The Effect of Travel on Baseball Performance

Abiel Fattore
University of Colorado Boulder
abiel.fattore@colorado.edu

Brandon Zink
University of Colorado Boulder
brandon.zink@colorado.edu

Cameron Connor
University of Colorado Boulder
cameron.connor@colorado.edu

Problem Statement/motivation

We are going to observe the effect of both distance traveled and the number of time zones crossed on offensive and defensive performances across many advanced statistics. We hope to see whether traveling eastward or westward or traveling in general has a negative impact on player statistics. For offensive, we will be looking at statistics from runs scored, batting averages (BA), on base percentage (OBP), slugging (SLG), and isolated power (ISO). For defensive we will be looking at statistics from runs allowed, earned runs average (ERA), home runs per nine (HR/9), strikeouts per nine (K/9). Any notable observations could somehow be used by MLB teams to figure out how to combat any negative impacts when traveling or if there are any positive impacts they could use that to their advantage. Also, if allowed, the observations could help gamblers make more informed decisions on the teams they bet on. Our main motivation is to get meaningful results and expand on previous studies.

Literature Survey

Previously, there has been one study that focused on the effect of traveling on a team's statistics, "How jet lag impairs Major League baseball performance" by Ravi Allada [1]. The study looked at 46,535 MLB games that were played between the years of 1992 and 2011. The study looked at both offensive and defensive statistics on team winning percentages and individual player statistics.

Comparing the home and away performance, the study had concluded that MLB teams that had athletes traveling eastward had a decline in performance while teams that had athletes traveling westward had no significant change. Basically, Allada is saying that jet lag had a greater negative impact when traveling from the west to the east. Extending from that, the decline in performance when traveling eastward had to do more with defensive performance.

Data Set

We will be using the Retrosheet Events data, from http://www.retrosheet.org/. Having more than three million data objects, the Retrosheet Events has data on every single AB in the MLB since 2000. Some of the attributes are what player was at what position, was there an LF error, game identification, away team identification, how many strikes, how many outs, who was pitching, who was batting, what type of throw, what type of het, did it count as and AB, etc. The attributes types are either going to be symmetric binary such as if the team was away or home, or numeric. There are more useless columns, useless meaning that we won't be using them in our analysis, than there are useful columns. We only want the columns that are relevant offensive/defensive information. There is also some sparsity. Some of the objects have missing data.

Proposed Work

Since we have our problem statement, the first thing that will be done will be cleaning the data with Jupyter Notebook. We will apply data reduction, more specifically, dimensionality reduction. We aren't interested in columns like what player was at a certain

[1] Rachael LallensackJan. 23, 2017 , 5:00 PM et al. 2017. Jet lag puts baseball players off theirgame. (December 2017). Retrieved March 5, 2018 from http://www.sciencemag.org/news/2017/01/jet-lag-puts-baseball-players-their-game

position, so any useless column/attribute will be purged completely. In the columns that we do want, we will want to remove the objects with incomplete data so that we don't run into any errors processing the data. Also, we will make sure that all the data in each column are of the same type. We also want to get rid of any noisy data. We will do this by clustering relevant data (attributes) and removing the outliers. We could semi-supervise the data but that wouldn't be efficient due to the millions of data points. In the end we will have a new dataset consisting of only relevant columns filled with meaningful data.

After cleaning we will aggregate the play-by-play statistics, using SQL, into game-by-game statistics. So, by applying appropriate queries, we will take the list of every AB and turn it into a list of every game with the specific data that we want.

We will then process the dataset. Here we will calculate the distance traveled/time zones crossed by the MLB teams using the home latitude and longitude, the away latitude and longitude, and the radius of the earth and then plugging this information into a distance formula. We will be getting the longitude/latitude values by google searching since the stadiums are always going to be in the same places. Those values will be put into a new dataset along with the team name, team id, and time zone. We can then apply the distance formula to the appropriate columns. We will aggregate statistics, using Pandas, by looping through every game in the dataset and place them in a data frame based on the distance/time zone information and then calculate the previously mentioned statistics. Then we will sort them into the appropriate areas. This is where we differ from the previous study as Allada only looked at the basic statistics while we go on to evaluate more advanced statistics that include the basic statistics. This will hopefully give us more information on the effects of travel in baseball.

Evaluation Methods

Once we are done aggregating the data and computing the statistics we should have a chart where we have various offensive and defensive statistics so that we can analyze them.

We will be using Hypothesis Testing such as the chi-squared test to evaluate our results. With our chart where the x-axis represents the number of time zones crossed and the y-axis represents the distance traveled by the team like, for example,

		Time Zones						
		-3	-2	-1	0	1	2	3
Distance (miles)	<250							
	250-500							
	500-1000							
	1000+							

we will be able to evaluate our results. Each of the empty cells will have the appropriate statistics that we calculated. We will then use the chi-squared test to compare the statistics to the MLB average, giving us p values for each statistic. We can observe if there are significant changes in offensive/defensive performance. Based on the previous study, we should expect to see a decline in player performance as a team travels towards the east and insignificant impacts as a team travels west.

We will use probability models to see if the differences we find are relevant and interpret in game effects by team (do the Yankees have a noticeable advantage over Mariners since those athletes tend to travel less).

Tools

SQL/MySQL Workbench will be used by converting the dataset from play-by-play statistics into game by game statistics.

Python/Jupyter Notebooks will be used for any math analysis or any analysis in general. Numpy/Pandas

libraries will be extremely useful for processing the data. Also, it will be used to clean the dataset in the ways we described earlier.

The Google Maps API may be used to find latitude and longitude values for when calculating distance traveled/time zones crossed.

Milestones

2/27/2018

Get data – Find a couple of datasets and agree on one dataset for the project.

3/6/2018

Clean data – Clean the data. Get rid of incomplete and noisy data and, also, irrelevant attributes.

3/20/18

Categorize the data into distance/hour – Using our clean/transformed dataset we will categorize the data into distance/hour.

4/3/2018

Compute basic and advanced statistics for each category – Using Jupyter Notebooks we will process the data and compute advanced statistics for offensive/defensive and organize them into a chart.

4/10/18

Statistics tests on different sections – We will compare each statistic in each different section of the chart to the MLB average.

4/24/18

Interpret and find follow up analysis based on findings – Interpret our final results and draw a conclusion on what our findings say about the effect of travel on player performance.

Summary of Peer Review

A great question that was asked during our presentation was how we were going to analyze the data. Our answer to was to use the chi-squared test but we weren't too sure. Furthering our understanding we will be using the chi-squared test because it does provide the best way to analyze our final chart. Also, that question extended to how we are going to apply our analysis. We hadn't really thought about the too much until then. We responded with seeing if west coast MLB teams started at a disadvantage since they had to travel more in terms of deciding where to go. That also got us thinking about how betting decisions could become more accurate based on the analysis.

It came to our attention that using the Google Maps API could be problematic since we might not be able to get all the data we need in our allotted time. We will probably have to find a different way to get that information. The best was way we right now would be to google the location data since, as previously mentioned, the stadiums do not move.