures

Chapter 1

Introduction

Major League Baseball is the fourth most popular sport in the United States, with over 20 million viewers during the 2017 World Series alone [9]. One of the most noteworthy, memorable, and rarest aspects of the game is the home run. However, just in the past several years, the number of home runs during a game has substantially increased causing the MLB in 2017 to commission an independent study to investigate why this increase is observed now, when the game of baseball has been around since 1846. Using positional data of the ball collected from Doppler radar systems installed in MLB stadiums in 2015, this study suggests there has been a change in aerodynamic properties of balls, specifically a reduced drag coefficient. Changes in the ball since 2015, most notably the seam height, may account on average for a nearly 6ft. increase in batted ball distance, leading to the increased rate in home runs [1].

As this phenomenon has arisen within professional baseball, so has the appearance of notable "unlucky" or "lucky" pitchers. When a pitcher is about to throw, they grab a ball, examine it, and physically sense its weight in their hands. They then use intuition created from both playing baseball for their lifetime and being an expert in the sport to decide if it is a "good" or "bad" ball. In other words, they will decide if they want to use that ball to pitch or not. A bad ball is one more likely to have a home run scored off the pitch, where a good ball is one that is less likely to have a home run scored. This research seeks to investigate if the phenomenon of "lucky" pitchers are ones who routinely see fewer home runs scored off their pitch by unknowingly (or perhaps knowingly) self-selecting for low drag balls. The potential correlation between unluckiness and low drag balls is investigated by analyzing the trajectory flight path of balls pitched across 4 different years in the MLB, calculating the respective drag coefficients, characterizing the relatively high or low drag characteristic of the ball, then matching that information with pitchers in 2019 as evidence to determine if certain pitchers are routinely getting home runs scored off them on account of ball selection. This information could inform pitching strategies and even drive game-time coaching tactics or decisions.

Chapter 2

Baseball Aerodynamics

2.1 Forces

As a baseball moves through the air, it experiences three different forces that dictate its trajectory: gravity, lift (Magnus force), and drag (Figure 2-1). Gravity is independent of where the ball is located within space. It always points downwards, pulling the ball towards the earth at a constant acceleration. If the baseball were launched in a vacuum, the only force the ball would experience is gravity, rendering the computation of the distance the ball travels quite easy and solely dependent upon one constant force. However, the aerodynamics of a ball traveling through air is quite complex, as the ball not only is subject to gravity but also has to push air molecules out of the way to travel through it [1]. These forces are produced by the contact between the ball and the air and are dependent upon the velocity of the ball within space specifically with regard to the angle of the ball's flight path relative to the horizontal. The lift force describes how the ball moves normal to its flight path (Equation 2.1). ω is the angular velocity vector pointing out of the page, where the

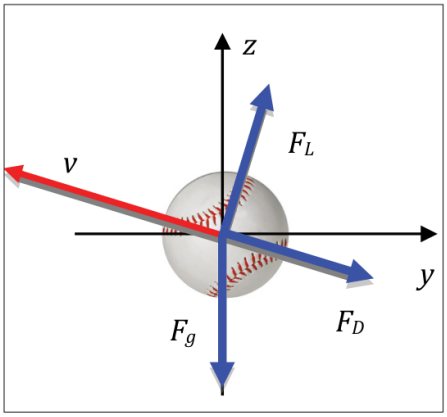

Figure 2-1: Forces acting on a baseball [5].

$\hat{t}$ vector is pointing in the direction of the baseball's flight path. Thus, the lift force is normal to that of the direction the baseball is traveling in (Figure 2-1).

$$F_L = \frac{1}{2}\rho A C_L V^2(\omega \times \hat{t}) \tag{2.1}$$

Comparatively, the drag force is in the opposite direction of the ball's flight path, describing how a ball slows along its path due interactions with the air (Equation 2.2). Thus, the drag force is associated with decreasing the carry of the ball, making it more difficult to fly through the air.

$$F_D = -\frac{1}{2}\rho A C_D V^2(\hat{t}) \tag{2.2}$$

2.2 Drag Force and Coefficient

Both drag and lift forces are commonly described in terms of their coefficients [3]. The drag coefficient characterizes the magnitude of the drag force on an object traveling through the air [3]. It is a unitless number and typically lies in the range of 0.25 - 0.5 for a baseball [1].

The drag coefficient is directly proportional to the drag force, and thus a greater drag coefficient means that there is a larger force of drag on the ball, making it more difficult to fly through the air, and thus it will travel a shorter distance (Equation 2.2). Comparatively, a smaller drag coefficient indicates a smaller force of drag on the ball, and less resistance is present from the air on the ball when the ball is mid-flight.

The drag force is also dependent upon the cross sectional area of the ball. A greater cross sectional area indicates that there are more air particles that need to be moved in order for the ball to travel through the air. More air particles mean there is a greater force of the air against the baseball, resulting in a larger drag force.

2.3 Recent Changes in MLB Drag Coefficients

The analysis conducted by the MLB Committee revealed that the primary reason for the increase in home run rates beginning in 2015 is that the ball carries longer for the same initial conditions. An initial hypothesis revolved around increased temperatures and humidity due to rising global temperatures, leading to lower air density and reduced drag. However, even at fixed temperatures in temperature-controlled stadiums, increased home run rates were observed [1, 2]. Data shows that the increase in home run rate between 2018 and 2019 was due in part to a change in aerodynamic properties of the baseball. They observed a notable decrease in drag coefficients, likely correlated to decreasing seam heights by about one-thousandth of an inch causing the ball to fly up to 6 feet farther and creating the increase seen in home runs in the

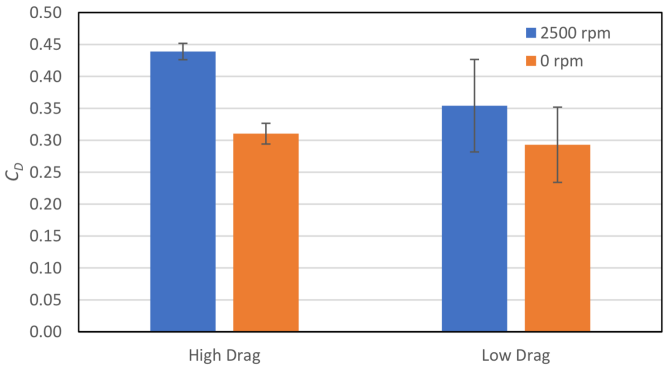

Figure 2-2: Effect of spin on drag coefficient [1].

.

year of 2019, [1].

2.4 Accounting for Aerodynamic Inconsistencies

The aerodynamic flow over a baseball is complex, resulting in drag properties that depend not only on seam height but also on spin rate, spin axis, seam orientation, application of mud, and possibly other factors not yet identified [1]. While the baseball can be modeled as a sphere flying through the air, its irregular shape causes it to stray from the expected model in a variety of ways. For example, the irregular shape of the baseball results in a relatively large variation in lift and drag, complicating the study of aerodynamic effects [2].

2.4.1 Effect of Spin and Speed on C_d

Drag is sensitive to ball speed [1]. Spherical sport balls can experience a phenomenon known as the "drag crisis", when a ball's drag coefficient decreases quickly with increasing ball speeds. At Reynolds numbers just above 10^5, the aerodynamic drag force on a sphere drops sharply as the flow begins to become turbulent in the boundary layer [7]. For baseballs, this "drag crisis" can occur at speeds which are typical for pitched or batted balls, when the laminar flow of air in a boundary layer near the ball begins to separate and become turbulent [4]. The effect of this turbulence in the boundary layer of air around the ball reduces the size of the turbulent wake behind the ball, and reduces the drag force. Thus, balls with increasing linear velocities show decreased drag coefficients. Changes in the drag regime occur when the Reynolds number exceeds about 10^5. The Reynolds number (Re) as seen in 2.3 is defined by the diameter of the baseball (d), the kinematic velocity of the baseball relative to the air (V) and the viscosity of the air (ν).

