Comparing Batted Ball Location Data and Offensive Success in Major League Baseball

Chris Silos
EM 224: Informatics and Software Development
May 9th, 2020

I pledge my honor that I have abided by the Stevens Honor System
Chris Silos

1. Overview of Research Question:

Since its creation in the late 1800s, baseball has always been a sport of statistical data, with nearly every pitch able to generate multiple different data points. However, throughout the course of the last decade, new forms of data collection have started to transform the game by creating an entirely different source of data, “batted-ball” data, which can be utilized outside of traditional baseball statistics. In the past, data generation largely relied on the outcome of a play, such as a single, out, or home run. These basic data points allow for the calculation of traditional baseball statistics, such as batting average, on-base percentage, or slugging percentage. On the other hand, batted-ball data utilizes advanced sensors to create new data points, such as the speed at which the ball exits the bat (exit velocity), the angle at which the ball leaves the bat (launch angle), or the exact distance and location where the ball lands on the field.

This report will analyze a combination of traditional and batted-ball data to provide insight into the importance of which side of the field the batter hits the ball, and how it could potentially be utilized to increase the overall success of a team. Specifically, it will cover the importance of run differential (the amount of runs a team scores vs the amount of runs they allow) when predicting the winning percentage of a team. Then, a comparison of batted-ball data between the 2018 and 2019 seasons of the Minnesota Twins, as well as a case study of Twins’ player Max Kepler, will show that changes in batted-ball data could explain incredible improvements made in both individual and team performance.

2. Dataset Description:

In total, five datasets were used in this analysis. Four sets come from BaseballSavant, a site that downloads game files from MLB Advanced Media. BaseballSavant is directly linked to the official website of Major League Baseball, and is widely regarded as the most complete source of MLB batted-ball data. The final dataset was retrieved from FanGraphs, a similar site to BaseBallSavant, but with more of an emphasis on traditional statistics, rather than batted-ball data.

The first four datasets are pitch-by pitch logs of the entire 2018 and 2019 seasons of the Minnesota Twins, and pitch-by pitch logs of the 2018 and 2019 seasons of Max Kepler. The datasets for the 2018 and 2019 Twins contain 24,403 lines, and 24,493 lines, respectively. The datasets for Max Kepler’s 2018 and 2019 seasons contain 2,352 lines, and 2,161 lines, respectively. These sets are in exactly the same format as the previous ones. In all four of these datasets, each line correlates to a single pitch, and 63 data points are recorded for each pitch. For this analysis, however, only the columns correlating to the coordinates of the location of the hit, the distance of the hit, the outcome of the hit, the trajectory of the hit, the exit velocity of the hit, and the handedness of the batter were analyzed. Each of these categories will be explained more thoroughly later in the report. The last dataset is a collection of standard statistics for each MLB team in the 2019 season. It only contains 31 lines, with each line correlating to an individual team. It contains 24 columns, but only columns regarding amount of wins, amount of losses, runs scored, and runs allowed were used in this analysis.

3. Dataset Provenance:

The following steps were taken for the data.

1. Individual player datasets were downloaded from BaseballSavant using the "Search" tool
2. 20+ individual player files were merged using the VBA module in Excel for the two larger datasets.
3. Basic manual data cleaning was performed on the Excel files in order to fill-in empty columns with "n/a", and remove any excess commas within cells.
4. The required libraries were imported in order to work with the data
5. A function was defined to produce a "spray-chart" for each individual dataset (spray-charts will be explained more thoroughly later in the report)
6. The function contains an interactive feature that allows the user to select which data they want to view on the spray chart, as well as certain statistics regarding their choice.
7. Dual-bar graphs were created in order to show comparisons between the 2018 and 2019 seasons of both the Minnesota Twins and Max Kepler.
8. A final dual bar graph was created in order to compare the actual vs predicted winning percentage of MLB teams in the 2019 season.

4. Methodology:

A graph was generated to show the accuracy of predicting a team's winning percentage using run differential. Then, a function was defined to produce a "spray chart" for various offensive outcomes from the 2018 and 2019 seasons of the Minnesota Twins and Max Kepler. Cartesian hit location coordinates were converted to polar coordinates, and used in combination with distance values to produce polar coordinates for each hit. These coordinates were plotted onto a quarter-sphere, roughly the same shape as a baseball field. Means were used to calculate certain batted-ball percentages from the Minnesota Twins and Max Kepler from 2018 and 2019 to provide an explanation for their incredible offensive improvements.

