USING A MARKOV CHAIN MODEL TO OPTIMIZE PITCHING SUBSTITUTION STRATEGY

ZACH BECKER

Abstract. In the new age of data-driven baseball, it is more important than ever to take advantage of the modeling tools available to gain an edge over opponents. This paper uses Markov chain models as well as linear and logistic regression in order to compare various managerial pitching strategies. Using a dataset formed from Retrosheet.org's gamelogs, transition matrices are constructed and then used to determine the frequency and resulting value of each transition. There is no single pitching metric to always adhere to as a manager making a substitution, but it is revealed that the 2018 Colorado Rockies utilized a mix of advanced Sabermetrics and basic pitching stats to significantly outperform their estimated win total.

Contents

1. Introduction

I believe that the importance of managerial pitching decisions throughout a baseball season is underestimated by the average spectator, and by finding the best possible substitution strategy, a team is able to win far more close

Date: 2023-04-28.

This document is a senior thesis submitted to the Department of Mathematics and Statistics at Haverford College in partial fulfillment of the requirements for a major in Mathematics.

ballgames throughout the course of a season. This thesis will compare several strategies by diving into the Colorado Rockies 2018 season, and using data for substitutions based on a variety of pitching metrics.

The Sean Lahman baseball database, and game logs from Retrosheet.org are crucial for investigating the data, since they allow for me to carry out a process of data wrangling to get a clean dataset in R. Additionally, with access to FanGraphs' collection of all offensive and defensive statistics as well as performance metrics for each team in recent years, I will be able to use R to determine parameters in a Markov Chain Model. This model will allow me to determine which substitution strategies had the most success in various situations throughout the 2018 Colorado Rockies season.

2. Core Hypothesis

There are plenty of stats and metrics to measure a pitcher's performance, but it is most useful to investigate the stats that do the best job at taking out chance and/or the performance of the defense behind the pitcher. In the long run, we would guess that the best predictor of a pitcher's performance, and the substitution strategy that leads to the most beneficial transitions for the pitcher, would either be the handedness orientation of the pitcher/hitter matchup, or an advanced stat that does not rely on the defensive performance or alignment on the play.

Thus, through this thesis I will compare several ways of measuring pitcher performance (which will be the stats used for various substitution strategies) and determining which way is the best way for predicting wins. My prediction is that the most optimal strategy, when used, will have helped the Rockies win several close games. My model for each substitution strategy will directly estimate wins as well as positive transition results through a run expectancy matrix so that I can compare the results. The run expectancy matrix, which will be explained later, helps to determine which transitions are beneficial for the defense, and can help track approximately how many runs each strategy could save.

In 2018, the Rockies won 8 more games than they were estimated to, given their total runs scored and total runs allowed. Bill James' Pythagorean Expectation for Wins equation, given by $Wpct = \frac{R^2}{R^2+RA^2}$ is a commonly used equation for estimating how well a team performed in a season. R represents the total runs scored by a team, and RA represents the total runs allowed by the team. This means that the Rockies won several close games thanks to their pitching staff performing exceptionally well in clutch situations, and I want to determine how much each strategy contributed to this. The red dot in the graph below represents the Rockies, and displays

how they outperformed just about every other National League team in 2018 when it comes to estimated versus actual win totals.

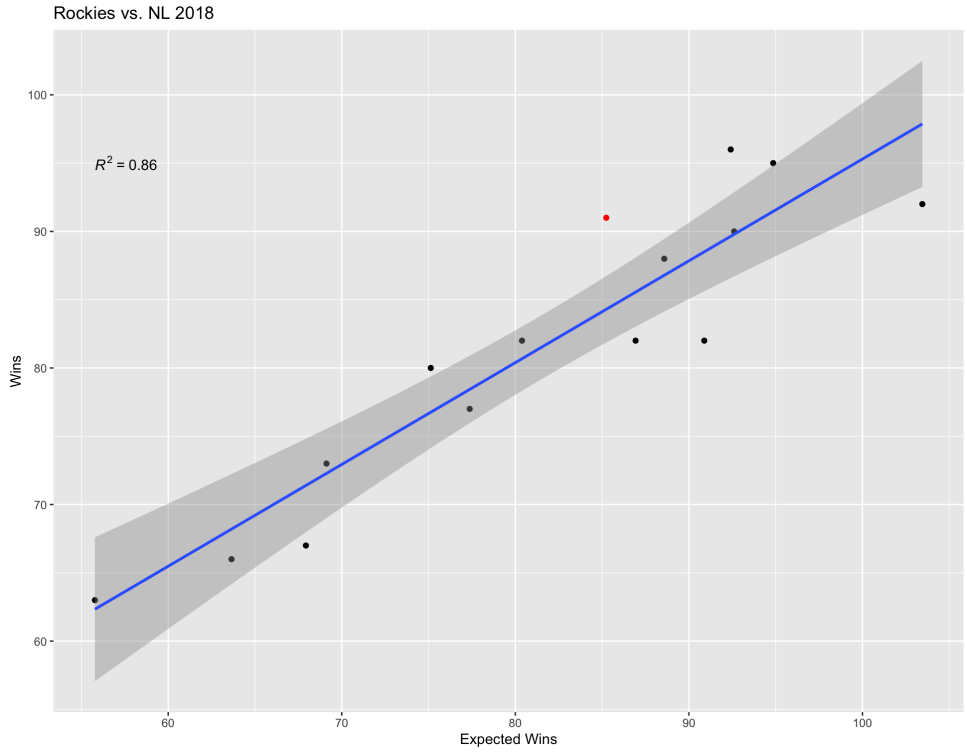

FIGURE 1. 2018 Pythagorean Expectation.

My process will consist of detailing how I got my data into usable form, describing what my Markov chain models will look like, and then diving into how I will use my model to test the hypothesis. I give a thorough description of the substitution strategies I am working with, as well as defining one of the main features of my model which is the "transition value" I derived from the 2018 MLB run expectancy matrix. I then include my regressions and results in order to determine the best way for a manager to make sub decisions for relief pitchers.

3. Data Collection

The process of putting together a clean dataframe in R where each row represents a play in the Rockies 2018 season was a quite exacting task. I was able to download a file of all the plays from the full 2018 MLB season, but needed

to install a github package so that I could have access to the full dataset in R. From there, I also loaded in the Sean Lahman baseball database into R and was able to merge the data in order to create a single dataframe that contained necessary biographical information about the players.

From there, I utilized the dplyr package in R to get the data in my desired format: the games where the Rockies were either the home or away team, the plays where the Rockies were on defense, and only the plays that resulted in a transition so that I could eliminate rows that correlated to things such as pickoffs. I still needed to create several columns, specifically columns that described the current state before the play occurred, the resulting state, whether a sub was made before the play and if so, what the substitution strategy was, the pitcher for the previous play, the transition value of the play which will be discussed later on, whether the play was beneficial or not for the defense, and whether the play occurred in a game where the Rockies won or lost.

Using the mutate function and several case statements in dplyr, I was able to get my full dataframe that consisted of the 6,205 plays that occurred in 2018 when the Rockies were on defense. From there, I needed to create a couple more dataframes in order to gather the details for determining what constituted each specific substitution strategy. I created dataframes that described the number of "high leverage" situations each Rockies pitcher had in 2018, and also the transitions that were most common in late-game scenarios. I describe "high leverage" situations here as the plays that start with two runners in scoring position, i.e., runners on second and third base or bases loaded, and late-game situations as any plays in the 7th inning or later.

The purpose of including the pitchers with the most high leverage appearances is to get a sense of the usage rate for the Rockies' pitchers when in important situations. Futhermore, the number of high leverage situations a pitcher appears in helps us to understand which pitchers did the best in clutch situations, assuming that the pitchers with more success in high leverage situations would be substituted into the game more than the pitchers who did poorly.