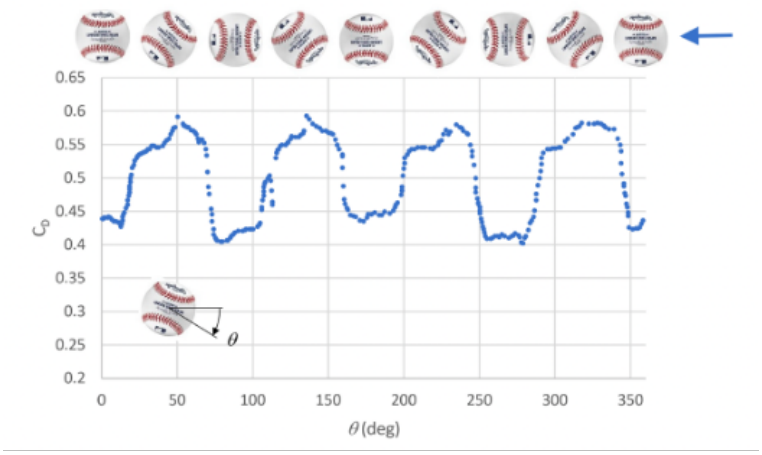

Figure 2-3: Spin dependence on drag coefficient [9].

$$Re = \frac{Vd}{\nu} \tag{2.3}$$

Drag is also dependent on the angular speed and rotation of the ball. As seen in Figure 2-2, in comparing a ball with spin and a ball with no spin, it's clear those with increasing spin have larger drag coefficients [1]. Additional research has shown that the change in the positions of the baseball seams as the ball rotates through the air might explain the inconsistent, oscillatory behavior in the drag coefficient (Figure 2-3).

Chapter 3

Analysis Technique / Methodology

To determine if the change in the drag coefficient leading to an increase in home run rates has led to pitchers consistently getting or not getting home runs scored off of their throw, the estimated drag coefficients of batted balls are calculated and matched to the pitcher who threw it. As the effect of drag on carry is far greater than the effect of lift on the carry of the ball, the sole focus of this study is on calculating the drag coefficient [1]. The strategy for this research is as follows:

1. Derive the equations of motion
2. Take the derivatives of equations of positional data
3. Calculate the drag coefficient
4. Match the drag coefficient of the batted ball to the pitcher
5. Find presence or absence of correlation between high or low drag balls and specific pitchers

3.1 Data Used

The XYZ positional data used in this study are from the Statcast data collected from Doppler radar systems in the MLB ballparks, provided by the MLB during the seasons of 2016-2019. The Doppler radar systems were initially installed in 2015, but the Committee found that data from the year of 2015 was anomalous [1], and acknowledged that the system underwent many adjustments and recalibrations as the season progressed. Thus data from the years 2016-2019 were used in calculating the drag coefficient of balls, and the year 2019 was matched to players, due to the focus on the drag analysis conducted in 2019 by the MLB committee as well as the available pitcher data from 2019 sent from the MLB.

The coordinate system is typical in the Statcast system, with the origin at the

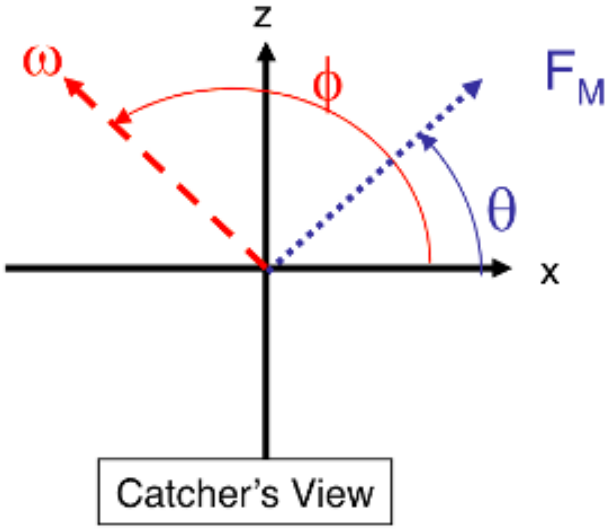

Figure 3-1: Coordinate system from the perspective of the catcher [8].

home plate, $\hat{y}$ that pointed towards the pitcher, and $\hat{z}$ points vertically upward. The x-axis points to the catcher's left and right, and this coordinate system was used throughout the extent of this research [8].

3.1.1 Drag Calculation

This analysis begins by looking at the sum of the forces on the ball, using Equations 2.1 and 2.2 to account for F_L and F_D respectively:

$$\vec{F} = m\vec{a}$$
$$\vec{F_L} + \vec{F_D} + m\vec{g} = m\vec{a} \tag{3.1}$$

The instantaneous acceleration at each point in time point can be a direct measure of a in Equation 3.1, and can be calculated by taking the derivative of the velocity V (Equation 3.2), with V calculated directly from the XYZ Statcast positional data (3.3).

$$\frac{dV_i}{dt} = \frac{V_{i+1} - V_{i-1}}{t_{i+1} - t_{i-1}} \tag{3.2}$$

$$V_i = \sqrt{v_{x,i}^2 + v_{y,i}^2 + v_{z,i}^2}. \tag{3.3}$$

Acquiring $\frac{dV}{dt}$ from positional data and combining it into Equation 3.1 leads to a governing equation of motion (Equation 3.4) where where the only unknowns are C_L and C_d [1].

$$m\frac{d\vec{V}}{dt} = \overrightarrow{F_L} + \overrightarrow{F_D} + m\,\vec{g} \tag{3.4}$$

Next, the sum of the forces can be broken down into the sum of the forces along each axis. In isolating the forces on the ball solely in the direction of the drag force, this eliminates the need to calculate for the lift force and lift coefficient as the lift force is perpendicular and has no magnitude or contribution of force in the direction of ball's flight path. Taking the direction of each of these expressions into account, gravity (g) is now represented in both magnitude and direction in our calculation by the vector (0, 0, -g) [1]. Equations 3.5 and 3.6 show the breakdown from extracting the forces only in the direction of the ball's flight path.

$$m\frac{dV}{dt}(t) = F_D(t) - mg\sin\theta(t) \tag{3.5}$$

$$m\frac{dV}{dt} = -\frac{1}{2}\rho A C_D V^2 - mg\sin\theta. \tag{3.6}$$

A new variable θ is introduced as a reflection of the projection of the gravitational force in the direction of the tangent vector to the ball's flight path. Theta can be computed by examining the angle created by the projection of the ball in the XY plane and in the Z plane.

$$\theta_{i+1/2} = \arctan\left(\frac{z_{i+1} - z_i}{\sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}}\right) \tag{3.7}$$

$\theta_{i+1/2}$ lies between subsequent time coordinate values as it is calculated with positional values immediately next to each other. To shift the vector values of θ onto the correct time values as the positional coordinates, the average of the value of θ with one half step forward and one half step backwards is averaged in equation 3.8.

$$\theta_i = \frac{1}{2}(\theta_{i+1/2} + \theta_{i-1/2}). \tag{3.8}$$

A vector of C_D values associated with each point in time can now be calculated with the above information.

$$C_{D,i} = (-g\sin\theta_i) - \frac{dV_i}{dt}) * \frac{2m}{\rho A V_i^2} \tag{3.9}$$

Values of parameters used Equation 3.9 are as listed below [1]:

- Air density (ρ) = 1.225 kg/m^3
- Baseball radius (r) = 3.69 cm

- Cross-sectional area of the baseball ($A = \pi r^2$) = 42.776 cm^2
- Gravitational acceleration (g) = 9.8 m/s^2

3.1.2 Selection of Trajectory

The extraction of the drag coefficient begins with identifying and extracting the pitch, then selecting for the rising trajectory of the batted ball (Figure 3-2, 3-3, 3-4). In Figure 3-2, the trajectory of one data file is shown along with the first point in time shown in red. Figure 3-4 represents the changing XYZ coordinates over time from the play in Figure 3-2. The small (<1.5 second) gaps in time are representative of when the ball hits the ground and/or when a player takes time to grab the ball and throw it to a different player. To simplify calculations, this research only considered points from after the ball was batted up to the trajectory peak.

As seen in Figure 3-2, one Statcast trajectory file does not only contain the pitch and the batted ball. Rather, it contains information from the entire play from when the ball leaves the pitchers hand, until the end of the play, (likely at first base in this example). The pitcher mound and coordinates stay consistent throughout the trajectory files. However, the first data point in time collected does not always begin when the ball leaves the pitchers hand like it does in Figure 3-2. Additionally, most trajectory files do not consist of only the pitch and rising part of the trajectory. The Doppler radar system at times can accidentally capture trajectory segments from the previous play or segments within the next play. This presents obstacles in the extraction of the drag coefficient: if the incorrect part of the play is selected that is from a different play (either the one just before or the one just after), it can result in a drag coefficient from a ball different than the one that is pitched. Thus, the assumption that first data point collected in time is representative of when the ball leaves the pitchers hand cannot be made, and care must be taken to correctly identify the beginning of the batted ball as it will be later tied to a pitcher's name and performance. A Matlab program was created to identify the beginning of pitch independent of time, and then select for the rising part of the trajectory (Figure 3-5). Additional information on pitch identification parameters and the selection process employed within this study and Matlab program are located in Appendix A.