5. Structure:

The first graph shows the accuracy of the Pythagorean Winning Percentage in predicting the actual winning percentage of an MLB team. Since this formula is based entirely on run differential, it highlights the importance of scoring large amounts of runs in order to have a successful season. Next, spray charts will show hit location data for the 2018 and 2019 seasons of both the Minnesota Twins and Max Kepler. They will show that there is a heavy correlation between which side of the field the ball is hit, and the amount of home runs a batter can produce, thus increasing their overall run production. Finally, dual bar graphs will show that the Twins' and Max Kepler's offensive improvements may possibly be attributed to an increase in pull percentage. This is interesting because pull percentage is often overlooked in comparison to other batted-ball metrics, such as exit velocity and launch angle (will be explored more in-depth in section 7).

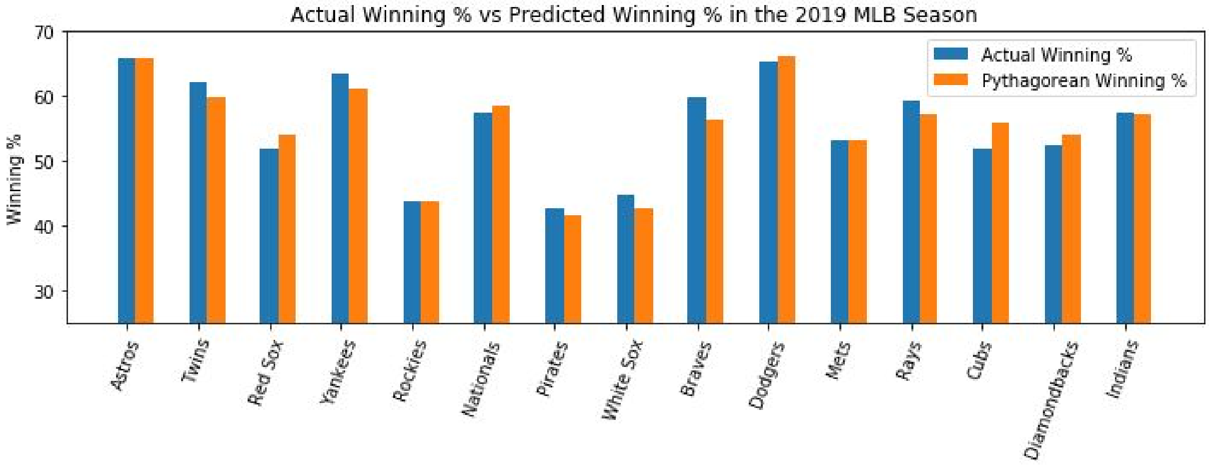
6. Pythagorean Winning Percentage

Run differential is defined as the difference between the amount of runs a team scores and the amount of runs they allow. For example, a team that scores a large amount of runs but allows very few will have a high run differential, while a team that scores few runs but allows many runs will have a poor run differential. Now, it may seem obvious that this is an important factor when predicting the success of a team, but the extent to which it is important is often understated.

The Pythagorean Expectation is a sports analytics formula developed to estimate the winning percentage of a team. Winning percentage is defined as the amount of games a team wins divided by the total amount of games they play. The Pythagorean Expectation is shown below.

$$\frac{\text{runs scored}^{1.83}}{\text{runs scored}^{1.83} + \text{runs allowed}^{1.83}}$$

In baseball, where there are hundreds of different statistics that could be used to predict winning percentage, only two are needed to predict it with incredibly high accuracy. The accuracy of the Pythagorean Expectations for fifteen MLB teams are shown below.



```
Average Difference: 1.73%
```

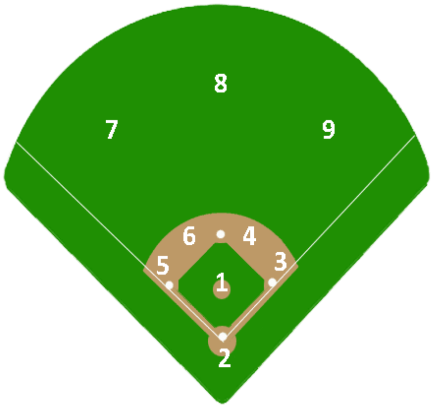
As you can see, the Pythagorean Expectation is usually within 1.73% of the actual winning percentage, which means that it is only about three games off in a 162 game MLB season. This is incredibly accurate for using only two basic statistics, and highlights the importance of scoring as many runs as possible in order to have a successful season. This information is relevant as it shows the value of the home run, the most efficient way to score runs in baseball.