As far as the most common late game transitions, this helps us understand how the Rockies managed to win so many close games. The transitions that happen towards the end of the game impact the expected winning percentage the most for each team. Thus, we can get a better sense of the full data set by seeing how the Rockies performed when it mattered most. The late game transitions are also important, because by the 7th inning, a pitching substitution has typically already been made, and another substitution may

be considered for every subsequent at bat. This is because relief pitchers typically do not pitch more than a couple innings per game, with the majority of relief pitchers only coming in to face a handful of hitters per appearance. Relief pitchers are often "specialists" that have a high velocity fastball or excellent off-speed pitches, but do not have the stamina to pitch several innings like starting pitchers do.

game_id	inn_ct	outs_ct	resp_bat_hand_cd	resp_pit_hand_cd	resp_pit_id	base1_run_id	base2_run_id	base3_run_id	event_tx	current_state	next_state	sub_made	sub_strategy	transition_value	win_or_loss
ANA201808270	1	0	L	R	graxj003	NA	NA	NA	8/L	1	9	N/A	N/A	-0.24	0
ANA201808270	1	1	R	R	graxj003	NA	NA	NA	8/L	9	17	No	N/A	-0.18	0
ANA201808270	1	2	R	R	graxj003	NA	NA	NA	9/F	17	25	No	N/A	-0.11	0
ANA201808270	2	0	L	R	graxj003	NA	NA	NA	7/L	1	9	No	N/A	-0.24	0
ANA201808270	2	1	R	R	graxj003	NA	NA	NA	8/L	9	17	No	N/A	-0.18	0
ANA201808270	2	2	R	R	graxj003	NA	NA	NA	53/G	17	25	No	N/A	-0.11	0
ANA201808270	3	0	L	R	graxj003	NA	NA	NA	8/F	1	9	No	N/A	-0.24	0
ANA201808270	3	1	R	R	graxj003	NA	NA	NA	K	9	17	No	N/A	-0.18	0
ANA201808270	3	2	L	R	graxj003	NA	NA	NA	7/F	17	25	No	N/A	-0.11	0
ANA201808270	4	0	L	R	graxj003	NA	NA	NA	59/L	1	2	No	N/A	0.41	0
ANA201808270	4	0	R	R	graxj003	calhk001	NA	NA	SB2	2	3	No	N/A	1.23	0
ANA201808270	4	0	R	R	graxj003	NA	calhk001	NA	S3/BG.2-3	3	6	No	N/A	0.63	0
ANA201808270	4	0	R	R	graxj003	fletd002	NA	calhk001	S7/L+.3-H;1-2	6	5	No	N/A	0.75	0
ANA201808270	4	0	L	R	graxj003	troum001	fletd002	NA	BK.2-3;1-2	5	7	No	N/A	1.49	0
ANA201808270	4	0	L	R	graxj003	NA	troum001	fletd002	HR/8/F.3-H;2-H	7	1	No	N/A	0.49	0
ANA201808270	4	0	R	R	graxj003	NA	NA	NA	7/F	1	9	No	N/A	-0.24	0
ANA201808270	4	1	R	R	graxj003	NA	NA	NA	5/P7LF	9	17	No	N/A	-0.18	0
ANA201808270	4	2	L	R	graxj003	NA	NA	NA	63/G	17	25	No	N/A	-0.11	0
ANA201808270	5	0	R	R	graxj003	NA	NA	NA	43/G	1	9	No	N/A	-0.24	0
ANA201808270	5	1	L	R	graxj003	NA	NA	NA	63/G	9	17	No	N/A	-0.18	0
ANA201808270	5	2	L	R	graxj003	NA	NA	NA	D9/L	17	19	No	N/A	0.22	0
ANA201808270	5	2	R	R	graxj003	NA	calhk001	NA	43/G	19	25	No	N/A	-0.33	0
ANA201808270	6	0	R	R	graxj003	NA	NA	NA	HR/7/L	1	1	No	N/A	1.00	0
ANA201808270	6	0	L	R	graxj003	NA	NA	NA	13/G	1	9	No	N/A	-0.24	0
ANA201808270	6	1	R	R	graxj003	NA	NA	NA	53/G	9	17	No	N/A	-0.18	0
ANA201808270	6	2	R	R	graxj003	NA	NA	NA	53/G	17	25	No	N/A	-0.11	0
ANA201808270	7	0	L	R	graxj003	NA	NA	NA	31/G	1	9	No	N/A	-0.24	0
ANA201808270	7	1	R	R	graxj003	NA	NA	NA	53/G	9	17	No	N/A	-0.18	0
ANA201808270	7	2	L	R	graxj003	NA	NA	NA	58/L	17	18	No	N/A	0.13	0
ANA201808270	7	2	L	L	mcgej001	youne003	NA	NA	59/G+.1-2	18	21	Yes	handedness	0.22	0
ANA201808270	7	2	R	L	mcgej001	calhk001	youne003	NA	K	21	25	No	N/A	-0.46	0
ANA201808270	8	0	R	R	ottaa001	NA	NA	NA	W	1	2	Yes	baa	0.41	0
ANA201808270	8	0	L	R	ottaa001	troum001	NA	NA	S7/L.1-2	2	5	No	N/A	0.61	0
ANA201808270	8	0	R	R	ottaa001	ohtas001	troum001	NA	W.2-3;1-2	5	8	No	N/A	0.77	0
ANA201808270	8	0	R	R	ottaa001	martj007	ohtas001	troum001	9/F9LF/SF.3-H;2-3	8	14	No	N/A	-0.09	0
ANA201808270	8	1	L	R	ottaa001	martj007	NA	ohtas001	K	14	22	No	N/A	-0.69	0
ANA201808270	8	2	L	R	ottaa001	martj007	NA	ohtas001	SB2	22	23	No	N/A	1.06	0
ANA201808270	8	2	L	R	ottaa001	NA	martj007	ohtas001	W	23	24	No	N/A	0.17	0
ANA201808270	8	2	L	R	oh---s001	cowak001	martj007	ohtas001	SB/L.3-H;2-H;1-3	24	22	Yes	era	1.77	0
ANA201808270	8	2	L	R	oh---s001	youne003	NA	cowak001	W.1-2	22	24	No	N/A	0.23	0

FIGURE 2. Sample of Rockies 2018 Dataframe.

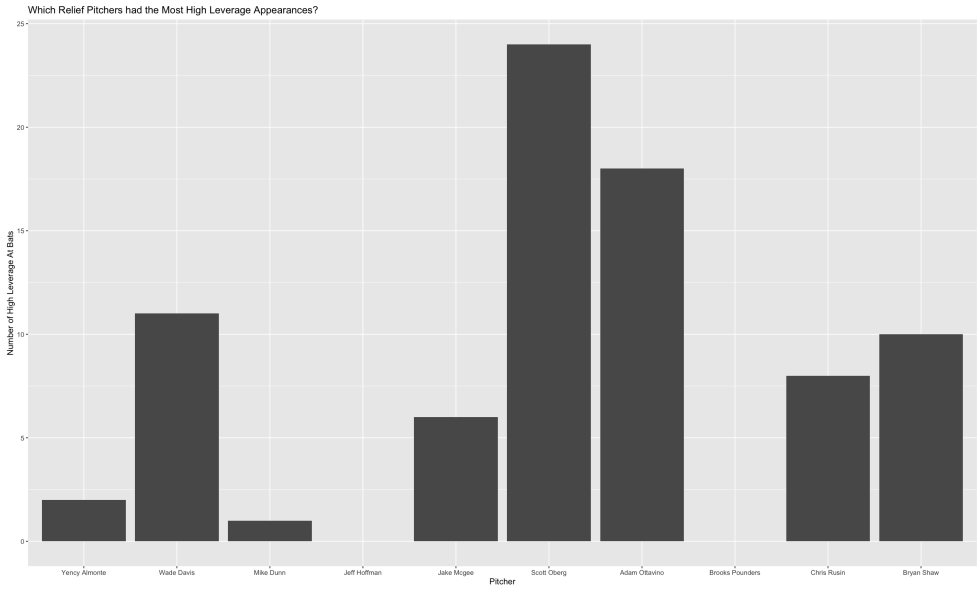

FIGURE 3. High Leverage Appearances.

