Understanding the Relationship Between Weather Conditions and Home Run Rates in the MLB

by

Tyler Ashoff

Submitted to the Department of Mechanical Engineering
in partial fulfillment of the requirements for the degree of

Bachelor of Science in Mechanical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018

Signature redacted

Author .....
Department of Mechanical Engineering
May 16, 2018

Signature redacted

Certified by .....
Anette Hosoi
Associate Dean of Engineering
Thesis Supervisor

Signature redacted

Accepted by .....
Rohit Karnik
Undergraduate Officer
Associate Professor of Mechanical Engineering

Understanding the Relationship Between Weather Conditions and Home Run Rates in the MLB

by

Tyler Ashoff

Submitted to the Department of Mechanical Engineering
on May 16, 2018, in partial fulfillment of the
requirements for the degree of
Bachelor of Science in Mechanical Engineering

Abstract

This observational study explores the relationship between home run rates and weather conditions, both on game day and over the preceding weeks. Data were collected from ESPN and Weather Underground for over 36,000 games between the 2003 and 2017 seasons. These consisted of game statistics and 59 weather variables. Random Forests was used to determine which set of these variables were important predictors of home run rates. Humidity was found to be the most important weather variable for predicting home run rates. The data suggest that a change of game day humidity from 100% to 0% can increase home run rates by 27% and ball travel by $15ft$.
For access to the data, please visit tylerashoff.com.

Thesis Supervisor: Anette Hosoi
Title: Associate Dean of Engineering

Acknowledgments

Thank you to Dr. Peko Hosoi and Dr. Colin Fogarty for guidance and support throughout this research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

During the 2015 season, Major League Baseball (MLB) noticed a record high number of home runs. The cause of this spike was unclear and an investigation into its cause began. The restitution of the ball itself and the aerodynamics of the ball were among the suspected factors contributing to the spike, and are the primary interest in this study. Weather relates to these factors by changing the moisture content of the ball and the properties of the air respectively. To study these effects, game statistics were collected from ESPN and weather conditions were collected from Weather Underground. A full list of variables can be found in chapter 2.

Understanding the effect of weather conditions on home run rates can help inform decisions about ball storage and game strategy. Ensuring that equipment is properly standardized will help keep games fair, and understanding favorable conditions will help players realize their full potential.

1.1 Motivation for Suspicions

1.1.1 Coefficient of Restitution

The coefficient of restitution (C.O.R) is a measure that describes the elasticity of an object. Formally, it is the ratio of the relative speed of two objects prior to collision to the relative speed after collision, $C.O.R = (v_{2_0} - v_{1_0})/(v_{2_f} - v_{1_f})$. Changing the

C.O.R of a baseball could change the ball bat interaction significantly. A higher C.O.R results in more energy being transferred to the ball and a longer distance of travel. Conversely, a low C.O.R results in increased energy dissipation and a shorter distance of travel.

Experiments have shown that a ball stored in 100% humidity has an expected travel $28ft$ shorter than one stored at 0% humidity (Kagan & Atkinson, 2004). The results of this experiment can be found in Figure 1-1. To put this change in perspective, a $14ft$ decrease results in about a 25% increase in home runs, in a park with a $380ft$ back wall (Nathan, Smith, Faber, & Russell, 2010). This calculation was based on an estimation that for every percent change in home run distance, there is a 7% change in the probability of hitting a home run (Adair, 2002).

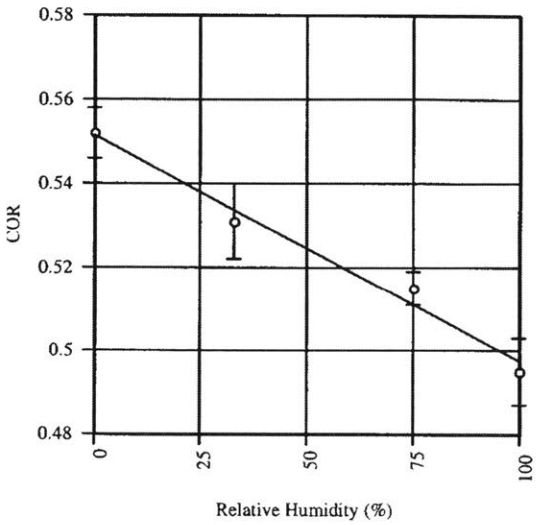


Figure 1-1: COR vs. Relative Humidity graph reproduced from (Kagan & Atkinson, 2004)

This difference is significant in baseball, and is similar to the reason aluminum bats are not allowed in major league play. Analysis of the C.O.R of balls and bats found that an aluminum bat increased ball travel by $30ft$ (Adair et al., 1995). Similarly, the difference between the park with the longest average distance to the back wall and the park with the shortest, is $37ft$. This means that changing how the ball is stored may have as big of an impact on home runs as changing bat material or moving parks entirely.

