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.

One study published in 2008 used regression analysis to model the effects of four factors (altitude, temperature, humidity, and barometric pressure) on the flight of a baseball or softball under different atmospheric conditions [Bahill, A. Terry & Baldwin, David & S Ramberg, John, 2009]. The study was able to show that on a typical July afternoon, altitude clearly has the biggest impact on ball flight accounting for 80% of the variability between ball flight in different stadiums. The study directed most of its attention towards the force acting opposite of the relative motion of a moving object which is known as the object's drag force. The study determined that the drag force of a ball travelling through the air is weakest in Denver and strongest in San Francisco due to the altitude differences of both cities. This would suggest that if two balls were to be hit with an identical force and direction in both Denver and San Francisco, the ball in Denver would travel further under similar weather conditions. This conclusive information is already a good indication of altitude certainly providing an unfair advantage for teams playing at higher elevations. However, altitude is not the only factor affecting the ball flight of a batting hit for a given stadium. In fact the study also mentions that Coors field, which is the stadium with the highest altitude, has a high probability of experiencing high winds opposite to the direction of typical ball flight in a game. This factor decreases the number of home runs in the stadium and nearly compensates for the stadium's high altitude. Therefore, it is extremely hard to determine if altitude provides teams with an unfair advantage in terms of ball flight in their home stadiums because of the all the other factors that may also affect the travel flight of a baseball in each stadium. For this reason, we wish to focus most of our attention on the statistics recorded for each MLB game between the years 2005 and 2015. Of course, we will also have to account for the different variables that can affect the travel flight of a ball in each stadium. However, our conclusions can certainly be used to either support or refute previous work done on the altitude effects of different stadiums in the MLB.

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

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

Bahill, A. Terry & Baldwin, David & S Ramberg, John. (2009). Effects of altitude and atmospheric conditions on the flight of a baseball. International Journal of Sports Science and Engineering.

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 2,111,526 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.

7.1 MILESTONES COMPLETED

Our first milestone of getting the data on all of our machines has been completed. We first had to download the .csv file. Once this was done, we then integrated this data into a dataframe in Pandas and a table in MySQL.

We have also finished cleaning our data, at least for now. Depending on where the project goes, we may end up removing more attributes or rows, or adding more back in. For now, we have removed several columns from the database that will not be needed for our project. Some examples are time_since_1900 (measures the time since 1900 to that game), game_id, bat_hand_cd (simply whether the batter was left or right handed), base runner IDs, pitch sequence transactions, ball and strike count, and more. Some of those are pretty self-explanatory as to why we won't need them. Some of them, however, could prove useful in other projects, but not for ours. For example, we don't need whether a hitter was left or right handed, as we are concerned with offensive statistics as a whole.

We also added the corresponding altitude of the location of each event in the database. By doing this, we will now be able to create graphs and charts comparing offensive stats to altitude. Because Retrosheet does not provide us with specific stats for each team, we have begun calculating offensive stats for each team as well. We are able to do this through the event code column in Retrosheet. The event code column has a number corresponding to what exactly happened in the at bat. So, for example, we have calculated the number of home runs that occurred at each stadium, as well as the number of total runs.

7.2 MILESTONES TO DO

One big milestone that we have left to do is to officially decide which statistics we are going to use in showing whether there is a correlation between offensive production and altitude. So far, we know we want to use total runs scored and total home runs at each stadium. But this is just a top level look. There is much more we can look at. For example, it has been shown that ground balls at stadiums with higher altitude will have a faster velocity. We could use this piece of information to compare the total number of errors at stadiums of high and low elevation. Additionally, it has been proved that pitchers cannot get a ball to break as much at higher elevations. Because of this, if they try to throw some form of curve ball, it might not curve as much as they are used to, leaving the ball in a better spot for the hitter to hit. This could lead to pitchers having worse pitching statistics at stadiums of higher elevations as well. So ERA and runs per 9 innings are two more stats that we will look at in the future.

Another future milestone we have is to integrate park factor into our dataset. While the park factor stat that the MLB has implemented does not use the same information we have, it would still be interesting to compare our results to results that have been found using a different method.

