Environment Independent Sabermetrics:

Modeling Baseball Pitchers Performances

Conor Walsh

https://www.linkedin.com/in/conorpwalsh4

conorpwalsh4@gmail.com

8 March 2025

Abstract

In any given moment, each team in Major League Baseball (MLB) is working to most effectively measure baseball players' ability to perform in order to inform salary and player acquisition decisions that ultimately will decide their team's success. This research isolates traditional player performance metrics from variables beyond their control, such as the ballpark in which they compete, as well as the strength of their opponent on any given day. Ballpark factor is empirically defined by a weighted linear least squares regression analysis, and opponent offensive strength is derived from a combination of a team's recent and full season's ability to create runs. Predictive modeling is performed using the newly defined metrics as features to provide evidence of their value in assigning small modifications to starting pitchers' ERAs.

Table of Contents

Table of Figures

Table of Tables

Table of Equations

1. Introduction

Baseball performance has been long been measured to assess appropriate salaries and ultimately construct a competitive team; however, the methods for analyzing performance have changed dramatically over the years. Consider the ways a player's offensive strength has been measured over time; Henry Chadwick is credited with introducing 'batting average,' defined by the number of hits divided by the number of at-bats, in the mid-to-late 1800's. (Thorn, 2013) However, this formula for batting average weights each hit the same and does not include outcomes such as walks. Over time, statistics such as On-Base Percentage (OBP) and Slugging Percentage (SLG) attempted to overcome the shortcomings of batting average by accounting for walks and the number of basses gained for a given hit. In present day there are more complete and mature metrics such as "Weighted Runs Created Plus" (wRC+) and 'Wins Above Replacement' (WAR) which both offer a more comprehensive review of a player's offensive contributions. One thing that has remained constant over time has been the problem statement for baseball statistician: How to best isolate a player's performance from environmental factors beyond the player's control. This analysis shares the same problem statement; In particular, this aims to quantify how ballpark factors and recent opponent performance can be used to augment an instantaneous measure of a baseball pitcher's performance. Finally, these new measures of a player's performance will be used as features for modeling a more accurate assessment of player performance.

2. Literature Review

2.2 Independent Pitching Statistics

The intuition of baseball fans alike suggests that all MLB games' unique environments are not equally challenging to perform in; There are many factors affecting the gameplay, such as each team's overall offensive and defensive proficiencies, injuries or rest days accounting for players missing from the starting lineup, weather, and the unique playing field. Baseball is unlike other major US sports leagues such as the National Football League (NFL) and National Basketball Association (NBA) in that their stadiums do not have to conform to very strict standards. Baseball stadiums differ in many ways including overall climate, distance to the outfield wall, height of the outfield wall, surface area of the out-of-bounds, colloquially called foul territory, to name a few. For the most accurate measurement reporting on player's performance statistics there should be an adjustment made to the observed statistics to account for some of these factors.

2.2.1 Park Factors

Park Factors was introduced to measure the differences among stadiums, as alluded to in the previous section. A simple formula often credited to have been popularized by ESPN is shown in Equation 1 and utilizes runs scored (RS) and runs allowed (RA) to isolate the effect the stadium might have on the performance of the team by keeping the team constant.

$$Park\ Factor\ (PF) = \frac{\left(\frac{RS_{home} + RA_{home}}{Games_{home}}\right)}{\left(\frac{RS_{road} + RA_{road}}{Games_{road}}\right)} \quad (1)$$

There have been several attempts to improve this formula since it was first introduced including using a weighted least squares regression analysis as introduced by Acharya. (Acharya

et al. 2008) The analysis in this research adapted Acharya's approach with the underlying assumption that the number of runs scored in each MLB game is a linear combination of the following factors: the offensive team's strength, the defensive team's strength, and the ballpark in which the game is played. Therefore, a system of linear equations $\boldsymbol{Ax = b}$ is defined where each game is modeled as a set of two linear equations; One equation represents the home team as the offensive team and the other represents the home team as the defensive team. Therefore, the matrix A in equation 2 is composed of $2n$*g rows and 3n columns where n represents the number of teams who competed in that season and g represents the number of games per team in a season. (Calzada, 2018)

$$\begin{bmatrix} O_i & D_i & P_i & \cdots & O_n & D_n & P_n \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ O_{2ng} & D_{2ng} & P_{2ng} & \cdots & O_{2ng} & D_{2ng} & P_{2ng} \end{bmatrix} x = b, x = \begin{bmatrix} o_i \\ d_i \\ p_i \\ \vdots \\ o_n \\ d_n \\ p_n \end{bmatrix} \quad (2)$$

