Abstract

A study about sleep cycles[1] found that there were improvements in pattern recognition, reaction speed, and muscle contraction strength during a peak performance window of a human's biological cycle. We sought to investigate if travel between time zones had any effect on a game's outcome due to this biological performance peak. We chose Major League Baseball to study this effect as they play the most games a year with the shortest duration between games. We examined the effect on team performance, as measured in wins and losses, of visiting teams traveling from time zones different than that in which their home games are played. Our analysis involved building several logistical regression models in order to see if our target variable, Minutes Past Biological Noon (MPBN), was a predictor for the likelihood of the home team winning or losing.

Introduction

An important concern among fans and professional sports leagues is in maintaining a high degree of fairness in play across all teams and players in the league. If travel was a predictor of outcomes and player performance, bettors and fantasy owners would find it invaluable to know who to bet on or start in their fantasy leagues. Cote's study concerning biological rhythms found that people perform best on mental and physical acuity tests in the late afternoon as measured with respect to their body clock. However, it did note that: "Because international competition is itself arousing, laboratory findings may not indicate any real advantage of competitive performance." The problem we wish to investigate is what effect, if any, traveling across time zones has on team performance in Major League Baseball. Our hypothesis is that a player's biological circadian rhythm results in them having an interval of peak performance during the day, and that those teams which have more players than their opponents playing during this peak performance interval will demonstrate a slight edge in winning percentage over their opponents. If such an effect can be shown it could serve to aid those looking for an edge in forecasting a game's outcome and would be of keen interest to professional sports leagues so that they could work to assure fairness of play between competing teams.

Data Preparation

Our dataset consisted of the outcomes of all regular season MLB games played during the 2012, 2013, 2014, and 2015 seasons. This dataset consisted of data merged from three different data sources: baseball-reference.com[2], the mlbgames[3] python module, and stevetheump.com.[4]

---

[1] Cote, K. (2012). "Sleep, Biological Rhythms, and Performance" in Encyclopedia of Human Behavior (0-12-375000-8, 978-0-12-375000-6), (p. 435).

[2] Sports Reference LLC. "2015 Major League Baseball Team Statistics and Standings." Baseball-Reference.com - Major League Statistics and Information. http://www.baseball-reference.com/.

baseball-refererence.com is a freely available website that maintains detailed baseball statistics going back to the inception of the game of baseball in the late 19th century. Among the data that is maintained on this site are tables consisting of the game-by-game results for every team, every season. Our study used these tables as the primary source of data on game results, as the tables included columns that we could either directly use as explanatory variables or that we could use in computations to derive other needed variables. Each table is provided separately on this site for each team for each season. The relevant columns in this table to our study was the game date, the name of the opposing team, the result of the game (win or loss), the cumulative win/loss record of the team on that day, the duration of the game, and the win or loss streak that the team was amid on that day. The table for each team provides a chronologically ordered list of the results of each game played by the team for that season. The site provides a useful link to download each table contents as a CSV (comma separated values) text file. Thus, to obtain our complete data, each table of results for each team and for each season was downloaded separately, then imported and concatenated together in Microsoft Excel.

A critical variable that was missing from the baseball-reference.com data was the game start time. Due to the nature of our focus on circadian rhythm, this was a critical value for our study. This value was obtained for each game using a lookup to data archived for the MLB Gameday web service. This service is a publicly exposed and freely available API that serves XML formatted game data primarily for MLB's proprietary Gameday application, which provides web multimedia coverage of games for subscribers to MLB's service. In addition to this API, an open-source python module, mlbgame, can be used to access this data from a python program. The module provides a python code interface to this API by which scripts can query baseball data. The data that is queried is cached locally by the mlbgame module, preventing unnecessary internet bandwidth.

For the purposes of this study, a python script was written which accepted as input the concatenated CSV file described above. The main purpose of this script was to obtain the start time for each game in the data set and convert this to a numerical value that could be used in a regression model. Our python script thus read in each line in the exported CSV file and for each line, it performed a lookup of that game in the mlbgame module using the date of the game and the home team. The mlbgame module provided the game start time for each game in EDT (Eastern Daylight Time) in the format *(hours):(minutes)*, and our script converted this to a value of *minutes past local noon* by parsing this value, and doing the necessary *hours* * 60 + *minutes* calculation. To determine local noon with respect to EDT, each home city's local time zone was looked up from a separate file that our team produced.

Our code performed a further refinement on the game start time metric by introducing the value *Minutes Past Biological Noon (MPBN).* This value is the number of minutes past a team's "biological" noon (12:00 noon in the visiting team's home time zone) that the game started. To calculate this value, our script calculated the time zone differential (TZ_{diff}) between the EDT (the

---

[3] Zach Panzarino, Python MLB Game Module, computer software, version 2.2.1, Python Package Index, July 30, 2016, https://pypi.python.org/pypi/mlbgame/2.2.1.

[4] Steve Orinick, "Steve O's Baseball Umpire Resources," Steve O's Baseball Umpire Resources, 2016, http://www.stevetheump.com/.

time zone for which all game times were reported) and the team's *home time zone* (the time zone of their home city), and used this number, which could be positive or negative, in the following formula:

For visiting teams, this value should capture the effect of time zone offset that this team would be feeling at game time, since TZ_{diff} will be different than that of the home team. MPBN was calculated in our data set for both home and visiting teams for each game, and each was used as a separate explanatory variable in our models. They were used as the primary explanatory variables for our hypothesis that travel disruptions to circadian rhythm affects overall team performance as measured in wins and losses.

Further, two additional variables were calculated in the python script which we hypothesized might affect the models using MPBN. These variables were *Prev_Games_In_Series* and *Prev_Games_In_Time_Zone*. In major league baseball, it is common for teams to play each other in a sequence of games played on consecutive days typically lasting two, three, or four days. Such a sequence is referred to as a *series*. It seemed plausible to us that the longer a visiting team spent in each city, the more that that team would become accustomed to a different time zone. In the first iteration of our python script, as we processed the games for each team in chronological order, we calculated the number of consecutive days that the same two opponents had played. This running total was the Prev_Games_In_Series variable. Later, a refinement was made to our models, and Prev_Games_In_Series was dropped from our models in favor of a new variable: Prev_Games_In_Time_Zone. The reason for this switch is that Prev_Games_In_Time_Zone should have better explanatory power, since in baseball most teams play away series in multiple cities consecutively, in what are called *road trips*. Since it is common for teams to organize their road trips to visit cities that are geographically close, it is common that a visiting team spends many days in a foreign time zone playing multiple opponents. Additionally, for teams that embark on a road trip where the first opponent is located in a city in the same time zone as their home city (say, Boston and New York, for example), there is no time zone adjustment. In such cases, the Prev_Games_In_Series does not encode the correct adjustment effect while Prev_Games_In_Time_Zone does.

