Modeling Plate Discipline from the KBO to MLB

By: Ben Howell

Email: Ben.Howell@utexas.edu
Twitter: Benhowell71
Website: benhowell71.com

Contents

Abstract

The main goal of this project is to model how plate discipline, measured by BB%, K%, and the resulting BB/K ratio, translates from the Korean Baseball Organization (KBO) to Major League Baseball (MLB). In addition, this project will examine the differences between the playing environments of the KBO and MLB. This project primarily uses data that was manually charted by the author and hosted in the KBO Wizard, an R Shiny app created by the author.

Major League Baseball is becoming increasingly diverse, with 30% of Opening Day roster spots filled with players from outside of the United States. Most of these players hail from Latin America, but a few, including notables such as Choo Shin-soo (Texas Rangers), Yu Darvish (Chicago Cubs), Shohei Ohtani (Los Angeles Angels), and Kwang Hyun-kim (St. Louis Cardinals), hail from Asian countries. MLB clubs have a vested interest in finding the best players from these countries as they've shown they've been able to contribute positive value to an MLB team.

Challenges exist when projecting from these leagues to MLB given the different playing environments and talent levels. Previous attempts to quantify the differences between foreign leagues and MLB, from Alex Chamberlain, Justin Perline, and Clay Davenport, used box score stats to explore the difference between leagues. This project uses data from MLB's StatCast system and pitch-by-pitch data charted for the KBO Wizard (approximately 30,000 pitches) to explore those differences. The KBO Wizard and this project's calculation of Estimated xwOBA for the KBO were presented at the Ohio State Sports Analytics Conference on November 13th, 2020 and was the Research Winner in the Baseball category.

Using a paired data set from a player's last season in the KBO/MLB and first in the opposite league, this project will examine how player tendencies and approach changed from one league to the other. To model BB% and K%, the project will use statistics like Swing%, O-Swing%, and O-Contact% charted for the KBO Wizard to predict MLB plate discipline primarily through a stepwise regression.

Introduction

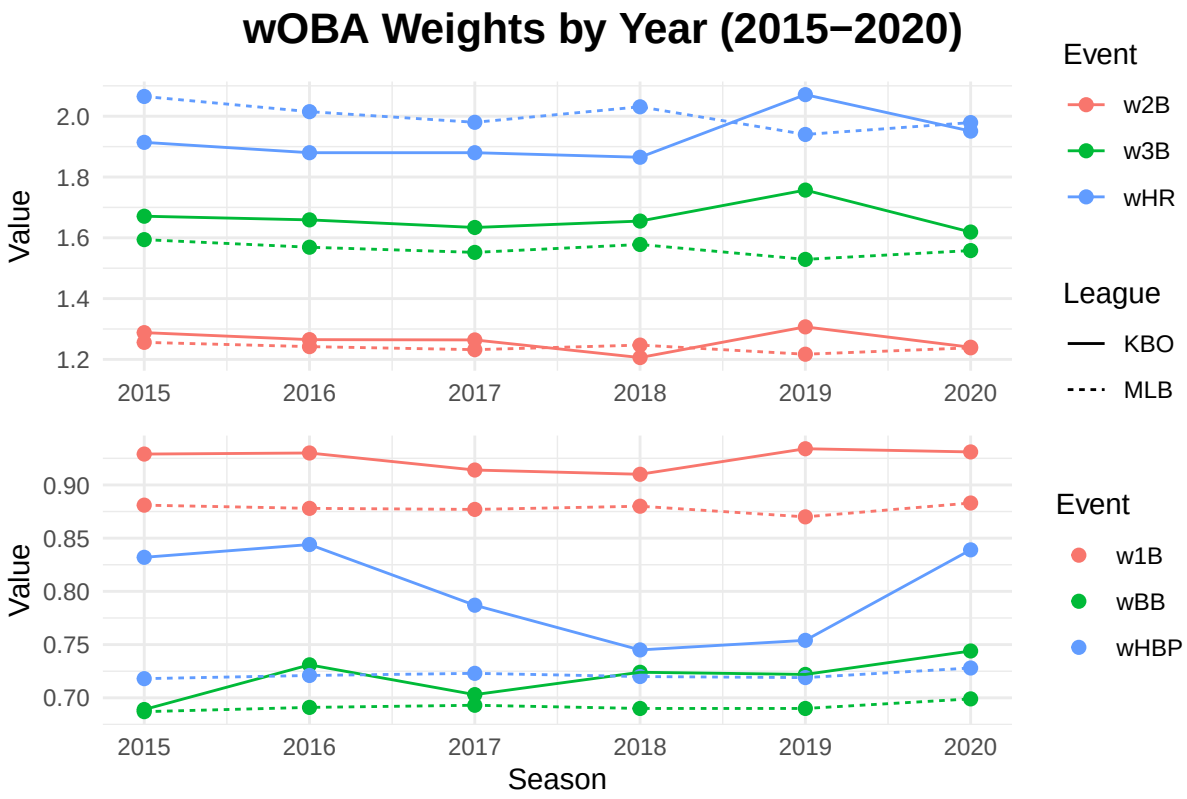
The 2020 season was a landmark season for the KBO. Due to COVID-19 and the delay of the MLB season, ESPN began broadcasting KBO games. Interest in the KBO spiked and has increased as multiple players made the jump from KBO to MLB following the 2020 season.

Chris Flexen signed with the Seattle Mariners. Kim Ha-seong signed with the San Diego Padres, and Na Sung-bum was posted. The future looks bright, with youngsters Lee Jung-hoo (MLB ETA 2023) and Kang Baek-ho (MLB ETA 2024) developing into strong prospects at the plate.

However, Korean-born hitters from the KBO have had mixed results when making the transition to MLB. Park Byung-ho was a below-average MLB hitter after posting a wRC+ of 184 in the KBO. Kang Jung-ho and Kim Hyun-soo produced at an above-average level when they were on the field in MLB. There have also been recent transplants from MLB who resurrected their career in the KBO before finding more success in MLB, such as Eric Thames and Darin Ruf.

The question that faces hitters (and pitchers) looking to jump from the KBO to MLB is the level of competition that they're facing. For hitters, the question centers around whether or not they can succeed against MLB velocity. This project looks to project BB% and K% from the KBO to MLB. Baseball will continue to be a global sport, but playing environments differ from league-to-league, making it crucial that we are able to accurately project MLB performance between leagues.

I used wOBA weights to compare the different playing environments in the KBO and MLB. These weights indicate how much certain events were worth for that season. The higher the weight, the more the event is worth. MLB weights were pulled from FanGraphs' GUTS! page while the KBO weights were derived from the FanGraphs 2015-2020 KBO leaderboards, using the `limSolve` package.



Weights derived from FanGraphs

One thing immediately pops up; the KBO spike in wHR, w3B, and w2B in the 2019 season before returning to their historic levels in 2020. The KBO has been searching for the optimal mix of HRs and offense in the game; one way they've approached the problem has been by altering the baseball (Kim, 2019) and widening the strike zone for the 2017 season (Kim, 2018).

In 2018, the KBO baseball was juiced, and HR numbers soared to record rates. Over these six seasons, five of the ten highest single-season HR totals came in 2018, devaluing them. While HR's and 2B's were at their lowest value in 2018, it wasn't by much.

The KBO changed course in 2019, de-juicing the baseball. Offensive production plummeted; Park Byung-ho led the league with 33 HRs in 2019, the 24th-most from these six seasons. Suddenly, HRs were at a premium and worth an incredible amount, as were triples and doubles. This drastic change has led to some interesting fluctuations in performance, which may have hurt the perception of certain hitters in the KBO who were interested in returning to MLB.

Walks, hit-by-pitches, and singles are consistently more valuable in the KBO than MLB. I believe that much of this is a result of the talent disparity in the KBO; some of the best players, like Kim Ha-seong or Chris Flexen, are MLB-caliber players, while you also have players who might not make a Single-A roster. In such an environment, getting on-base to let those excellent hitters do their job is one of the most valuable things that a low-end hitter can do. KBO wOBA weights bear a resemblance to Division I baseball weights, an environment with a similar talent disparity to the KBO (Driveline).