It is also worth mentioning that experiments like the one conducted by Kagan & Atkinson, store balls in humidity for varying lengths of time. In the following discussion game day humidity is primarily used. This is justified because game day humidity is strongly related to the humidity over the previous days.

1.1.2 Air Density

Air density effects the movement of the ball both before and after the interaction with the bat. Before the hit, the ball can move more if the air density is higher. This means the pitcher can better control the curve of a pitch. After the hit, higher air density will impede the ball's flights and result in shorter travel. It is also important to mention that dry air is more dense than wet air. Meaning as humidity decreases the air becomes less dense. Further information about the calculation of air density can be found in Section 2.2.3.

Chapter 2

Data Collection

2.1 Web Scraping

Using Beautiful Soup, a web scraper was built to collect game statistics and weather data on MLB games over the 2003-2017 seasons. Over the 15 years sampled, information on 36,731 games was collected. The majority of these games come from the regular and post season schedule, however some exhibition games are included in the data as the scraper did not differentiate between them. The rate of sampling per year was higher in recent years and lower in earlier years. Table 2.1 provides a breakdown of the games played from 2003 to 2017.

Season	Games Played
Regular Season	2,430/season
Post Season	41/season (2003-2011) 43/season (2012-2017)
Total	22,239 (2003-2011) 14,838 (2012-2017)
Grand Total	37,077

Table 2.1: Number of Games Played

2.2 Collected Variables

2.2.1 Game Statistics

The scraper accessed the ESPN website to check if, on any given day between March and November of each season, a team played a home game. If a game was played, game statistics were saved - see Table 2.2 for a complete list.

Variables Collected
Date
Home Team
Home Runs by Visiting Team
Home Runs by Home Team
Hits by Visiting Team
Hits by Home Team

Table 2.2: Raw Game Statistics

2.2.2 Weather Data

On each day a game was played, the scraper accessed Weather Underground and gathered weather data from the airfield nearest to the ballpark at which the game was played. In some cases, the nearest airfield did not have sufficient data, in these cases the nearest airfield with complete data was used. The collected data included weather conditions from game day and past conditions. The past conditions are an average of each variable type over a given time period - see Table 2.3 for a complete list. For example, the two week temperature high is the average of each day's high temperature over two weeks leading up to, but not including, game day.

The type and period are subsets of each variable - the daily low, mean, and high temperatures were averaged over each of the periods listed - the historical two day, five day, ten day, and two week periods do not include game day. The variables include, temperature,the air temperature measured in degrees Fahrenheit, humidity, the percent relative humidity, dew point, the difference between the real air temperature and the fully saturated air temperature in degrees Fahrenheit, and sea level pressure, measured in inches of mercury, is a correction of the station pressure to sea level by taking into account elevation and temperature dependencies. This correction

makes comparison of pressures across locations easier.

Variable	Type	Period
Temperature ($°F$)	Mean, Low, High	Game Day, Two Day, Five Day, Ten Day, Two Week
Humidity (%)	Mean, Low, High	Game Day, Two Day, Five Day, Ten Day, Two Week
Dew Point ($°F$)	Mean	Game Day, Two Day, Five Day, Ten Day, Two Week
Sea Level Pressure ($inHg$)	Mean	Game Day, Two Day, Five Day, Ten Day, Two Week

Table 2.3: Raw Weather Variables

2.2.3 Geographic and Combined Variables

The geographic and combined variables in this section were not gathered by the scraper. Rather they were hard coded into the collection process - see Table 2.4 for a complete list.

The geographic variables are attributes of the ballparks themselves. Elevation is the Elevation from sea level of the ball park. Outfield range is the average of the distance from home plate to left field, center field, and right field (Spirito, 2013).