To examine the time zone adjustment effect in models including MPBN, we needed to account for other factors in each game datum that might affect win or loss outcome. One obvious factor affecting the outcome of a game is the quality of the teams participating. One would expect "better" teams to win in head-to-head contests with "lesser" teams. Our study utilized two explanatory variables to account for team quality: *Differential_Winning_Percentage*, and *Differential_Team_Payroll*. The python script produced Differential_Winning_Percentage for each game by computing separately the home and visiting teams' winning percentages on the day of the game and taking the difference, home winning percentage minus visiting winning percentage. The individual team percentages themselves were calculated by making a first pass through the entire data set and building a dictionary keyed by team and game date, with a value of the winning percentage for the team on that date. The dictionary values were computed by parsing the "W-L" raw character data column from the baseball-reference.com export, which had the character format *wins-losses*, and computing *wins/(wins + losses)*. The Differential_Team_Payroll variable was calculated as the home team's total team payroll minus

the visiting team's payroll, measured in millions of U.S. dollars. All team payroll data was provided by a static CSV file downloaded from stevetheump.com. This data consisted of total team payroll on opening day of each season and was sourced by this website from "documents obtained from the MLB Players Association, club officials and filed with Major League Baseball's central office." The python script loaded the contents of this CSV file in a dictionary keyed by team and year, and for each game looked up the corresponding salary values for each opponent and computed the difference.

By using the preparation steps described above, all the explanatory variables in our study were numerical. The response variable for each item in our data set was the outcome of the game: a home win or loss. The python script encoded this result in the variable *Home_Win* as a binary variable, with 1 indicating the home team won the game and 0 indicating a loss.

A data cleaning step that we undertook in our study was to remove duplicate game entries. Unless removed, each home game would appear elsewhere in our data set as some other team's road game. To prevent the presence of these duplicates from muddying our regression computations, prior to model generation in our R scripts we removed the duplicates from the data frame under consideration by removing all away games, which were indicated by the presence of a "@" character in the *Home_Away* column.

Data Summary

Since 19th century, Major League Baseball (MLB) has become one of the most popular sports in the United States and developed into multi-billion-dollar industry[5] with TV money and multi-million dollars of players. Since becoming one of the main sports in the United States, a lot of strategies and systems have been developed for various reasons based upon statistical analysis. As with any professional sport, the probability of winning the games in Major League Baseball depends on various factors characterized by either external or internal influences. For example, Major League teams could experience up to three-time zone changes on a flight to an away game, the overall effect of disrupted circadian rhythms to the whole team theoretically could impact results and change the results of a season.

Initially, there were a total of 21 columns for our dataset. We altered the base data set to isolate significant features for our final data set as we wanted to explore four significant explanatory variables for our final model: Biological Offsets, Win Percentage, Payroll, and Game Duration. These variables were either already components of our original dataset or computed through Python/R scripts which were explained above.

[5] Maury Brown, "Major League Baseball Sees Record $9 Billion In Revenues." SportsMoney. December 10, 2016.http://www.forbes.com/sites/maurybrown/2014/12/10/major-league-baseball-sees-record-9-billion-in-revenues-for-2014/#6975e8c96cb2.

Year	Date	Tm	Opp	Home/Away	W/L	R	RA	Run Diff	Record
2012	Sat May 7	ARI	STL	Away	L	6	9	-3	14-16
2012	Sun May 8	ARI	STL	Away	L	1	6	-5	14-17
2012	Mon May 9	ARI	STL	Away	L	2	7	-5	14-18
2012	Weds May 11	ARI	SFG	Home	W	5	1	4	15-18
2012	Thurs May 12	ARI	SFG	Away	L	2	5	-3	15-19
2012	Fri May 13	ARI	SFG	Away	L	3	7	-4	15-20

Diff Win %	D/N	Home Team MPBN	Visiting Team MPBN	Diff Team MPBN	Game Duration (Minutes)	Home Pay	Away Pay	Pay Diff
0.466667	N	460	520	-60	204	74.28483	110.3009	-36.016
0.451613	N	460	520	-60	147	74.28483	110.3009	-36.016
0.4375	N	460	520	-60	177	74.28483	110.3009	-36.016
0.454545	N	460	400	60	150	74.28483	117.6207	-43.3359
0.441176	N	370	310	60	190	74.28483	117.6207	-43.3359
0.428571	D	130	70	60	163	74.28483	117.6207	-43.3359

Biological Offsets

Sports are no stranger to statistical analysis. From armchair quarterbacks to the ones setting the betting lines in Vegas, knowing or predicting the outcome of games is a very important to a large swath of the viewing public. Our analysis decided to target baseball and the effect of changing time zones on game outcomes. Baseball games are the most frequent of American major league sports with roughly 162 regular season games a year between 30 teams, though the distribution is heavily weighted to the Eastern time zone. The league also is broken up into conferences based on geographic location of the teams and so as one would expect the average offset of the biological times between teams on average is around 0. The way we calculated this offset is based on the away teams' home location time zone subtracted from the home team's time zone and then converted this into minutes. We also considered game start times to get an overall +/- Minutes Past Biological Noon (MPBN) along with the overall offset.