4. Transition States and Transition Matrices

Definition 4.1. A *stochastic process* is a collection of random variables $X_t, t \in T$ that are defined on the same probability space, where $T \subset \mathbb{R}$. [12]

We say that a Markov chain is a stochastic process that has the Markov property, meaning that the probability distribution of future states only depends on the current state, independent on the occurrences prior to the current state. In baseball, when a play ends, a new state results, but the play that comes after only depends on the situation that results from the original play, and none of the plays before that one. Thus, we can easily fit the flow of a baseball game into this stochastic process. This is a general diagram that helps show what the process of a Markov Chain looks like:

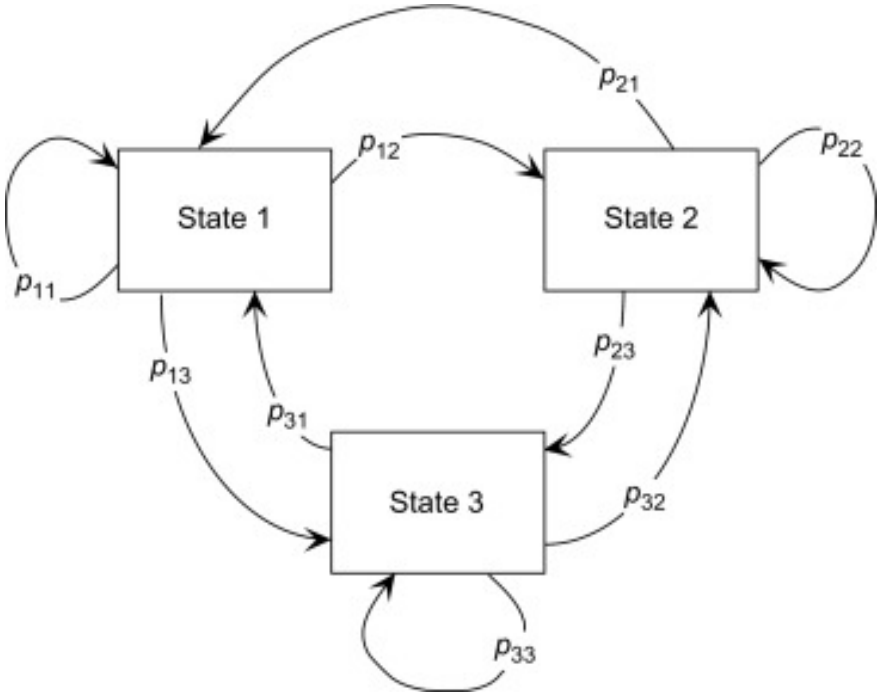

FIGURE 4. Markov Chain Overview.

This is an illustration of a basic Markov chain that has 3 states. Each state can transition to either of the other two states, or back to itself. A baseball diagram would have 28 states, but according to the rules of baseball, some states are not able to transition to any of the other states. For example, it would not be possible to go from state 1 to state 17, since that would require getting two outs on a play where there are no runners on base.

The Markov Chain Model unlocks valuable insights for those investigating the probabilities of each possible transition throughout the course of a baseball game. On the offensive side, there exist 28 states that are made up of combinations of runners occupying various bases and number of outs. Each play in a baseball game corresponds to a transition, meaning that the situation before the play occurs (which bases are occupied and how many outs there are) will be the current state, and the resulting situation after the play happens will be the following state. We refer to the states that do not result in 3 outs and thus the end of the inning as non-absorption states, or *transient* states. It is impossible to leave an absorption state, meaning that $p_{ii} = 1$, given that p_{ii} is an absorption state. The transient states are illustrated here:

(0,0)	(1,0)	(2,0)	(3,0)	(12,0)	(13,0)	(23,0)	(123,0)
(0,1)	(1,1)	(2,1)	(3,1)	(12,1)	(13,1)	(23,1)	(123,1)
(0,2)	(1,2)	(2,2)	(3,2)	(12,2)	(13,2)	(23,2)	(123,2)

The first element in the parentheses represents the runners that are on base, with 1 representing a runner on first base, 12, representing runners on first and second base, 123 representing a bases loaded scenario, and so on. The second element in the parentheses represents the number of outs. The states are numbered going across the rows, so (0,0) is state 1, (123,0) is state 8, (0,1) is state 9, (0,2) is state 17, etc. State 25 represents 3 outs with 0 runs scored, state 26 represents 3 outs with 1 run scored, state 27 is 3 outs with 2 runs scored, and state 28 is 3 outs with 3 runs scored. This is an exhaustive list of the possible states in accordance to the rules of baseball as it is not possible to decrease the number of outs in an inning, a runner will not backtrack on the basepaths during a play, 4 runs cannot score on an inning ending play, etc. The absorption states will result in the end of that current half inning, and the start of the next half inning, unless the game comes to an end. We are brought back to state 1, which represents 0 outs and no runners on base.

Definition 4.2. Let $X_1, X_2, ...$ be a sequence of random variables defined on a common probability space Ω. We say that S is the state space. Then, $X_1, X_2, ...$ is called a *Markov process* with state space S if $P(X_{n+1} = s_{n+1} \mid X_n = s_n, ..., X_2 = s_2, X_1 = s_1) = P(X_{n+1} = s_{n+1} \mid X_n = s_n)$ holds for any $n = 1, 2, ...$ and any $s_1, s_2, ..., s_{n+1}$ with $s_k \in \mathrm{S}\ for\ 1 \leq k \leq n + 1$. [12]

Because the probability of moving to a future state depends only on the current state of the game, we use a discrete-time Markov chain, defined by: $P(x^{n+1} = x_j | x^n = x_i, x^{n-1} = x_k, ..., x^0 = x_l) = Pr(x^{n+1} = x_j | x^n = x_i) = P_{i,j}$. We have that $p_{i,j} \geq 0$, and the sum of the probabilities for all transitions $P_{i,j} = 1$.

A transition from a state x^n to x^{n+1} will be described by our P_n matrix below, and is defined as $x^{n+1} = x^n P_n$. We are not able to quantify the probability differences when considering situational hitting, so for simplicity, I will leave it out of the Markov chains. In baseball, situational hitting typically refers to trading an out in order to advance or score a runner on the play. Examples of trading an out to move a runner would be bunting with a runner on first base to move them to second base, hitting a groundball to the right side of the field with a runner on second base to move them to third base, or hitting a flyball with a runner on third base so that the runner would be able to score a run by tagging up on the play. Situational hitting would increase probability of specific transitions due to standard offensive strategy, such as a transition from state 12 to state 17, or state 3 to state 12. These

transitions, for example, represent a sacrifice fly with one out, and a play where the runner on second advances to third, respectively.

The square, stochastic matrix that results from the transition matrix is defined as follows:

$$P_n = \begin{bmatrix} A_0 & B_0 & C_0 & D_0 \\ 0 & A_1 & B_1 & E_1 \\ 0 & 0 & A_2 & F_2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The three 8x8 A matrices will allow for the movement of baserunners, while keeping the number of outs constant (the A_0 matrix represents 0 outs, A_1 represents 1 out, and A_2 represents 2 outs). The two 8x8 B matrices have all their elements increase the number of outs by 1, and assume that baserunners will not advance on an out for simplicity. The 8x8 C matrix elements increase outs by 2, and allow for movement of baserunners. Matrices D, E, and F are 8x4 and add the number of outs needed to end an inning. The 1 matrix in the bottom right is 4x4, with the elements being absorption states and represent the states with 3 outs. Lastly, the matrices represented by zeros represent states that are not possible in accordance to the rules of baseball. [2]

We are able to create these transition probability matrices mentioned above for each pitcher through the retrosheet and Lahman data, as well as statistics that are available on FanGraphs. An interesting way to visualize how often each transition happened in the Rockies 2018 season is through a raster plot, which shows every possible transition from current to future state.

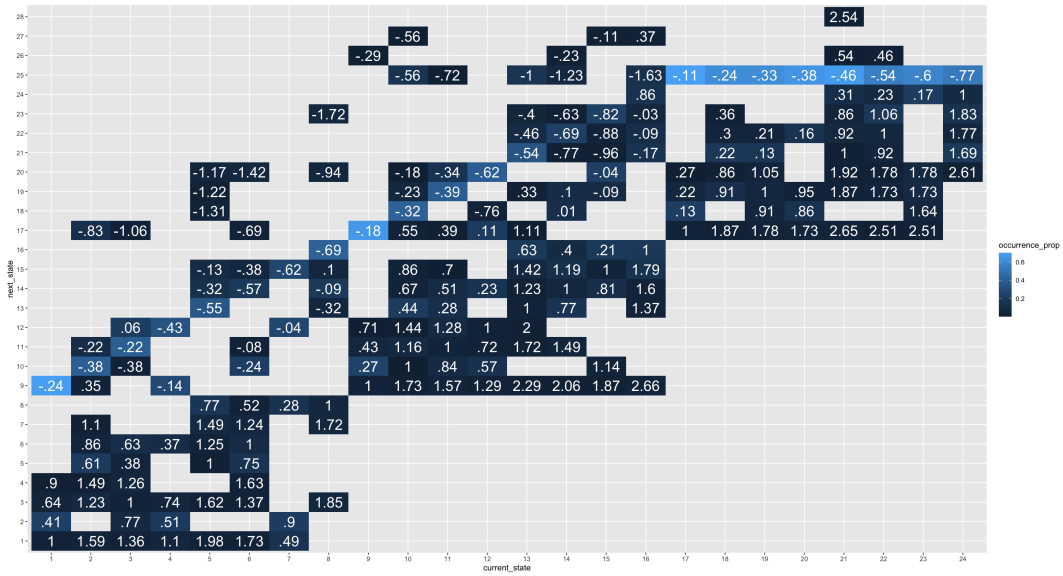

FIGURE 5. How often each transition occurred in Rockies 2018 season.

