Using Markov Chain Models to Find Optimal Pitching Substitution Strategies

Zach Becker
Haverford College

Friday 28th April, 2023

Topic Overview

The goal on the defensive side of a baseball game is to minimize runs allowed to the other offense. Managers have to make decisions throughout the game in order to ensure that they are maximizing the number of pitcher vs. hitter matchups that are advantageous for the pitcher. I am seeking to find out the best way to maximize these advantageous matchups by looking at all the plays in the Colorado Rockies 2018 season where the Rockies were on defense.

Data Collection

The process of putting together a clean dataframe in R where each row represents a play in the Rockies season was quite tedious. I downloaded a file that consisted of all the plays from the full MLB season in 2018, but needed to install a github package so that I could have access to the full dataset in R. I also loaded in the Sean Lahman database and was able to merge the data to create a single dataframe that also contained necessary biographical information about the players.

From there, I did lots of work with the dplyr package in R to get my dataframe cleaned up. I added a handful of columns using the mutate function and several case statements, and ended up with my full dataframe that consists of the 6,205 defensive plays from the Rockies 2018 season.

Data Collection

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	grayj003	NA	NA	NA	8/L	1	9	N/A	N/A	-0.24	0
ANA201808270	1	1	R	R	grayj003	NA	NA	NA	8/L	9	17	No	N/A	-0.18	0
ANA201808270	1	2	R	R	grayj003	NA	NA	NA	9/F	17	25	No	N/A	-0.11	0
ANA201808270	2	0	L	R	grayj003	NA	NA	NA	7/L	1	9	No	N/A	-0.24	0
ANA201808270	2	1	R	R	grayj003	NA	NA	NA	8/L	9	17	No	N/A	-0.18	0
ANA201808270	2	2	R	R	grayj003	NA	NA	NA	53/G	17	25	No	N/A	-0.11	0
ANA201808270	3	0	L	R	grayj003	NA	NA	NA	8/F	1	9	No	N/A	-0.24	0
ANA201808270	3	1	R	R	grayj003	NA	NA	NA	K	9	17	No	N/A	-0.18	0
ANA201808270	3	2	L	R	grayj003	NA	NA	NA	7/F	17	25	No	N/A	-0.11	0
ANA201808270	4	0	L	R	grayj003	NA	NA	NA	S9/L	1	2	No	N/A	0.41	0
ANA201808270	4	0	R	R	grayj003	calhk001	NA	NA	SB2	2	3	No	N/A	1.23	0
ANA201808270	4	0	R	R	grayj003	NA	calhk001	NA	S3/BG.2-3	3	6	No	N/A	0.63	0
ANA201808270	4	0	R	R	grayj003	fletd002	NA	calhk001	S7/L+.3-H;1-2	6	5	No	N/A	0.75	0
ANA201808270	4	0	L	R	grayj003	troum001	fletd002	NA	BK.2-3;1-2	5	7	No	N/A	1.49	0
ANA201808270	4	0	L	R	grayj003	NA	troum001	fletd002	HR/8/F.3-H;2-H	7	1	No	N/A	0.49	0
ANA201808270	4	0	R	R	grayj003	NA	NA	NA	7/F	1	9	No	N/A	-0.24	0
ANA201808270	4	1	R	R	grayj003	NA	NA	NA	5/P7LF	9	17	No	N/A	-0.18	0
ANA201808270	4	2	L	R	grayj003	NA	NA	NA	63/G	17	25	No	N/A	-0.11	0
ANA201808270	5	0	R	R	grayj003	NA	NA	NA	43/G	1	9	No	N/A	-0.24	0
ANA201808270	5	1	L	R	grayj003	NA	NA	NA	63/G	9	17	No	N/A	-0.18	0
ANA201808270	5	2	L	R	grayj003	NA	NA	NA	D9/L	17	19	No	N/A	0.22	0
ANA201808270	5	2	R	R	grayj003	NA	calhk001	NA	43/G	19	25	No	N/A	-0.33	0
ANA201808270	6	0	R	R	grayj003	NA	NA	NA	HR/7/L	1	1	No	N/A	1.00	0
ANA201808270	6	0	L	R	grayj003	NA	NA	NA	13/G	1	9	No	N/A	-0.24	0
ANA201808270	6	1	R	R	grayj003	NA	NA	NA	53/G	9	17	No	N/A	-0.18	0
ANA201808270	6	2	R	R	grayj003	NA	NA	NA	53/G	17	25	No	N/A	-0.11	0
ANA201808270	7	0	L	R	grayj003	NA	NA	NA	31/G	1	9	No	N/A	-0.24	0
ANA201808270	7	1	R	R	grayj003	NA	NA	NA	53/G	9	17	No	N/A	-0.18	0
ANA201808270	7	2	L	R	grayj003	NA	NA	NA	S8/L	17	18	No	N/A	0.13	0
ANA201808270	7	2	L	L	mcgej001	youne003	NA	NA	S9/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	S8/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: Sample of Rockies 2018 Dataframe.