The number of runs scored per game, b, is weighted to reflect the number of outs recorded during the game. This allows for a modeling games where the bottom half of the 9th inning is not played when the home team is winning, as well as games that end in a walk-off hit. Each equation is therefore a summation of 90 variables and their coefficients, as well as a constant value. The summation of any given row in the matrix $\boldsymbol{A}$ should always equal 3, with a 1 in the indices representing the offensive team, the defensive team, and the ballpark.

2.2.2 Opposition Recent Success

A casual baseball fan knows that sometimes a team "gets hot," meaning they are performing better than their expected performance. E.g. in 2024 The New York Mets narrowly avoided missing the postseason by clinching while on the road against a division rival, and then they won their first playoff series with late game heroics that included coming back from a 2-0 deficit in the final inning of the final game. This earned them the opportunity to play the Philadelphia Phillies in the National League Division Series (NLDS), and although the Phillies performed better in the regular season, the Mets won the series in a decisive 3 games to 1 fashion. It is clear the Mets team performing in October was a stronger opponent than perhaps they were in May of that year when they won only 9 of 28 games; this serves as an example to motivate a metric valuing a team's recent performance.

The Elo rating system, named after chess master Arpad Eldo, was a system designed in 1960 to measure chess players ratings over time by updating them based on weighted performance scores. (Wikipedia, 2025) A player with a higher Elo rating is expected to perform better than a player with a lower rating, and after each game the losing player's rating is adjusted downward while the winning player's score is adjusted upward. In this way they use the results of a zero-sum game to adjust their performance assessment of players. This system has been adapted to assessing baseball team's strengths over the course of a season. (FiveThirtyEight, 2022) Their model attempts to predict baseball game outcomes by accounting for "home-field advantage, margin of victory, park and ERA effects, travel, rest and — most importantly — starting pitchers." (FiveThirtyEight, 2022) The analysis in this research does not wish to predict individual game outcomes; however, it will aim to capture the essence of the Elo ratings by

approximating a team's strength at any given time based on an assessment of the recent measured success in scoring runs. Although this research uses runs scored, traditional metrics such as a team's SLG or OBP may be used in a similar manner to measure the opponent's strength.

3. Data

The data used for this analysis was primarily collected from two sources including: Retrosheet, a non-profit website with historical MLB box scores, and the Lahman Baseball Database, which is a relational database containing pitching, hitting, and fielding statistics for Major League Baseball from 1871 through 2023. (Sean Lahman, 2025) The data sourced from Retrosheet contains traditional box score values for previous MLB seasons. This analysis uses a subset of the data available including home team, visiting team, runs scored by each team, and hits recorded by each team.

The Lahman Database is updated yearly, and at the time of collection for this analysis the 2023 statistics are the most recent available. It is freely available online in multiple formats including CSVs, or SQL files compatible with MySQL. For this research, the CSVs were used in conjunction with SQL scripts to define the schema and populate a PostgreSQL database, which was connected to the python environment through the psycopg2 library. There are five primary Lahman Database tables used in this analysis which are tabulated in Table 1. In addition, full descriptions of the data are available via a 'readme,' which is published with the data.

Table 1: Lahman Database Tables

Table Name	Description
People	Player names and biographical information
Teams	Yearly stats and standings
Batting	Batting statistics such as AVG, SLG, OBP, OPS
Pitching	Pitching statistics such as ERA, WHIP

3.2 Data Preparation & Cleaning

The author populated a database in PostgreSQL, a common relational database system, utilizing the CSV files sourced from the Lahman Database as well as modified SQL scripts obtained via a pubic GitHub page. (Nycum, 2016) Each CSV files represents one of twenty-seven unique tables. The game log information is parsed from txt files and stored in memory in the Python environment. Due to minor discrepancies of team name abbreviations between data sources, data cleaning steps were taken to preserve the relational nature of the data. The discrepancies are caused my Retrosheet's unique identifiers for the current MLB teams being different due to the granularity necessary to store data from teams who may have relocated or rebranded. For more information, please see Appendix A.

