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

1 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 in general has a negative impact on player/team 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). Additional statistics include BB/9 and H/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.

2 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.

[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

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.

This is like what we are trying to figure out except we are not checking to see if the direction of travel has an effect on baseball statistics but just if traveling in general has an effect.

3 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.

4 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 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. If we find that we are deleting too much data then we will resort to automatically filling in incomplete cells. We could fill in those cells with the average value of the column. 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 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 and team id. 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 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.

After getting our result, we will calculate the adjusted run value for each team for each season to see the number of wins that are lost or gained based on travel. So, because of traveling, did teams lose games, or gain wins, or both. Also, we will calculate the adjusted win percentage. We want to see what would have been each team's win percentage each season if they had not traveled. This will be done using the Pythagorean winning percentage formula.

With these last two calculations we will then be able to come to a conclusion(s) on traveling and baseball performance. An example would be that if we found that a team that traveled during the season had worse statistics than statistics based on no traveling, we could say that traveling had a negative impact on baseball statics. This is an oversimplified example but gives a view on what type of conclusion we are going for.

5 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 T-test to evaluate our results. The x-axis of our chart, which is represented as Chart 1, represents the average statistics for teams at certain distances from their home stadium, the p values for those distances, and the MLB average. The y-axis will be the various calculated batting and pitching statistics. The distances also represent the different time zones.

Once we fill out the average statistics for each distance we will use the T-test to calculate p values to later determine whether our calculated statistics are significant or not. Based on the previous study, we should expect to see a decline in performance as a team travels because they will be traveling east or west. Since we are calculating more statistics we will be able

to find more proof. Or we might find other observations.

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).

5.1 Chart 1 - Comparison of statistics from at each distance to the MLB average.

	MLB Averages	< 500 miles p val	< 500 miles avg.	500 - 1000 miles p val	500 - 1000 miles avg.	1000 - 2000 miles p val	1000 - 2000 miles avg.	> 2000 miles p val	> 2000 miles avg.
Batting R/9									
Batting H/9									
Batting BB/9									
Bating K/9									
Batting BA									
Batting OBP									
Batting SLG									
Batting ISO									
Pitching H/9									
Pitching BB/9									
Pitching K/9									
Pitching ERA									
Pitching ISO									

6 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.

7 Milestones

2/27/2018

Get data – Search for interesting data sets and agree upon one to use for our project. Also, make sure that these datasets fit the project requirements.

3/6/2018

Clean data – Clean the data. Get rid of incomplete and noisy data and, also, irrelevant attributes. If too many objects are purged due to incomplete data then we may fill them in automatically.

3/20/18

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

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. Based on the p values we will see if the difference of the two averages are significant or not.

4/14/18

Calculate adjusted run value for each team – We will calculate the adjusted run value for each team for each season using the results from the previous milestone.

4/20/18

Calculate adjusted win percentage – Calculate the adjusted win percentage for each team for each season to see what their win percentage would be if the team did not travel.

4/22/18

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.

7.1 Milestones Completed

Got data – We got our dataset. The dataset, Retrosheet Events, has data on every single AB in the MLB since the year 2000. It has fits the project requirements by having more than three million data points.

Cleaned data – We cleaned the dataset by removing irrelevant attributes, unnecessary characters, and removing any objects with incomplete data. We did not automatically fill in any empty cells because the data set still had more than a million data points to work with after purging, still fitting the project's requirements.

Calculated proposed statistics – We calculated BA, OBP, SLG, ISO, ERA, ISO, etc.

Categorized our calculations - Organized our calculated statistics into our chart to be able to begin evaluating and reaching a conclusion.

Statistics testing – Using the T-test we were able to compute p values for each or the statistics at each distance/time zone. Each cell of the chart is now filled and we can compare are calculations to the MLB averages.

7.2 Milestones Todo

Calculate adjusted run value for each team – Given our filled-out chart we still need to do some work to come to a concrete solution. This should be simple calculations that shouldn't take us too long.

Calculate adjusted win percentage – Just like the previous todo milestone we still need to do some work

to come to a concrete solution. This is a simple calculation that shouldn't take too long to compute.

Interpret and find follow up analysis based on findings – After all calculations are done we can come to a concrete conclusion with proof and provide other perspectives on the effect of travel on baseball statistics.