Geographic	Combined
Elevation ($Feet$)	Total Home Runs (-)
Outfield Range ($Feet$)	Total Hits(-)
	Home Runs per Hit (-)
	Air Density (kg/m^3)

Table 2.4: Geographic and Combined Variables

The combined variables are combinations of the variables in Table 2.2 and 2.3 and

are calculated per game. Total home runs and total hits are the total of the home runs or hits by both teams. Home runs per hit is the ratio of home runs to hits.

Air density, D ($\frac{kg}{m^3}$), depends on the Mean Temperature, T ($°C$), Dew Point, T_d ($°C$), and Sea Level Pressure, P (Pa), on game day and is derived as follows (Shelquist, 2012). The appropriate conversions were made from the raw data into appropriate the units.

$$D = \left(\frac{P_d}{R_d T}\right) + \left(\frac{P_v}{R_v T}\right) \tag{2.1}$$

$$R_d = \frac{R}{M_d} = 287.05 \tag{2.2}$$

$$R_v = \frac{R}{M_v} = 461.495 \tag{2.3}$$

where, R (universal gas constant) = 8314.32, M_d(molecular weight of dry air) = 28.964 ($\frac{g}{mol}$), and M_v(molecular weight of water vapor) = 18.016 ($\frac{g}{mol}$).

The pressures P_d (Pa) and P_r (Pa) are defined as:

$$P_d = P - P_v \tag{2.4}$$

$$P_v = \frac{ES0}{p^8} \tag{2.5}$$

where, ES0 (saturation vapor pressure over water at $0°C$) = 610.78 (Pa).

The dimensionless function p incorporates the effects of the dew point temperature and can be estimated as:

$$\begin{aligned} p = c_1 + T_d(c_2 + T_d(c_3 + T_d(c_4 \\ + T_d(c_5 + T_d(c_6 + T_d(c_7 + T_d(c_8 \\ + T_d(c_9 + T_d(c_{10})))))))) \end{aligned} \tag{2.6}$$

where, coefficients can be found in Table 2.5.

$$
\begin{aligned}
c_1 &= 0.99999683\ (-) \\
c_2 &= -0.90826951 \times 10^{-2}\ (\tfrac{1}{^\circ C}) \\
c_3 &= 0.78736169 \times 10^{-4}\ (\tfrac{1}{^\circ C^2}) \\
c_4 &= -0.61117958 \times 10^{-6}\ (\tfrac{1}{^\circ C^3}) \\
c_5 &= 0.43884187 \times 10^{-8}\ (\tfrac{1}{^\circ C^4}) \\
c_6 &= -0.29883885 \times 10^{-10}\ (\tfrac{1}{^\circ C^5}) \\
c_7 &= 0.21874425 \times 10^{-12}\ (\tfrac{1}{^\circ C^6}) \\
c_8 &= -0.17892321 \times 10^{-14}\ (\tfrac{1}{^\circ C^7}) \\
c_9 &= 0.11112018 \times 10^{-16}\ (\tfrac{1}{^\circ C^8}) \\
c_{10} &= -0.30994571 \times 10^{-19}\ (\tfrac{1}{^\circ C^9})
\end{aligned}
$$

Table 2.5: Coefficients

Chapter 3

Discussion

The data analysis is split into two main categories: short term and long term. Short term trends deal directly with individual games to understand how weather conditions effect the home run rates. Long term trends deal with average weather conditions at each percentile of the home run per hit metric, this will be explained more fully in section 3.3. Random Forests^{TM}, identified key parameters in both the long term and short term trends. Using these key parameters efforts were made to develop models for home run rates based on weather conditions.

3.1 Random Forests^{TM}

RandomForests (RF) was used to develop a non-parametric model for the weather conditions and home run rates. Because the relationships between the weather variables are unknown, RF was especially attractive. RF works by creating an ensemble of trees, and the forest's prediction is the average of the trees in this ensemble (Grömping, 2009). Each tree is constructed by testing a random set of observations and creating a sequence of nodes using about a third of the variables each time.

Once the forest is complete, the variable importance is determined by comparing each variable's %IncMSE. This is computed by passing a variable down each tree and recording the number of correct classifications, and doing the same for a randomly shuffled version of the variable. The average of the difference between these two

scores is the score for each variable (Breiman, 2001). Essentially, this determines how effective the variable is compared to a random variable across all the trees. A high %IncMSE signals that changing this variable has a large effect on model's predictive power. A low %IncMSE signals that the variable is closer to a random variable.