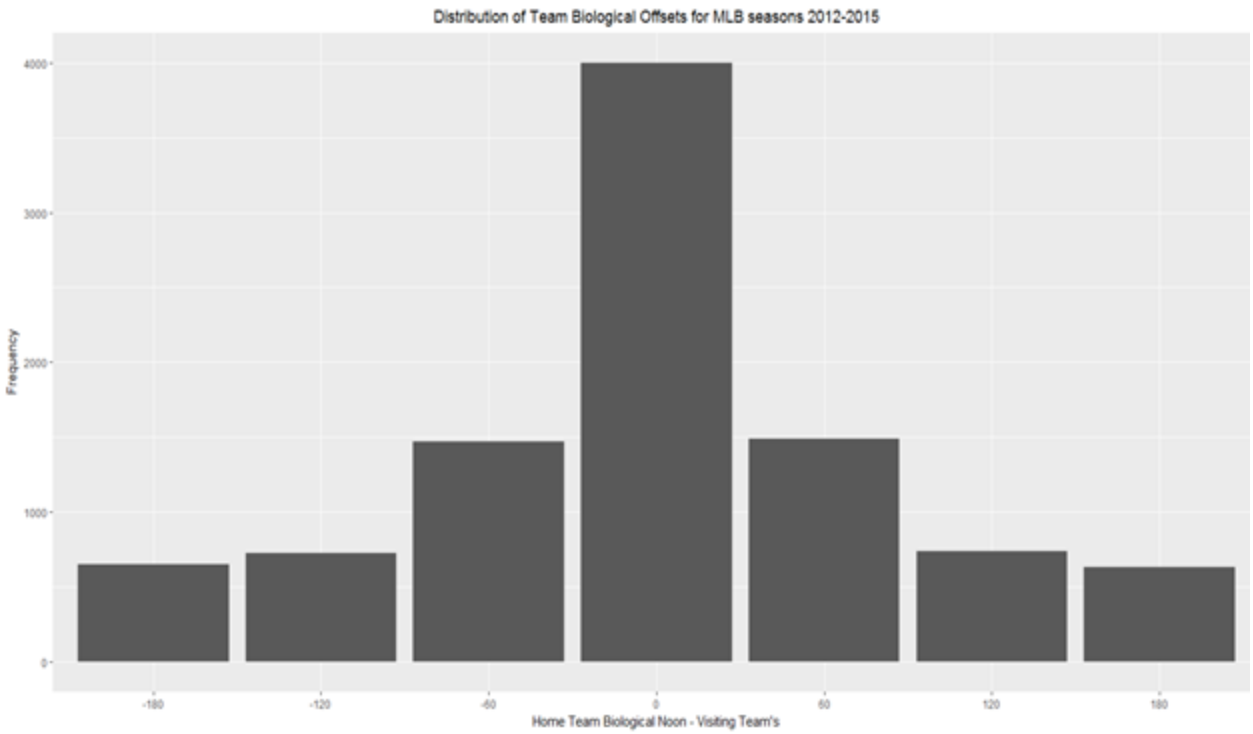

Figure 1: Histogram of Biological Offsets

Since America has four time zones the maximum and minimum offsets are between -180 and 180 minutes. From the histogram above you can see much of games are played with no offset between the home and away teams. The Frequency is the total games played based on time zone offset (with duplicates removed.)

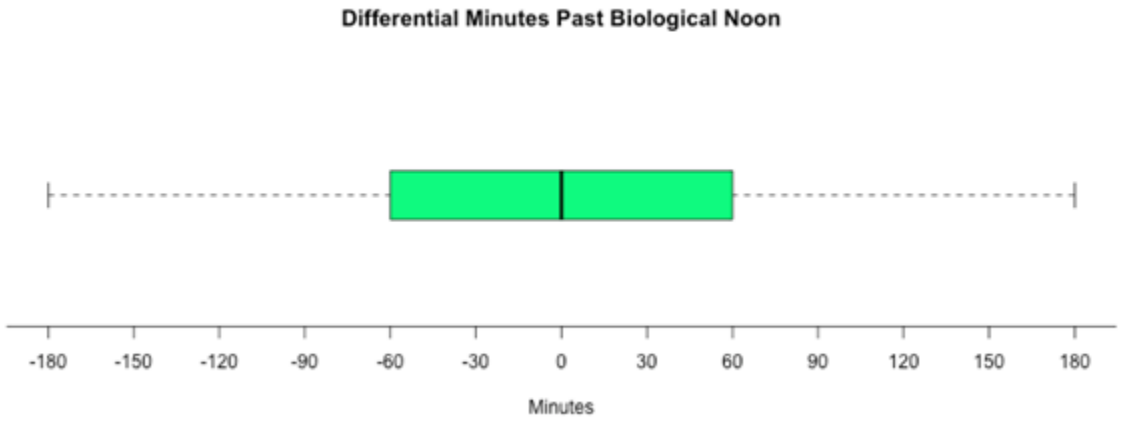

Figure 2: Box and Whisker Plot of Offsets

The offsets are pretty evenly distributed with a six-number summary as such:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-180.00000	-60.00000	0.00000	-0.03088	60.00000	180.00000

The one take away from this data is that slightly more games are played with a negative offset when measuring from their biological noon since the mean is slightly below zero.

Win Percentage

When we compare the win loss percentage we actually find that home teams won less from our analysis than what MLB predicts as the home field advantage, which they predict at 54.2%[6]. Our data gives home teams a 53.6%-win advantage over visiting teams when we do a raw comparison between game outcomes and total played though MLB used game data from the past decade instead of the smaller timeframe of our data. Our data might be closer to MLB's numbers if we could access older data from their API.

When we break down win percentage based on time zone offset we get a bit more interesting breakdown of how the time zone differences affect the overall and per offset win percentage.

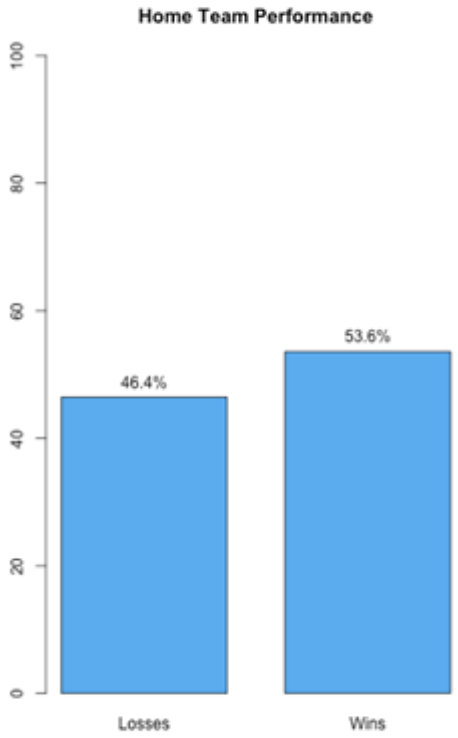

Figure 3: Overall Home Team Win/Loss %

Offset (Min)	-180	-120	-60	0	60	120	180
Total Win %	0.03705	0.040349	0.077817	0.220586	0.087287	0.038600	0.033968
Total Loss %	0.02995	0.034379	0.073700	0.191147	0.066289	0.037776	0.031085

[6] Phil Rogers, "All-Star Game Victory Brings Important Edge to World Series," Major League Baseball, July 15, 2014, http://m.mlb.com/news/article/84750574/phil-rogers-all-star-game-victory-brings-important-edge-to -world-series/.