8 Result So Far

8.1 Updated Chart 1

	MLB Averages	< 500 miles p val	< 500 miles avg.	500 - 1000 miles p val	500 - 1000 miles avg.	1000 - 2000 miles p val	1000 - 2000 miles avg.	> 2000 miles p val	> 2000 miles avg.
Batting R/9	4.469147	0.980672	4.469785	0.185479	4.505821	0.330036	4.495314	0.001634	4.363489
Batting H/9	8.867632	0.576710	8.851423	0.002866	8.958878	0.955658	8.869281	0.001258	8.747855
Batting BB/9	3.159069	0.862545	3.156151	0.133418	3.132522	0.884506	3.161555	0.043506	3.202651
Bating K/9	6.903588	0.801893	6.909313	0.056095	6.857684	0.008663	6.964527	0.139710	6.860518
Batting BA	0.251447	0.568042	0.251081	0.001180	0.253635	0.786985	0.251270	0.001975	0.248911
Batting OBP	0.315477	0.723249	0.315239	0.077472	0.316729	0.880163	0.315374	0.105661	0.314083
Batting SLG	0.395052	0.459431	0.395966	0.025486	0.397960	0.980203	0.395083	0.000051	0.388665
Batting ISO	0.143605	0.111501	0.144885	0.394182	0.144325	0.799237	0.143813	0.000167	0.139755
Pitching H/9	9.408321	0.037480	9.343602	0.000757	9.519189	0.100543	9.460520	0.000151	9.257242
Pitching BB/9	3.529970	0.097254	3.498981	0.154305	3.502003	0.190263	3.554947	0.016006	3.587730
Pitching K/9	6.695536	0.103370	6.733333	0.001323	6.617616	0.201703	6.725580	0.897956	6.699317
Pitching ERA	4.261071	0.655474	4.272210	0.221306	4.293226	0.335458	4.285616	0.000308	4.146361
Pitching ISO	4.581866	0.214000	4.602778	0.347773	4.565336	0.872696	4.579127	0.751365	4.575079

Our results are given by this chart. The green highlighted values represent positive changes in baseball statistics while the red highlighted values represent negative changes in baseball statistics. The cells with red filling represent p values, calculated from our T-test, that are less than 0.05. This means that the changes at that distance on batting or pitching statistics are not significant. If the cell is not filled in with red then the changes at that distance on the batting or pitching statistic are significant.

At less than 500 miles from home stadiums, teams had positive changes in Batting R/9, K/9, SLG, and ISO. There were negative changes in Batting H/9, BB/9, BA, and OBP. For pitching, teams had positive changes in K/9, ERA, and ISO while having negative impacts in BB/9 and H/9.

All the changes were significant excluding the Pitching H/9.

At between 500 and 1000 miles from home stadiums, there were positive changes in Batting R/9, H/9, BA, OBP, SLG, and ISO while there were negative changes in BB/9 and K/9. There were positive changes in Pitching H/9 and ERA while there were negative changes in BB/9, K/9, and ISO.

The changes that were not significant were Batting H/9, BA, and SLG and Pitching H/9 and K/9.

At between 1000 and 2000 miles from home stadiums, there were positive changes in Batting R/9, H/9, BB/9, K/9, SLG, and ISO while there were negative changes in BA and OBP. There were positive changes in Pitching H/9, BB/9, K/9, and ERA while there were negative changes in ISO.

The change that was not significant was just Batting K/9.

At greater than 2000 miles from home stadiums, there were positive changes in Batting BB/9 while there were negative changes in the rest. There were positive changes in Pitching BB/9 and K/9 while there were negative changes in the rest.

The changes that were significant were Batting K/9 and OBP and Pitching K/9 and ISO.

Looking at these results, there were more significant results in pitching and batting performances when teams were less than 500 and between 1000 and 2000 miles away from their home stadium. When teams were greater than 2000 miles away from their home stadiums there were very few significant changes. Of the significant changes for each of the distances, there were seven positive changes in the less than 500 miles column, four in the 500-1000 miles column, nine in the 1000-2000 miles column, and one in the greater than 2000 miles column. This is interesting because it would be expected that as distance increased or decreased that statistics would solely be negatively or positively impacted. Here we see that certain statistics are negatively impacted at certain distances while being positively impacted at others.