Before providing a detailed analysis on the methodologies, it is important to understand where each experiment is contained within the overall research. Therefore, figure 4 depicts a process flow diagram intended to be used as a visual aid when reviewing individual experiments.

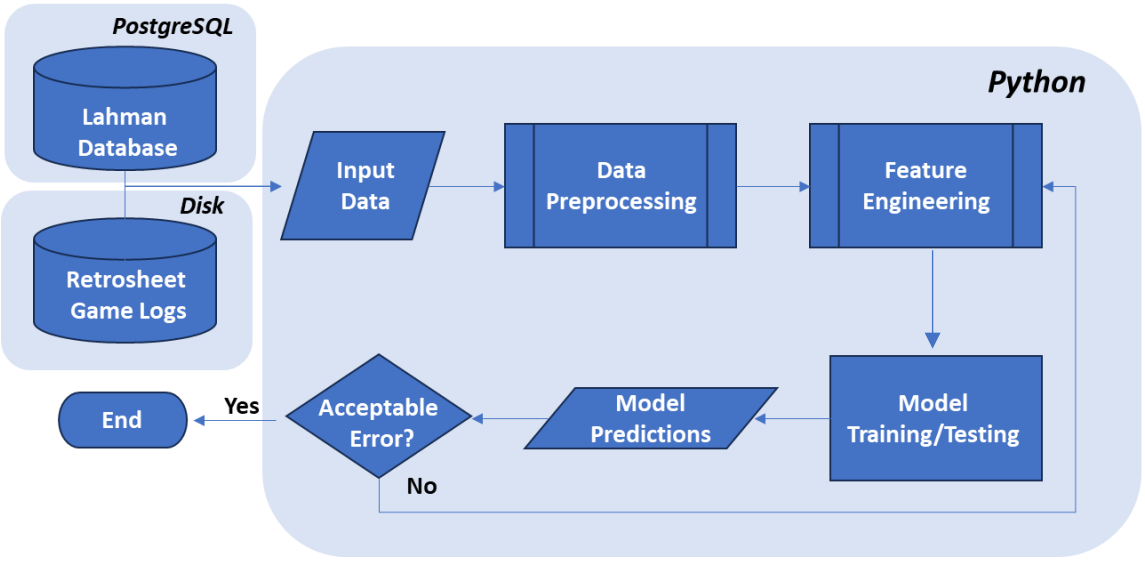

Figure 1: Analysis Flow Diagram

4. Methods

Each of the following experiments and methods were performed in Python using various libraries documented on the author's GitHub page.

4.2 Weighted Least Squares Regression, Park Factor

As stated in section 2.2.2, this analysis assumes that the number of runs scored in each MLB game can be modeled as a linear combination of three factors: the offensive team's strength, the defensive team's strength, and the ballpark in which the game is played. A linear system of equations was defined using all games played in the 2022 and 2023 MLB regular seasons, and then was solved for via the weighted least squares model from the 'statsmodels' library. Because the output vector *b* represents the number of runs scored by a team, the vector *x* represents the number of runs produced because of the offense, defense, and park respectively.

To assess the results of the regression, the values were compared to Baseball Savant's 2023 Park Factor estimates (Baseball Savant, 2025). Baseball Savant scales their metric to an average of 100, and in the case of the 2023 data was in the range of 93 – 113. Therefore, both Baseball Savant's metric as well as this analysis' metric were normalized via min-max normalization to compare their values on the same scale, zero to one. The table below shows the 5 smallest and 5 largest differences between the different normalized park factor metrics.

Table 2: Park Factor Comparison

Team	Difference (%)
COL	0.00
LAA	0.07
TOR	0.25
SDP	1.29
NYY	1.31
PIT	15.74
HOU	17.27
WSN	17.32
ATL	19.21
KCR	22.73

The results are mixed, which can be expected due to this research's park factor approximation account for only three variables.

4.3 Game Difficulty Assessment

Section 2.2.2 also introduced the concept of baseball teams "getting hot," or performing exceptionally well in recent observations. In this section, a unique score is derived for each game measuring the difficulty of playing that opponent at that time, given the opponents' recent success in scoring runs. The motivation for this metric is to use it to augment the starting pitcher's performance assessment based on the difficulty of facing a particular team on that day.