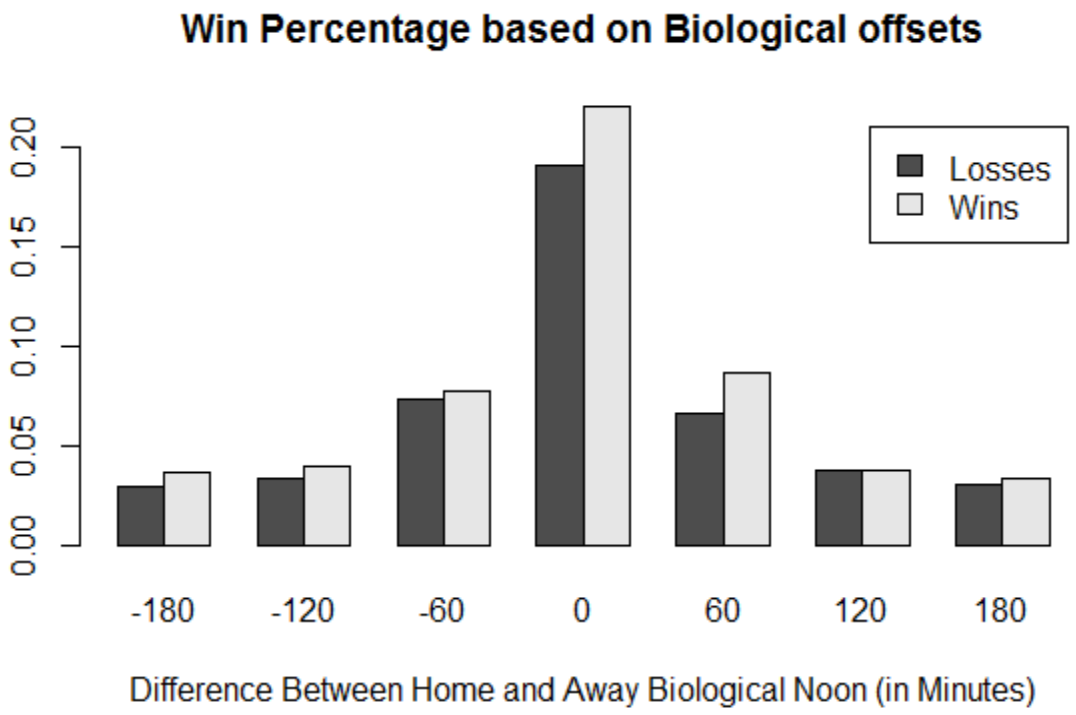

Figure 4: Distribution of Win/Loss % by Offset

This breakdown shows that at no point does the home team ever lose its advantage no matter what the time zone offset is, though at some points the win percentage is much higher and others the margin slims considerably. This can be seen when we examine the marginal distribution of this data:

Offset	**-180**	**-120**	**-60**	**0**	**60**	**120**	**180**
Win %	0.552995	0.539944	0.513587	0.535750	0.568364	0.505390	0.522151
Loss %	0.447004	0.460055	0.486413	0.464250	0.431635	0.494609	0.477848

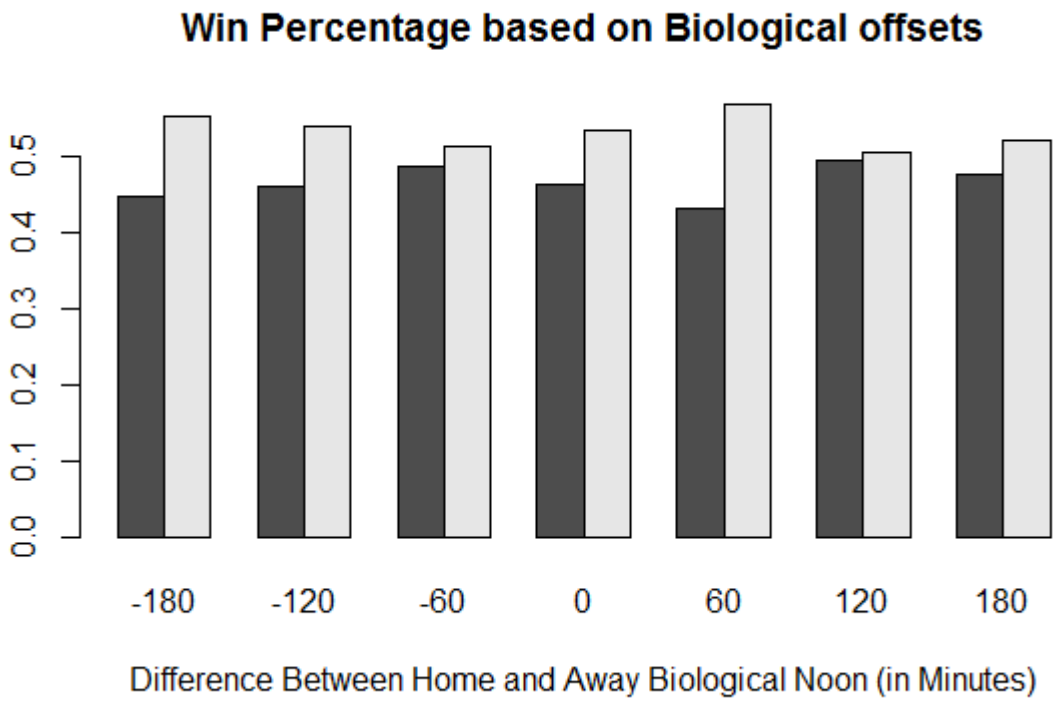

Figure 5: Marginal Distribution of Win/Loss % by Offset

The legend for this chart is the same as above with light gray meaning wins and dark gray meaning losses. This marginal distribution analysis proves interesting because we see that -120 and 0 categories are roughly the same as the overall win percentage, the -180 and 60 offsets are a few percentage points higher than the average win percentage, and the remainder are lower than the average win percentage with the 120-offset having only a half a percentage difference between the win and loss percentages for the home team. What is interesting is that these offsets don't cluster and are not on opposite ends of the distribution, rather the greatest and smallest difference in win percentage are offsets that are closest to one another.

Payroll

Another approach we took was to test whether a large variation in payrolls among the teams is correlated with winning. Team payroll is a distinct feature in MLB compared to other sports which have a salary cap. In the MLB teams are free to spend as little or as much on players as the franchise can afford. For large market teams like the New York Yankees and Los Angeles Dodgers they often outspend smaller or rebuilding teams at a significant amount. For example, in 2013 the Houston Astro's spent 22 million on players while the New York Yankees spent 228 million, a factor of more than 10. As players come to the end of a contract and enter free agency, teams who have a lot of cash and need quality players at that position will often outbid smaller clubs to buy these players thus inflating their payrolls and ideally increasing the talent on their team.

The MLB payroll data was collected from www.stevetheump.com/Payrolls.htm, and from our analysis we find that there is a correlation between teams who spend a greater amount on players and their overall win percentage. This variable being significant seems reasonable for baseball, though in other sports this may no longer prove valuable as teams' payroll varies by much smaller margins.