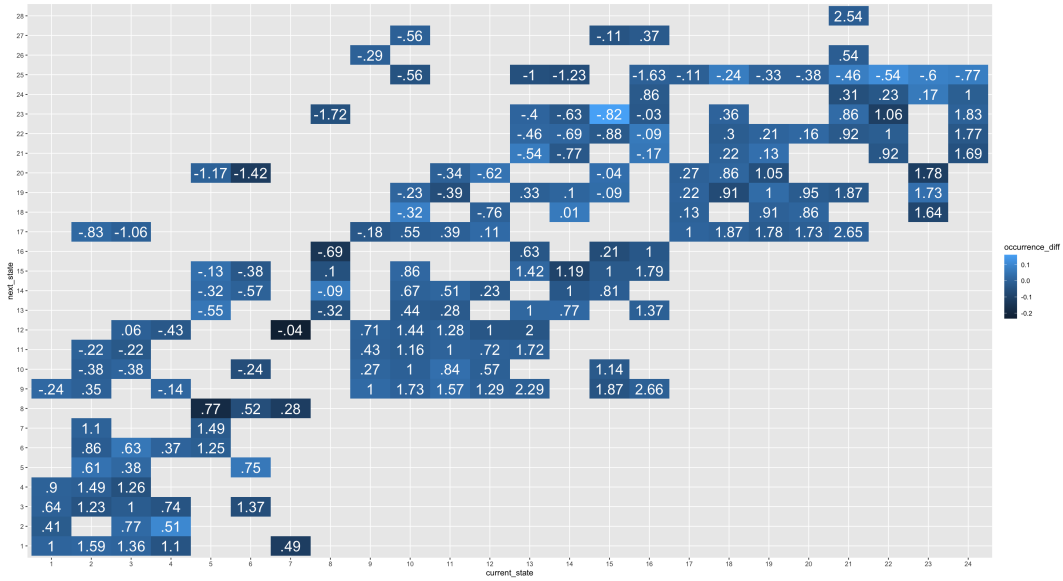

FIGURE 6. Most Common Late Game Transitions.

To understand how the Rockies won so many close games, it is helpful to bring back up the late game transitions described in the previous section. This plot shows the difference in occurrence probabilities for late game transitions versus all transitions during the Rockies 2018 season. The transition values described in each rectangle will be described in detail later on, but the more negative the value, the better the transition for the pitcher.

In order to describe the structure of the 28x28 square, stochastic matrix, we will show the structure of the several smaller matrices that help make up the full matrix given by:

$$P_n = \begin{bmatrix} A_0 & B_0 & C_0 & D_0 \\ 0 & A_1 & B_1 & E_1 \\ 0 & 0 & A_2 & F_2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

For the A matrices, A_0 keeps the outs at 0, A_1 keeps the outs at 1, and A_2 keeps the outs at 2. If we take the ordered pair (O,R), where O represents the number of outs, and R represents the runners on base, we see that the A matrices take the form:

$$\begin{bmatrix} P_{HomeRun} & P_{Single}+P_{Walk} & P_{Double} & P_{Triple} & 0 & 0 & 0 & 0 \\ P_{HomeRun} & 0 & 0 & P_{Triple} & P_{Single}+P_{Walk} & 0 & P_{Double} & 0 \\ P_{HomeRun} & P_{Single} & P_{Double} & P_{Triple} & P_{Walk} & 0 & 0 & 0 \\ P_{HomeRun} & P_{Sinlge} & P_{Double} & P_{Triple} & 0 & P_{Walk} & 0 & 0 \\ P_{HomeRun} & 0 & 0 & P_{Triple} & P_{Single} & 0 & P_{Double} & P_{Walk} \\ P_{HomeRun} & 0 & 0 & P_{Triple} & P_{Single} & 0 & P_{Double} & P_{Walk} \\ P_{HomeRun} & P_{Single} & P_{Double} & P_{Triple} & 0 & 0 & 0 & P_{Walk} \\ P_{HomeRun} & 0 & 0 & P_{Triple} & P_{Single} & 0 & P_{Double} & P_{Walk} \end{bmatrix}$$

P_{Walk}, P_{Single}, etc. represent the probability of walking, hitting a single, etc. and is calculated by $\frac{Walks}{AtBats}$, $\frac{Singles}{AtBats}$, etc. The entries in the A matrices are filled by the ways that one state is able to transition to a different state without adding an out. For example, the first column is only home runs because that is the only transition that results in the bases being cleared with no outs added, and the fourth column is only triples because that is the only way to end up with a single runner on third with no outs resulting on a play. For simplicity, we assume that a runner on second or third base will score on a single, a runner on first base will end up at third base on a double, and a runner on first base will end up at second base on a single.

For the B matrices, B_0 represents transitions from no outs to one out, and B_1 represents transitions from one out to two outs. The 8x8 B matrices are rather simple once you are able to calculate P_{Out} for each current state. The

matrices take the form of a P_{Out} vector of length 8 multiplied by the identity matrix of size 8. This gives us P_{Out} along the diagonal and zeros for the other entries.

The C matrix increases the outs from 0 to 2, meaning that it includes the double-play transitions. In the most basic form, it looks like:

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ P_{DoublePlay} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ P_{DoublePlay} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ P_{DoublePlay} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

We say, for simplicity, that the probability of a double play with states 2,3, or 4 will be the same among them all, and the p's are estimated based on seasonal averages for the pitchers. There are several entries in the C matrix that can be populated, such as the points at (5,4), (8,7), or (6,1), but we are not going to assume the end base location of runners that were on base at the start of the play and did not end up being tagged/forced out on the play. This variance with the baserunning will create far too many possible transitions to handle.

D matrices represent the transitions that result from a triple play. They are 8x4 matrices, similar to the E and F matrices, which represent double plays and normal outs to the end inning, respectively. If there are two runner on base when a triple play occurs, we know that no runs will score, but if the bases are loaded, it is a possibility that the runner on third could score before the three outs are recorded, so we can fill in those entries accordingly. E matrices are similar, and we know they will have 0 entries in row 1 since there needs to be a runner on base for a double play to occur. For the F matrices, we have that any out will result in a transition to an absorption state, and the number of runs scored on the play can range from 0 to 3. The 1 matrix in the bottom right designates the absorption states, and we know that we cannot have a transition with a current state of 3 outs.

5. Using Model to Test Hypothesis

The Markov model will allow for me to use transition states as control variables in my model that will predict win outcome based on substitution strategy. This will allow for me to remove confounding factors such as a substitution during a lopsided game, a high leverage situation, etc.

In the bigger picture, I will want to test how substitution strategy affects a team's probability of winning the game. I will be able to do this using a logistic regression model and controlling for confounding variables such as the score of the game and the inning in which a substitution is made. Furthermore, I will be able to use regression in order to determine just how good of a transition is likely to come after each substitution strategy is used. This will come from a process of using transition values that I will describe later on, and then seeing how high or low the resulting estimates will be for the associated regression.

Because I have been able to gather data for the whole 2018 Rockies season, I will compare the Markov Chain models for each of the substitution strategies I am investigating. The specific models will tell me how much the strategy is able to predict whether or not the Rockies won the game in which the strategy was used, as well as the likelihood that the strategy leads to a positive transition for the pitcher. After seeing the predictions, I will be able to compare and determine how the Rockies could have better optimized their bullpen usage in 2018.

To validate my findings and to make some predictions for how much each strategy contributes to winning and the frequency of certain transitions, I run logistic regressions in R.

6. Substitution Strategies

It can be argued that there are far more substitution strategies than the ones that I am diving into in this thesis, but I believe that these are the 6 most strategies that managers would want to take into account before making a substitution. A pitcher's xwOBA, xFIP, and Soft Contact % do a great job at eliminating confounding variables that come with defensive positioning and performance, while BAA and ERA are the two most common stats for evaluating pitcher performance. I am saying that each of these metric correspond to a substitution strategy. For example, if the manager puts in the pitcher with the best xFIP or Soft Contact %, that move would correspond to that specific strategy. The strategy that is a bit unlike the others is the handedness orientation of the pitcher (and hitter).

pitcher	handedness	pitcher_type	xwoba	xfip	soft_contact	baa	era
Adam Ottavino	R	Reliever	0.23	3.13	20.1	0.158	2.43
Wade Davis	R	Reliever	0.247	3.63	14.8	0.185	4.13
Seunghwan Oh	R	Reliever	0.262	4.05	11.3	0.209	2.53
German Marquez	R	Starter	0.282	3.1	17.5	0.241	3.77
Scott Oberg	R	Reliever	0.291	2.83	16	0.213	2.45
Tyler Anderson	L	Starter	0.296	4.21	20.9	0.248	4.55
Kyle Freeland	L	Starter	0.3	4.22	20	0.24	2.85
Jon Gray	R	Starter	0.307	3.47	16	0.266	5.12
Chris Rusin	L	Reliever	0.309	4.25	21.1	0.268	6.09
Antonio Senzatela	R	Starter	0.319	4.43	20.1	0.266	4.38
Jake McGee	L	Reliever	0.336	4.41	11.1	0.285	6.49
Chad Bettis	R	Reliever	0.343	4.76	20.5	0.265	5.01
Bryan Shaw	R	Reliever	0.352	4.35	14.9	0.313	5.93

FIGURE 7. Rockies 2018 Pitching Stats.