Transition States and Transition Matrices

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.

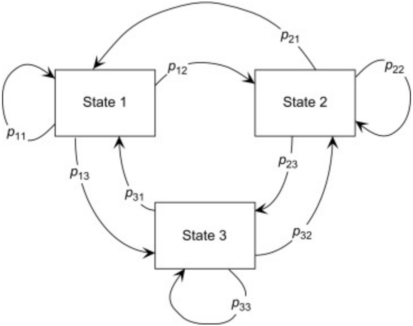

Figure: Markov Chain Overview.

Transition States and Transition Matrices

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.

On the offensive side, there exist 24 states that are made up of combinations of runners occupying various bases and number of outs. We refer to the states that do not result in 3 outs and thus the end of the inning as non-absorption states.

(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)

Transition States and Transition Matrices

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:

$$Pr(x^{n+1} = x_j | x^n = x_i, x^{n-1} = x_k, ..., x^0 = x_l)$$
$$= \Pr(\mathrm{x}^{n+1} = x_j | x^n = x_i) = P_{i,j}, \text{ and } \sum P_{i,j} = 1.$$

The square, stochastic matrix that results from the transition matrix is given by:

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

Transition States and Transition Matrices

A matrices:

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

Transition States and Transition Matrices

For the B matrices,they 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 outs from 0 to 2, meaning that it includes the double-play transitions. They take the form:

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

Transition States and Transition Matrices

D, E, and F matrices are 8x4. D matrices represent transitions resulting from a triple play, E matrices represent double plays to end the inning, and F matrices represent normal outs to end the inning. The 1 matrix in the bottom right designates the absorption states, and we cannot have a transition with a current state of 3 outs.

Substitution Strategies

I am considering 6 different metrics for my substitution strategies: xwOBA, xFIP, Soft Contact %, BAA, ERA, and handedness. These metrics are a mix of some very basic and common pitching performance stats such as Earned Run Average and Batting Average Against, and some much more advanced stats, such as Expected Weighted On-Base Average, and Expected Fielding Independent Pitching.

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: Rockies 2018 Pitching Stats.

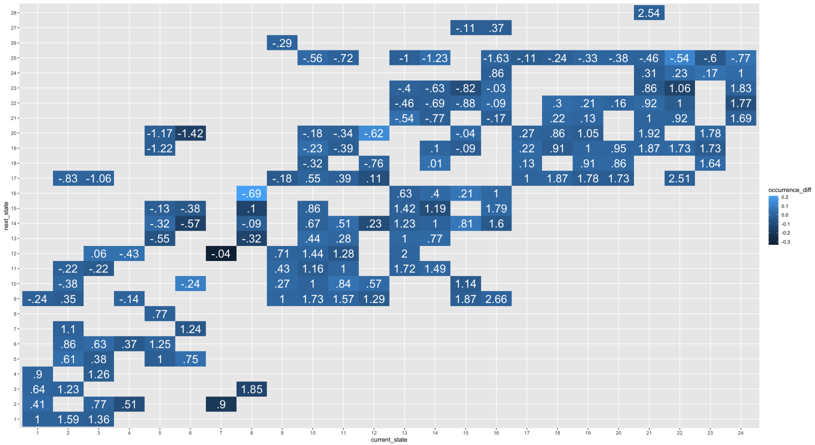

Figure: Most Common Transitions for the 2018 Rockies Pitchers with the 6 Best Expected Weighted On Base Averages.