7. Batted-ball Data

7.1 What is Batted-ball Data?

Advanced sensors are capable of tracking a multitude of new metrics that did not exist even as recently as 10 years ago. Currently, two of these metrics, exit velocity and launch angle, have taken the world of baseball by storm, allowing hitters to track exactly what happens every time they make contact with the baseball. Exit velocity is defined as the speed at which the ball leaves the bat of the hitter. For example, more powerful hitters will have higher exit velocities than weaker hitters. Launch angle is defined as the angle at which the ball leaves the bat. A launch angle that is either too high or too low will result in either a poor fly ball or a ground ball, which are not ideal for producing runs. Ideally, a hitter would want to have a high exit velocity, as well as a slightly above average launch angle in order to produce better hits, such as doubles and home runs, instead of weaker hits, such as singles. Because of this, many teams across all levels of the sport have been modifying their hitting methodologies in order to improve these two metrics.

However, an often overlooked subset of data is hit-location data. This data allows you to plot the exact location of a batted ball onto a spray chart. A spray chart is a bird's eye view of a baseball field, with each hit plotted to its respective location. These charts allow for the calculation pull percentage, straightaway percentage, and opposite field percentage. In order to understand these metrics, we first need to understand the terms "pull," "straightaway," and "opposite field."

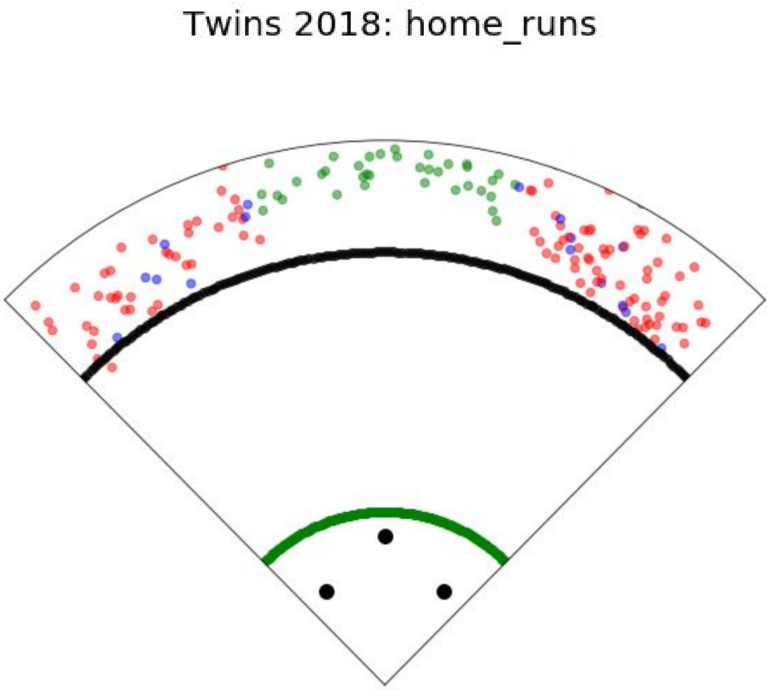


Pictured above is a baseball diamond. Each number on the diamond corresponds to a certain position. Now, a batter can either be right-handed or left-handed. If he is right handed, then he would stand on the left side of home plate in the figure above. If he is left-handed, then he would stand on the right hand side of home plate in the figure above. For a right-handed hitter, the left side of the field (towards the numbers 5, 6, and 7) is considered his "pull" side. The right side of the field (towards the numbers 4, 3, and 9) is considered his "opposite field" side. The middle portion of the field is considered "straightaway." For left-handed hitters, the "pull" and "opposite field" sides are switched, while "straightaway" remains the same.

Using this knowledge, plotting the location of each hit allows for the calculation of how many hits the batter pulled, how many they hit to the opposite side, and how many they hit straightaway. These values allow for the calculation of the pull percentage, opposite field percentage, and straightaway percentage of a hitter.