3.2 Extracting the Drag Coefficient

In Figure 3-4, the graph on the left shows the identification of the XY projection of the pitch. After being identified, it is subsequently removed along with any data following the trajectory peak. This ensures the correct segment of the rising trajectory of the batted ball is used in the calculation of the drag coefficient, shown in the graph on the right in Figure 3-5.

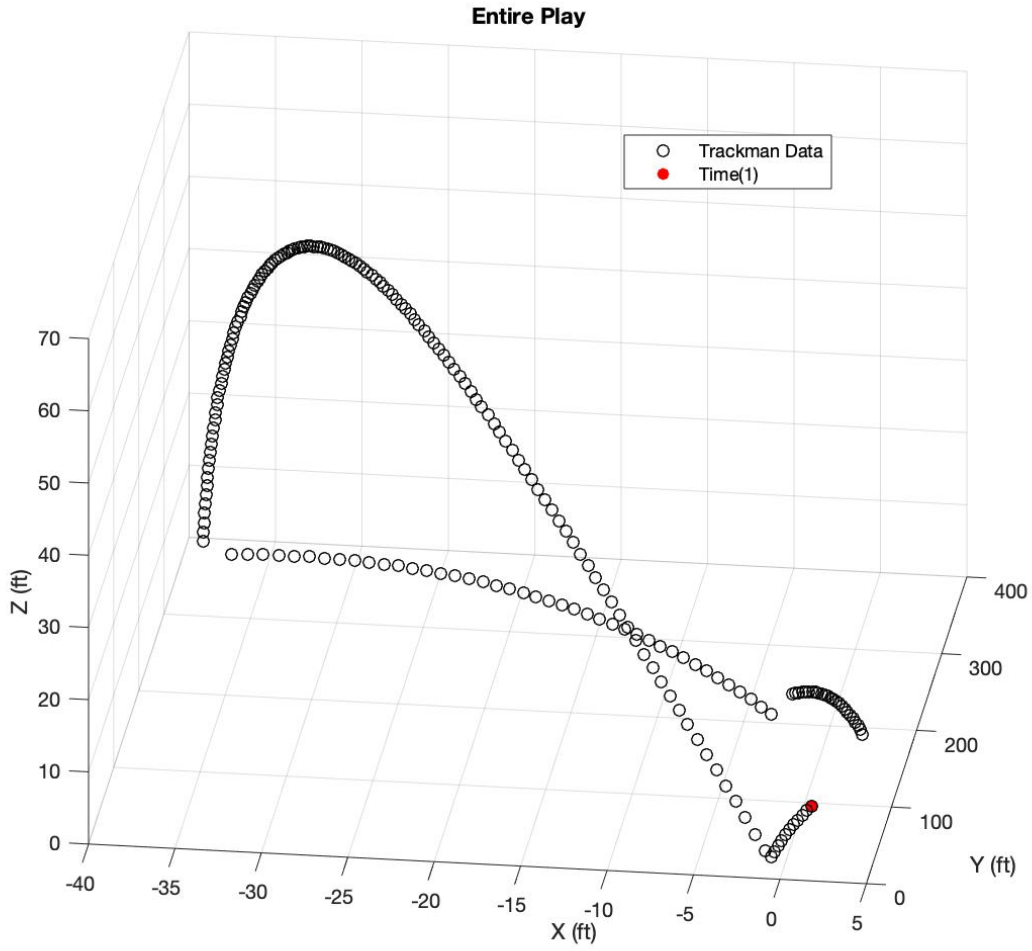

Figure 3-2: Example play from one Statcast trajectory file.

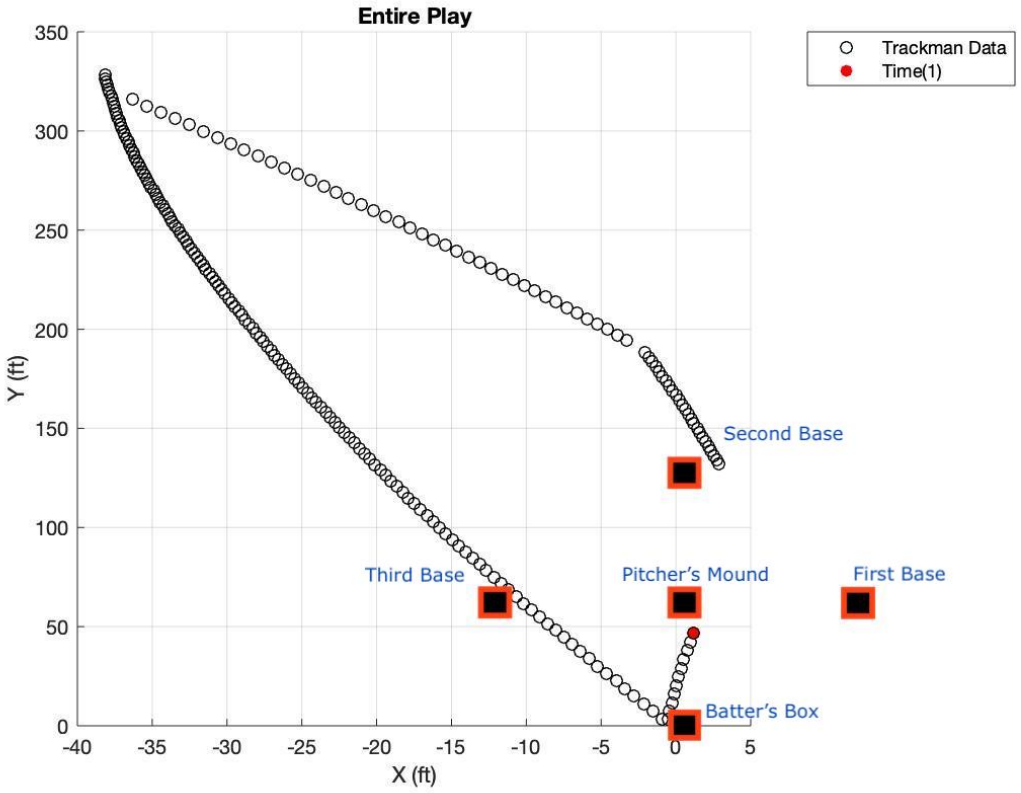

Figure 3-3: Orientation of trajectory in Figure 3-2 on baseball diamond.

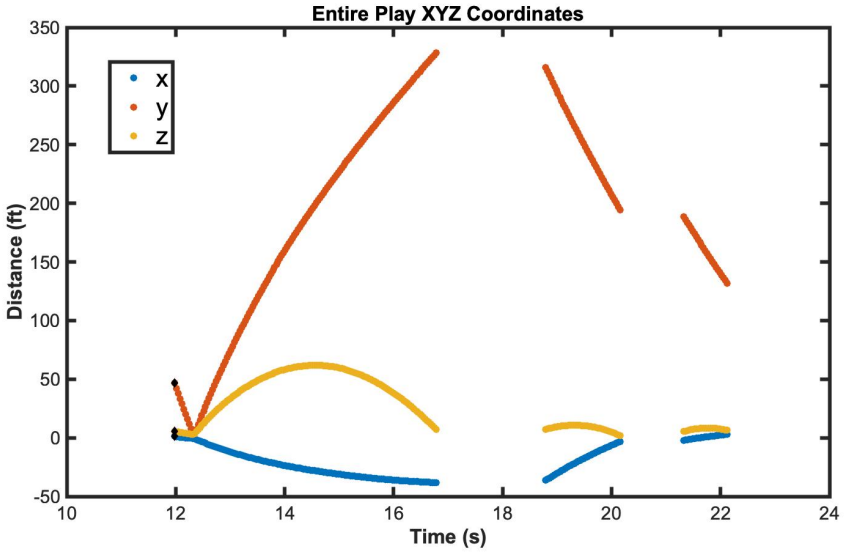

Figure 3-4: Progression of XYZ trajectory coordinates from Figure 3-4 over time.

