Paper 2875-2018

Simulating Major League Baseball Games

Justin Long, Slippery Rock University; Brad Schweitzer, Slippery Rock University;
Christy Crute Ph.D, Slippery Rock University

ABSTRACT

The game of baseball can be explained as a Markov Process, as each state is independent of all states except the one immediately prior. Probabilities are calculated from historical data on every state transition from 2010 to 2015 for Major League Baseball games and are then grouped to account for home-field advantage, offensive player ability, and pitcher performance. Using the probabilities, transition matrices are developed and then used to simulate a game play-by-play. For a specific game, the results give the probability of a win as well as expected runs for each team.

INTRODUCTION

Mathematics and statistics are becoming incorporated into the game of baseball. The main points of focus in this research are the teams that are playing, the significance of a team's home-field advantage, and each player's performance. Each game can be modeled entirely using a sequence of random events where each event only depends on its immediate predecessor, namely, Markov Chains. Data from 2010 to 2015 is used to find the probabilities of transitioning from one state to the next. These transitions are further refined by creating two matrices per team; one for home performance, and one for away. To increase accuracy, players are then categorized into eight classes based on their weighted on-base average (wOBA). To incorporate the varying performance of pitchers, scalar matrices are developed by classifying pitchers in a similar manner. These scalar matrices are applied to the batter transition matrices to create a set of final matrices. These matrices are then used to create a final simulation. By walking through these events and keeping track of how many runs are scored, our model can output a team's overall win-percentage.

RELATED WORK

Baseball is a huge area of research in the realm of predictive analytics. A variety of topics have been researched including winning advantages, attendance, and injuries. While our model focuses on predicting winning outcomes of each game, it is necessary to address other research topics that have been done in this field.

Kyle and Ken Lin developed the concept of winning advantage (WA), which represents the chance of winning that is attributed to the state itself and does not include the chance of winning that is attributed to the players [1]. Plays that contribute more to the team winning the game will have a greater WA than a play that has little or no impact. For example, a home run in the 9th inning when the game is tied will have greater WA than getting an out. For ranking players, their winning advantage score (WAS) is the cumulation of the WA of each play they were involved in. While an innovative concept, it does not give a fair score to players that do not have a chance to be a part of game winning plays: however, this could be an alternative measure to group players by in our own model.

With a two-stage Bayesian model, Tae Young Yang and Tim Swartz successfully created a model to predict winners of each game [2]. They combined the past winning percentages of two teams, overall batting ability, and starting pitcher’s ERA into a single measurement which was used to determine the strength of a team at a given point in time which is known as a team’s relative strength. This model also included a variable to account for home field advantage. They clearly stated that all other variables not relating to relative strength and home field advantage contribute, but not significantly enough to make a difference to the outcome. The two stage Bayesian model includes a stage where the probability of winning is based off a beta distribution with parameters stemming from relative strength and home field advantage. The second stage uses a Bernoulli distribution to calculate outcome. To add an extra level of variation, a one stage Bernoulli distribution was created to account for the difficulty of modeling outcomes.

To interpret these models easily, Markov chain algorithms were used. This model was capable of predicting future game outcomes as well.

A study titled Predicting the Final Score of Major League Baseball Games, written by students at Stanford makes use of data similar to what will be incorporated into our model [3]. The students created a model that makes use of both teams starting lineups, as well as the current state of the game and details about the field. Their research focuses more on simulating each at-bat and determining the most likely outcome, to then try and predict the final outcome of the game. The model does not take into consideration that players may perform differently under different conditions; it places more emphasis on batter and pitcher match-ups.

Keith Woolner published an article that focuses on the creation of a model to forecast the number of runs a team will score per inning [4]. The created model accurately predicts the probability of a team scoring a certain number of runs per inning based solely on a team's average number of runs per game and the number of times a team scored a certain number of runs per game. The model is accurate within 2\% of the actual result. It is also worth noting that the predicted runs were then used to output the likelihood of each team winning the game.

A study conducted by Bruce Bucket and Elliotte Rusty Harold is closely related to the research we are interested in [5]. Their research focused on evaluating the performance of baseball teams and the individual players themselves using a Markov chain method. They use deterministic models to calculate runner advancement along with pitching and defensive variables. The model is able to predict run distributions and solve for variation in expected number of runs using various transition matrices. Bucket and Harold also used their model to optimize batting lineup.

In the industry of major league sports, baseball is the most advanced in terms of statistical analysis methods, and accordingly, there are studies and proposed models to focus on nearly every aspect of the sport; however, there is room for improvement. Each article focuses on a specific part of the game, whether it be predicted runs per inning, individual player performance, pitcher matchups or overall team performance; the models don't bring together every significant factor. Whereas they take a narrower focus, we aim to incorporate a wide range of parameters in order to accurately and reliably forecast the probability of a team winning a game, while still maintaining the precision that the respective authors did when they isolated their variables.