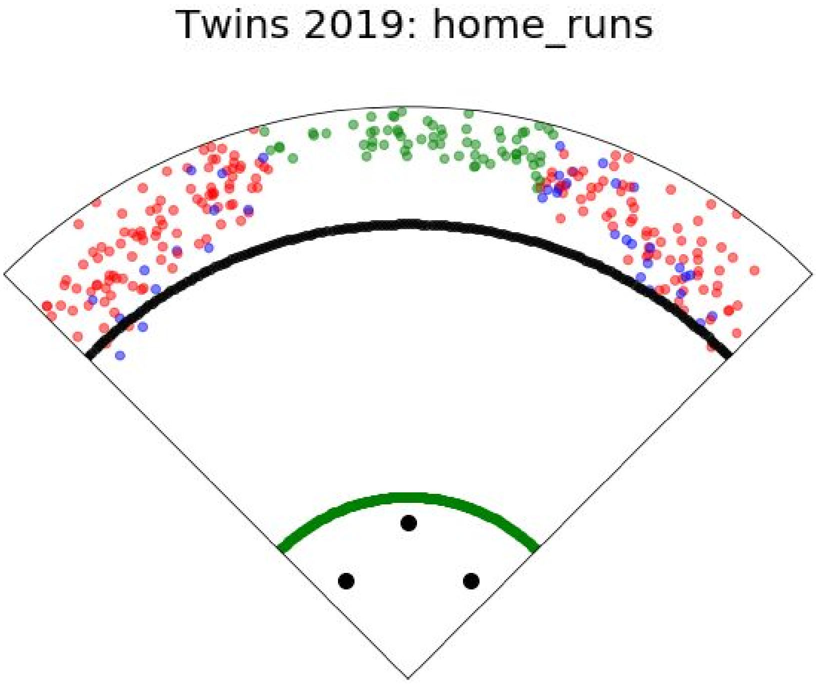
5.2 Minnesota Twins

I chose to analyze batted-ball data of the Minnesota Twins due to their significant increase in performance between the 2018 and 2019 seasons. In 2018 they finished with a record of 78 wins and 84 losses. However, in 2019, they finished with a record of 101 wins and 61 losses, and secured a top spot in the playoffs. They also went from a subpar offensive season in 2018, to breaking the all-time single season home run record in 2019.



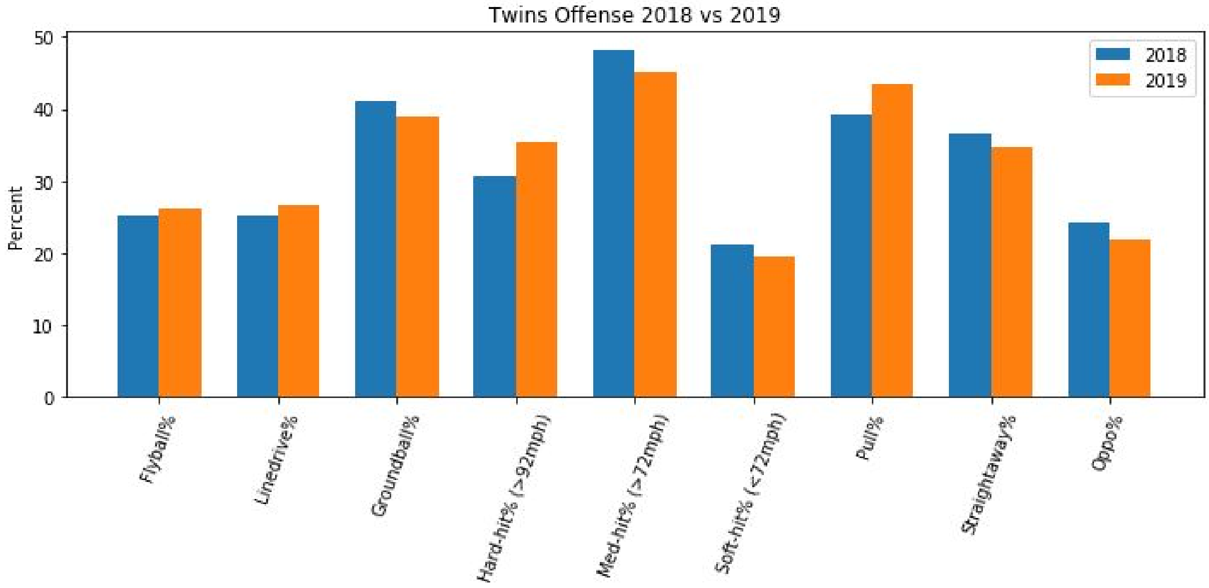
```
Total home_runs : 160
home_run pull percentage: 67.5
home_run straightaway percentage: 22.5
home_run opposite field percentage: 10.0
```