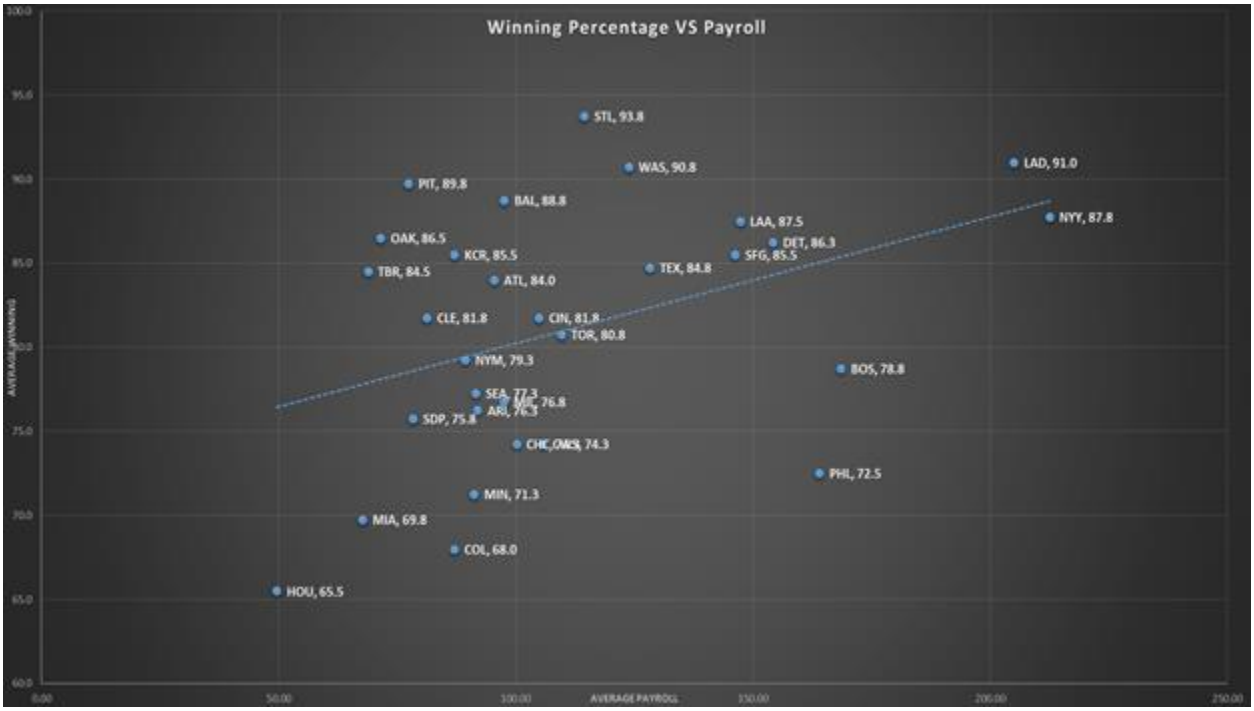

Figure 6: Win % vs Payroll

Moreover, the teams who qualified for the World Series between 2012 – 2015 tend to have higher payroll compared to other teams which did not qualify for the postseason. For instance, San Francisco Giants have been in top 10 MLB teams in payroll between 2012 and 2015, (Rank: 8->6->7->5) and have won two World Series Championships in 2012 and 2014 respectively. Though teams who have the highest payrolls are not necessarily guaranteed to win as in Figure 6 you will notice the Philadelphia Phillies who between 2012-2015 had one of the highest payrolls yet did not find success due to money promised to underperforming and players past their prime.

Game Duration

From our dataset, we converted the game time to minutes to easily calculate how the game duration affects the game results in both home and away games. Regarding game duration, the partial correlation coefficient for game duration was seen to be negative and very highly statistically significant ($p < 2 \times 10^{-6}$ for almost all models), thus meaning that shorter games favor the home team winning. This is consistent with an understanding of the different ways that baseball games can be won, namely that home teams are guaranteed to receive "last bats" and that a game is won immediately when the home team takes the lead in the ninth inning or later.

Conversely, if the visiting team takes the lead in the ninth inning or in extra innings, the home team must bat for their half inning, thus resulting in a longer game time.

Model Building

The response variable for all models in our study was Home_Win, a binary variable indicating whether the home team won or lost a particular game. Therefore, all models that we considered in this study utilize the logistic regression procedure, whereby a model predicts the probability of a success (home win) or failure (home loss). Following the standard procedure, we evaluated the significance of a model coefficient using the Wald-test, the null hypothesis being that the coefficient was zero. For evaluating the quality of our models, we partitioned our data set into test and training sets, randomly assigning 20% of the data to the test set and using the remaining 80% as the training set. The model was built using the training set, and we then computed the classification accuracy of the model using the test set, using a threshold value of 0.536, corresponding to the computed home field advantage that home teams enjoyed in our data.

The first model that we considered included only the explanatory variables directly related to our hypothesis: visiting and home team MPBN:

Home_Win ~ Visiting_Team_MPBN + Home_Team_MPBN

Coefficient values, Wald-Z values and the corresponding z-test p values are shown in Table 1.

Variable	**Estimate**	**Z - value**	**p - value**
Visiting_Team_MPBN	9.634e-05	0.412	0.68068
Home_Team_MPBN	-9.335e-05	-0.353	0.72414

Table 1 – Model including only MPBN variables

It is apparent that neither of the MPBN variables pass the Wald test, which agrees with the correlations results seen with MPBN and Home_Win.

Given that there was no direct effect seen for MPBN in our data, we considered the possibility that its effect, while possibly still present, was being drowned out in the noise that could otherwise be explained by other explanatory variables, so for our second model, we included all explanatory variables. The results for this model are shown in Table 2:

Home_Win ~ Differential_Win_Pct + Differential_Team_Payroll + Home_Streak + Away_Streak + Home_Team_MPBN + Game_Duration_Minutes + Prev_Games_In_Timezone + Visiting_Team_MPBN

Variable	**Estimate**	**Z - value**	**p - value**
Differential_Win_Pct	0.4840	3.176	0.00149

Differential_Team_Payroll	1.080e-09	2.818	0.00483
Home_Streak	-0.003724	-0.351	0.72546
Away_Streak	-0.01507	-1.390	0.16463
Home_Team_MPBN	4.225e-05	0.141	0.88813
Game_Duration_Minutes	-0.01330	-14.999	< 2e-16
Prev_Games_In_Timezone	0.002774	0.902	0.36714
Visiting_Team_MPBN	3.630e-07	0.001	0.99892

Table 2 – Model including all explanatory variables

The variables in this model which passed the Wald test were Differential_Win_Pct, Differential_Team_Payroll, and Game_Duration_Minutes. The classification accuracy of this model (based on a 0.536 threshold) was 0.6105.

Neither Home_Streak nor Away_Streak were significant in the second model. Because these variables were not related to this study's hypothesis about game time and were present only as "quality of team" measures to assist in reducing noise, we removed them. Prev_Games_In_Timezone also showed no significance by itself, so for the time being we removed it and ran a third model, the results for this model are shown in Table 3:

Home_Win ~ Game_Duration_Minutes + Differential_Win_Pct + Differential_Team_Payroll + Home_Team_MPBN + Visiting_Team_MPBN

Variable	**Estimate**	**Z - value**	**p - value**
Game_Duration_Minutes	-0.001331	-16.868	< 2e-16
Differential_Win_Pct	0.5775	4.355	1.33e-05
Differential_Team_Payroll	1.285e-09	3.761	0.000169
Home_Team_MPBN	1.194e-04	0.442	0.658734
Visiting_Team_MPBN	-1.016e-04	-0.424	0.671389

Table 3 – Model including all variables discovered significant and MPBN

The third model had a classification accuracy of 0.6147 when run against the test data, showing an improvement over the second model due to removing statistically insignificant variables. However, the MPBN variables remained insignificant.

