Altitude and Its Effect on MLB Park Factor

Kevin Nguyen
University of Colorado Boulder
Boulder, CO
Keng3875@colorado.edu

Samuel Busser
University of Colorado Boulder
Boulder, CO
sabu0818@colorado.edu

Blaise Page
University of Colorado Boulder
Boulder, CO
blpa1200@colorado.edu

Phillip Arsenian
University of Colorado Boulder
Boulder, CO
phar0872@colorado.edu

1 PROBLEM STATEMENT/MOTIVATION

Since its creation in 1993, Coors field has notoriously been one of the most hitter friendly parks in the major leagues due to the high altitude. The Rockies have tried to combat this by storing baseballs in a humidifier. However, there are several parks across the league that are also considered hitter friendly that are not at a high elevation.

Figure 1: Humidors in Coors Field to keep the baseballs from getting dry

Our goal is to look at teams and players stats throughout the stadiums played in the MLB (pitching, hitting, and fielding) and try to determine how big of an impact altitude has on players stats and the game of baseball in general.

2 LITERATURE SURVEY

This idea of the park affect at Coors Field and many other stadiums across the league has been debated for a long time, thus there is lots of work on this topic already. However, most of this work deals with teams at a general level, and does not consider the effects of altitude specifically.

$$PF = \frac{\left(\frac{Run\ Scored(Home) + Runs\ Allowed(Home)}{Number\ of\ Games\ at\ Home}\right)}{\left(\frac{Run\ Scored(Road) + Runs\ Allowed(Road)}{Number\ of\ Games\ on\ the Road}\right)}$$

Figure 2: Current Park Factor Formula

The MLB has instituted a statistic called park factor. However, this statistic does not focus on the altitude of the stadium or anything local to the stadium in general. Instead, it is calculated by comparing a team's runs scored in games at home versus a team's runs scored in games on the road. We will differentiate from this, as we will focus on

team's offensive, defensive, and pitching performances at different altitudes.

Since the turn of the century, statistics and data analytics has become an increasingly important aspect of the game of baseball (famously coined "moneyball"). This kind of thinking has revolutionized the game, but again the work done in moneyball is not the work we will be doing. The moneyball revolution emphasized on base percentage and prioritizing players who get on base in any way possible. While we may use on base percentage as one of our statistics to evaluate the difference altitude makes in the game, we will not solely be looking at this one stat.

Park Factor Stat:
http://www.espn.com/mlb/stats/parkfactor

Moneyball: The Art of Winning an Unfair Game by Michael Lewis

3 PROPOSED WORK

Data Collection: Each member in our group has downloaded the dataset from the retrosheet database on our respective machines.
Link to database: http://www.retrosheet.org/

Preprocessing: Our project requires us to omit all attributes and events that are *not* useful for determining how altitude affects the offense in a ballpark. These attributes include (but are not limited to) For example, a steal is an example of a specific event in the retrosheet database. This type of event will not be helpful for our project, so we are safe to remove these events. We can also do this for defensive interference, caught stealing, pickoffs, wild pitches, passed balls, balks, and several more. In addition to removing specific events, we can also remove certain attributes from each event that we do keep. Each event has a large number of attributes, many of which are much too specific to be useful for us. Some examples of these attributes are the handedness of the pitcher and hitter, what players were playing each position in the field, which players were on base, what player was on deck, and many more. We also have to account for NULL entries in the dataset. These are very few and far between, as RetroSheet has done a good job of providing a complete database. Because there are so few NULL entries, we can simply remove these entries and not have to worry about losing significant data. Doing all of this will help our results look much cleaner and easier to handle. More importantly though, it will greatly improve the speed of any calculations we will be doing with the data.

Process for derived information: As we have stated, Retrosheet contains every event from every game. However, it does not contain player or team statistics. This is where the bulk of our project lays. We will have to go through every event and compute the statistics ourselves. For example, if we are looking for all the home runs that occurred at Coors Field, we would have to go through every event that had the event code of a home run, and had a home team ID of Colorado. We will have to do this for every statistic that we wish to compute.