Beyond the distribution of talent, what spurs this approach? KBO hitters have a different approach to hitting. It's not MLB where some people believe that the "three true outcomes"

approach has taken over the sport. While the KBO does have some high-strikeout, high-power hitters, they generally play a "fundamental" style of baseball that is focused on getting on-base. So, how is this idea reflected in the approach of KBO hitters?

Table 1: MLB Averages vs Charted KBO Averages and Confidence Interval

Metric	MLB	Charted KBO	Lower 95% CI	Upper 95% CI	Contains MLB?
Approach					
Swing%	46.6	47.1	46.5	47.7	Yes
Whiff%	24.5	22.2	21.5	22.9	No
F-Swing%	28.3	29	28	30	Yes
Results					
GB%	45.3	51.2	49.9	52.5	No
K%	21.8	17.4	17.1	17.7	No
BB%	8.3	9.4	9.2	9.6	No

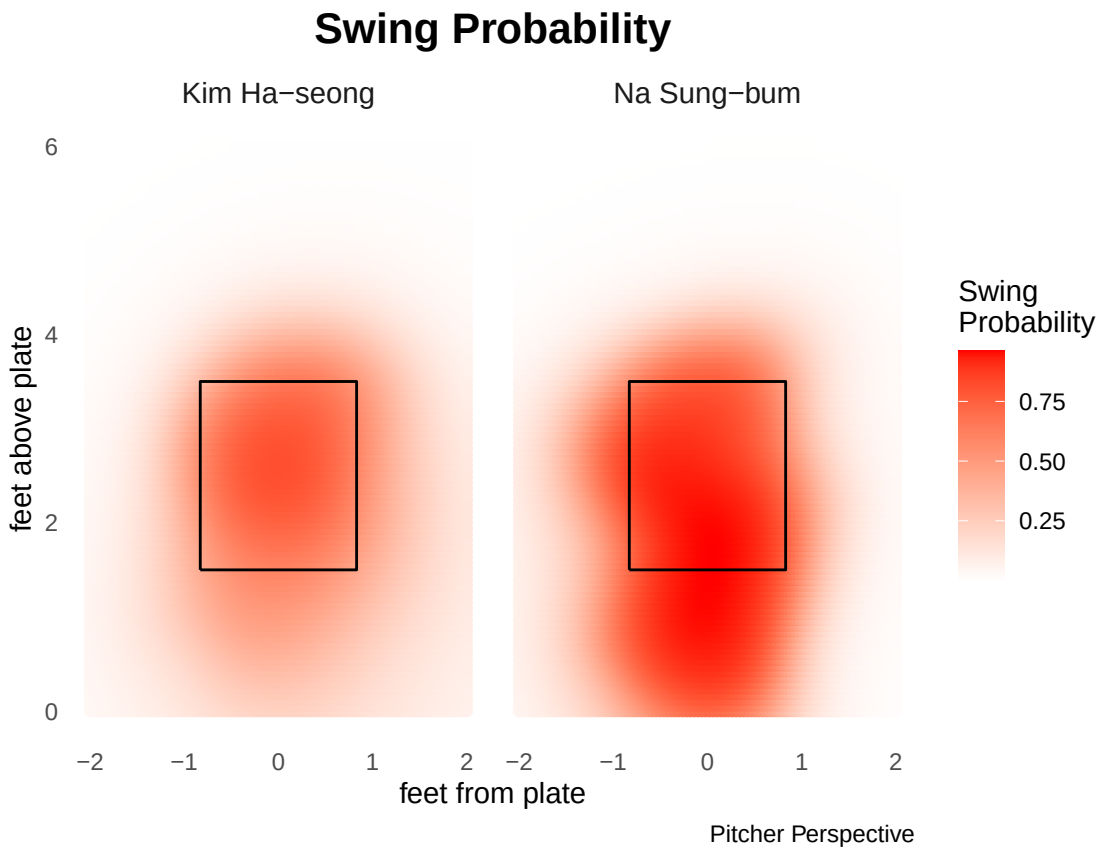
It boils down to one concept: hit for contact. After charting nearly 30,000 pitches from the KBO this season, I expected the KBO average swing rate to be below the MLB average, but the average KBO swing% came out to 47.1% versus the MLB average of 46.6%. That's a small difference and, with a 95% confidence interval, it's not a significant difference. The same holds for first-pitch swing%; the KBO F-Swing% of 29% is not significantly different from MLB's F-Swing% of 28.3%.

Where the two leagues start to differ is in terms of results-based stats. Whiff%, GB%, K%, and BB% are all statistically different between the two leagues.

KBO hitters swing at a similar rate as MLB hitters, but they're more focused on putting the ball in play. They don't view a single as a "worse" outcome than an HR; many hitters focus on singles and getting on-base because that's what they're good at. With a more controlled swing designed to hit the ball where the defense isn't, KBO hitters display more adaptability at the plate, leading to fewer swings-and-misses and consistently fouling off pitches at the plate. That complex combination results in a higher swing% and BB%, yet lower whiff% and K% in the KBO than MLB.

Why does this matter? Even though the KBO takes a different approach to hitting and baseball, there is still significant variation from player to player, as seen in Kim Ha-seong and Na Sung-bum. Kim Ha-seong possesses above-average plate discipline and patience, with a BB/K of 1.10, an SwStr% of 6.3%, and a 43.5% Swing%. Na Sung-bum resembles the sluggers that we typically see in MLB, with a 0.33 BB/K, a 16.5% SwStr, and an absurd 61.4% Swing%.

Using a `Generalized Additive Model` trained on the data collected for Kim and Na this season, 510 and 360 pitches respectively, I looked to see if they swung at or took a pitch based on its location. The decision to swing or take a pitch is mutually exclusive and collectively exhaustive. The probability of a swing falls in the `binomial` family and varies by the location of the pitch. Using a smoothing function in the `gam` model and a grid of 10,000 points, I created a representation of swing probability for Kim Ha-seong and Na Sung-bum.



Kim and Na have drastically different approaches at the plate. Kim is patient and measured, with the highest probability of a swing coming over the heart of the plate where he can do the most damage. Na swings at anything, whether it's in the zone or in the dirt.

Over the past few years, six hitters have gone from the KBO to MLB. Two of the six, Eric Thames and Darin Ruf, began their career in MLB and transitioned to the KBO after struggling to find steady at-bats. Steady playing time fueled improvements, and they returned to MLB where they've performed well.

Of these six hitters, each saw a decrease in wRC+ and BB/K from the KBO to MLB, although the degree of this reduction varied by hitter. Two hitters stand out: Park Byung-ho and Hwang Jae-gyun.

Park Byung-ho and Hwang Jae-gyun are notable as the two hitters on this list who posted an MLB wRC+ below league average. Park saw his K% spike from 26% in the KBO to an untenable 33% in MLB as he struggled mightily with MLB velocity, with a wOBA of 0.149 against 95+ mph pitches (Choi). Hwang kept his BB% stable at 8.8%, but his BB/K dropped from 0.75 to 0.33, and he posted an abysmal 26 wRC+. Hwang Jae-gyun outperformed his career averages in 2016 and hasn't reached that level since returning to the KBO.

Among the hitters who successfully made the KBO to MLB transition, a trend starts to emerge. Kim Hyun-soo, Eric Thames, and Darin Ruf managed to keep their MLB BB/K rate from falling too far below 0.50. Kang Jung-ho saw his BB/K fall to 0.28, but his K% stayed at 21% from the KBO to MLB. In his second MLB season, his BB% climbed back to 9.7% with a 21.4% BB% and a 0.46 BB/K ratio.