We next considered models including interaction terms with the Visiting_Team_MPBN variable. Under the assumption that visiting teams may grow accustomed to a time zone the longer they play games in that time zone, we introduced a term, Visit_MPBN_Prev_Games_In_Timezone, which was computed as Visiting_Team_MPBN * Prev_Games_In_Timezone. We then built a new model using this variable with the same variables as in the third model. We also included

Prev_Games_In_Timezone to account for any shift in the model function if the interaction term were found to be significant. The results for this model are in Table 4.

Home_Win ~ Game_Duration_Minutes + Differential_Win_Pct + Differential_Team_Payroll + Prev_Games_In_Timezone + Home_Team_MPBN + Visiting_Team_MPBN + Visit_MPBN_Prev_Games_In_Timezone

Variable	**Estimate**	**Z - value**	**p - value**
Game_Duration_Minutes	-0.01328	-16.832	< 2e-16
Differential_Win_Pct	0.5744	4.332	1.48e-05
Differential_Team_Payroll	1.264e-09	3.694	0.000221
Prev_Games_In_Timezone	0.003849	0.779	0.435997
Home_Team_MPBN	9.796e-05	0.361	0.718324
Visiting_Team_MPBN	-5.510e-05	-0.214	0.830570
Visit_MPBN_Prev_Games_In_Timezone	-2.024e-06	-0.141	0.887764

Table 4 – Model 3 with Visiting_MPBN & Prev_Games_In_Timezone interaction term

The interaction term, Visit_MPBN_Prev_Games_In_Timezone, was not seen to be significant. The classification accuracy for this model was 0.6141, about the same as that for model 3.

We then considered another interaction term, Visit_MPBN_Game_Duration_Minutes, being the product of Visiting_Team_MPBN and Game_Duration_Minutes. The rationale for this term was to examine if perhaps a visiting team's time zone adjustment effect may have a correlative effect when considered with game duration. In other words, perhaps the time zone adjustment effect is primarily seen with longer or shorter games. The results for this model are in Table 5.

Home_Win ~ Game_Duration_Minutes + Differential_Win_Pct + Differential_Team_Payroll + Home_Team_MPBN + Visiting_Team_MPBN + Visit_MPBN_Game_Duration_Minutes

Variable	**Estimate**	**Z - value**	**p - value**
Game_Duration_Minutes	-0.01251	-8.235	< 2e-16
Differential_Win_Pct	0.5785	4.363	1.28e-05
Differential_Team_Payroll	1.287e-09	3.764	0.000167
Home_Team_MPBN	1.221e-04	0.452	0.651596
Visiting_Team_MPBN	3.604e-04	0.456	0.648324
Visit_MPBN_Game_Duration_Minutes	-2.548e-06	-0.613	0.539594

Table 5 – Model 3 with Visiting_MPBN & Game_Duration_Minutes interaction term

The Visit_MPBN_Game_Duration_Minutes variable is seen not to be significant. This model had a classification accuracy of 0.6009.

Finally, for completeness sake if nothing else, we produced a model consisting of all significant variables from model 3 and both interaction terms with Visiting_Team_MPBN (Table 6):

> *Home_Win ~ Game_Duration_Minutes + Differential_Win_Pct + Differential_Team_Payroll + Prev_Games_In_Timezone + Home_Team_MPBN + Visiting_Team_MPBN + Visit_MPBN_Prev_Games_In_Timezone + Visit_MPBN_Game_Duration_Minutes*

Variable	**Estimate**	**Z - value**	**p - value**
Game_Duration_Minutes	-0.01251	-8.237	< 2e-16
Differential_Win_Pct	0.5754	4.340	1.42e-05
Differential_Team_Payroll	1.266e-09	3.698	0.000217
Prev_Games_In_Timezone	0.003816	0.774	0.439105
Home_Team_MPBN	1.007e-04	0.371	0.710658
Visiting_Team_MPBN	3.909e-04	0.492	0.623067
Visit_MPBN_Prev_Games_In_Timezone	-2.014e-06	-0.141	0.888169
Visit_MPBN_Game_Duration_Minutes	-2.462e-06	-0.593	0.553412

Table 6 – Model 3 with both interaction terms

As with all previous models we analyzed, all terms involving MPBN were not significant to the model. This model had a classification accuracy of 0.6052.

The three variables that our analysis *did* find contributed to an explanation of win likelihood, Game_Duration_Minutes, Differential_Win_Pct, Differential_Team_Payroll, were analyzed individually against our data. For Game_Duration_Minutes, a logistic regression model including only it as explanatory variable and Home_Win as response was created (Beta1 = -0.0133069, p < 2e-16, classification accuracy of the test set 0.6017). A logistic plot of this model's predictive results is shown in Figure 7. The red line displays the predicted probability of Home_Win for different values of Game_Duration_Minutes. The histograms above and below show the distribution of actual wins and losses in the training set. The horizontal dotted line indicates the threshold value used for classification accuracy in all our analyses, and corresponds to the home-field advantage. The vertical solid line shows the Game_Duration_Minutes value (x) for which the predicted logistic fit would equal the threshold value.

The same analysis was done for Differential_Win_Pct (Beta1 = 0.66726, p = 2.30e-07, classification accuracy = 0.5293, Figure 8) and Differential_Team_Payroll (Beta1 = 0.0015506, p = 2.63e-06, classification accuracy = 0.5114, Figure 9)