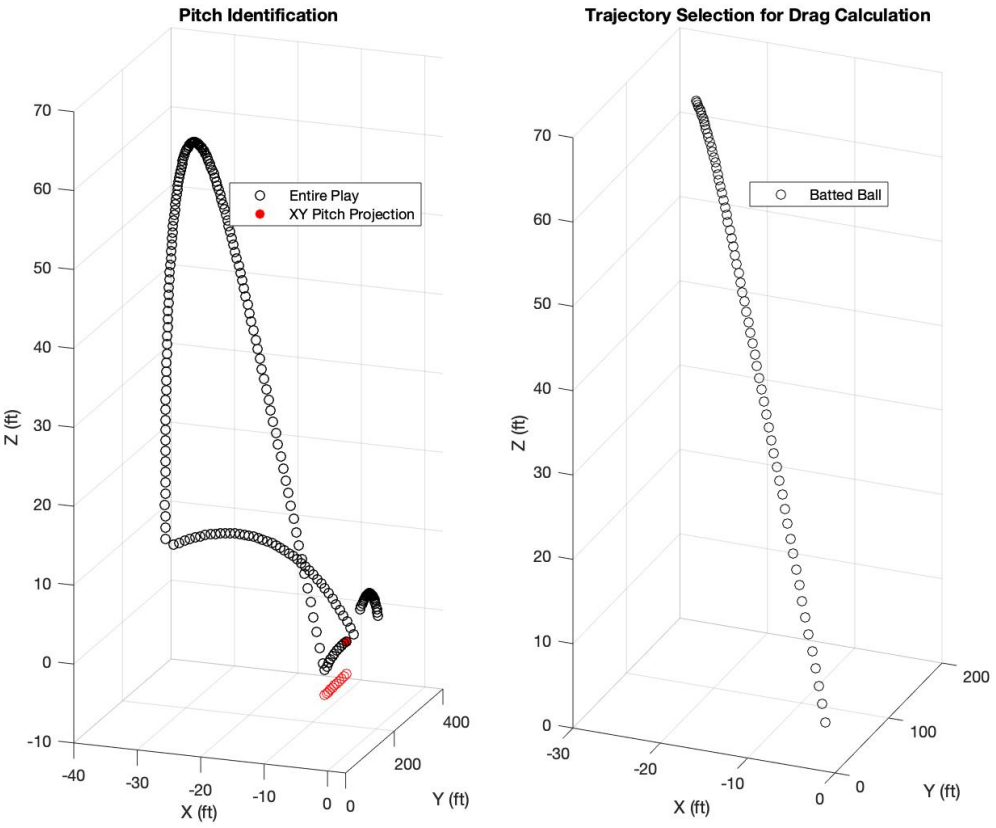

Figure 3-5: Selection of rising trajectory of batted ball.

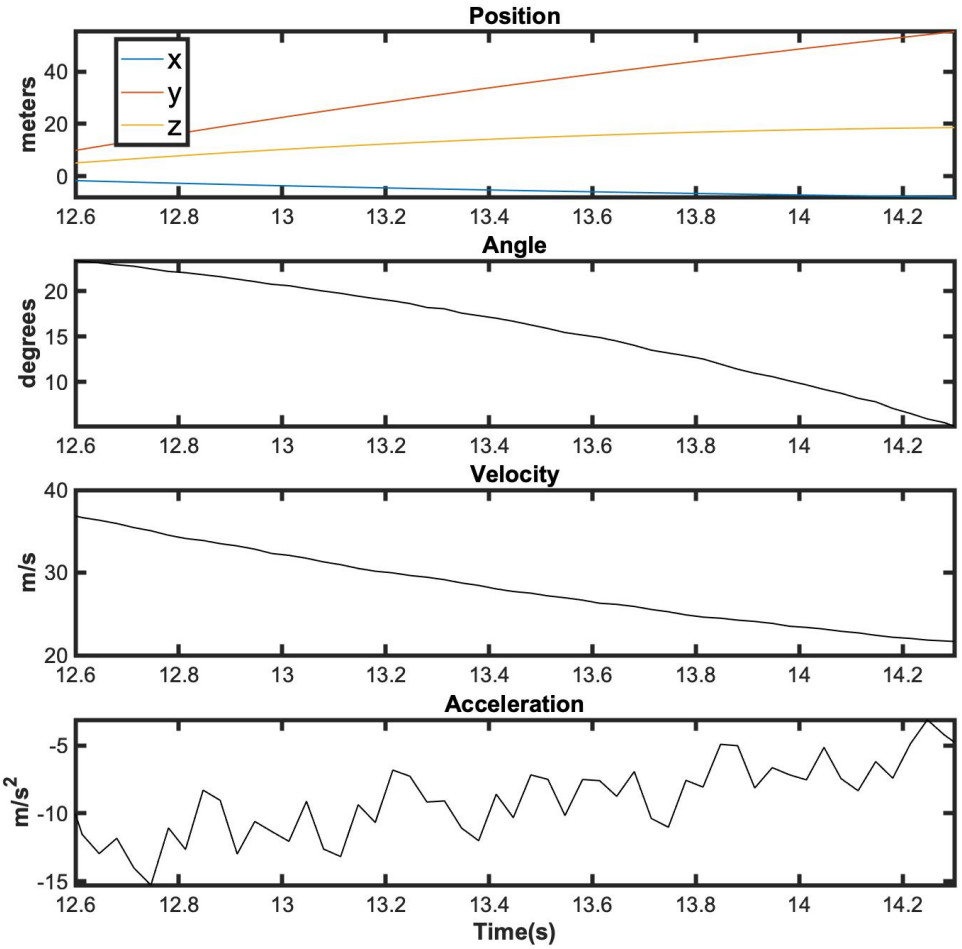

Figure 3-6: Instantaneous position, angle, velocity, and acceleration over time of the rising ball with respect to the horizontal.

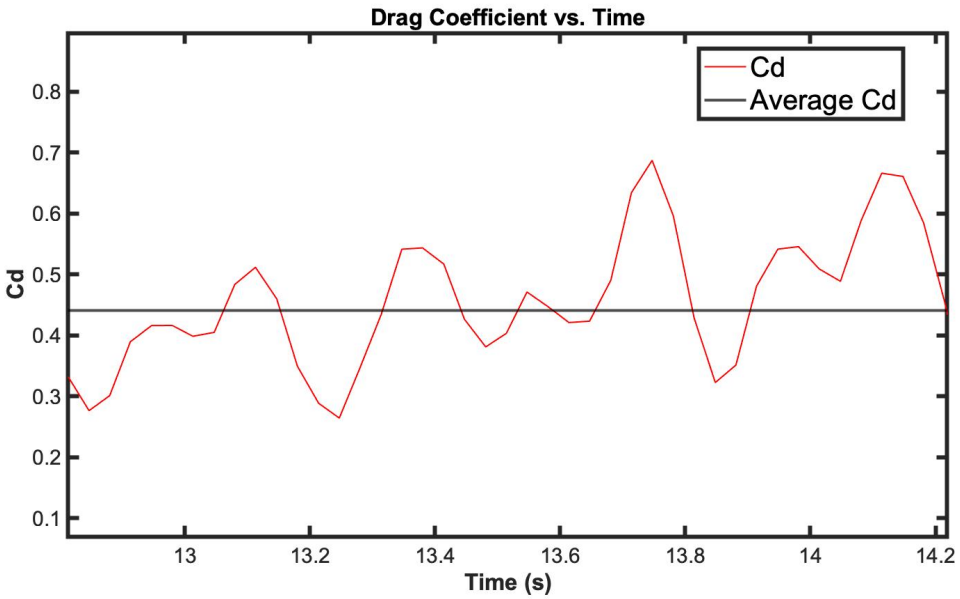

Figure 3-7: Plot of drag coefficient over time.

It was noticed that the drag has an oscillatory behavior in time as seen in Figure 3-7. However previous studies have acknowledged this phenomenon as well and connected it back to the rotation of the laces in the air that changes the airflow around the ball. The cause of the oscillations seen in the drag coefficient in this study may be a result of the changing airflow around the ball [9]. To find a value of C_d independent from this oscillatory behavior, the average C_d over time was taken as the representative value of the drag coefficient in time for each batted ball (Figure 3-7).

3.2.1 Spin Correction

The Statcast data for spin is truncated at 3500 RMP in early 2015 data sets [1]. To remove this bias, as the committee did in their report, only trajectories with spin values under 3500 RPM and above 1900 RPM were considered in this report. Knowing that spin increases the value of the drag coefficient (Figure 2-2), steps are taken to remove the dependency of the drag coefficient on spin.