3.2 Short Term

The short term trends were developed using the raw data. Each game was a single data point and the goal was to find weather conditions that could predict home run rates on any given game day. Because these trends are based on individual games, they are more susceptible to other game day factors like specific players and coaching decisions. The long term analysis in Section 3.3 attempts to mitigate these effects through aggregation. This analysis is important because it offers insight to how coaching decisions can be made in real time which will be discussed more in Section 4.

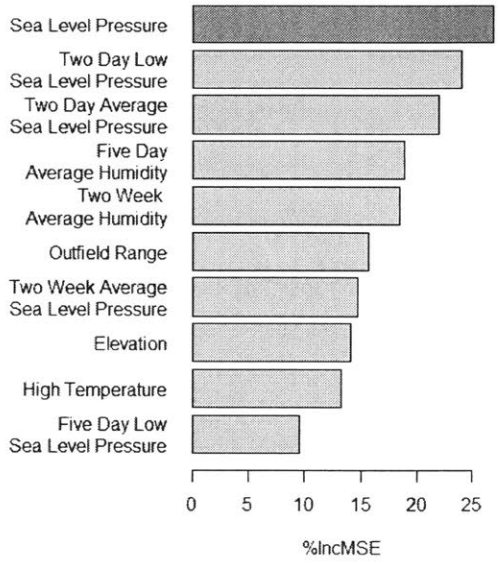


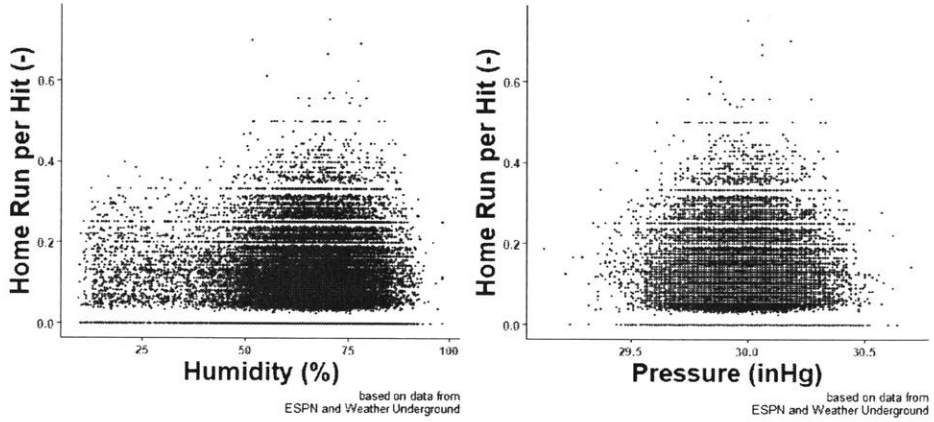
Figure 3-1: Top Ten Short Term Variables

3.2.1 The Model

The model created using RF found that by using all of the gathered weather conditions, 9.86% of the variation in home runs rates was explained. The RF parameters used in the short term model are mtry:10 and ntree:200. The top ten most important variables for short term trends can be seen in Figure 3-1. They consist of variations of pressure, humidity, outfield range, elevation, and temperature. The non-weather related variables are good sanity checks. It makes sense that the further away the

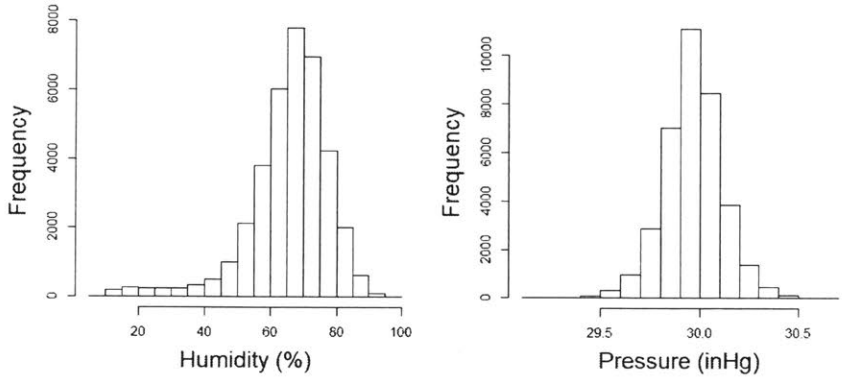
back fence, the fewer home runs will be hit in any given game.