Finally, our last significant milestone remaining is to perform all of the data science correlation calculations on our data to see if there really is a correlation between offensive production and altitude. This will include hypothesis tests, computing correlation coefficients, and r-squared values.

8 RESULTS SO FAR

With the data set we have cleaned so far, we have been able to create different visual representations to compare the altitude of each stadium with the statistics of that stadium in between the years 2005 and 2015. The first diagram we created was a bar chart that mapped the frequencies of home runs to each teams' home stadium. This bar chart allows us to visualize the variations present in the number of home runs scored at each stadium. While this diagram only shows home runs, it gives us some insight into our problem already. Before we saw the results, one would expect Coors field to have the most home runs since it is at the highest elevation. As can be seen in the graphs below, Colorado is in the top ten, but it does not have the most home runs hit. The team that had the most home runs hit in their stadium was the Cincinnati Reds, with the New York Yankees and Baltimore Orioles close behind. Interestingly, the elevation of these three cities are 482', 280', and 480' feet respectively. These are all miniscule compared to the mile high elevation at Coors Field, and yet they all have a higher home run rate than Coors Field does. Again, this is only one stat that we have looked at so far, so it is not conclusive. But home runs are a pretty significant stat in the altitude park factor argument, so the fact that the highest altitude does not have the highest park rate is very interesting.

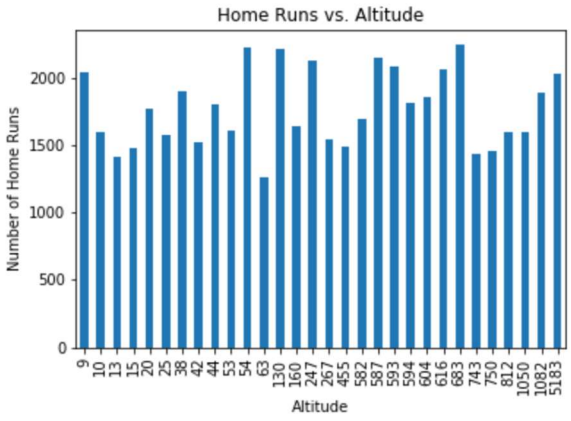

Figure 3: Bar chart showing number of home runs hit at different altitudes

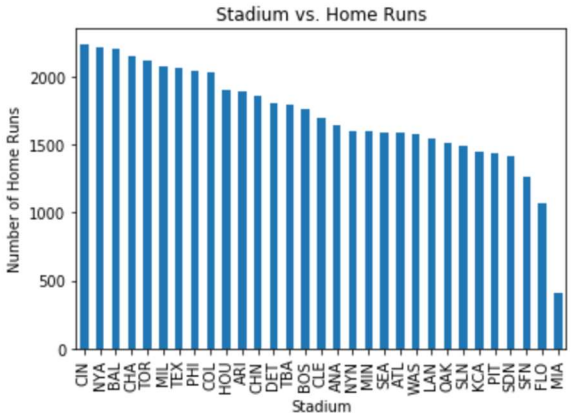

Figure 4: A bar chart showing number of home runs hit in each stadium

Our next step was to create a diagram that would help us better isolate the effects of elevation on the number of home runs hit at different altitudes. Therefore, we created another bar chart that compared the number of home runs hit at specific elevations. From this graph we were able to see that most the home runs were definitely hit at regions of higher elevations. However, the impact of altitude on the number of home runs hit was definitely more subtle than originally expected and there were certainly many exceptions to the general trend. For example, the stadium with the lowest altitude (New York Yankees) recorded more home runs than the stadium with the highest altitude (Colorado Rockies). There are also many stadiums with high altitudes that reported less home runs than average and many stadiums with low altitudes that reported more home runs than average. Because altitude did not seem to create a huge difference in the number of home runs recorded, we are unable to use this bar chart to reach any conclusions. However, this only shows that there are other statistics that have to be accounted for before determining the effects of altitude on offensive statistics.