DATA

Data was sourced from retrosheet.org, fangraphs.com, and sealahman.com [6,7,8]. Retrosheet is an extensive baseball database that contains play-by-play data for every game from 1920 to the present. This historical data was used to create probabilities for state transitions, as well as providing most of the data needed to calculate a player's wOBA. The Lahman Baseball Database, created by award-winning database journalist Sean Lahman, was among the first of its kind to provide historical baseball data. It was used to provide the remaining data needed to calculate wOBA and additional information on players for a more effective batting lineup selection. Fangraphs, a popular baseball statistics and analysis website, was the source of weights for the wOBA calculation. Fangraphs was also used to determine each pitcher's SIERA.

METHODOLOGY

ASSUMPTIONS

In order to simplify the model, the following assumptions were made:

- The batter begins each state transition, i.e. stealing, pickoffs, bulks, etc. are not considered.
- No delays or injuries take place.
- Field conditions are constant and insignificant.
- Starting lineup remains constant.

- Pinch hitters and pinch runners are not used.
- No runs are scored on a 3rd out play

STATES AND TRANSITIONS

Being a Markov chain, baseball is a series of transitions from one state to another. In general, there are a total of eight base-states from which players move in and out; Table 1 below shows these states, with each "x" representing a player on the given base. For a clarifying example, state 3 gives runners on 1st and 2nd base and no one on 3rd. Although there are 21 possible candidate actions the batter can cause (e.g. single, double, triple, walk, etc.), we are only interested in the result of said action: not the action itself. For example, if the game is currently in state 2 and ends in state 4, which action occurred is irrelevant. Of relevance is that it ended in state 4. But transitioning between these 8 states does not account for runs and outs, so a separate matrix must be used to account for the runs and outs that occur during a play.

	Bases		
State	1st	2nd	3rd
0	-	-	-
1	x	-	-
2	-	x	-
3	x	x	-
4	-	-	x
5	x	-	x
6	-	x	x
7	x	x	x

Table 1. Basic Base States

TRANSITION & RUNS MATRIX

In order to account for runs and outs, the 8 possible base states are combined with having either 0, 1, 2, or 3 outs. This combination results in 32 base-out states (which we will be referred to as states throughout the paper). With these new states, runs can be inferred by certain transitions in the matrix. For example, a transition from state 3 with 0 outs to state 0 with 0 outs will yield 3 runs for the team.

To represent state transitions easily, a 32x32 matrix is created, with rows and columns representing start and end states, respectively. Each cell of the matrix contains the probability of the transition from a start state to end state. Given a start and end state, probabilities are found by taking the number specific transitions and dividing by the total number of transitions to any end state. This is given by Equation 1 below:

$$P(end\ state \mid start\ state) = \frac{\#\ of\ transitions\ from\ start\ state\ to\ end\ state}{\#\ of\ transitions\ from\ start\ state\ to\ any\ end\ state}$$

Equation 1. Probability of Transitions

Thus for a given transition matrix T, $T_{i,j}$ will give the cell containing a transition from starting state i to the ending state j, so $T_{5,6}$ represents the cell containing the probability P(1_1|1_0) which is starting in state 1 with 0 outs and ending in state 1 with 1 out. An example of an arbitrary transition matrix is given below in Table 2.

Transition Matrix		End State			
		0_0	0_1	. . .	7_3
Start State	0_0	P(0_0 | 0_0)	P(0_1 | 0_0)	. . .	P(7_3 | 0_0)
	0_1	P(0_0 | 0_1)	P(0_1 | 0_1)	. . .	P(7_3 | 0_1)
	.	.	.	.	.
	.	.	.	.	.
	.	.	.	.	.
	7_2	P(0_0 | 7_2)	P(0_1 | 7_2)	. . .	P(7_3 | 7_2)

Table 2. Transition Matrix

Since we are now working with base-out states, runs can be inferred by certain transitions in the matrix. A fixed number of outs and runs need to occur for certain transitions. Thus, a play can result in a certain number of outs, the same number of runs, or some combination of the two, as long as they all have equivalent sums. The transition matrix considers the number of outs that occurred on the play, so the number of runs that occurred on the play can then be derived. The number of runs that occur on each state transition are stored in a runs matrix, which is used in the simulation to determine the number of runs scored by each team.

HOME-FIELD ADVANTAGE

In order to consider home field advantage, data from each team is separated between their home and away games to calculate the probabilities for the transition matrices. Thus, the data is partitioned by team, and then further by when they played at home versus away. These matrices were created from the most recent five years of events to account for roster changes each year.

PLAYER ANALYSIS

To factor in a player's offensive ability, weighted on base average (wOBA) is used. This measure gives a unique weight to all possible batting outcomes, and combines them into one statistic for a measurement of a player's overall offensive contribution. The weights, which are calculated from historical data by Fangraphs, correspond to how much a given batting event adds to run production. It follows that, a double's weight is greater than a single's, a triple's weight is greater than a double's, etc. The formula used for computing a batter's wOBA is given by Equation 2 below.

$$wOBA = \frac{(w_1 * NIBB) + (w_2 * HBP) + (w_3 * 1B) + (w_4 * 2B) + (w_5 * 3B) + (w_6 * HR)}{Plate\ Apperances}$$

Equation 2. Calculation of wOBA

The variables used in the calculation are in Table 3.