Table 2: KBO to MLB Stats (FanGraphs)

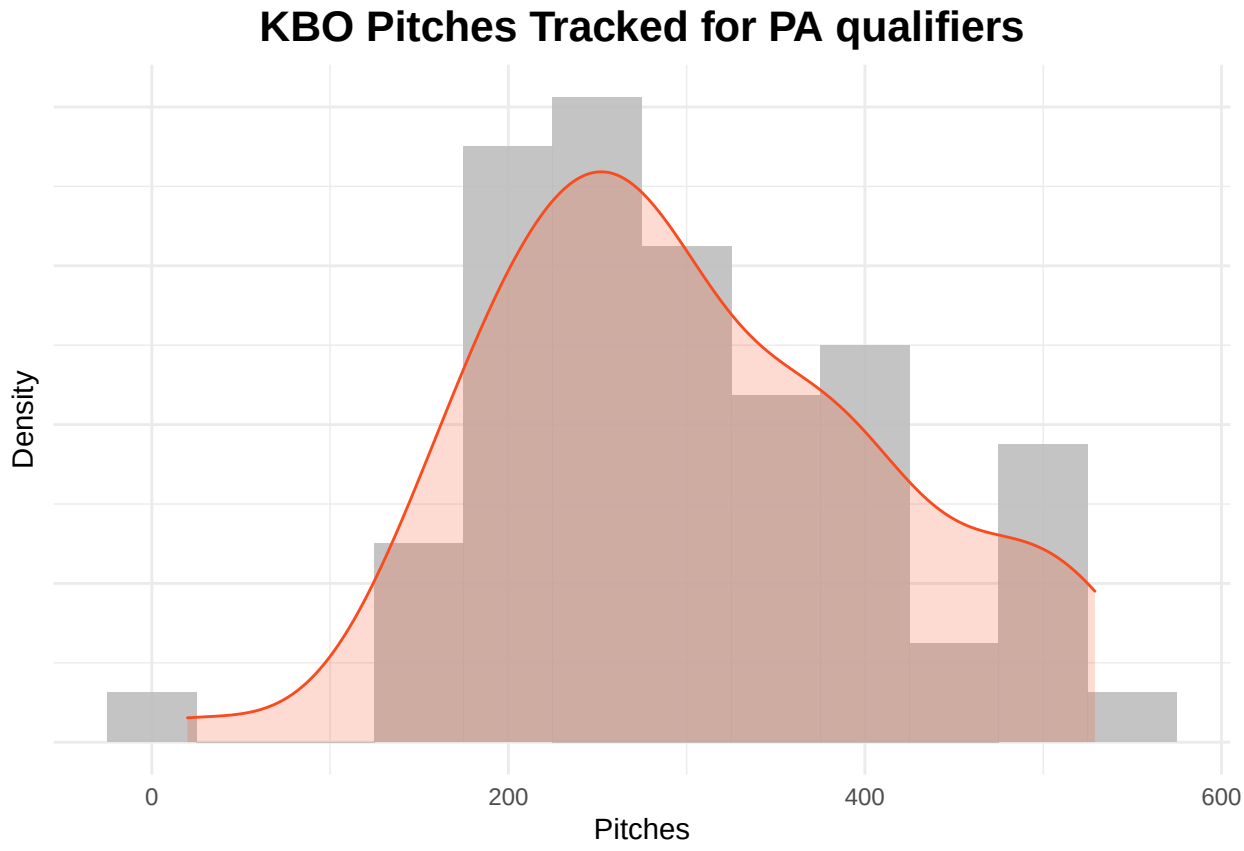
	KBO Stats					MLB Stats				
Player	**Season**	**BB%**	**K%**	**BB/K**	**wRC+**	**Season**	**BB%**	**K%**	**BB/K**	**wRC+**
Kang Jung-ho	2014	13.6	21.2	0.64	189	2015	6.0	21.2	0.28	128
Kim Hyun-soo	2015	16.0	10.0	1.60	148	2016	10.4	14.7	0.71	121
Park Byung-ho	2015	12.5	25.9	0.48	184	2016	8.6	32.8	0.26	79
Eric Thames	2016	14.0	19.5	0.72	165	2017	13.6	29.6	0.46	125
Hwang Jae-gyun	2016	8.8	11.8	0.75	134	2017	8.8	26.3	0.33	26
Darin Ruf	2019	14.1	15.3	0.92	149	2020	13.0	23.0	0.57	141
Na Sung-bum	2020	8.4	25.3	0.33	155	NA	NA	NA	NA	NA
Kim Ha-seong	2020	12.1	10.9	1.11	141	NA	NA	NA	NA	NA

The hitters who succeeded in MLB all posted a KBO BB/K above 0.64 and did so over multiple seasons, indicating their talent level. This cursory look puts a damper on Na Sung-bum's outlook while Kim Ha-seong's still looks promising. This analysis is similar to that done by Alex Chamberlain (Cuban National Series) and Justin Perline (Nippon Professional Baseball).

Data

This project pulled data from multiple sources to complete its KBO to MLB projections. MLB advanced and plate discipline stats from 2016-2020 were downloaded from FanGraphs. Results-based stats for the KBO were also downloaded from FanGraphs, using both their Standard and Advanced leaderboards.

One of the biggest factors limiting in-depth analysis of the Korean Baseball Organization and its players is the lack of publicly available TrackMan data or a StatCast-type system. The lack of publicly-available pitch-by-pitch spurred the creation of the KBO Wizard, a free-to-use R Shiny application that was created in order to host data from 29,736 pitches that were hand charted by the author. Using a separate R Shiny tool, the author charted pitch and BBE locations, assigned pitches types, and pitch-by-pitch results to start bringing KBO pitch-by-pitch data to the public. The data for the KBO Wizard was charted from ESPN/Twitch broadcasts, while lineups were pulled from myKBO.com.



The KBO Wizard data was not charted for every game, and was focused on the best pitchers and teams; the incomplete sample of pitches collected per player could be a concern when looking at swing-based characteristics. However, of our 64 PA qualifiers, only one had fewer than 100 charted pitches (Park Chan-do, who saw 20 pitches). Another 11 hitters saw from 137-193 pitches, while the remaining 52 hitters saw upwards of 200 pitches that were charted.

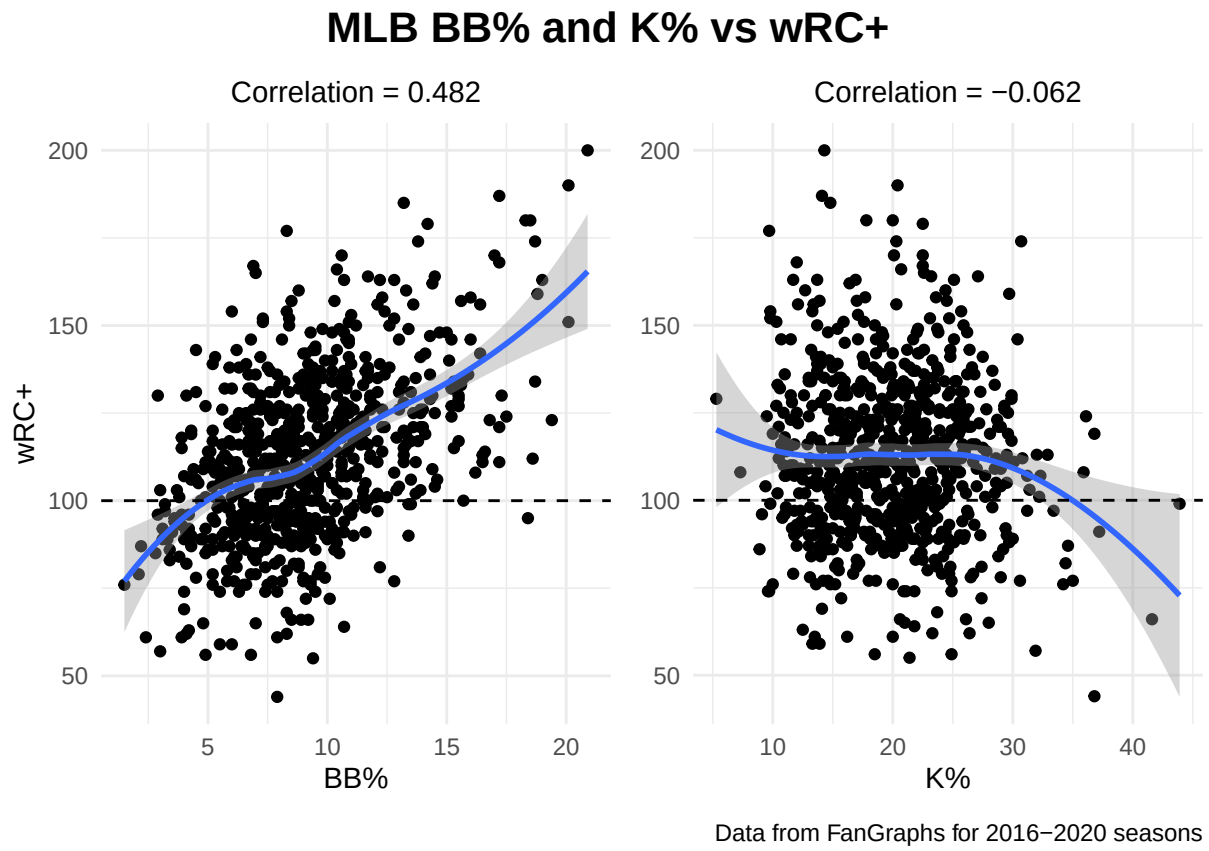
Previous MLB research indicates that swing rate stabilizes around 50 PAs and contact rate stabilizes around 100 PAs using a 50 PA increment (Carleton 2007). For Kim Ha-seong, 143 plate appearances and 510 pitches were charted while 107 plate appearances and 360 pitches were charted for Na Sung-bum. While these cut offs are for MLB hitters, I am comfortable using the charted swing/contact rates for Kim and Na since they crossed the 100 PA sample size.