Two averages were calculated for the opponent: a moving average of runs scored in the previous six games, as well as an average of runs scored for the entire year. Six games was chosen because it approximates two series in baseball, which are traditionally 3 to 4 games against a particular team; however, this is a configurable parameter for future studies. Together these metrics should represent both the challenge of facing a good team who has performed well overall, as well as a 'hot' team who is performing well lately. This analysis weighted these scores defining the moving average to be twice as important as the overall average; however, this is also configurable and there is opportunity to learn the appropriate weights, which will be discussed in section 8. In figure 2, the Philadelphia Phillies 2023 season's difficulty scores are shown as an example with a subset of the games displaying the Phillies starting pitcher for that day. E.g. a score of 5 would represent 5 runs expected to be score by the opponent based on their averages across the entire season as well as the previous six games. The first six games of the season were removed from the plot as the moving average requires 6 games.

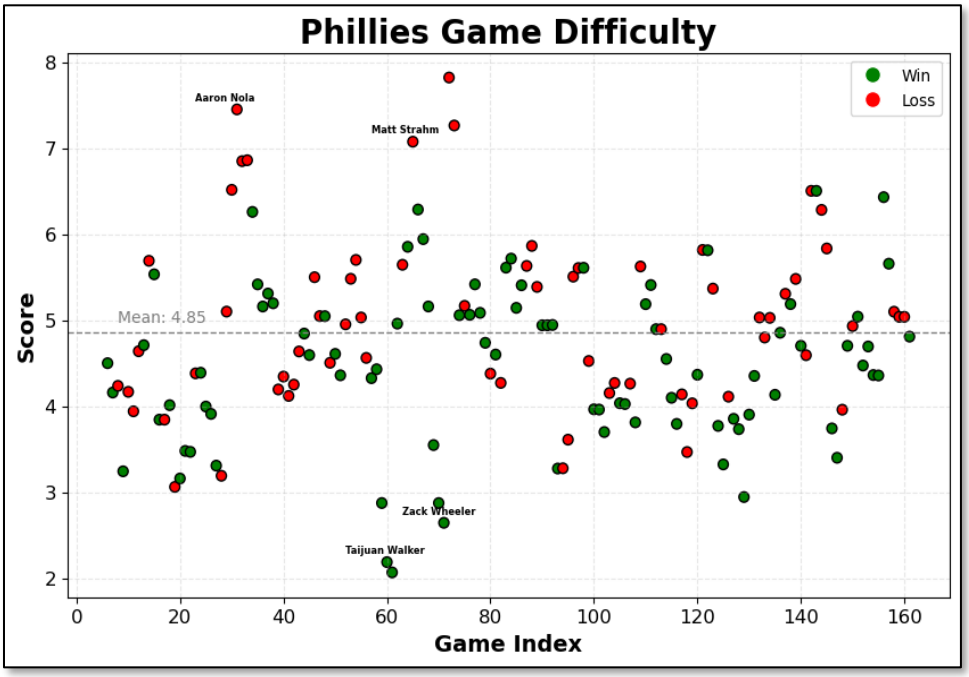

Figure 2: Philadelphia Phillies Game Difficulty

It is clear from the scatterplot that the Phillies won many of the easiest games of the year and lost some of the most difficult games of the year. In fact, by declaring outlier games to be when the opponent's difficulty score is beyond two standard deviations of the mean, there are seven outlier games total in 2023: three games on the lower end, and four games on the higher end. Each of the three lower end outliers were won by the Phillies, and each of the four upper end outliers were lost by the Phillies. Of the three lower end outliers, two were started by Zach Wheeler and none were started by Aaron Nola; Of the four upper bound outliers, none were started by Zach Wheeler and two were started by Aaron Nola. Each of these pitchers started 32 games for the Phillies in 2023, so these games represent a small subset of their starts; however, this illustrates how a starting pitcher's performance metrics can be influenced by the difficulty of the games in which they pitch. Figure 3 depicts the distribution of 2023 Phillies game difficulty score separated by wins and losses. Unsurprisingly, the mean difficulty of those games the Phillies lost was higher than the mean for those games they won.