In 2018, the league average xwOBA was .313, the league average xFIP was 4.15, Soft Contact % was 18%, BAA was .248, and ERA was 4.15. We see that just about every Rockies relief pitcher that was used consistently in non-lopsided games had these metrics better than the league average benchmark. As far as Soft Contact %, the best Rockie pitchers according to the other statistics did not all have stellar quality of contact. For example, Scott had a very low Soft Contact %, but performed well in all the other statistical categories, while a pitcher such as Chris Rusin had a high Soft Contact %, but had poor stats otherwise, causing the Rockies to not insert him into many high-leverage situations. Some of these exceptions may be explained by the high strikeout rate pitchers that the Rockies had in 2018. This means that a pitcher will earn himself lots of strikeouts, but is prone to giving up many hard-hit balls when the batter is able to put the ball in play.

In baseball, it is almost always favorable for the defensive team to have a left-handed pitcher to face a left-handed hitter and a right-handed pitcher face a right-handed hitter. This is because most off-speed pitches that break horizontally will curve away from the hitter during the trajectory towards home plate in matchups where the hitter and pitcher are both left or right-handed. Thus, Major League Baseball teams will typically construct their batting lineup at the start of each game in a way that includes their best performing players and several opposite-handed matchups. Being able to have a staff with a mix of right-handed pitchers and left-handed pitchers can be exceptionally beneficial for a team when in close games. When investigating

batting splits against right and left-handed pitchers, nearly all hitters in the MLB perform better against pitchers of the opposite handedness orientation. Thus, managers will want to utilize their pitchers that have the best same-handedness statistical splits in key moments when they need to get one or two specific outs. In turn, simply maximizing same-handedness matchups with pitchers that are not the most reliable pitchers on the staff will be unlikely to result in more wins than a typical substitution strategy.

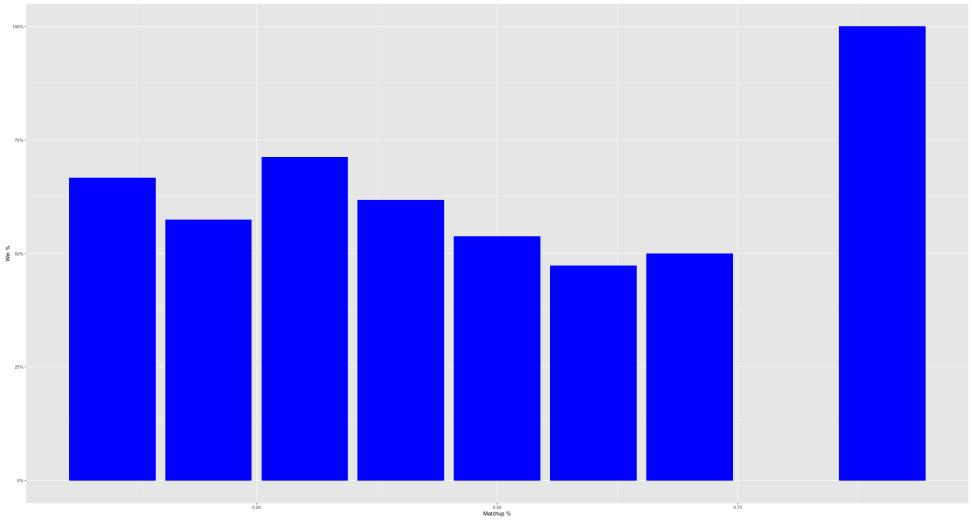

FIGURE 8. Wins and Losses by Same-Handedness Matchup %.

We see that, for the Rockies 2018 games, they actually ended up winning more when they had fewer same-handedness matchups in the game. Thus, we can conclude that while a handedness-based substitution may help for getting a specific hitter out, it should not be the only strategy used throughout a ballgame.

Expected Weighted On-Base Average (xwOBA) essentially measures how often a batter facing the specific pitcher is expected to get on base given the launch angle and velocity averages against the pitcher that year. It is a metric that has gained popularity in the past few years since it can calculate "the average run value of every batted ball for a hitter (or allowed by a pitcher), and adds in defense-independent numbers."[9] It is one of the most comprehensive stats since it eliminates variables outside of the pitcher's control that could affect the stat, such as defensive alignment or performance (which xFIP also does, but presents the stats in a different way). Once the

ball has left the hitter's bat, it captures the probability of a batter being able to reach base, given how hard the ball was hit and its direction.

Another interesting way to look at xwOBA is by considering only the balls that are put into play. This statistic is called $xwOBAcon$, meaning the expected weighted on-base average of balls where contact is made, and is a mix between xwOBA and soft contact %. The MLB Technology Blog notes that $xwOBA_{con}$ is based on the exit velocity of the baseball off the bat, the launch angle of the hit off the bat, and the sprinting speed of the hitter. A few years ago, the blog published the graph shown below, displaying the relationship between these variables.

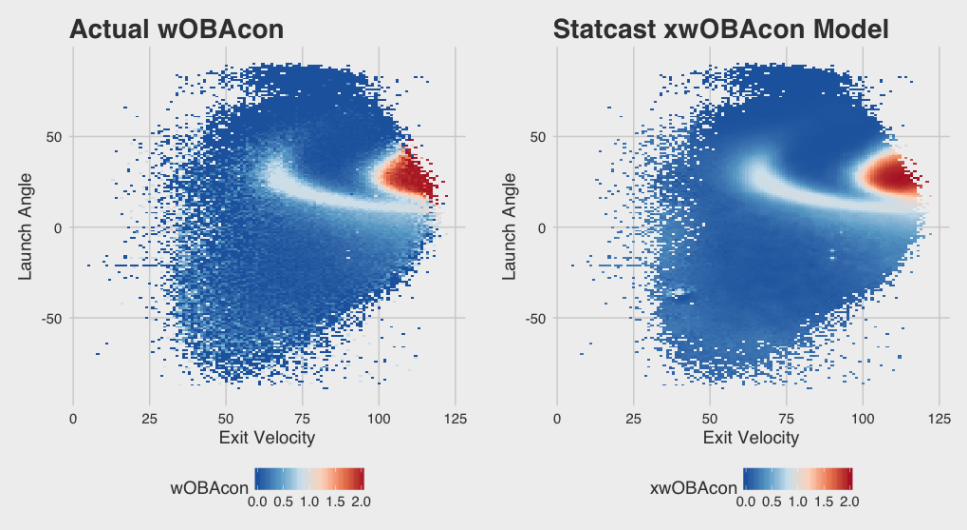

FIGURE 9. MLB Technology Blog's $xwOBA_{con}$ visualization. [7]

The similar stat, Expected Fielding Independent Pitching (xFIP), is a stat that has been more historically used to measure a pitcher's performance in comparison to xwOBA. It is a great predictor for a pitcher's ERA and seeks to "isolate a pitcher's contribution to run prevention."[9] It essentially captures the projected home run rate for each pitcher. It differs from Fielding Independent Pitching (FIP) in that it takes into account each pitcher's HR/FB rate, which is the number of home runs that they give up divided by the total amount of fly balls given up, rather than total number of home runs allowed. Due to some ballparks being more hitter or pitcher friendly, some fly balls that would result in outs in stadiums such as Comerica Park in Detroit may be home runs in a stadium such as Coors Field in Denver. Thus, xFIP does

a better job than FIP at capturing how skilled a pitcher is at keeping the ball in the ballpark.

The stat is calculated by taking the FIP constant and adding the number of flyballs multiplied by the league home run to flyball rate times a constant of 13, adding 3 times the number of walks and hit-by-pitches, subtracting 2 times the number of strikeouts, and then dividing it all by the total number of innings pitched. The FIP constant is calculated by taking the league average ERA, subtracting 13 times the league home run rate, adding 3 times the league walk and hit-by-pitch rate, and then subtracting the league strikeout rate over the league average number of innings pitched. The constants 13, 3, and 2 are used to bring FIP and xFIP to around what ERA would be. The exact calculations for why these are chosen are not publicly available, but they do change each year based on the other league average rates in the equations.