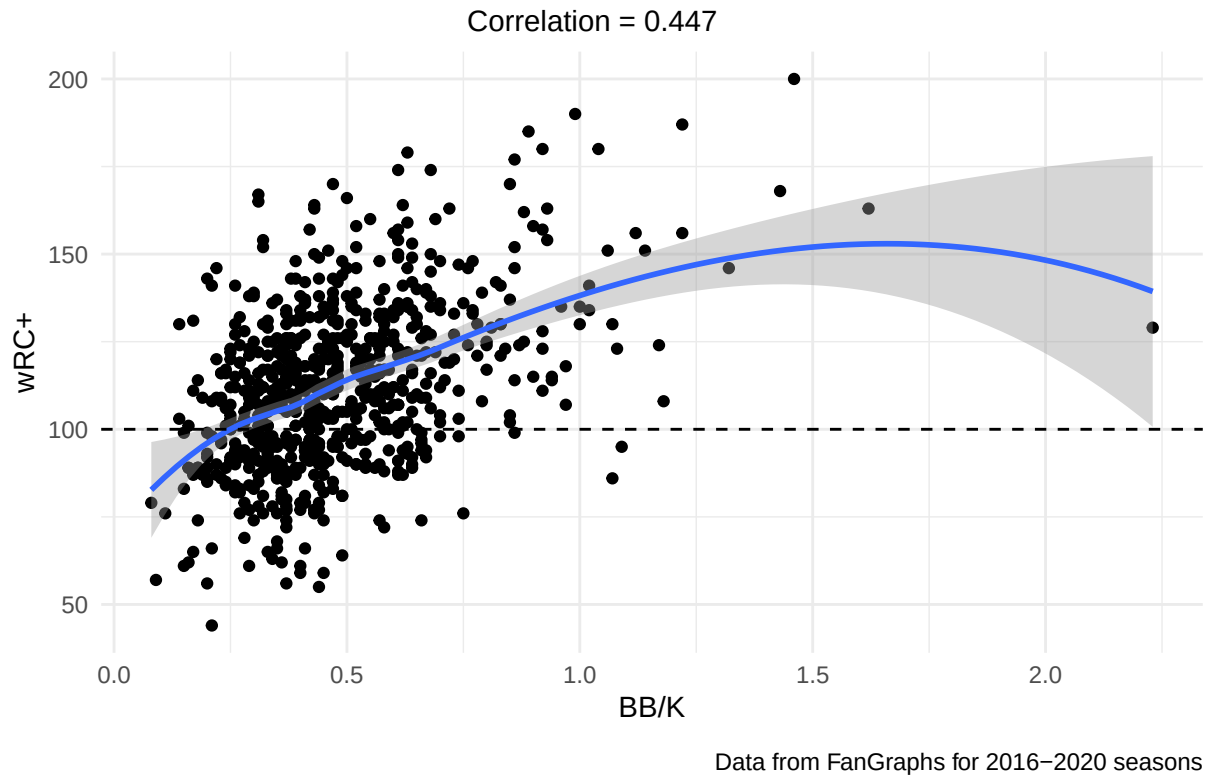
Why Is Plate Discipline Important?

Using MLB data from 2016 to 2020, I identified that BB% and BB/K are important factors for MLB success as a hitter (measured by wRC+). BB% has a correlation coefficient of 0.48 with wRC+ and BB/K has a 0.45 correlation coefficient. Batted ball characteristics certainly influence wRC+, but this project is focused on projecting MLB plate discipline measures from KBO data.

It is certainly possible for hitters to walk infrequently and remain above-average MLB hitters, but having a strong BB% and/or BB/K ratio provides a solid floor. Over the past five MLB seasons, there has only been one instance of a hitter posting a BB% over 15% and a wRC+ below 100 (Carlos Santana in 2020).

K% does not share a meaningful relationship with wRC+, with a correlation coefficient of -0.06. While BB% provides a solid floor for MLB hitters, an extreme K% (above 35%) starts to hamper a hitter's ceiling. Of all the seasons with a K% above 35% from 2016-2020, Willy Adames' 124 wRC+ in 2020 is the highest mark.





With BB% and K% in mind as the plate discipline metrics that we're most interested in, how do we project them? I will use the projected BB% and K% figures to produce a BB/K ratio. Using our MLB data, I created a correlation matrix to examine the relationship between various swing/contact metrics and our plate discipline metrics of BB% and K%.

Table 3: Correlation Matrix for MLB Stats

	BB/K	BB%	K%	SwStr%	O-Swing%	Z-Swing%	Swing%	O-Contact%	Z-Contact%	Contact%	Zone%
BB/K	1.00	0.69	-0.48	-0.60	-0.61	-0.36	-0.59	0.40	0.36	0.48	0.04
BB%	0.69	1.00	0.20	-0.14	-0.76	-0.30	-0.71	-0.18	-0.20	-0.12	-0.17
K%	-0.48	0.20	1.00	0.76	-0.04	0.13	-0.01	-0.84	-0.83	-0.88	-0.27
SwStr%	-0.60	-0.14	0.76	1.00	0.44	0.61	0.53	-0.82	-0.84	-0.94	-0.43
O-Swing%	-0.61	-0.76	-0.04	0.44	1.00	0.54	0.91	0.02	-0.04	-0.15	-0.31
Z-Swing%	-0.36	-0.30	0.13	0.61	0.54	1.00	0.82	-0.29	-0.37	-0.39	-0.35
Swing%	-0.59	-0.71	-0.01	0.53	0.91	0.82	1.00	-0.08	-0.16	-0.22	-0.20
O-Contact%	0.40	-0.18	-0.84	-0.82	0.02	-0.29	-0.08	1.00	0.73	0.92	0.29
Z-Contact%	0.36	-0.20	-0.83	-0.84	-0.04	-0.37	-0.16	0.73	1.00	0.91	0.30
Contact%	0.48	-0.12	-0.88	-0.94	-0.15	-0.39	-0.22	0.92	0.91	1.00	0.43
Zone%	0.04	-0.17	-0.27	-0.43	-0.31	-0.35	-0.20	0.29	0.30	0.43	1.00

Swing and Contact Metrics

- BB/K - *Walks divided by strikeouts*
- BB% - *How many plate appearances ended in a walk*
- K% - *How many plate appearances ended in a strikeout*
- SwStr% - *Swings and misses over total pitches faced*
- O-Swing% - *How often a hitter swung at pitches outside of the zone*
- Z-Swing% - *How often a hitter swung at pitches inside the zone*
- Swing% - *How frequently a hitter swung*
- O-Contact% - *How often a hitter made contact on pitches outside of the zone*
- Z-Contact% - *How often a hitter made contact on pitches inside of the zone*
- Contact% - *How often a hitter made contact*
- Zone% - *How many pitches were thrown in the strike zone against a hitter*

I am using two categories of variables: swing metrics and contact metrics. Swing metrics correlate well with BB%, while contact metrics correlate strongly with K%. These two categories serve as the basis of our BB% and K% models.