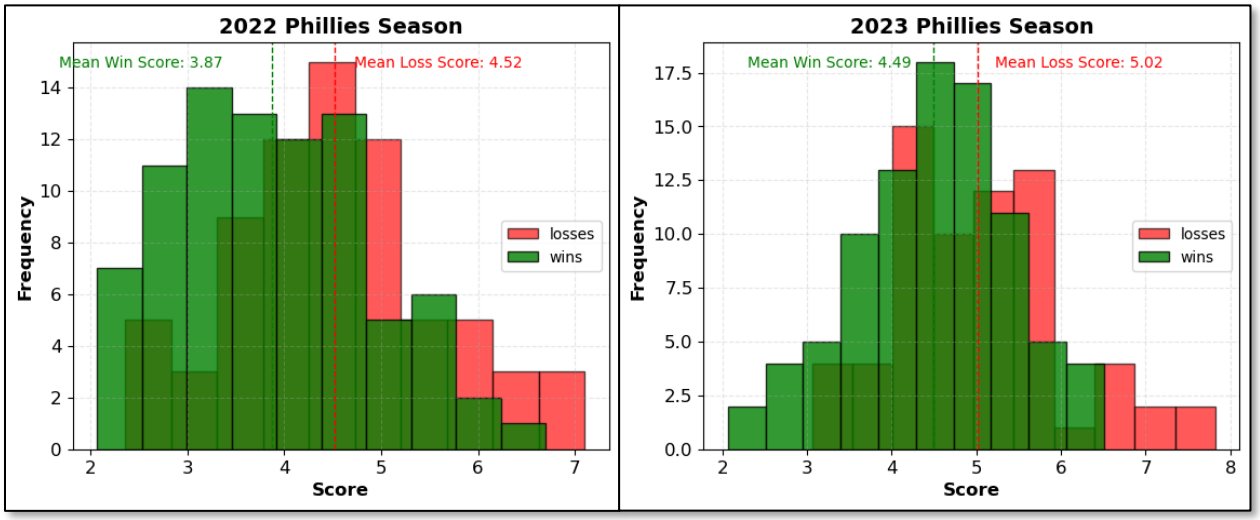

Figure 3: Philadelphia Phillies Game Difficulty Distributions

4.4 Isolating ERA Deviations from Uncontrolled Factors

Once again, this analysis uses the Philadelphia Phillies as an example to show how methodology could be applied to the MLB at large. Philadelphia's home field has a normalized park factor score of 0.365, which indicates it is more favorable to the pitchers than the batters; Table 3 shows some of the statistics collected for the following analysis. Unsurprisingly, the park factor measurements for pitchers who share a home field are close in value due to approximately half of the games pitcher for each of them being in the same ballpark.

Table 3: Philadelphia Phillies Starting Pitcher Metrics

	2022			2023		
Pitcher Name	**Avg Game Difficulty Score**	**Park Factor Norm**	**ERA**	**Avg Game Difficulty Score**	**Park Factor Norm**	**ERA**
Aaron Nola	4.18	0.34	3.25	4.77	0.35	4.46
Ranger Suarez	4.33	0.32	3.65	4.83	0.37	4.18
Zach Wheeler	4.25	0.34	2.82	4.58	0.32	3.61

This analysis used the observed changes in park factor and game difficulty score for each pitcher in combination with their 2022 ERA to predict their 2023 ERA. Pitchers included in this analysis were first limited by having started at least seven games in both seasons, which accounted for nearly 150 unique pitchers. In addition, knowing there are many other contributions to a pitcher's ERA that are not modeled here, the data was further reduced to only those pitchers whose change in ERA from 2022 to 2023 was within one standard deviation of the distributions mean. The reason for this is because the expectation is that large changes in ERA are expected to be a result of more impactful changes such as injuries or a change in the

effectiveness of the pitchers themselves due to velocity, spin rate or other; this analysis aims to extract small changes in ERA for an otherwise steady pitcher due to park factor and opponents.