The ratio of League Average Home Run rate divided by Flyball% for 2018 was 12.7%. This means that for every fly ball hit during the 2018 season, 12.7% of them wound up as homers. The constants 13, 3, and 2 in the xFIP formula are derived from another algorithm that is not publicly available, but it allows for xFIP to capture a value similar to ERA when adding in the FIP constant by placing various weights on HR/FB% and walk or hit-by-pitch to strikeout rates. BB stands for total walks, HBP stands for total Hit by Pitches, K stands for strikeouts, and IP stands for innings pitched. In the formula to find the FIP constant, LeagueERA is the league average ERA, LgHR is league Home Run rate, LeagueBB is league walk rate, and so on. The constant is a tool to bring the league average FIP up or down to match the league average ERA. [9]

Similar to the xFIP calculation, the FIP constant used the 13, 3, and 2 constants to transform the constant into a number that is more usable and similar to ERA. The FIP constant changes every year, as the league Earned Run Average of course changes from year to year. For the 2018 season, the FIP constant was 3.134 across the MLB, and an above average xFIP would be 3.5 or less. Historically, the best pitchers of all time had xFIPs between 1.25 and 2.

Soft Contact % is an interesting way to measure how effective a pitcher is at minimizing hard contact. It only takes into account batted balls, so it will not measure things such as strikeouts or walks. A weakly hit ball will not always result in a positive transition for the pitcher, but it is certainly much more likely. Similarly, a ball that is hit very hard will not always end up being a bad transition for the pitcher, but it is far more likely. The exact algorithm for what is to be considered soft, medium, and hard contact is not

publicly available, but it would certainly be heavily influenced by the exit velocity off the bat. Baseball Info Solutions (BIS) is the one to capture the data and uses hang time, location of the batted ball, and trajectory in order to determine the quality of contact.

The most common stat used to evaluate pitcher performance is Earned Run Average (ERA), which measures the number of earned runs that a pitcher will allow on average per nine innings. It is the most basic pitching stat, which is found by taking $9 * \frac{EarnedRuns}{InningsPitched}$.

Another basic stat that is very common in evaluating pitcher performance is Batting Average Against (BAA). This simply measures the batting average that hitters have against a specific pitcher. A batting average is calculated by taking $\frac{Hits}{AtBats}$.

By looking at the raster plots for these 5 stats, we can get a preliminary sense of how often each transition happens using each of the substitution strategies. It is similar to the general raster plot shown in the Markov Chain Model section, but only considers the transitions made after a substitution using each respective strategy. Similar to the late game transitions raster plot, I am comparing each of the substitution strategies to the full 2018 Rockies season to see the transition states in which the pitchers excelled for each strategy. Thus, if the rectangle is a light shade of blue, the transition happened more often with that strategy, and if the rectangle is darker colored, then that transition happened less often. The transition values are listed for each transition that happened for each sub strategy.

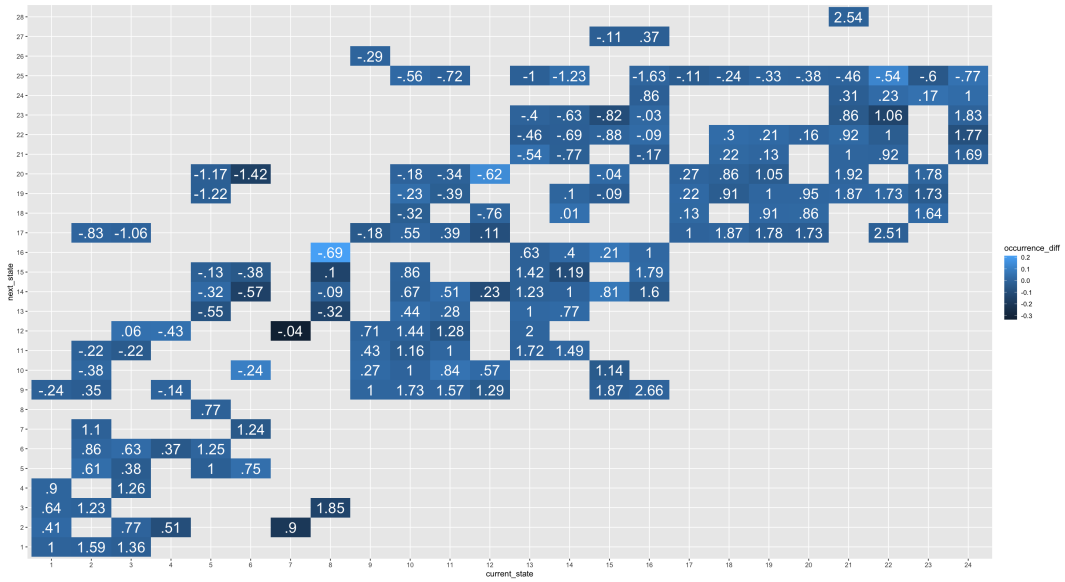

FIGURE 10. Most Common Transitions for the 2018 Rockies Pitchers with the 6 Best xwOBA's.

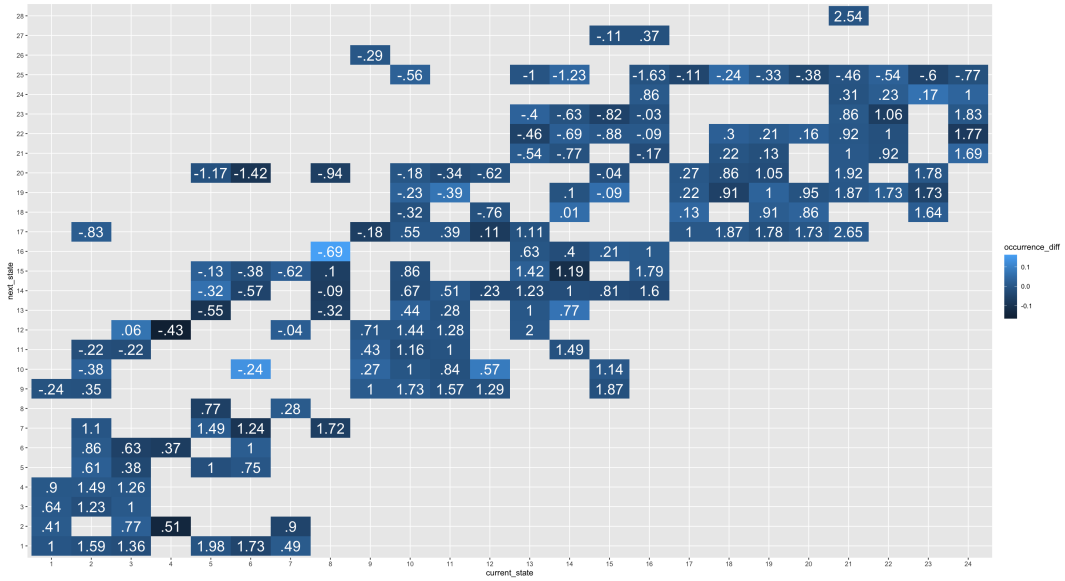

FIGURE 11. Most Common Transitions for the 2018 Rockies Pitchers with the 6 Best xFIP's.

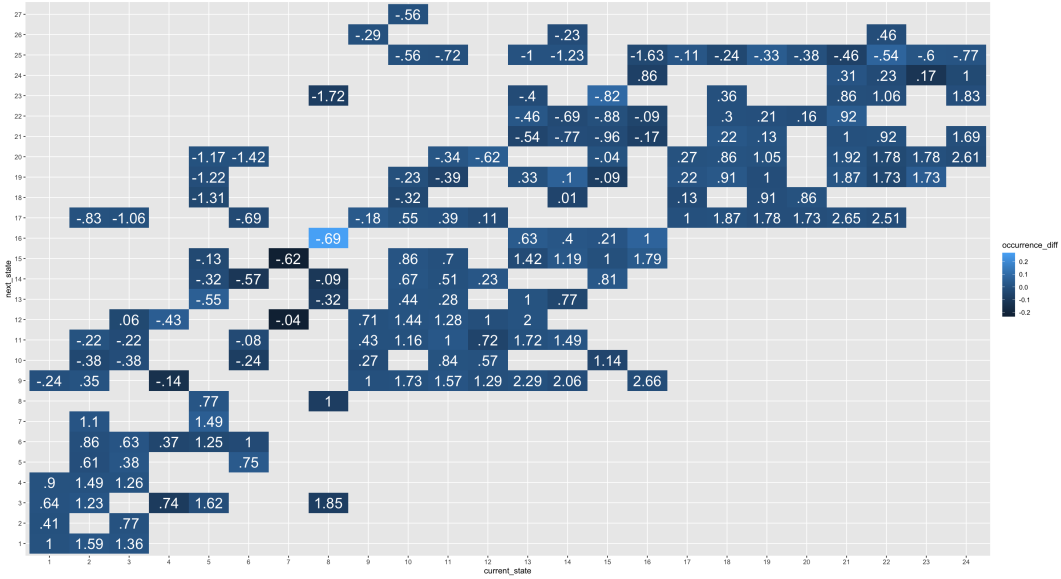

FIGURE 12. Most Common Transitions for the 2018 Rockies Pitchers with the 6 Best Soft Contact Percentages.