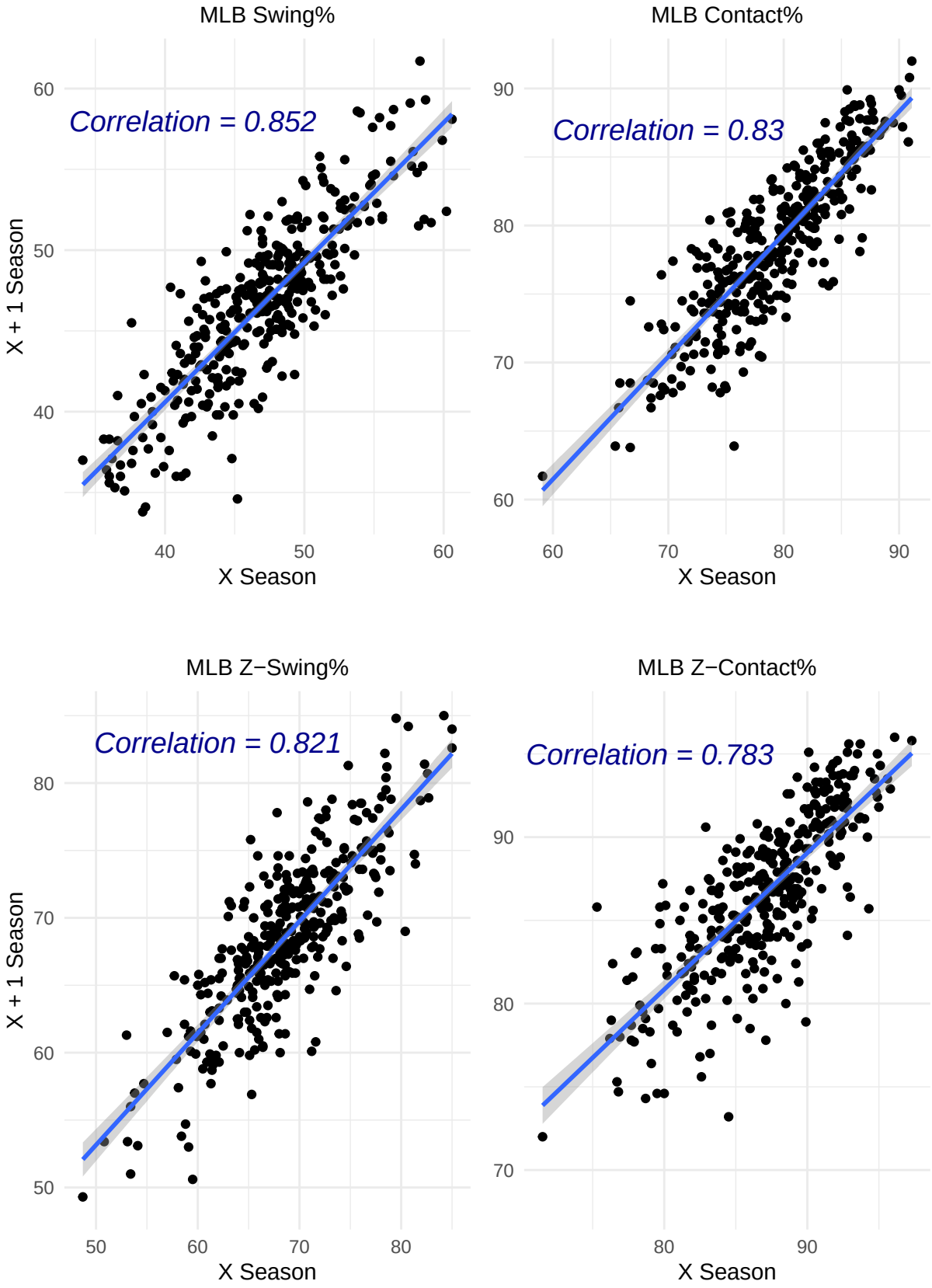
O-Swing% has a -0.76 correlation with BB%, Swing% has a -0.71 correlation, and Z-Swing% has a -0.30 correlation. The less you swing, especially at pitches out of the strike zone, the more likely you are to draw a walk.

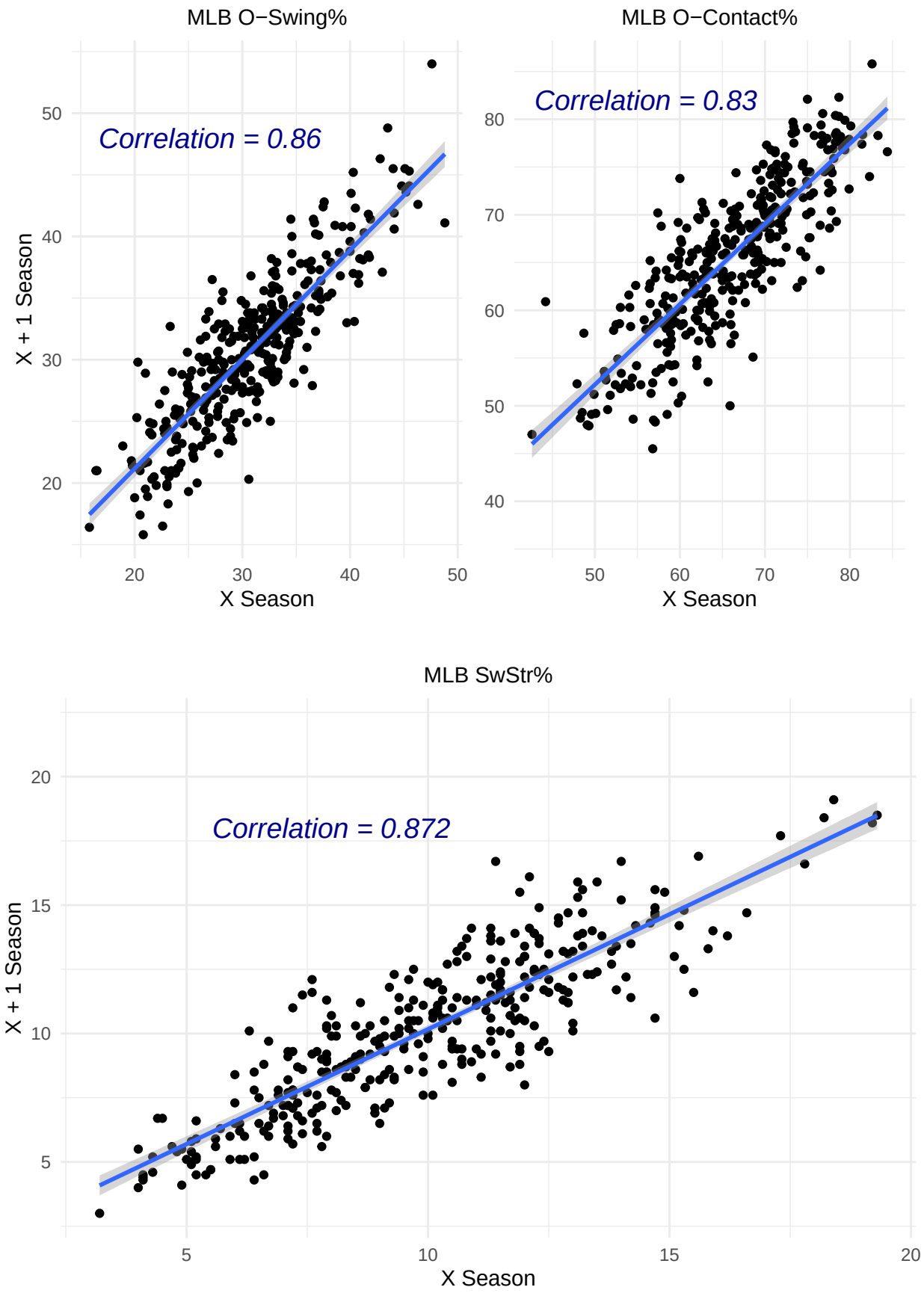
Contact% has a -0.88 correlation with K%, O-Contact% has a -0.84 correlation, and Z-Contact% has a -0.83 correlation. The more contact that you make, the less you'll strikeout. While none of the swing metrics have a relationship with K%, SwStr% has a 0.76 correlation as the combination of Swing% and Contact%.

The goal of this project is to model MLB plate discipline based on the player's approach at the plate. One assumption that I made for this project is that swing decisions are up to the hitter; hitters must CHOOSE to swing the bat and, while inputs from the pitcher (release point, arsenal, etc.) influence that decision, the pitcher cannot force a swing.