3.2.2 Graph Sampling



(a) Home Run per Hit vs. Five Day Average Humidity

(b) Home Run per Hit vs. Game Day Sea Level Pressure



(c) Five Day Average Humidity

(d) Game Day Sea Level Pressure

Figure 3-2: Short term home run rates

Figure 3-2 shows scatter plots of two important variables for short term trends along with their histograms. Both scatter plots represent the variability of game day predictions discussed earlier in Section 3.2. Rather than predicting home run rates on game day, these graphs show favorable conditions for high home run rates.

Both five day average humidity and sea level pressure have unique distributions, but it is the deviation from these that is interesting. The two week average humidity, seen in Figure 3-2a, is leptokurtic, with an excess kurtosis of 3.36 and is highly left

skewed, with a skewness of -1.38. The two week high sea level pressure, seen in Figure 3-2b, is also leptokurtic, with an excess kurtosis of 0.96 and is approximately symmetric with a skewness of -0.039. These results can be seen in Figures 3-2c and 3-2d.

Essentially this means that for the humidity, more data reside in the left tail than would be expected from a normal distribution. The sea level pressure distribution varies from normality similarly, in that more data reside in the left tail, but much more modestly than that of the humidity.

By inspection of the graphs it appears that the distribution in Figure 3-2a has a heavier left tail than would be expected even from the underlying distribution of humidity, and the distribution in Figure 3-2b exhibits a slightly heavy right tail, the opposite of the underlying sea level pressure distribution. This relationship makes physical sense as well. An inverse relationship between humidity and pressure is described in Equation 2.5, and is realized here in the changing tail weights.

This intuition deserves further analysis and formalization. However, because the long term trends captured more variation of the home run rates, this investigation placed more attention on that analysis.

3.3 Long Term

The long term trends were developed by aggregating the data points by their respective home run rates. Bins were created for the home run rates rounded to the second decimal place and each weather condition was averaged within these bins. This method was deemed 'long term' because the averaging reduces the effects of non-weather relate variables, such as individual pitchers or batters. This is in contrast to the 'Short Term' analysis in Section 3.2 which sought to find relationships for individual games which were more susceptible to these other factors.

3.3.1 The Model

The model created using RF found that by using all of the gathered weather conditions, 61.98% of the variation in home run rates was explained. The RF parameters used in the short term model are mtry:10 and ntree:2500. From this model the most important variables were largely dependent on humidity. A list of the top ten most important variables for long term trends can be found in Figure 3-3. With the exception of Outfield Range, which has served as a much needed sanity check throughout this investigation, the top ten variables consist of variations of humidity and dew point. Dew point is a measure of how close the air is to saturation and is related to humidity and temperature. These results point strongly towards a relationship between home run rates and moisture in the air.