In this spray chart, red dots correspond to home runs that were pulled, blue dots correspond to home runs that were hit to the opposite field, and green dots correspond to home runs that were hit straightaway. Red and blue dots on both sides of the field is a result of right-handed and left-handed hitters having different pull and opposite field sides. The black arc represents a wall located 350 feet away from home plate. The green line and black dots at the bottom of the graph represent the standard dimensions of a baseball infield.



```
Total home_runs : 296
home_run pull percentage: 62.8
home_run straightaway percentage: 24.7
home_run opposite field percentage: 12.5
```

Both of these spray charts show that the large majority of home runs from the Twins are pulled. Knowing the importance of scoring runs as efficiently as possible, it is reasonable to infer that pulling the ball can have a significant impact on scoring more total runs.

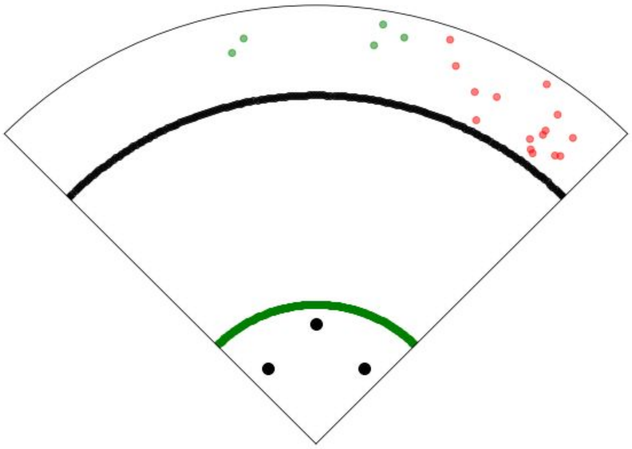


Using other values found in the dataset, we are able to calculate several more batted-ball metrics. As we can see here, in 2019, the Twins hit more fly balls, and had a higher average exit velocity than they did in 2018. This fits quite nicely with our earlier reasoning that many teams are trying to increase both their exit velocity and launch angle. However, we also see that the Twins had a significant increase in their pull percentage, and a significant decrease in their opposite field percentage, which is often overlooked.

5.3 Max Kepler

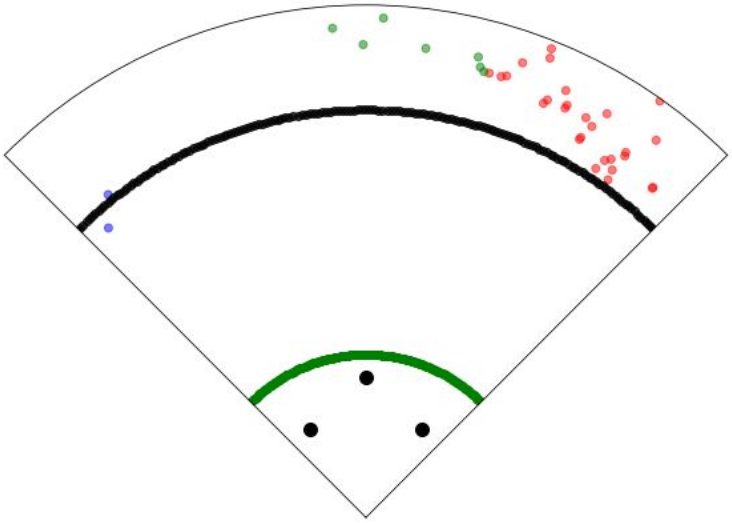
Max Kepler is a player on the Minnesota Twins who had a significant improvement between the 2018 and 2019 seasons. He nearly doubled the amount of home runs he hit (20 to 36), and even received votes for Most Valuable Player in 2019. His individual batted-ball data reveals information quite similar to that of the Minnesota Twins.