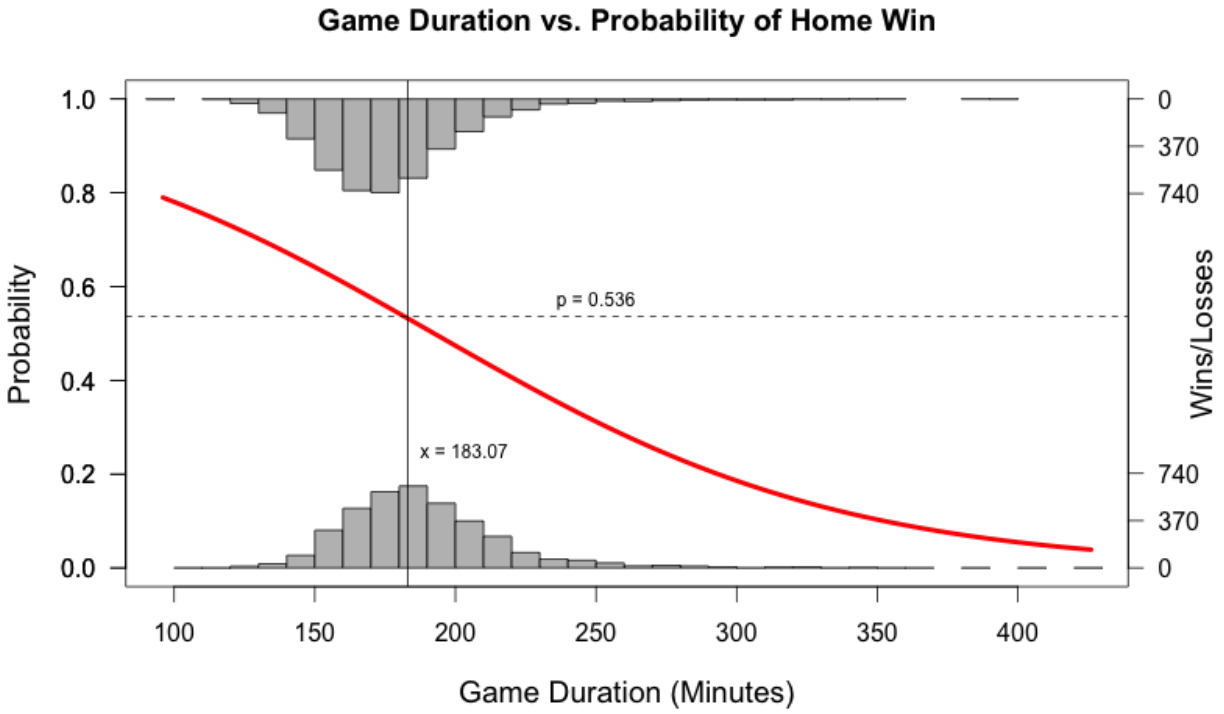

Figure 7: Game Duration Logistic Plot

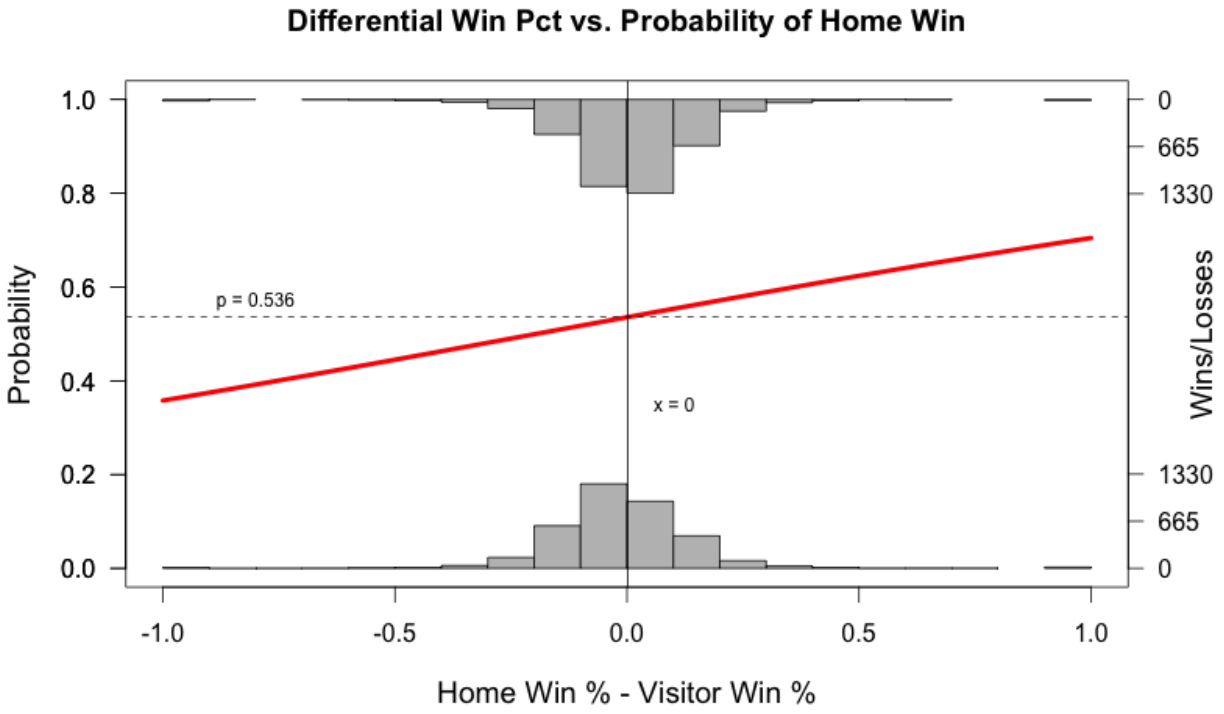

Figure 8: Win Percentage Logistic Plot

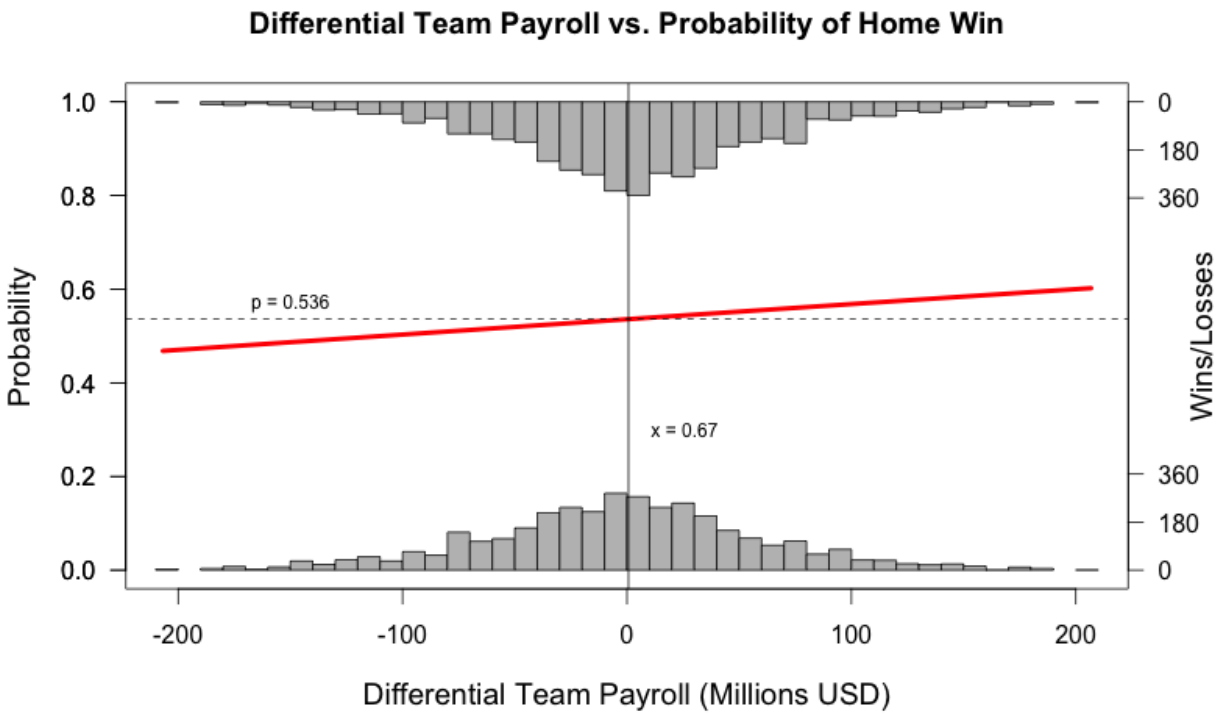

Figure 9: Payroll Logistic Plot

Analysis

> Was the target variable easy or hard to model? What does that tell us? What are the practical applications?