As seen in Figure 3-7, there is a linear relationship between the spin of the ball and the drag coefficient in the years 2016-2019. The drag coefficient increases for increasing rates of spin. Thus, an increased value of spin makes the ball more difficult to fly through the air. A decreased value of spin is correlated with a smaller drag coefficient value, showing that for smaller values of spin, the ball can fly farther through the air. These linear relationships demonstrate how there is a dependency of spin on the calculation of the drag coefficient. It's important to remove the dependency of spin on the drag coefficient to enable the comparison of drag coefficients of balls with

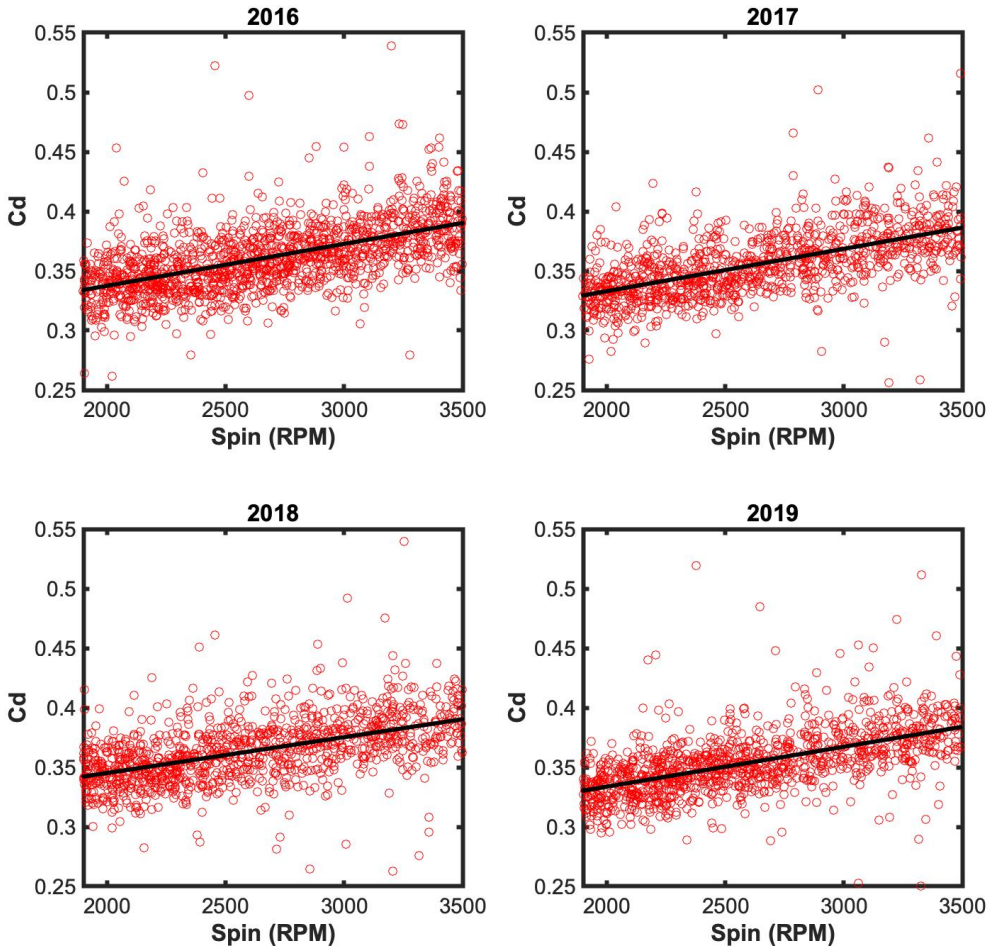

Figure 3-8: Plots of the average drag coefficient across 4 years against corresponding values of spin.

different spins.

As confirmed by the study commissioned by the MLB in 2017, the drag coefficient is found to be well approximated by a linear function of spin as seen in Equation 3.10, enabling for a streamlined extraction of the added value of the spin to the drag coefficient [1]:

$$C_d = m * spin + C_{d0}. \tag{3.10}$$

This can be done simply by subtracting from the average drag found in each of the data files, the slope of the line of best fit (Figure 3-7) multiplied by the measured value of spin (Equation 3.2).

$$C_{d0} = C_d - m * spin \tag{3.11}$$

The y-intercepts (the new average spin-independent C_d values across all the data points for each year in Figure 3-8) for 2016 - 2019, respectively, are 0.2672 ± 0.0078, 0.2619 ± 0.0081, $0.285 \pm .0076$, $0.2667 \pm .0077$.

The removal of the additional spin contribution from the drag enables for the calculation of an average Cd for each year, as the drag coefficient values are now normalized to a horizontal line.

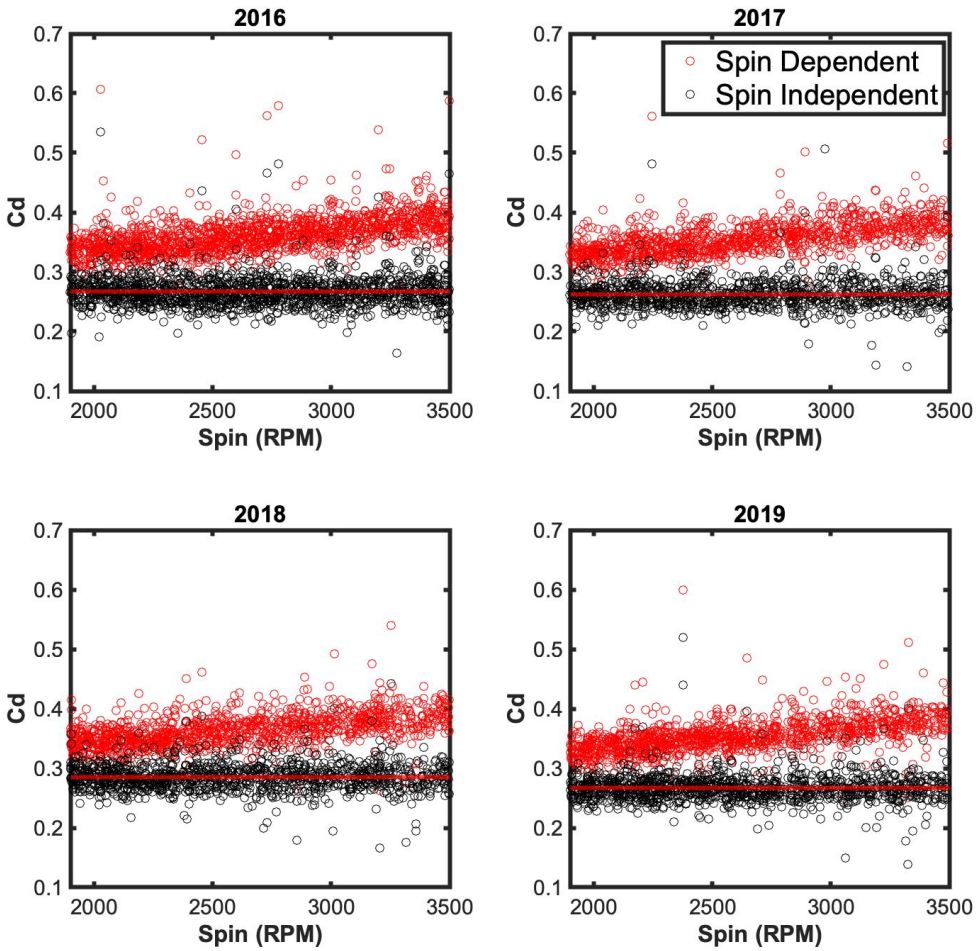

Figure 3-9: Conversion of spin-dependent to spin-independent C_d values using Equation 3.2.

Chapter 4

Results

The value of the drag coefficient was initially thought to be inversely proportional to statistics descriptive to baseball performance, with higher drag balls thought to be correlated with less home runs against the pitcher.

There were two initialized hypotheses around the drag of the baseball: (1) pitchers were routinely selecting for either high or low drag balls, as there has been spoken observations around "lucky" or "unlucky" pitchers routinely picking up characteristically high or low drag balls, respectively, without necessarily being "better" or "worse" pitchers that season. (2) High drag balls were thought to be indicative of better performance, and lower home runs scored against them.