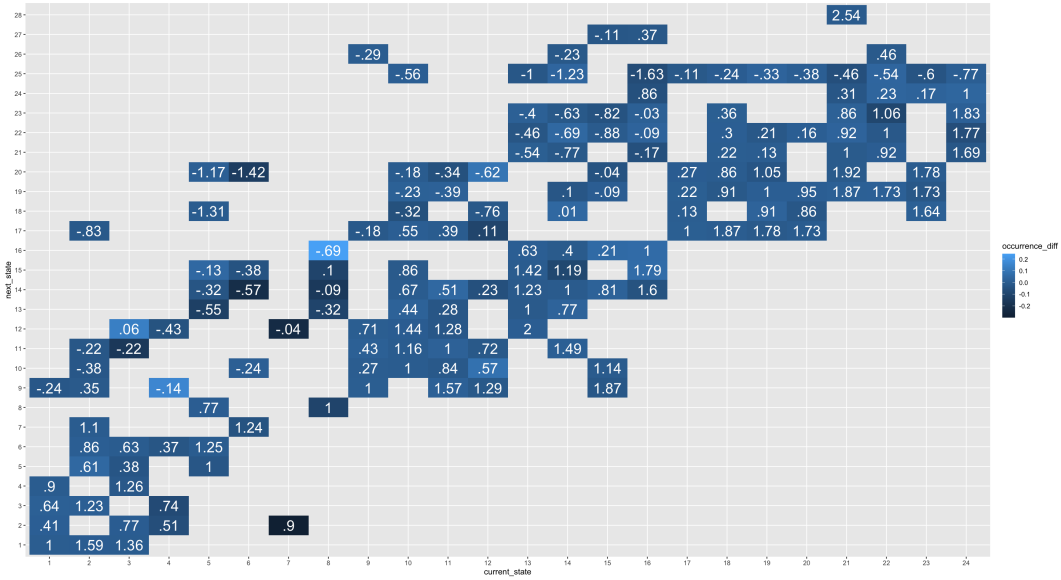

FIGURE 13. Most Common Transitions for the 2018 Rockies Pitchers with the 6 Best Batting Averages Against.

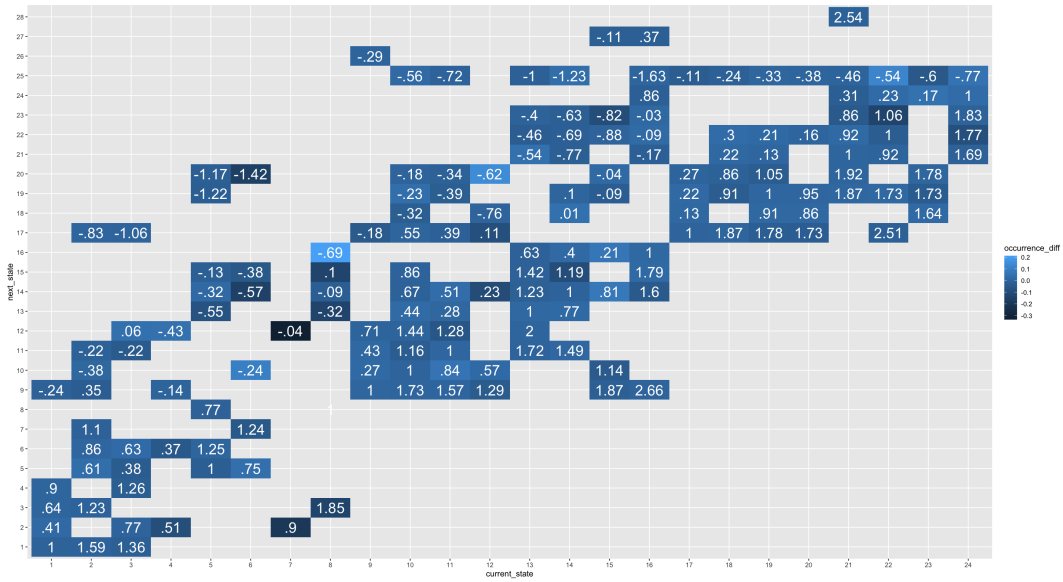

FIGURE 14. Most Common Transitions for the 2018 Rockies Pitchers with the 6 Best ERA's.

7. Assigning Values to Transitions

We have determined which transitions are "good" and which transitions are "bad", but to better understand just how good or bad a pitcher performed after entering a game, I use a run expectancy matrix to create a numerical metric for measuring each transition value. The run expectancy matrix for Major League Baseball this season is given by:

$$\begin{array}{c} \\ \text{0 outs} \\ \text{1 out} \\ \text{2 outs} \end{array} \begin{array}{c} \begin{array}{cccccccc} 000 & 100 & 020 & 003 & 120 & 103 & 023 & 123 \end{array} \\ \begin{pmatrix} 0.53 & 0.94 & 1.17 & 1.43 & 1.55 & 1.80 & 2.04 & 2.32 \\ 0.29 & 0.56 & 0.72 & 1.00 & 1.00 & 1.23 & 1.42 & 1.63 \\ 0.11 & 0.24 & 0.33 & 0.38 & 0.46 & 0.54 & 0.60 & 0.77 \end{pmatrix} \end{array} \quad [10]$$

Each entry in the run expectancy matrix corresponds to the expected number of runs that will be scored in the rest of the inning for the situation. For example, if there is one out and no runners on base, that corresponds to the leftmost entry in the middle row, which is 0.29. Thus, there was on average 0.29 runs scored in an inning from this situation during the 2018 MLB season. These entries represent the same transition states for our Markov Chain model.

The value that I assign each transition comes from the number of runs scored on the play plus the difference in run expectancy values from the current state to the next state. For example, if the bases are loaded with 0 outs (state 8) and the play results in an out with no runs scored, leaving the bases loaded with 1 out (state 16), the value for that play would be the run expectancy for state 16 minus the run expectancy for state 8 ($1.63 - 2.32 = -0.69$). An example that includes a run scoring on the play could be transitioning from state 12, which is a runner on third base with one out, to state 17, which is no runners on base and two outs. A run must score on the play, which may happen through a sacrifice fly or an RBI groundout, but the run expectancy for the rest of the inning will decrease. Thus, our calculation for the value of this transition will be $1+0.11-1.00 = 0.11$. A transition value of 0 represents a play that is neither advantageous nor disadvantageous for the defense, a negative transition value helps the defense, and a positive transition value hurts the defense.

8. Modeling and Regression

Because my goal is to predict how much each substitution strategy contributes to a win or loss, as well as which strategy leads to the most positive transitions, doing a few regressions is a perfect method to analyze the data. I am able to take the binary outcomes of win and loss, as well as consider the transition value to see which strategy leads to the most beneficial transitions. I am essentially asking the two questions "How much did each substitution strategy impact the probability of having a positive transition on the following play?" and "How much did each substitution strategy impact the probability of the Rockies winning a game?"

In order to model which substitution strategy contributes the most to wins, I can simply use the binary outcomes of win versus loss in a logistic regression. We have defined the states in the Markov chain and of course have the independent and dependent variables of substitution strategy and win/loss, respectively. I will want to control for the current score of the game, the current state that is inherited by a new pitcher, and the inning at the point of substitution. This would prevent cases where a pitcher comes in during a lopsided game and their performance does not correlate to the final result of the game, or a substitution very early in the game which may not be very indicative of the final result of the game. Also, I will include the substitutions where the relief pitcher has better stats than the previous pitcher (considering the substitution strategy metric correlated with that relief pitcher). For example, I will want to investigate a relief pitcher substitution for a player with a better ERA than the starting pitcher and how that affects the following transitions and overall contribution to win/loss.

Logistic regression assumes that each outcome will be random, and follows a Bernoulli distribution. If Y is the outcome of an occurrence, we have $Y \sim Bern(p)$ where p is the expected proportion of "success". p will be modeled as a linear function of the covariates. We will have that the outcome variable of win or loss as a Bernoulli random variable. Logistic regression uses log odds, which convert the model to a likelihood based model. Essentially, this regression is a linear model for predicting log odds.