Our target variable was MPBN, which was calculated for both the home and away teams, as well as the difference between the two. We were able to calculate this by making assumptions about the visiting team's previous origin of travel and computing the time zone difference between the previous origin and the location the game was being played. While this value was not difficult to compute given the available data for Major League Baseball, we did find that we needed to take care in coming up with our assumptions about visiting teams' travel origins. A first attempt at our data simply assumed that for each game, the visiting team was feeling as if they were in their home time zone. In other words, we made to allowance for previous games played by the visiting team in the same time zone. Our second attempt utilized the number of games played by the two teams in a series, thus allowing for the fact that visiting teams could be expected to adjust to a time zone the longer they were in it. Finally, we computed the visiting team's travel origin by examining the location of the game played previous to the one in question. Since our data was sorted in order of games played by each team, we were able to compute this value fairly easily.

One small difficulty in computing MPBN was that it relied on the game's start time. While our data had scheduled start times for each game, we did not have actual first pitch times. For most games, there is not a significant difference between these times, however, in cases of rain delayed games or other circumstances, there may be games where these values would differ widely. Still, we felt that the prevalence of delayed start games in our data was not frequent enough to warrant concern.

Despite the relative ease in computing accurate values for this variable, under no circumstance were we able to find that MPBN in any form or any of its interaction terms were significant predictors. Since it was never significant in any model and added nothing to predictive power given multiple modeling and filtering methodologies it shows us that there appears to be no significant competitive advantage based around the player's proximity to their competitive biological peak.

The practical application of this finding is that Major League Baseball does not appear to need to concern itself with correcting for such a discrepancy in its rules or scheduling. If, for example, our study had found that east coast teams faced a non-insignificant disadvantage when playing west coast teams (or vice-versa), then in the interest of maintaining fairness, the effect would need to be addressed by either adding extra days to adjust or staggering trips so that teams only traveled by single time zones when possible.

What were the most meaningful explanatory variables? Are they what you expected?

Our models found that the difference in payroll between teams was a significant predictor for game outcomes. Similarly, we found that a team's record, which was calculated as overall win percentage, was significant and when we used post game data we found that game duration was a significant predictor in the outcome of each game.

Since the p-value for payroll difference was, $p < .01$, we can claim that there is a higher probability of winning for those who have invested in bringing in and retaining great players. As discussed earlier this variable being significant seems reasonable for baseball with the vast disparities between team payrolls.

The difference between team records proved significant where $p < .01$ for both respectively. The fact that a team’s win/loss record helps determine game outcome seems pretty straight forward. A team that wins more often versus a team that loses more often would logically be a decent predictor at a game’s outcome. It might have been a more interesting finding if these variables proved not to be significant.

Regarding game duration, the partial correlation coefficient for game duration was seen to be negative and very highly statistically significant ($p < 2 \times 10^{-16}$ for almost all models), thus meaning that shorter games favor the home team winning. The only problem using this highly significant variable is that it acts in real time and so is unavailable before the game begins so predicting the final game outcome only using this in a model should be highly cautioned.

Generally speaking, the variables that proved to be significant were understandable in the context of Major League Baseball and sports in general. Some terms were borderline significant and likely need to be followed up on to determine whether or not a refinement may help push them closer to being relevant and significant in future models.

Does your models tell us something about the domain you have chosen?

The effect that we chose to study, and did not find significant, has been written about before for Major League Baseball[7], and similarly for the National Football League[8], so the lack of it in our model means that the effect previously published, which already only affected 20% of games in a season, may be less of an effect than shown. Even when we isolated games to the first day of travel there was never a significant effect that was present in our data let alone when we compared it against the entirety of a season. Certainly this challenges the findings of the previous study and proves that we should question how accurate the testing methodologies of these sports articles are when doing only a high level analysis of win/loss records for teams. It likely means that there were other factors that were better predictors than travel for the outcome of the games which were not accounted for.

As stated in the article regarding biological rhythms, that author opined that competition was already stimulating and thus may not show the same performance boost that they found. Since the effect that we sought to measure, circadian rhythm, is a biological effect, it seems plausible that the biological requirements of the sport could be a very significant determining factor in whether circadian rhythm affects performance in that sport. One advantage baseball athletes may have against this effect is that they are generally free to take pauses or time-outs much more frequently. One might expect that a baseball athlete feeling the effects of time zone difference might take longer or more frequent pauses between play. Therefore, it is plausible that the speed and physical requirements of baseball could mitigate the competitive advantage attributed to travel. Still while this study does lend some credence to the statement made by the author at least as the MLB is concerned as a sport and the effect of one's biological rhythm on it.

Conclusions

Our most important finding was that MPBN was not significant in predicting win-loss likelihood. We reached this conclusion after making several attempts using different models and mutations of the variable as seen above. Cleaning our data and adding, removing, or replacing other explanatory variables from our models still produced no significant results. Thus, we are led to conclude that win likelihood is not significantly affected by the difference of the biological peak created by the change in time zones of the visiting team in baseball.

---

[7] Recht, Lawrence D; Lew, Robert A; Schwartz, William J (1995). "Baseball Teams Beaten By Jet Lag". Nature. 377 (6550): 583.

[8] Scott Kacsmar, "Discomfort Zone: The Impact of Travel on NFL Games," FootballNation.com, April 18, 2012, http://www.footballnation.com/content/discomfort-zone-impact-travel-nfl-games/14372/.

Should we seek to follow up on this project some of the things we would like to better account for would be improved data gathering and analysis. One issue we discussed was the need to obtain more accurate start times for MLB data. We felt this was important so we would be able to account for games that are postponed from their scheduled start times. Another consideration we had was to use a weighted model for considering the winning percentage explanatory variable. We thought this might be needed to prevent early season values for this variable from overly influencing the model when predicting a game's outcome. Early in the season, with few games having been played, this value swings significantly with each win or loss and the difference between home and away teams can be much larger than every subsequent game as the season progresses. We felt its use in our model as a "quality of opponent" measure is diminished unless these early season values are compensated for correctly.

Future studies should also include a more accurate measure of the visiting team's place of origin in determining time zone effect. Our current treatment does not take account for travel days, for example, where the visiting team may have an extra day to adapt to a new location before playing a game there. Further time related additions may include a variable that considers the game end time and start time of consecutive games. This variable would give the total time between games which could help isolate the time zone effect in games where there was less time to rest when the team changes time zones.