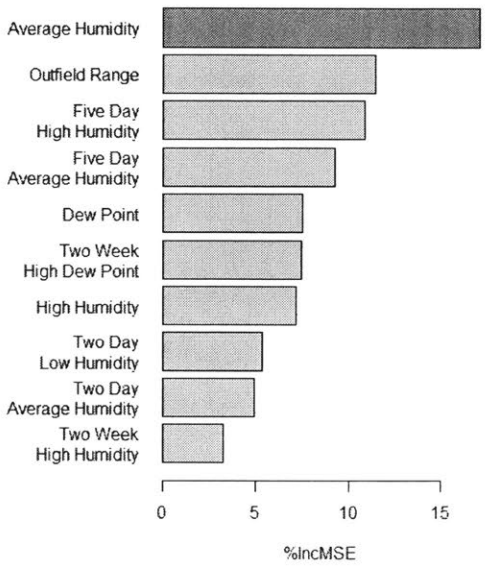


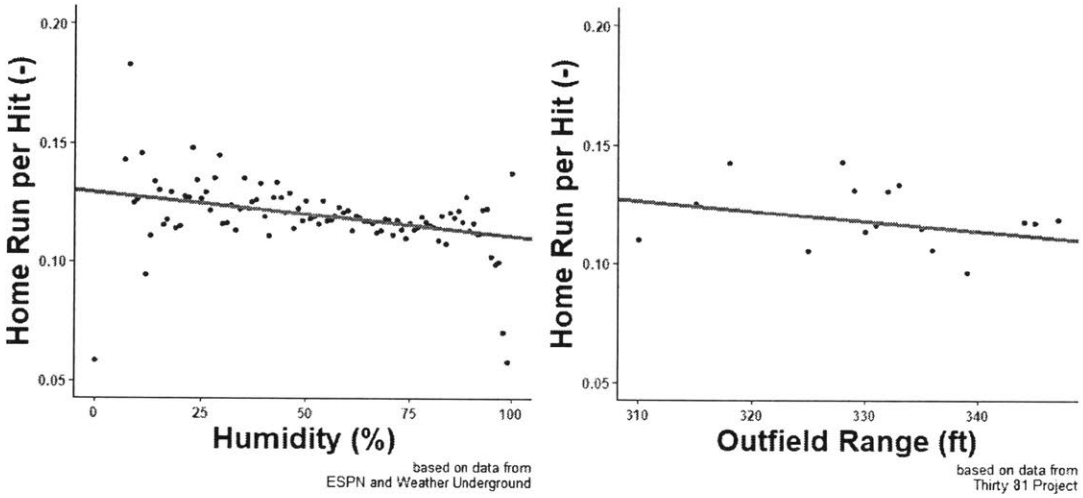
Figure 3-3: Top Ten Long Term Variables

3.3.2 Graph Sampling

Figure 3-4 shows a linear regression for the top two variables for long term trends. Figure 3-4a shows that as game day humidity increases the expected home run rate decreases with slope -0.018 ± 0.007. This relationship corresponds to a predicted 27% increase in home runs as humidity varies from 100% to 0% - Figure 4-3 shows how this change compares to other variables. Using the $380ft$ wall from Section 1.1.1, this results in a predicted $14.7ft$ increase of ball travel. This change in travel is about half of the increase predicted by Kagan & Atkinson. Similarly to the trends seen in Figure 3-2a, the home run rates are higher in dry conditions, ie. when the air is more

dense. This is an important note and will be discussed further in Section 4.

Figure 3-4b shows that as park size increases the expected home run rate decreases with slope -0.0004 ± 0.0007. This corresponds to a 10% decrease in home run rates from a home run distance of $310ft$ to $347ft$. The uncertainty on the slope is large, however it is worth mentioning that the upper bound is consistent with Adair's estimation as discussed in Section 1.1.1.



(a) Home Run per Hit vs. Average Game Day Humidity $R^2 = 0.24$

(b) Home Run per Hit vs. Outfield Range $R^2 = 0.094$

Figure 3-4: Scatter plots with weighted linear regression lines: Long term home run rates

Chapter 4

Conclusions

This investigation found that, out of the 59 variables tested, humidity related metrics showed the strongest relationship to home run rates in both the short term and the long term analysis. These variables are especially interesting because of their relationships to the COR.

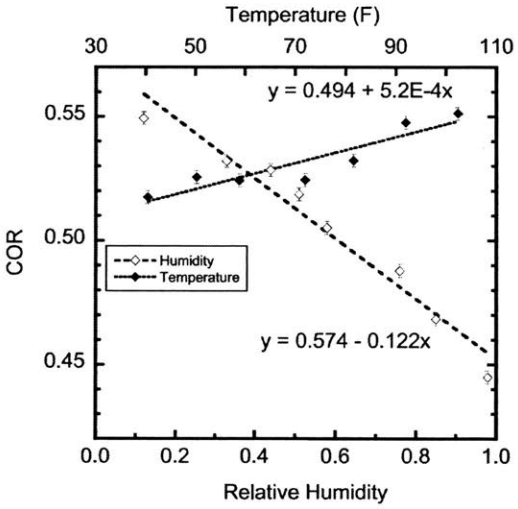


Figure 4-1: COR vs. Relative Humidity graph reproduced from (Nathan, Smith, Faber, & Russell, 2010)

Nathan, et al. measured the cylindrical COR of baseballs exposed to varying levels of humidity, the results can be found in Figure 4-1 (Nathan et al., 2010). These results show that as humidity increases the COR of the ball decreases with a slope of -0.122 ± 0.010. This is consistent with the results found in Figure 3-2a and 3-4a. These results offer an explanation as to why humidity has an effect on home run rates. Using the relationship between humidity and COR found by Nathan, et al., a relationship between home run rates and COR can be found - See Figure 4-2. This plot was created using the same aggregation method used for the long term trends, and the slope of the fit line for the home run per hit vs. COR is 0.157 ± 0.922.