Variable	Description
$W_1 \ldots W_6$	Batting Event Weight
NIBB	Non-intentional Walk
HBP	Hit by Pitch
1B	Single
2B	Double
3B	Triple
HR	Home Run

Table 3. wOBA Variables

Once each player's wOBA is found, they are grouped into one of 8 separate classes with the first class being the lowest performing players and class 8 being the best. Similar to the accounting for home-field advantage, the historical data is partitioned by wOBA class, and then used to calculate probabilities for the transition matrices.

PITCHER ANALYSIS

Pitchers were classed in a similar manner. Skill-Interactive ERA (SIERA) was used to classify pitchers into one of 7 classes. SIERA estimates ERA through a pitcher's walk rate, strikeout rate, and groundball rate. It attempts to eliminate factors that aren't controlled by the pitcher.

Following classification of the pitchers, scalar matrices are developed by first creating 7 transition matrices based off a pitchers SIERA. The matrices are then compared to a league average matrix to create the scalar matrix which is applied to a batter's transition matrix.

SIMULATION

To calculate the winning percentage of a team, simulations are performed in SAS. The steps are as follow:

1. Select home and away teams
2. Set starting line-up for each team
3. Choose number of games to be simulated
4. Set starting inning and score
5. Begin inning in state 0 with 0 outs
6. Game iterates play-by-play
7. Runs are recorded for certain transitions
8. Batting team changes when a 3rd out state is reached
9. If there is not a tie, the game ends at the end of 9th inning
10. Win is recorded for the respective team

At the start of an inning, the first batter's transition matrix is used to select the next state of the game. The next batter in the lineup will start in the previous end state and so on for the rest of the batters. In the case of a tie at the end of the ninth inning, the simulation will continue to run until either team is winning at the end of one of the additional innings. The game will also end if the home team is winning after the away team bats in the ninth inning. The simulation also records the difference between the home team and away team's score at the end of each game, and returns the average difference along with each team's win probability.

VALIDATION

For model validation, 10,000 simulations were run to mimic each of the 53 series in the Pittsburgh Pirates 2016 season. Since a full series was simulated instead of an individual game, the win percentage was multiplied by the number of games in the series to calculate expected wins. The simulation returned that the Pirates were expected to win 93 games, while they only won 78 games. Thus, there was a 9\% margin of error. This may be due to a variety of factors, as discussed in the conclusion section.

CONCLUSION

Our model does have a few drawbacks, as there is a discrepancy between the actual number of wins versus the simulated. For efficiency purposes, each series was simulated assuming that the starting lineup of each team remained unchanged, which is not always the case; usually, starting lineups vary game to game. This assumption also implies if a player is injured, his replacement will be a batter of the same wOBA class as him and that there were no pinch hitters or runners for either team. Relaxing this assumption could lead to better results.

The simulations for validation purposes were performed before pitchers were fully incorporated into the model. It's expected that incorporating pitchers will also reduce the margin of error.

The model also does not take into consideration the impact of winning or losing streaks, travel time, time between games, and batter fatigue. These factors are difficult to quantify and could be responsible for some of the variation in our results.

REFERENCES

[1] Lin, Kyle Y., and Ken K. Lin. "Winning Advantage and Its Applications." (2007).

[2] Yang, Tae Young, and Tim Swartz. "A Two-Stage Bayesian Model for Predicting Winners in Major League Baseball." *Journal of Data Science 2* (2004): 61-73. www.jds-online.com. 2004.

[3] Cserepy, Nico, Robbie Ostrow, and Ben Weems. “Predicting the Final Score of Major League Baseball Games.”

[4] Woolner, Keith. “An Analytic Model for Per-Inning Scoring Distributions.”

[5] Bukiet, Bruce, et al. A Markov Chain Approach to Baseball*. Operations Research*, vol. 45, no. 1, 1997, pp. 1423. JSTOR, JSTOR.

[6] Smith, David. *Retrosheet* www.retrosheet.org.

[7] Lahman, Sean *SeanLahman.com* www.seanlahman.com.

[8] *Baseball Statistics and Analysis* | *FanGraphs Baseball*, www.fangraphs.com

ACKNOWLEDGMENTS

Dr. Christy Crute, who has a strong background in the Operations Research field provided advisement.

Tom Tango, a database architect for MLB advanced media, provided guidance.

Data was sourced from retrosheet.org, fangraphs.org, and seanlahman.com.

CONTACT INFORMATION JUSTIN LONG

Your comments and questions are valued and encouraged. Contact the author at:

Justin Long
JBL1005@sru.edu

CONTACT INFORMATION BRAD SCHWEITZER

Your comments and questions are valued and encouraged. Contact the author at:

Brad Schweitzer
BXS1074@sru.edu

CONTACT INFORMATION CHRISTY CRUTE

Your comments and questions are valued and encouraged. Contact the author at:

Christy Crute
Christy.Crute@sru.edu