Max Kepler 2018: home_runs



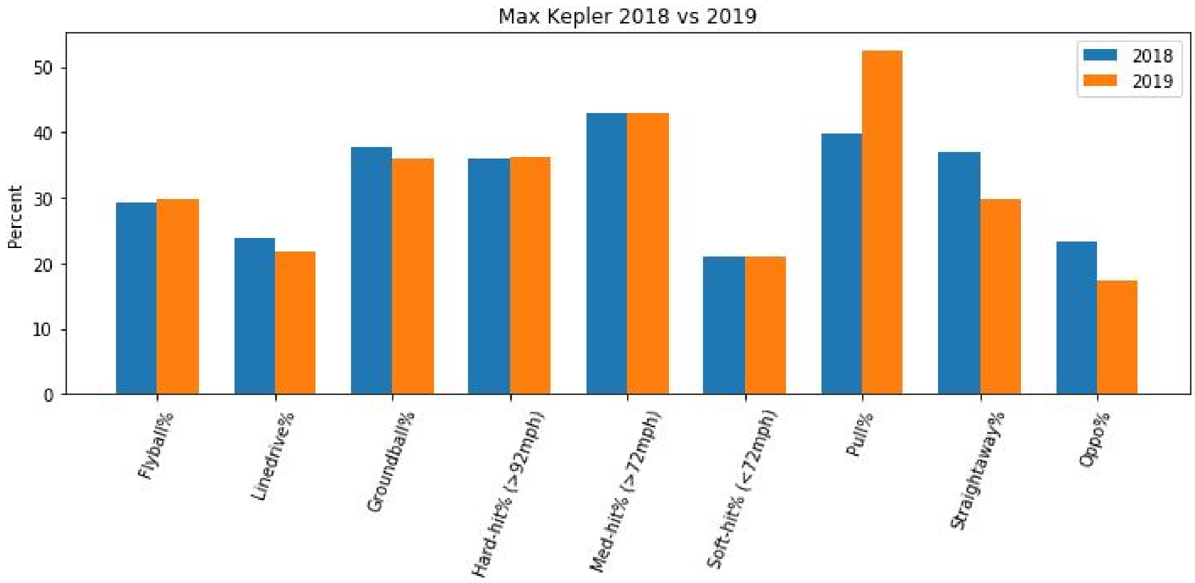
```
Total home_runs : 20
home_run pull percentage: 75.0
home_run straightaway percentage: 25.0
home_run opposite field percentage: 0.0
```

Max Kepler 2019: home_runs



```
Total home_runs : 36
home_run pull percentage: 75.0
home_run straightaway percentage: 19.4
home_run opposite field percentage: 5.6
```

Again, these spray charts show how nearly all his home runs are hit to his pull side, suggesting that pulling the ball results in more home runs.



Looking at more of his batted-ball metrics, we can find some more interesting information. Unlike the Twin's metrics, Kepler did not have any significant changes with stats relating to exit velocity or launch angle. However, he experienced a very large increase in pull percentage (over 10%), with a significant decrease in straightaway percentage and opposite field percentage. Since these were the only batted-ball metrics that changed from one season to the next, it is possible that they could explain his large increase in offensive success from 2018 to 2019.

6. Conclusions

Analyzing the 5 datasets, we reached the following conclusions:

1. Runs scored is an important factor when predicting a team's winning percentage
2. Over 60% of the Twin's home runs have been hit to the batters' pull side
3. The Twins experienced significant increase in pull percentage between the 2018 and 2019 seasons
4. 75% of Max Kepler's home runs were hit to his pull side
5. Max Kepler's pull percentage increased over 10% between the 2018 and 2019 seasons, while not seeing a significant change in other batted-ball metrics

These results indicate that hitting the ball to the pull-side of the field increases the likelihood of producing a home run, the most efficient method of scoring runs in baseball. Scoring more runs increases the predicted winning percentage of a team, which correlates strongly with actual winning percentage. Currently, the most widely used batted-ball metrics in baseball are exit velocity and launch angle, with significantly less emphasis placed on hit location data. In upcoming seasons, more teams may begin to utilize hit location data to complement other metrics in order to improve their overall offense.