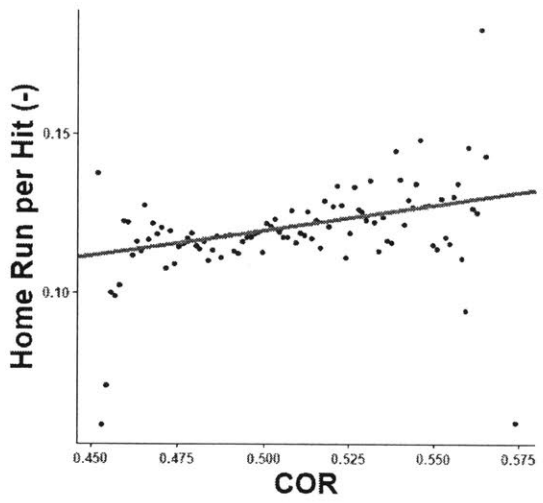


Figure 4-2: Home Run per Hit vs. COR

Figure 4-3 shows how each of the selected variables are expected to change home run rates. The change in home run rates due to bat type was calculated using Adair's estimates for changing bat material and home run rates discussed in Section 1.1.1 - a $380ft$ back wall was again used for these calculations. The changes due to COR, Humidity, and Dew Point used the collected home run data directly to find the percent change in home runs over the range of data collected. These ranges are 0.45 - 0.57, 100% - 0%, and $1°F$ - $81°F$ respectively. This data suggests that a change in humidity can change the number of home runs per game by up to 27%.

The baseball experiences four events during a home run sequence. First, the ball is pitched. Second, the ball travels towards the batter. Third, the bat hits the ball. Fourth, the ball flies over the back fence. A relationship between air density and home run rates was expected to be found. However, the data suggest that the typical changes in air density do not sway home run rates as much as other factors. This result, along with the relationship to the COR, suggest that environmental effects

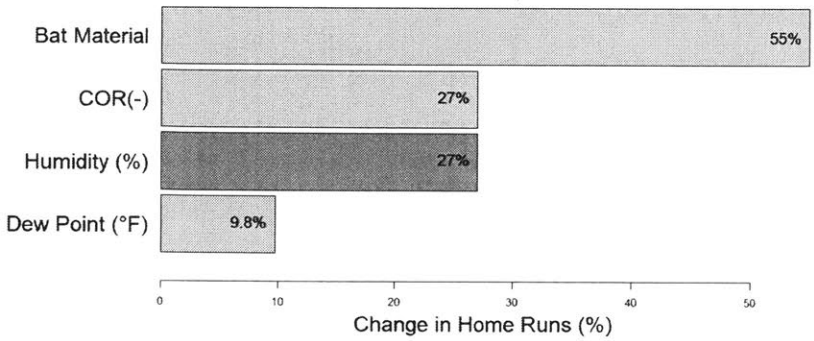


Figure 4-3: Change in Home Run Rates

have a larger effect on the pitcher and bat interactions with the ball than on the ball's time in the air. The drag on the ball in flight is a separate interaction: one that recent studies suggest may be linked to the manufacturing process.

An increase in the COR results in more energy conservation during the bat-ball interaction, and as discussed in Section 1.1.1, increased ball travel. There is also anecdotal evidence gathered in this investigation and mentioned by Nathan, et al., that balls stored at low humidity are described as slippery by pitchers. If the pitcher is not able to put as much spin on the ball, it might be easier to hit, thereby increasing home run rates. Both situations, if true, align with the data presented, however these interactions are outside the scope of this study.

References

Adair, R. K. (2002). *The physics ofbaseball.* HarperCollins, New York.

Adair, R. K., et al. (1995). The physics of baseball. *Physics Today*, *48*(5), 26–33.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, *63*(4), 308–319.

Kagan, D., & Atkinson, D. (2004). The coefficient of restitution of baseballs as a function of relative humidity. *PHYSICS TEACHER*, *42*(6), 330–333.

Nathan, A. M., Smith, L. V., Faber, W. L., & Russell, D. A. (2010). Corked bats, juiced balls, and humidors: The physics of cheating in baseball. *arXiv preprint arXiv:1009.2549*.

Shelquist, R. (2012). An introduction to air density and density altitude calculations. *Internet Survey, Visited on 25th of March*.

Spirito, L. J. (2013). Baseball's many physical dimensions [Computer software manual].