To test whether this assumption holds any weight, I looked at Year-over-Year correlations for these metrics using MLB data. What I found is that there is a strong, positive relationship between these seven variables in X Season and the variable in the X + 1 Season. Z-Contact% posted the lowest correlation, at 0.783, while the other six variables were upwards of 0.82.

With the strong relationship between these variables from one season to the next, I am comfortable proceeding with the assumption that swing decisions are inherent to the hitter and under their control.


Year−over−Year MLB Swing Correlations
MLB Swing%
Correlation = 0.852
MLB Contact%
Correlation = 0.83
MLB Z−Swing%
Correlation = 0.821
MLB Z−Contact%
Correlation = 0.783
X + 1 Season
X Season

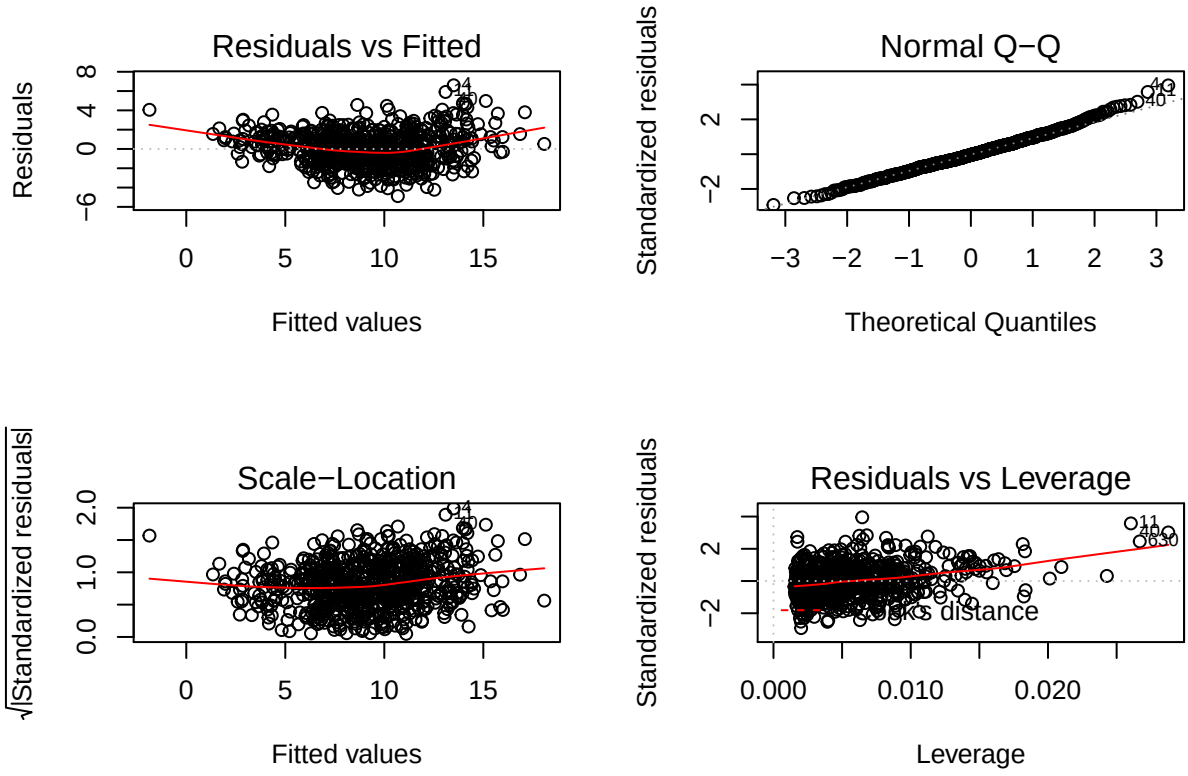

MLB O−Swing%
Correlation = 0.86
X + 1 Season
X Season
MLB O−Contact%
Correlation = 0.83
X Season
MLB SwStr%
Correlation = 0.872
X + 1 Season
X Season

Models

BB%

I used a stepwise regression with a backward step to model BB%, using `Swing%`, `O-Swing%`, and `Z-Swing%` as our inputs. Our model determined that each of these variables was statistically significant in determining BB%, producing an adjusted r-squared value of 0.75.

```
##
## Call:
## lm(formula = `BB%` ~ `O-Swing%` + `Z-Swing%` + `Swing%`, data = test_bb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8832 -1.0971 -0.0679  1.0169  6.5970
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.98966    0.76834   29.92   <2e-16 ***
## `O-Swing%`   0.34696    0.04058    8.55   <2e-16 ***
## `Z-Swing%`   0.65349    0.02962   22.06   <2e-16 ***
## `Swing%`    -1.48064    0.07001  -21.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.674 on 703 degrees of freedom
## Multiple R-squared:  0.7483, Adjusted R-squared:  0.7472
## F-statistic: 696.5 on 3 and 703 DF,  p-value: < 2.2e-16
```



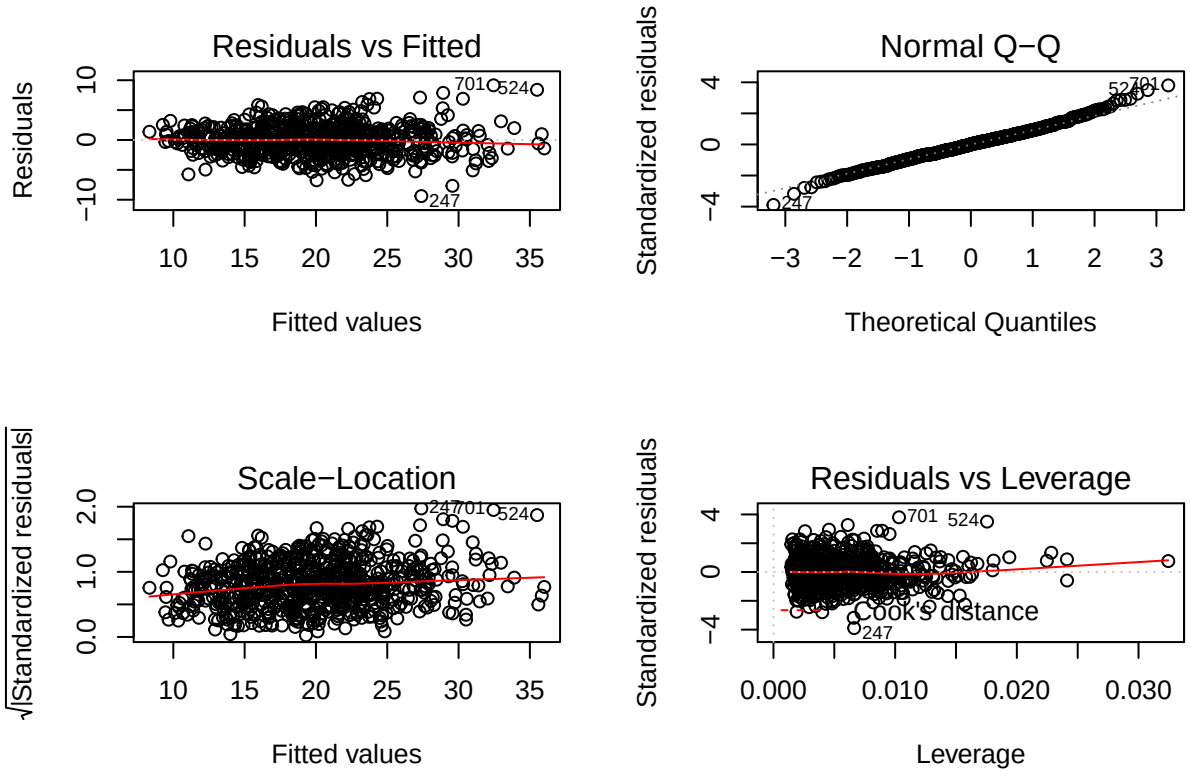
Season	Name	Team	PA	BB%	K%	wRC+	O-Swing%	Swing%
2017	Joey Votto	Reds	707	19.0	11.7	163	15.8	41.9
2020	Randal Grichuk	Blue Jays	231	5.6	21.2	112	37.3	51.8
2020	Hanser Alberto	Orioles	231	2.2	13.0	87	54.0	61.7

The residuals of our BB% Model are evenly distributed and randomly spread, but there are a few outlier values with high leverage. Joey Votto is known as one of the most patient hitters in MLB history and his 2017 O-Swing% of 15.8% was the lowest of the last five seasons. Hanser Alberto is an outlier on the other end of the spectrum with his 61.7% Swing% that leads this stretch of MLB. The Alberto and Grichuk seasons took place in the shortened 2020 season and it's possible they would swing less as the season continued, although they did clear the 100 PA barrier that Carleton lists as the baseline for swing contact metrics to stabilize (2007).