Design: We plan on collecting offensive, defensive, and pitching statistics for each ballpark, along with the elevation of each park. One statistic we will use will be home runs. This is a big one, as higher elevations are supposed to make the ball fly further, thus leading to home runs. We will also calculate the total number of runs scored in each park on average, as this is also a good indicator of the total offense in a game. We will also look at pitcher's combined ERA at each park. This will give us an idea of how many runs are allowed by pitchers. At high altitude, balls will not break as early as at low altitude, which can lead to a lot of easy to hit pitches. Fielding percentage is a defensive statistic we will look at. High altitude helps the ball fly further but also faster. Due to this increase in speed, it can be harder to field the ball, leading to more errors and lower fielding percentage. These are just a few of the statistics that we will calculate.

Evaluation: Once we have all the statistics we need, we can dig into our problem statement. To first prove that altitude has any affect at all on baseball games, we can perform a hypothesis test, with our null hypothesis being: Altitude has no effect on the game, and our alternate hypothesis

being: Altitude has an effect on the game. If we are able to reject the null hypothesis, then we can confirm altitude does effect the game. We can then use correlation analysis to determine which stats are most affected by altitude.

There could be several uses for this information if we do find a correlation between altitude and some of the stats. For example, if someone was looking to create a new team, they could consult our findings and choose a city with an according altitude. Teams could also use this information when drafting players. If, for example, we find that home runs and altitude are very strongly correlated, then a team like the Rockies should look for good home run hitters, as they play half their games at a very high altitude.

4 DATA SET

Our dataset is from the Retrosheet database. This database contains every single play that has occurred in every MLB game since 1989. In addition to what happened on each play, this database also contains how many outs there are, how many runners are on base, how many pitches were thrown in the at bat, what specific kind of pitch was used, where the ball was hit, who was at each fielding position, and much more.

Our dataset for the MLB events spans between 2005 and 2015 seasons. Our dataset contains approximately 2,097,152 usable objects, which include both players and teams. The dataset also contains 41 attributes for which the corresponding team/player statistics are generated.

5 EVALUATION METHODS

Our evaluation method will begin with comparing the altitude of each ballpark with the statistics we get from each ballpark We will see if there is a correlation between the two. We will also be using a null hypothesis to see if all the stadiums are similar in statistics, and we will test this using confidence intervals to see if we can reject the null hypothesis or fail to reject it.

6 TOOLS

The database that we will be using to derive our conclusions contains every single play in the MLB since 1989 including a variety of attributes that won't be helpful in exploring the park factor of a team. For this reason, one of the first tools we will be using is MySQL Workbench which will allow us to edit our database to consider only data that would be helpful in researching the effects of altitude in MLB statistics. This first step will make subsequent steps more manageable by eliminating a lot of the noise in our dataset and allowing us to more easily focus on the data that is more important for our project.

Our primary tool for our project will be a Jupyter Notebook which will allow us to manipulate our data and perform calculations through the use of python code and libraries. In order to easily manipulate our data, we will be using the PANDAS library. This library will allow us to smoothly load our dataset onto our Jupyter Notebook and will handle parsing our data into rows and columns. For most of our calculations, we will be using the NumPy library which offers additional functionality for large multi-dimensional arrays. Using NumPy supported functions will make our statistical analysis of our data much easier by allowing us to easily calculate important properties of our data such as the mean, standard deviation, and variance of specific attributes of our data. We will also use NumPy to run correlation tests and create confidence intervals to support our conclusions. We will be using matplotlib to visually represent our data through the use of different visualization tools such as boxplots, scatterplots, and histograms. These visualization tools will help us better support and explain our findings.

7 MILESTONES

Our first step in this project will be to get the data on each one of our devices. We have completed this milestone as of March 1st.

Our next step will to be clean the data, and remove all unnecessary columns and data. We will want this done by March 12th. Once we have that done, we will need to combine the events of each game and have them organized. This should be done by March 19th.

Lastly, we will have all the data we need, so now we need to run our correlation tests, and confidence intervals to get all the comparisons we need. This will need to be done by April 13th, and once that is done we will need to conclude on our data and evaluate what it means. This will be done by April 30th.

8 SUMMARY OF PEER REVIEW SESSION

One piece of advice from the peer review session that we found to be important is the previous work done in analyzing MLB statistics. The MLB relies heavily on statistical analysis for both the recruitment of players as well as in preparation to play opponents. By fault, there is already a lot of previous work in analyzing team and stadium statistics. We will continue to research the methods that different baseball franchises implement to evaluate the park factor, and we hope that this research will guide us to develop an optimal process for calculating the park factor.