However, the data indicates that pitchers do not appear to be selecting for high drag balls. There is no apparent correlation with statistics representative of pitcher performance and drag coefficients. As seen in Figure 4-1, there is no obvious correlation between pitchers who are selecting low drag balls and those that have a low hone run rate, as described by the statistic "HR/9". HR/9 is the number of home runs against the pitcher per 9 innings. A low value of HR/9 indicates that the pitcher has fewer home runs against them, and comparatively, a high value of HR/9 indicates a pitcher has more home runs scored against them. While pitchers such as "Pitcher 1" or "Pitcher 2" might colloquially claim that they select for high drag balls, there is no apparent correlation with their performance and that of the drag of the balls.

A few of the pitchers were selected from Figure 4-1 with average drag coefficients that fell outside of one standard deviation of the mean of all of the drag coefficients for all of the balls thrown, either higher or lower. However, only 4 pitchers of 151 had an average drag coefficient across the the different balls they threw to be outside of one standard deviation of the mean of all of drag coefficients from the balls thrown. In other words, only 2.65% of pitchers had average drag coefficient values that were farther than 68% of all of the drag coefficients calculated from the average drag coefficient across all balls thrown. For these results to be statistically significant and to be suggestive that pitchers are routinely selecting for high or low drag balls,

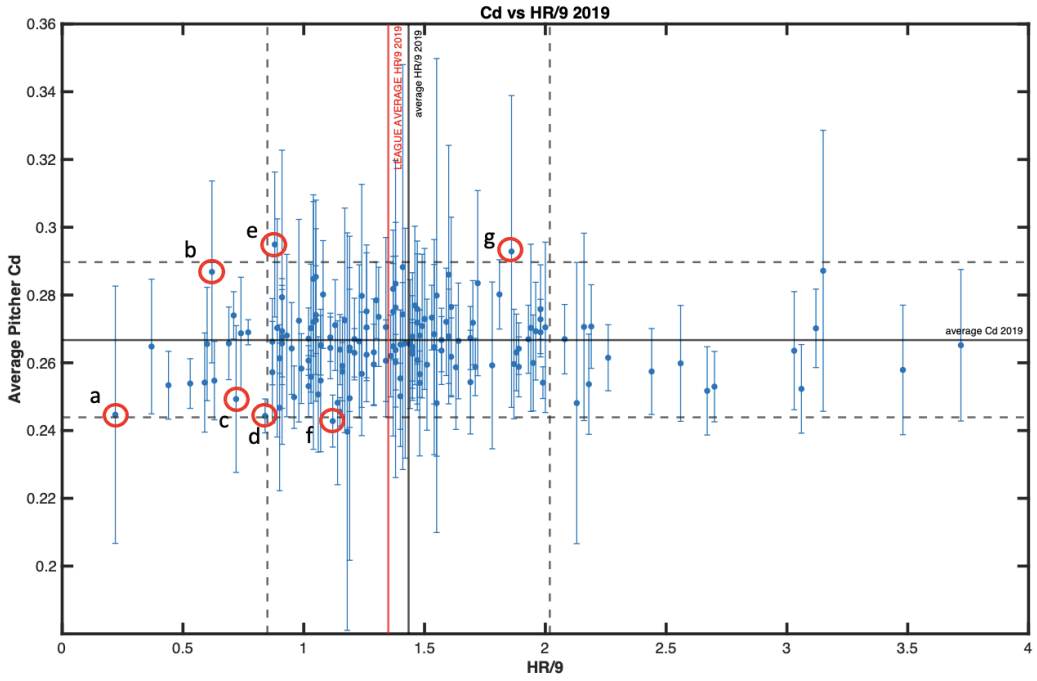

Figure 4-1: Plot of average C_d per pitcher vs. HR/9 statistic in 2019, denoting a few possible pitchers of interest.

or routinely seeing higher performance characteristics from consistencies in the drag of the balls chosen, we might expect to see it farther than 95% of all average drag coefficient values. However, because it is much less than that, we don't have evidence to suggest there is a characteristic similarity in high or low drag coefficients from the balls that this population of pitchers are selecting in the year 2019.

Two of these 4 pitchers, pitchers 135 and 145 were above one standard deviation of the population mean with average C_d values of 0.2949 and 0.2928 respectively and pitchers 8 and 129 were the only two with C_d values farther than one standard deviation below the population mean, with C_d values of 0.2428 and 0.2397. While this data doesn't yet have relations to their performance, there is no evidence to suggest that some pitchers are statistically choosing more higher or lower drag balls out of the current sample size on a regular basis in 2019. If we look at these notable athletes and see how their HR/9 scores compare, they do not fall outside of one standard deviation of the HR/9 statistics for this population of pitchers in 2019 as shown in Figure 4-2. This supports a notion that there is no evidence to suggest that even those pitchers who might be picking higher or lower drag balls than others within this group of pitchers have statistically significant differences in performance, either positive or negative.

ERA is another statistic of interest. It describes not only home runs against the pitcher across 9 innings, but all runs. This statistic was incorporated into the analysis

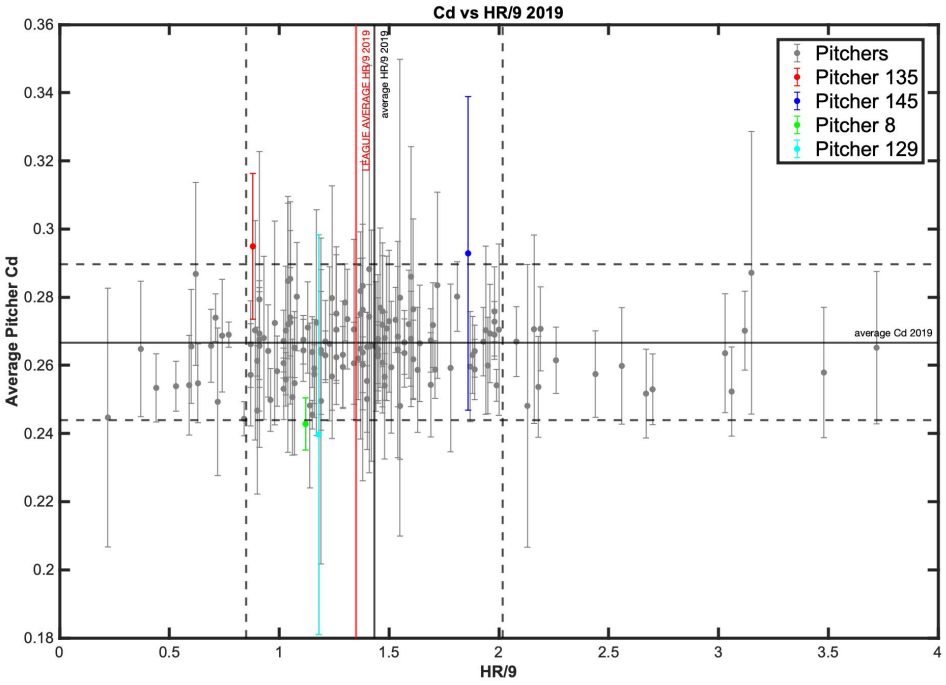

Figure 4-2: Plot of average C_d per pitcher vs. HR/9 statistic in 2019, noting the only 4 pitchers with average drag coefficients outside one standard deviation of the population mean.

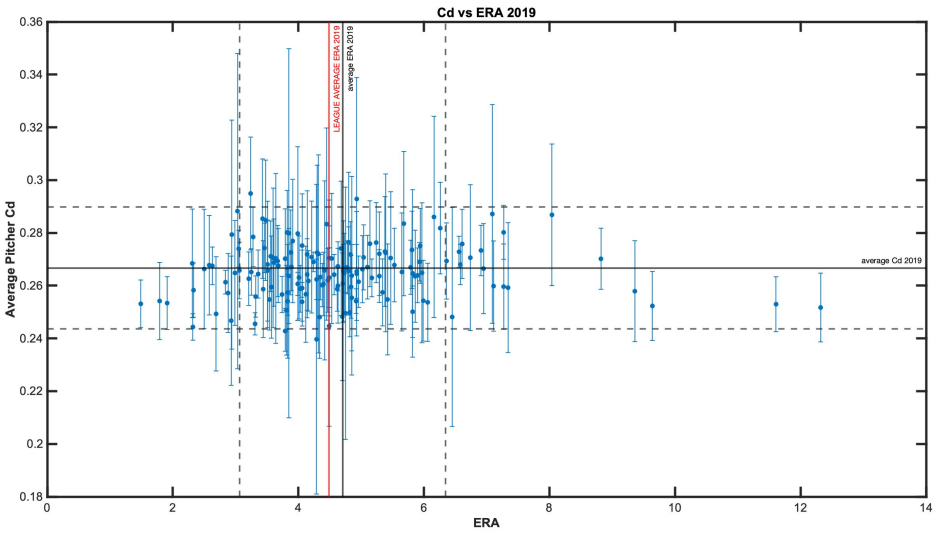

Figure 4-3: Plot of average C_d per pitcher vs. ERA in 2019.

to see if there was any correlation to any baseball play the pitchers has from their pitch and the drag coefficient of the ball. However, there is a similar amount of randomness existent in this analysis as the one in the HR/9, as seen in Figure 4-3.