K%

Our model for K% produced an adjusted r-squared of 0.81. This model used `O-Contact%` and `Z-Contact%` to measure the ability of the hitter to make contact on pitches in various locations. The interaction of `Contact%` and `Swing%`, was used to account for how frequently a hitter swung and how often those swings resulted in whiffs.

```
##
## Call:
## lm(formula = `K%` ~ `O-Contact%` + `Z-Contact%` + (`Contact%`:`Swing%`),
##     data = test_k)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3901 -1.5474  0.0427  1.4874  9.1488
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          90.5959721  1.8256183  49.625  < 2e-16 ***
## `O-Contact%`         -0.2869889  0.0165754 -17.314  < 2e-16 ***
## `Z-Contact%`         -0.5147312  0.0279965 -18.386  < 2e-16 ***
## `Contact%`:`Swing%` -0.0020598  0.0002489  -8.275 6.44e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.42 on 703 degrees of freedom
## Multiple R-squared:  0.8151, Adjusted R-squared:  0.8143
## F-statistic:  1033 on 3 and 703 DF,  p-value: < 2.2e-16
```



Our residual plots are normally and evenly distributed. It is worth noting that some of our residuals become quite large after our Predicted Value exceeds 30; this does make sense as our model is unlikely to predict extremely high values because those are so rare.

Results

Table 5: Kim/Na Projected MLB Stats

Hitter	BB%	K%	BB/K	Pred BB%	Pred K%	Pred BB/K
Kim Ha-seong	12.1	10.9	1.10	11.4	14.5	0.79
Na Sung-bum	8.4	25.3	0.33	3.6	21.2	0.17

Kim Ha-seong is projected for an 11.4% BB% and a 14.5% K%, resulting in a BB/K ratio of 0.79, a very favorable mark. A 0.79 BB/K would have ranked 14th in MLB for the 2020 season, between David Fletcher (125 wRC+) and Brandon Nimmo (148 wRC+). Na Sung-bum's projected 3.6% BB% would've been the third-lowest in MLB in 2020, and his 0.17 BB/K would have been the fourth-lowest in MLB, tied with Hanser Alberto (87 wRC+).

That's a solid projection; one that shows Kim's patience and Na's free-swinging ways. This projection works under the assumption that a hitter's swing attributes are inherent to the hitter and wouldn't change from the KBO to MLB. However, as seen with Park Byung-ho and the other Korean stars, things change when you go from facing KBO pitchers to MLB pitchers.

The average fastball velocity in the KBO was about 89 mph in 2020 whereas the average MLB velocity hovers around 93 mph, not to mention the differences in quality of secondary and breaking pitches; nobody in the KBO throws an Adam Ottavino-esque frisbee slider. That's a big adjustment for a hitter to make.

How to account for this adjustment and how the swing profiles of Kim and Na may change? I have swing data for the 2020 KBO season and swing data for MLB through the StatCast system. Since I don't have historical swing data for the KBO, I chose to examine how swing profiles changed for hitters who played in the KBO in 2020 and MLB for the 2018 or 2019 season.

Table 6: MLB Metric Divided By KBO Metric

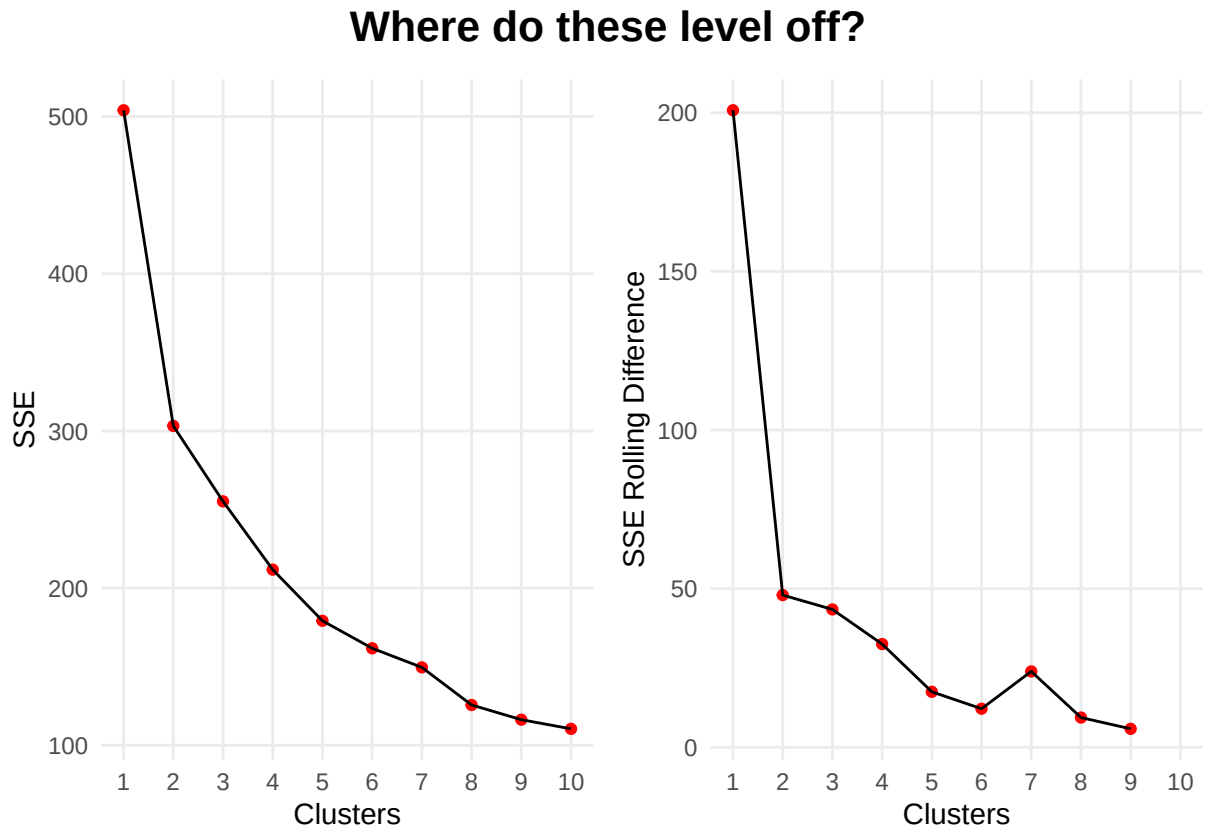
Hitter	wSwing	wContact	wSwStr	wOSwing	wOContact	wZSwing	wZContact
Aaron Altherr	1.10	0.87	1.54	1.06	1.04	1.01	0.79
Addison Russell	0.98	0.88	1.47	1.01	0.84	0.92	0.90
Brandon Barnes	1.05	1.24	0.77	0.99	1.56	1.10	1.04
Daniel Palka	1.09	0.88	1.46	1.13	1.00	1.07	0.84
Dixon Machado	1.05	0.93	1.64	1.11	0.86	1.03	0.94
Jose Miguel Fernandez	1.12	0.95	1.61	1.08	0.83	1.01	1.00
Preston Tucker	1.06	0.92	1.51	0.90	0.88	1.13	0.91
Average	1.06	0.94	1.27	1.04	0.96	1.04	0.91

Despite widening the player pool to include the 2018 or 2019 MLB season, I am still working with a seven-player sample. That's a small (and biased) sample to work with. However, I believe that we can still learn something from how their swing metrics and decisions changed from MLB to the KBO.