A Mulit-Layer Perceptron (MLP) was the model architecture chosen to predict the ERA; This includes an input layer of size three for the previous year's ERA value, year-over-year park factor

Layer (type)	Output Shape	Param #
input_layer_10 (InputLayer)	(None, 3)	0
Hidden_Layer_1 (Dense)	(None, 256)	1,024
dropout_20 (Dropout)	(None, 256)	0
Hidden_Layer_2 (Dense)	(None, 128)	32,896
dropout_21 (Dropout)	(None, 128)	0
Hidden_Layer_3 (Dense)	(None, 64)	8,256
Hidden_Layer_4 (Dense)	(None, 28)	1,820
Output_Layer (Dense)	(None, 1)	29

Figure 4: MLP Summary, ERA Prediction

change and the year-over-year game difficulty change, as well as an output layer of size one which represents the predicted ERA. Both the input and output ERA values were normalized to a range from 0 to 1, and the park factor and game difficulty features were normalized between-1 and 1. There are 4 hidden layers and two dropout layers, which represent the method of regularization. Empirical analysis provided motivation for a combination of the two activation functions: hyperbolic tangent and leaky ReLU. Leaky ReLU, like ReLU, is utilized in part to prevent vanishing gradient and converge more quickly, and hyperbolic tangent is used as the activation on the input features knowing two of the inputs contain negative values. Having measured the mean change in ERA among all pitchers in the training and testing datasets inclusive to be approximately 0.5, the values the model was defined to predict was the difference of the truth

2023 ERA value and the mean increase in ERA between 2022 and 2023. This was done to remove global effects that increased ERA in 2023 including rule changes to the MLB, which is briefly addressed in section 6.

5. Results

The 26 pitchers left out for testing the ERA prediction model are plotted below in figure 5. The results are shown from left to right in order of least to greatest error, where error is defined to be the difference between the predicted 2023 ERA and the measured 2023 ERA; the smallest error is less than 0.3%, and the greatest error is almost 34%.

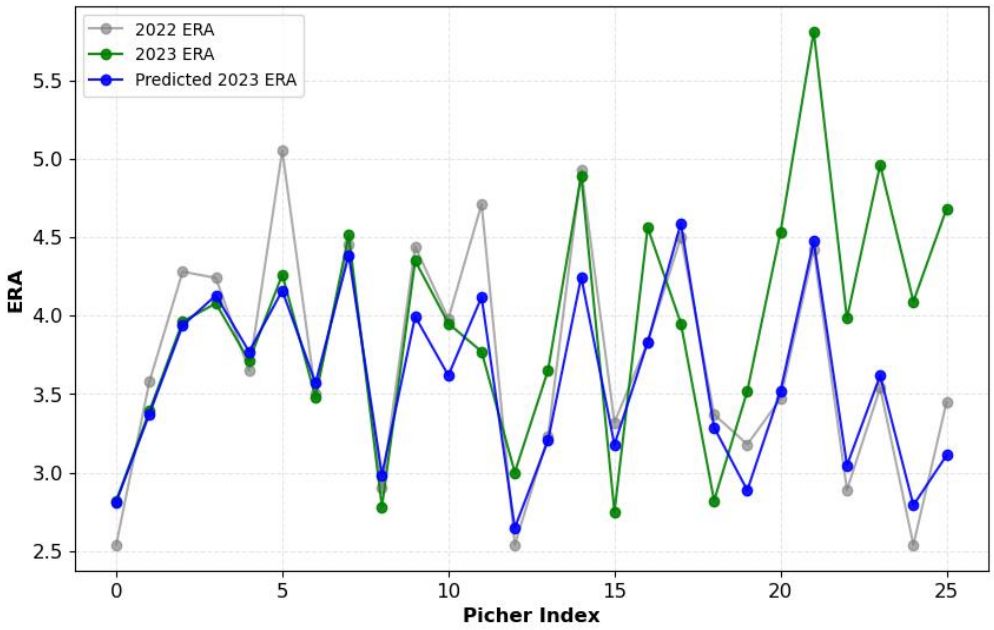

Figure 5: ERA Model Prediction Errors

When inspecting the larger errors in the test dataset, it becomes clear there is a correlation between error and total change in ERA. I.e. As the total change in year-over-year ERA scales in the positive direction, the larger the error in the model's prediction of that ERA. E.g., the largest error observed was for Garret Whitlock, who pitched for the Boston Red Sox in 2023 and observed a nearly 50% increase in his ERA. He dealt with more than one injury during the 2023

season that included hip surgery recovery and nerve issues in his elbow. (Couture, 2023) In addition, he recorded bottom of the league averages for 'batting average on balls in play' as well as barrel rate, which is effectively how well the batter makes contact with his pitches. (MLB.com, 2025) This information helps explain the large errors in predicted ERA; the difference in ERA due to ballpark score and game difficulty as measured in this analysis account for much finer changes in these metrics as compared to something like injuries. Figure 6 visualizes the correlation between the true change in ERA from 2022 to 2023 with error in the predicted 2023 ERA for the test dataset.