4.1 HR/9 - Handedness and League

Figures 4-1 and 4-2 show the HR/9 statistic for each player against that player's average drag coefficient in the 2019 season. Figure 4-4 depicts the different league these players are in (National vs. American) and Figure 4-5 shows the difference in handedness among pitchers. 47 out of 151 pitchers are left handed, and 104 are right handed. 66 pitchers are from the National League, and 85 out of 151 are from the American League. Similar to Figure 4-1, there is no apparent correlation between handedness or league in the magnitude of the drag coefficient or the HR/9 for each pitcher, and no evidence to suggest that in this study, handedness or league has any impact on either the average drag coefficient or on the HR/9 for pitchers.

4.2 Further Analysis

It should be noted that the error bars are quite high, and may be caused by high variance in C_d values from ball-to-ball, possibly due to surface roughness [1]. However, it is noteworthy that not all pitchers might want to select for high drag balls. They might want the ball to fly as fast as it can. As discussed in Section 2.4.1, with decreasing drag coefficients comes increasing speeds. This might not be advantageous

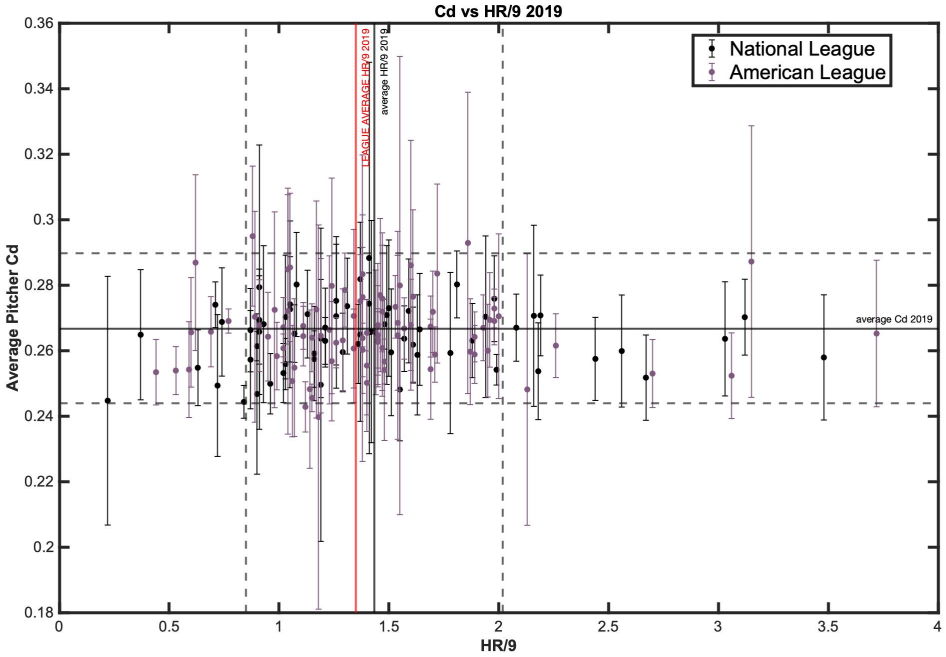

Figure 4-4: Plot of average C_d per pitcher vs. HR/9 in 2019 with focus on league.

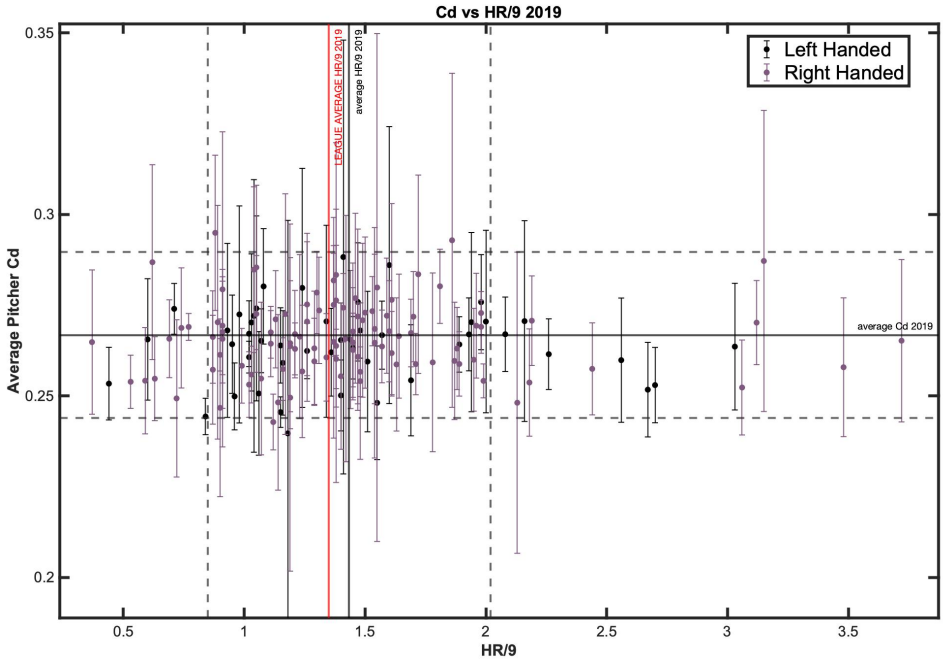

Figure 4-5: Plot of average C_d per pitcher vs. HR/9 in 2019 with focus on handedness.

for the batter, as increasing speeds of the pitch in conjunction with a lower drag coefficient might enable for a better opportunity to score a home run [1], but a ball that can fly fast through the air might be advantageous for the type of pitch the pitcher selects.

Appendix A

Extraction of Pitch

Statcast data provided by the MLB is collected through Doppler radar systems fastened to the top of the stadiums. The radar collects the XYZ coordinate points (ft) of each of the plays. These systems were first installed in 2015, and they're able to track everything on the field at all time, including balls in the wide range of 30 - 120 mph [6].

While the radar system is set to record each play at the start of the pitch, at times the radar will collect more in all of the years including 2019: either capture the end of the previous play before the pitch or capture the beginning of the next play. It is vital that the correct segment of the batted ball is chosen, because if not, it might not be that of the ball that was pitched with, lending to an inaccurate measurement of the drag coefficient. As the drag coefficient is linked to a specific pitcher and will be correlated with their performance, it's very important to select the accurate pitch. Thus, a conservative approach was taken to decide if a file should be used or not, as well as which segment of the data is selected for.

Central questions arose in selecting the correct and noticing repetitive inconsistencies data:

1. At what time does the pitch begin?
2. At what point in the play does the time begin?
3. Are there extra data points from the previous pitch and/or the following pitch?

It's impossible to answer question 1 and 2 initially as there is a lack of consistency around when time begins in the radar files. While it was initially assumed that time(1) would be the very beginning of the pitch where the play itself begins, it may be at the very beginning of the pitch, or in the middle of the batted ball, or in the middle of a play that is finishing just before the one of interest begins. With regards to question 3, if the beginning of the pitch can be located anywhere in the file, it can select for the correct trajectory amongst any extraneous trajectories regardless of