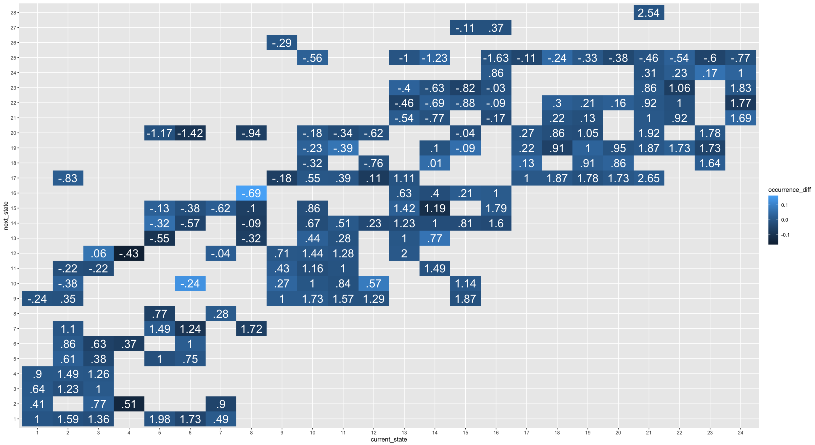

Figure: Most Common Transitions for the 2018 Rockies Pitchers with the 6 Best Expected Fielding Independent Pitching Stats.

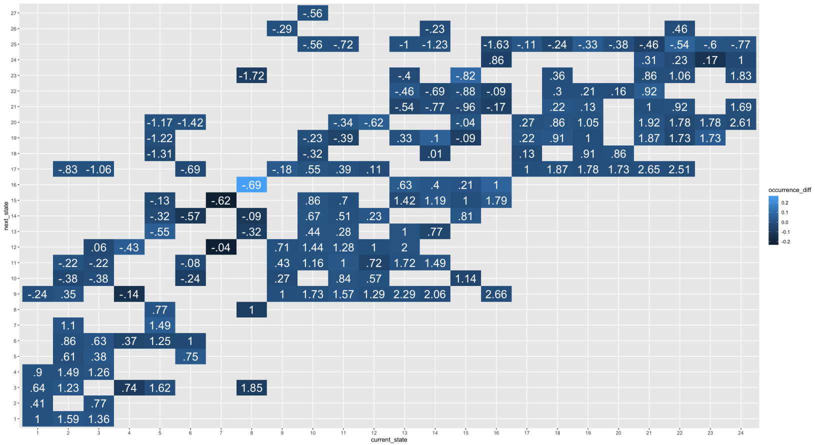

Figure: Most Common Transitions for the 2018 Rockies Pitchers with the 6 Best Soft Contact Percentages.