Logistic regression uses the logistic function, which is an S-shaped curve that does the job of turning the linear combination of the independent variables into a value between 0 and 1 representing probability. The logistic function can be written in the form:

$$p(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

If we have log odds and want to go to probabilities, we will use the logistic function stated above, but if we have probabilities and want to go to log odds, we will use the logit function. The logit of the probability is predicted by a linear model [14]. Given a p-value, our logit function is given by:

$$\text{logit}(p) = \ln(\frac{p}{1-p})$$

We know that from the earlier section on transition states that every $p_{i,j} \geq 0$, meaning that the probability of each transition cannot be less than 0, and of course, will not be greater than 1.

The first regression to look at is how each substitution strategy contributes to a positive transition. I will use a linear regression and my transition value metric that I calculated from the run expectancy matrix. I will not need to control for the score or for the current state at the time of the substitution since I am looking at individual transitions, predicting the value based on each sub strategy. A linear regression line takes the form $Y = a + bX$ where X is the independent variable (sub strategy) and Y is the dependent variable (transition value).

Strategy	Estimate	Std. Error	t value	Pr(>|t|)
xWOBA	-0.0600	0.0866	-0.693	0.488
xFIP	0.0936	0.0917	1.021	0.308
Soft Contact %	0.0549	0.0876	0.626	0.531
BAA	0.0511	0.0903	1.007	0.497
ERA	-0.0471	0.1199	-0.393	0.695
Handedness	0.0630	0.0763	0.825	0.410

The results of this transition value regression are very interesting, although not as statistically significant as the results for the win vs. loss regression, which we will see coming up. The regression considers the first batter faced after a substitution was made, rather than all the hitters that a pitcher faces after being subbed in, which leads to a sample size that may not be

remarkably large when considering just one season. However, we see that the estimates indicate that substitutions based on ERA and xwOBA are the two strategies that lead to the most success for the Rockies in 2018, with estimate values of -0.0471 and -0.0600, respectively. We see that ERA had the highest standard error at 0.1199, and handedness had the lowest standard error at 0.0763. Surprisingly, xFIP was the strategy that led to the worst transitions for the Rockies in 2018, despite xwOBA (a similar stat) being the best strategy. This may come from the fact that the Rockies play in an extremely hitter-friendly stadium, where pitchers with good home run to flyball rates will still give up a decent amount of home runs. We are able to see the variance in the data with the plot below, as well as roughly how many observations we have for each strategy.

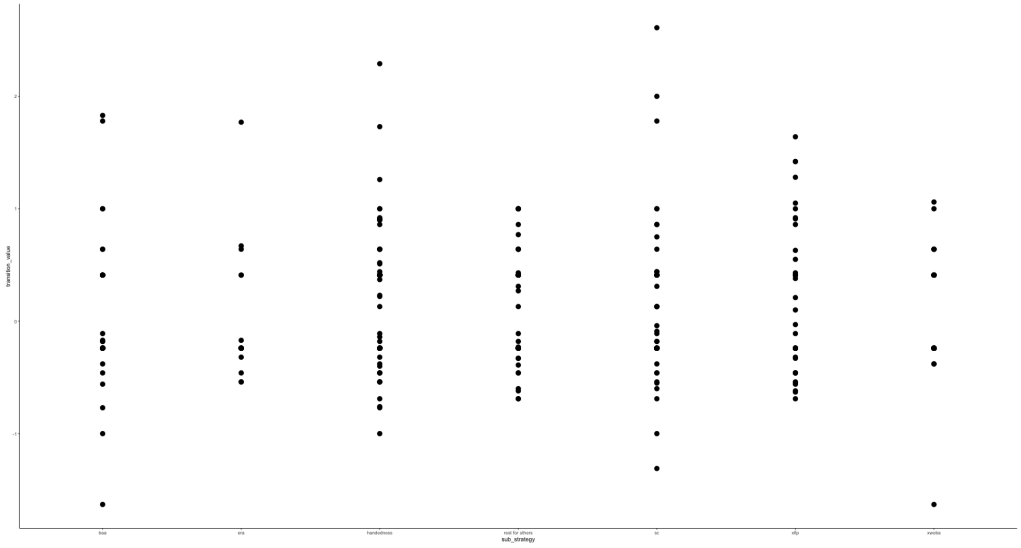

FIGURE 15. Linear Regression.

The next regression is a binary logistic regression where I want to see how each substitution strategy changes the probability of winning a game. This time, I will want to control for the score of the game when the substitution is made, the inherited current state, and the inning at which the sub is made. The results will not be nearly as significant if we don't control for the situations when a pitcher comes in during a lopsided game, where their individual performance will not affect the game's outcome, no matter how good or bad it is.

Strategy	Estimate	Std. Error	z value	Pr(>|z|)
xWOBA	0.9624	0.4484	2.146	0.0319
xFIP	-0.8352	0.3852	-2.168	0.0302
Soft Contact %	-1.0408	0.3757	-2.770	0.0056
BAA	-0.8551	0.4017	-2.172	0.0462
ERA	-0.7795	0.4893	-1.593	0.1111
Handedness	-0.6740	0.3258	-2.069	0.0386

Because this regression controlled for the score of the game, the state that the relief pitcher inherited, and the inning of the game, we get much more significant results and lower standard error values. We have a much greater sample size for this regression, and are able to see great results from this glm model. We see that xwOBA, handedness and ERA are the best strategies for the probability of winning a game, but with xwOBA as the clear best option. There is a .9624 log-odds increase in probability of winning the game when this strategy is used. After rest for others, Soft Contact % had the worst results for the Rockies in 2018 with log odds of -1.0408 for probability of winning a game.

Handedness once again had the lowest standard error at 0.3285, and ERA had the highest standard error at 0.4893. All the results are statistically significant except for ERA, but at 0.1111, it is certainly still important to consider those results.

9. Conclusion

We see that the two most successful strategies for the Rockies in 2018 were putting in the pitchers with the best xwOBA and the best ERA. It is significant that a new advanced, sabermetric stat in xwOBA did the best for the Rockies, but we also have that the most basic pitching stat, ERA, was a great measure of pitcher performance. Many baseball analysts have drifted away from the most basic stats such as ERA and BAA due to confounding variables such as defensive positioning and ability, so I find it extremely interesting that the regressions support the power of ERA.

I am interested in comparing these specific results to other Rockies' seasons or to other teams in 2018. Pitching strategy in Denver has always been a big debate due to the altitude causing Coors Field to be an extremely hitter-friendly park. I am surprised to see that Soft Contact % was the worst substitution strategy for the Rockies, since I imagined that would be a priority when playing at Coors Field. The results would likely be different when considering a different team, but for the Rockies I think this information could be very valuable when looking towards the future.

10. References

[1] Nobuyoshi Hirotsu. *Modelling a Baseball Game to Optimise Pitcher Substitution Strategies Incorporating Handedness of Players.* Economics, Management and Optimization in Sports. 2004.

[2] Max Marchi and Jim Albert. *Analyzing Baseball Data with R.* Chapman and Hall. 2018.

[3] Tom M. Tango, Mitchel G. Lichtman, and Andrew E. Dolphin. *The Book: Playing the Percentages in Baseball.* CreateSpace Independent Publishing Platform. 2014

[4] Momin Mehmood Butt. *Pythagorean Expectation in Sports Analytics, with Examples From Different Sports.* Towards Data Science. 2022.

[5] Lucas Calestini. *The Elegance of Markov Chains in Baseball.* Sports Analytics. 2018.

[6] Sam Sharpe. *An Introduction to Expected Weighted On-Base Average (xwOBA).* MLB Technology Blog. 2019.

[7] Pete Palmer. *Relief Pitching Strategy: Past, Present, and Future?* Society for American Baseball Research. Baseball Research Journal. 2018.

[8] Daniel Joseph Ursin. *A Markov Model for Baseball with Applications.* University of Wisconsin-Milwaukee. 2014.

[9] Craig Edwards. *FIP vs. xwOBA for Assessing Pitcher Performance.* FanGraphs. 2018.

[10] Jim Albert. *Summarizing a Runs Expectancy Matrix.* Exploring Baseball Data with R. 2020.

[11] Bruin, J. *Logit Regression.* UCLA Statistical Methods and Data Analytics. 2011.

[12] Sheldon Ross. *Stochastic Processes.* University of California, Berkeley. 1996.

[13] David Hosmer, Stanley Lemeshow, Rodney Sturdivant. *Applied Logistic Regression, Third Edition.* Hoboken: John Wiley Sons, Inc. 2013.

[14] Stephanie M. van den Berg. *Analysing Data using Linear Models.* 2022.