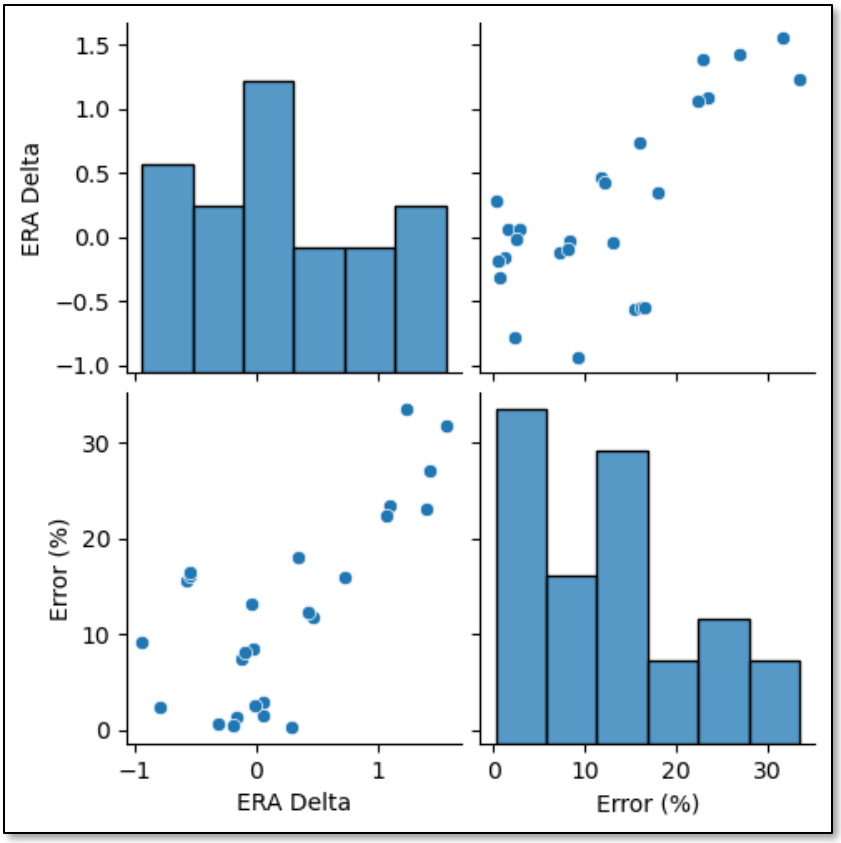

Figure 6: ERA Change vs Prediction Errors

6. Discussion

Although the exploratory data analysis provides evidence of recent opponent success and park factor having an effect in the result of baseball games, there is still a challenge in modeling

its effect in isolation from the multitude of other effects on a baseball pitcher's performance. E.g., there were several new rules implemented in the beginning of the 2023 season including: pitch clock timers, a ban on defensive shift alignments and an increase in base size. (Castrovince, 2023) Together these rules were designed to increase fan engagement by increasing the number of exciting plays and shortening the time of the game. Because the MLB devised these rules to create more exciting plays and consequently runs scored, the rules affected the pitchers' performance in a negative manner overall. As mentioned in section 5, there was a mean increase in ERA of 0.5 runs between the 2022 and 2023 season, in part due to these rule changes.

It is possible a more comprehensive data preprocessing analysis could lead to reduced errors in the model's predictions. For example, pitchers who were injured for an empirically defined amount of time may be removed from the dataset. The utility in this prediction would be for an MLB team to have a better understanding of how well its pitchers are performing based on how the difficulty in their appearances is trending. Therefore, there would be no use case where this is a necessary prediction after a player is injured.