Substitution Strategies

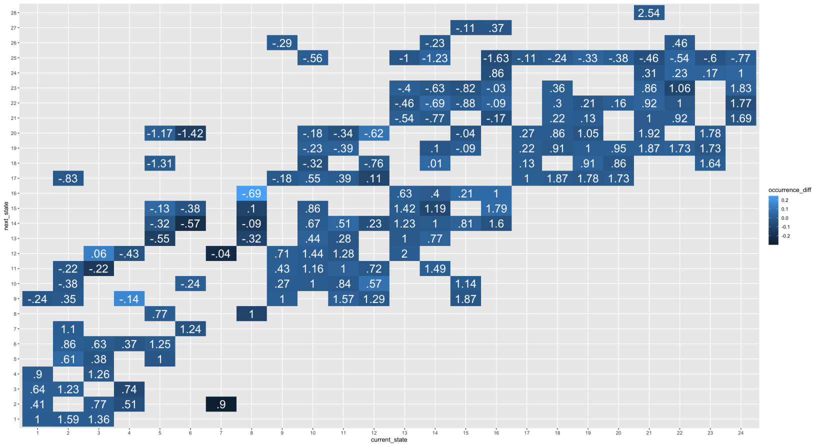

Figure: Most Common Transitions for the 2018 Rockies Pitchers with the 6 Best Batting Averages Against.

Substitution Strategies

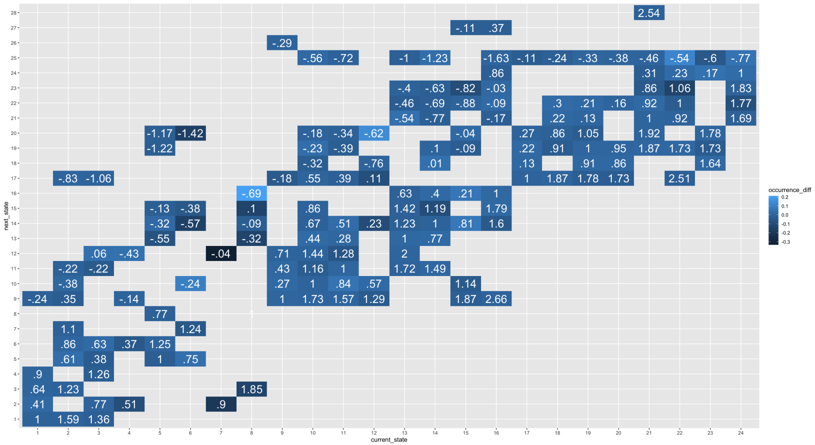

Figure: Most Common Transitions for the 2018 Rockies Pitchers with the 6 Best Earned Run Averages.

Assigning Values to Transitions, Modelling, and Regressions

In order to determine just how good or bad a transition is for the defense, I created a metric that I am calling the transition value. I calculate it by taking the number of runs scored on a play plus the difference in run expectancy values from the current state to the next state.

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

Assigning Values to Transitions, Regression Models

I am using logistic regression to model wins vs. losses, as well as a linear regression to model good vs. bad transitions given each substitution strategy. 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:

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

Assigning Values to Transitions, Regression Models

Transition value linear regression:

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

Assigning Values to Transitions, Regression Models

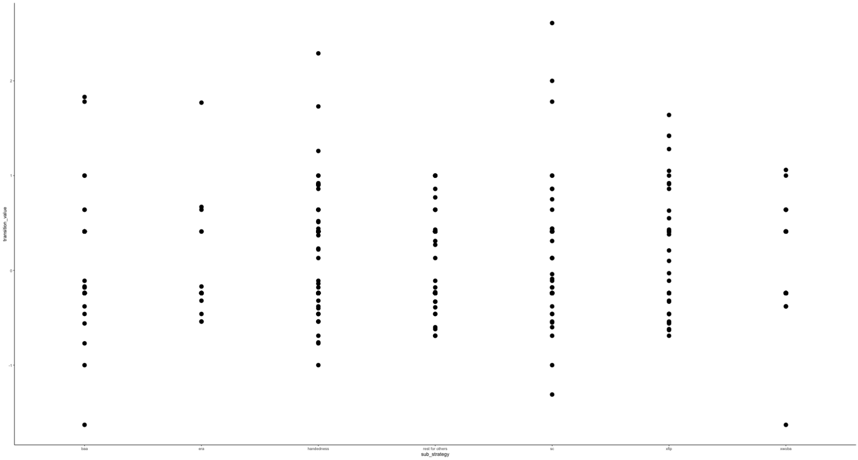

Figure: Linear Regression.

Win vs. Loss binary logistic regression:

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

Sources

[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.