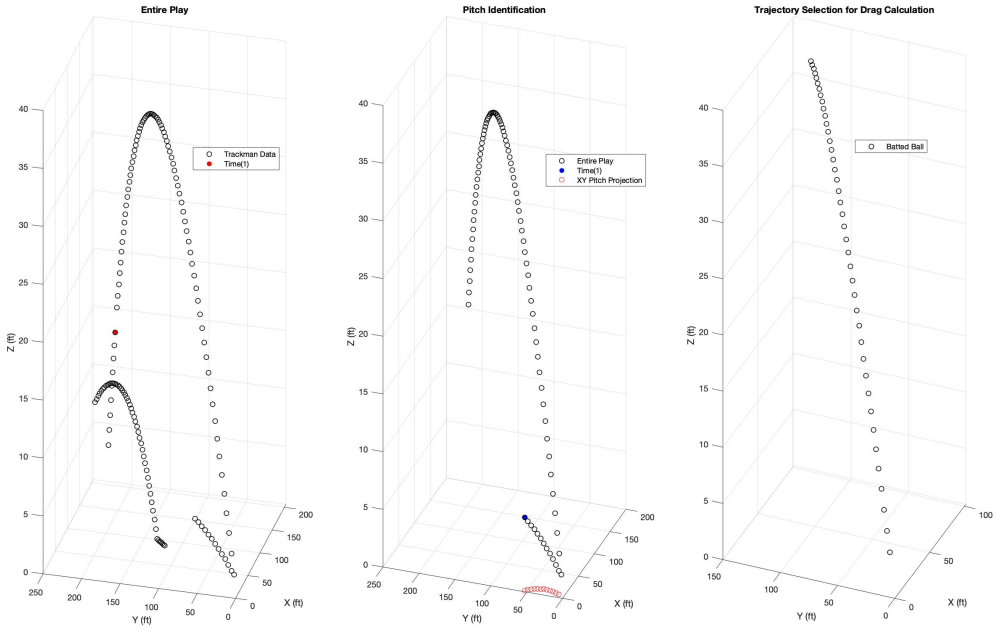

Figure A-1: Extraction of trajectory from sample inconsistent data

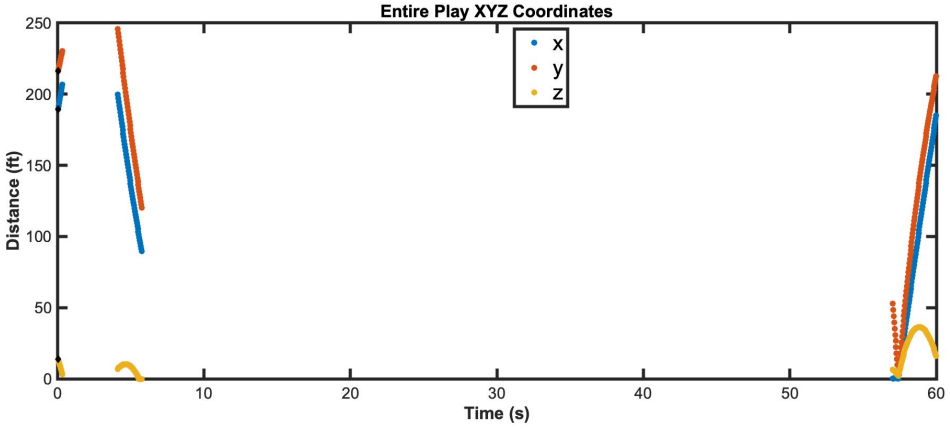

Figure A-2: XYZ trajectory data in time from sample inconsistent data

what extra data might lie before or after the pitch. This is the goal of the program and constraints created: to select for and extract the pitch associated with the rising trajectory of the batted ball.

However one initial assumption is made: the first pitch that is located is the pitch of interest. This assumption was made around the examination of over 4000 data files, where none contained greater than one complete trajectories of interest, where a trajectory of interest can be defined by a completed pitch + batted ball. Figure A-1 and Figure A-2 show an example of inconsistent data, where there is both extraneous data and the data didn't begin at the first point in the pitch, Additionally, there is great time between different aspects of the play, almost an entire minute, which acts as further evidence to suggest that this data did not come from the same play.

To preserve as many data files as possible and, a program was created to isolate the first pitch of the data file, and calculate the drag coefficient off of the subsequent batted ball. Parameters to characterize and locate the pitch are based off of 2 aspects of the flight path (Figure A-3): (1) the average number of points within a pitch and (2) the difference in the height (z) between the first and the last data point in the pitch. This accounts for characterizing the pitch in each of the three planes: (1) estimates the length of pitch in the X and Y planes, and (2) accounts for the change in height in the Z plane. The program also looks for one parameter not within the flight path, but the location of the baseball diamond – specifically the pitchers mound. It was recognized that the files where incorect data was selected, was if there was a play close to the ground from a previous play that was much like the trajectory of the pitch, just at a different location on the field. Thus, a paramter was accounted for to say that if along with the relative size of the pitch (in the XY plane as well as the Z plane) is an identification factor, then the location of these characteristics on the field need to also be a factor. Thus, the program will only select a pitch that is less than 75ft in the Y direction, limiting the presence of stray data from the outfield that could disguise itself as a pitch. The program then looks for windows of data that meets these parameters, and then shifts this window each subsequent point further in the trajectory data until a characteristic trajectory is found.

With regards to (1), the average number of points for all data sets within the pitch is 12 $\pm$1.3. As points cannot be partial, the range of 11-13 of points was selected to define the pitch, and the average change in height is between 4.44 ft and -8.4548 ft (one standard deviation of the mean of -2.02 ft) taking into account that the file could go backwards in time if time(1) is at the end of the positional data, in which the change from the first point to the last point could either be negative or positive depending on which way time leads. It is recognized that there may be changes in height that have the same values, that are much higher in the Z-axis, and not close to where the pitch would be near the ground. Thus, a third parameter was also introduced, which was to say if along with the change in height, if the first and last point are within one standard deviation of the average high between the located pitch, then that will

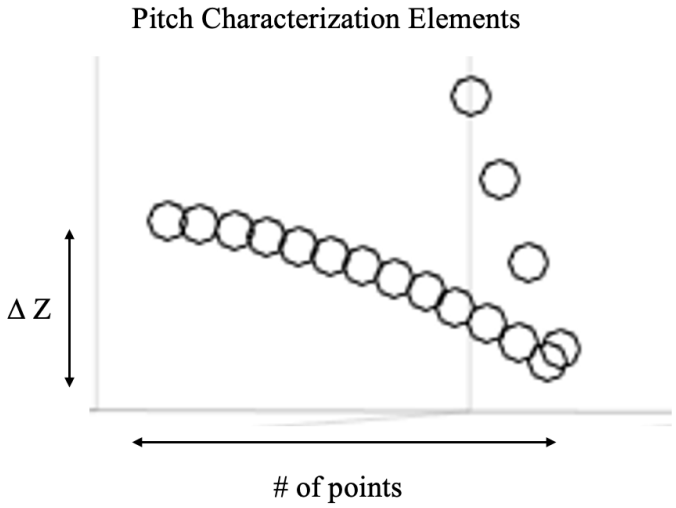

Figure A-3: Elements in pitch characterization

be another characteristic parameter of the pitch.

To summarize, if the number of points in the pitch lies within one standard deviation of the average of these different constraints described previously then the program selects that file to continue with. While the program only selects for 68% of the data files that fall within these constraints, it is done to ensure the best opportunity to choose for the right trajectory within these varying radar files. This is 2x the amount of files that are selected than previously, where any abnormality in the data would cause that entire file to be skipped over.

After finding the pitch, time = 1 is now set at the begging of the located pitch. It is understood that there is a range within the number of points of the pitch, so the program does not locate the beginning of the batted ball at the end of the selected pitch. Rather, the program looks for the abrupt change in slope that is indicative of the ball being struck off of the bat, and determines this as the batted ball.

Bibliography

[1] Bartroff J. Blandford-R. Brooks D. Derenski-J. Goldstein L.-Nathan A. Albert, J. and L. Smith. "Report of the Committee Studying Home Run Rates in Major League Baseball". 2018.

[2] Nathan A. Albert, J. and L. Smith. "Preliminary Report of the Committee Studying Home Run Rates in MLB". 2019.

[3] Aguirre-López M. A. Díaz-Hernández O. Hueyotl-Zahuantitla F. Morales-Castillo J. Escalera Santos, G. J. and F.-J. Almaguer. "On the Aerodynamic Forces on a Baseball, With Applications". *Frontiers in Applied Mathematics and Statistics Journal*, 4, 2019.

[4] C. Frohlich. "Aerodynamic Drag Crisis and Its Possible Effect on the Flight of Baseballs". *American Journal of Physics*, 55.

[5] D. Kagan and A. M.. Nathan. "simplified models for the drag coefficient of a pitched baseball". *The Physics Teacher*, 52.

[6] D. Kagan and A. M. Nathan. "Statcast and the Baseball Trajectory Calculator". *The Physics Teacher*, 55.

[7] Kensrud J. Lyu, B. and L. Smith. "The Reverse Magnus Effect in Golf Balls". *The Physics Teacher*, 55.

[8] A. Nathan. "Analysis of PITCHf/x Pitched Baseball Trajectories". 2008.

[9] L. Smith and A. Sciacchitano. "Baseball Drag Measurements in Free Flight". *Applied Sciences*, 12.