7. Conclusions

This analysis provided evidence of a strength of opponent score influencing the outcome of a game, as shown in figure 3 of section 4.3. Although the Philadelphia Phillies are used as an example, the same process can be extensible to other teams. In particular, when inspecting the only 7 games that fell outside two standard deviations of the mean score, the Phillies lost all those that fell above the upper bound and won all those that fell below the lower bound. This is evidence of some predictive value in this score; therefore, it can be used to augment pitcher's

metrics based on how difficult that game was based on factors beyond their control. In addition, the MLP defined in section 4.4 provided evidence that changes in ballpark factors and opponent success can be used to predict a change in ERA. For a pitcher who is not affected by larger influences such as injuries or loss of pitch effectiveness, this model can be used to predict small differences in player performance due to the environment in which the game is played.

8. Directions For Future Work / Analysis Limitations

There may be other model architectures or modifications this MLP that may produce better results. In addition, expanding the dataset and experimenting with further pre-processing may yield better results.

Overall, the author chose to use runs scored to define both park factor and opponent success; however, it is possible to use other traditional metrics or some combination of many metrics. This analysis assumed the running average opponent success was twice as important as a team's total season success when computing a 'game difficulty' score, but these weights can be learned as opposed to be predefined by the author. This may be achieved by defining a regression model using both values as features to predict the truth runs scored.

Finally, from an analysis pipeline perspective, there are possible improvements to make such as storing the Retrosheet game log data in a PostgreSQL as well. This would protect data integrity, increase accessibility, and remove the need to read and store the data in a python environment.

9. Data/Code Availability

All data and code utilized for this analysis can be made available upon request to the author.

References

Acharya, R. A., Ahmed, A. J., D'Amour, A. N., Lu, H., Morris, C. N., Oglevee, B. D., and Swift, R. N. "Improving Major League Baseball Park Factor Estimates." Journal of Quantitative Analysis in Sports 4, no. 2 (2008). https://doi.org/10.2202/1559-0410.1108.

Baseball Savant. "Statcast Park Factors | Baseballsavant.com." Accessed February 13, 2025. https://baseballsavant.mlb.com/leaderboard/statcast-park-factors?type=year&year=2023&batSide=&stat=index_wOBA&condition=All&rolling=1&parks=mlb.

Calzada, Daniel. DeepBall: Modeling Expectation and Uncertainty in Baseball With Recurrent Neural Networks. Master's thesis, University of Illinois at Urbana-Champaign, 2018. http://www.retrosheet.org/Research/Calzada/CALZADA-THESIS-2018.pdf.

Castrovince, Anthony. "New Baseball Rules for 2023 FAQ." MLB.com.. Last modified February 6, 2023. https://www.mlb.com/news/mlb-new-rules-for-2023-faq..

Couture, Jon. "Garrett Whitlock's Lost Year Mirrors That of His Red Sox." Boston.com, September 18, 2023. https://www.boston.com/sports/boston-red-sox/2023/09/18/garrett-whitlock-stats-2023-season-red-sox/.

FiveThirtyEight. "How Our MLB Predictions Work." Last modified May 2, 2022. https://fivethirtyeight.com/features/how-our-mlb-predictions-work/.

MLB.com.. "Barrel | Glossary." Last modified March 7, 2025. https://www.mlb.com/glossary/statcast/barrel..

Nycum, Brent. Conversion of the Lahman Baseball Database to PostgreSQL. GitHub, 2016. https://github.com/brentnycum/lahman-postgres/tree/master.

Sean Lahman. "Lahman Baseball Database." SeanLahman.com. Accessed January 17, 2025. http://www.seanlahman.com/.

Thorn, John. "Chadwick's Choice: The Origin of the Batting Average." Our Game. Last modified September 18, 2013. https://ourgame.mlblogs.com/chadwicks-choice-the-origin-of-the-batting-average-e8e9e9402d53.

Wikipedia contributors. "Elo rating system." Wikipedia, The Free Encyclopedia. Last modified January 2025. https://en.wikipedia.org/wiki/Elo_rating_system.

Appendix A

Team Name Information

Retrosheet contains game logs as far back as 1901; Since many teams have changed locations or team names over time, there are many more teams accounted for when compared to the team names returned from a query on the Lahman database from the years 2022-2023. Because there are so many extra teams, the team abbreviations do not exactly match what we expect today. E.g., there are unique 7 teams who played in Chicago since 1901. Today there are two: the Chicago Cubs, the Chicago White Sox. The abbreviation these days for the Cubs is “CHC”, but Retrosheet defines it to be “CHN”.