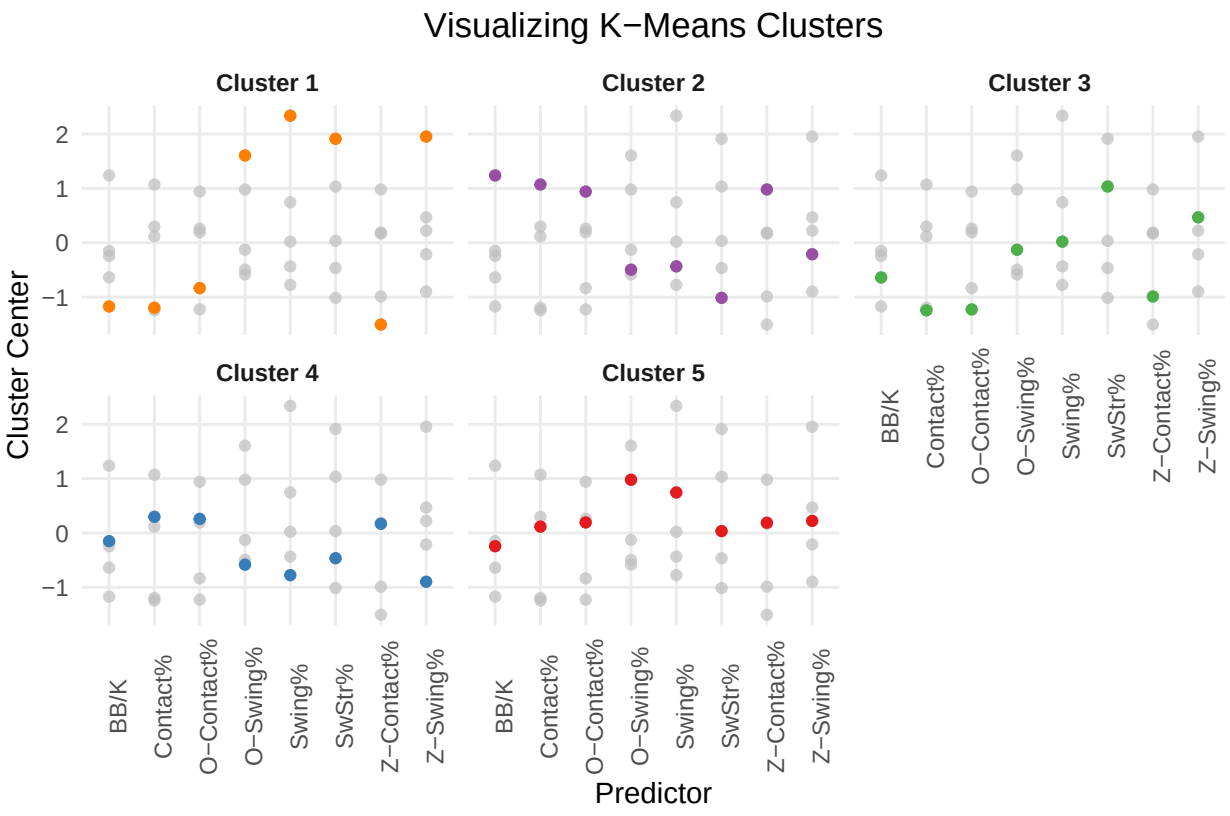
For most, it's a simple formula. They swung less and made more contact when they did swing in the KBO. Brandon Barnes is the only outlier, seeing his SwStr% increase from MLB to KBO, with most of his other metrics trending in the wrong direction from his last MLB stint.

How to apply this %-change from MLB to KBO to Kim Ha-seong and Na Sung-bum? One strategy is to use those average values as the adjuster for both Kim and Na. That approach has issues. Kim Ha-seong is not the same type of player as Daniel Palka. Na Sung-bum is not the same player as Dixon Machado. Adjusting Kim and Na's numbers with this average would be unfair because it doesn't account for different player archetypes.

To accurately adjust Kim and Na's tendencies from the KBO to MLB, I wanted to see how players similar to them changed when making that transition. To find which players are similar to Kim and Na I used a K-Means clustering algorithm to cluster our 64 qualified KBO hitters based on eight plate discipline measures.



I originally began this process by looking for 10 clusters for 64 KBO hitters. However, more than five clusters failed to make a meaningful difference in the rolling Sum of Squared Errors. Five clusters for 64 hitters should return a decent grouping of players in each cluster for our purposes.



Clusters of the 64 qualified KBO from FanGraphs data

Our K-Means cluster algorithm returned five different clusters of varying sizes and talent levels. Kim Ha-seong and Na Sung-bum landed in drastically different clusters, representative of how MLB clubs viewed them.

Kim Ha-seong landed in Cluster 2, home to the best plate discipline in the KBO. This is a great cluster to be in, with an average KBO wRC+ of 123. Four former MLB players, Kim Hyun-soo, Jose Miguel Fernandez, Dixon Machado, and Preston Tucker, are in this cluster, as are some of the best Korean-born bats, with Seo Keon-chang, Yang Eui-ji, and Lee Jung-hoo.

To create an adjustment for Kim Ha-seong as he transitions from the KBO to MLB, I used the average of Jose Miguel Fernandez, Dixon Machado, and Preston Tucker's translations since they were the three former MLBers most similar to Kim Ha-seong (I don't have Kim Hyun-soo's KBO swing data). For this exercise, I used an average of their translations and didn't weight by their MLB experience.

Table 7: Cluster 2 and KBO Stats

Cluster	Hitter	wRC+	BB%	K%	BB/K	Swing%	SwStr%
2	Yang Eui-ji	156	8.7	8.9	0.98	48.7	6.1
2	Preston Tucker	148	12.0	10.6	1.13	46.7	7.4
2	Park Suk-min	144	16.7	13.8	1.21	43.2	8.1
2	Kim Ha-seong	141	12.1	10.9	1.10	43.5	6.3
2	Kim Hyun-soo	139	10.2	8.6	1.19	42.9	6.0
2	Lee Jung-hoo	139	9.6	7.6	1.26	39.5	2.3
2	Jose Miguel Fernandez	138	8.7	6.3	1.38	45.5	4.9
2	Hur Kyoung-min	118	7.2	5.7	1.25	41.9	5.4
2	Kim Sang-su	117	11.7	11.7	1.00	40.7	4.5
2	Choi Won-joon	116	8.0	8.5	0.94	49.1	5.9
2	Seo Keon-chang	111	15.3	9.7	1.57	42.1	4.6
2	Dixon Machado	105	9.6	10.7	0.90	42.5	5.0
2	Jung Soo-bin	103	9.8	10.0	0.98	42.3	4.9
2	An Chi-hong	101	7.6	10.2	0.74	46.0	5.0
2	Lee Yong-kyu	99	12.1	7.4	1.64	48.2	1.6
2	Jang Sung-woo	96	8.4	14.1	0.59	47.9	4.3

Table 8: Cluster 1 and KBO Stats

Cluster	Hitter	wRC+	BB%	K%	BB/K	Swing%	SwStr%
1	Na Sung-bum	155	8.4	25.3	0.33	61.4	16.4
1	Kim Dong-yeop	125	6.4	17.5	0.37	59.3	16.6
1	Oh Ji-hwan	115	7.6	19.6	0.39	54.2	14.4
1	Park Chan-do	45	6.8	16.4	0.41	65.0	25.0

Na Sung-bum landed in Cluster 1 with three other hitters, Kim Dong-yeop, Oh Ji-hwan, and Park Chan-do. Park was the hitter who only saw 20 pitches, which is responsible for his inclusion; in his 20 charted pitches, he was a free-swinger.

Na is a very unique and aggressive hitter with a Swing% above 60%, and none of the transfers profiled close to him. Brandon Barnes of the Hanwha Eagles would likely be the closest, but his swing metrics transferred oddly from MLB to the KBO. In the end, I decided to adjust Na's numbers by using the average changes of all of the transfer players since there's just no one like him in the KBO; even in MLB, Hanser Alberto (who had a 0.17 BB/K) is the only hitter to post a Swing% higher than Na's 61.4%.

Table 9: Kim/Na Adjusted MLB Stats

Hitter	BB%	K%	BB/K	Pred BB%	Pred K%	Pred BB/K
Kim Ha-seong	12.1	10.9	1.10	11.4	14.5	0.79
Kim Ha-seong (C2)	12.1	10.9	1.10	9.2	19.9	0.46
Na Sung-bum	8.4	25.3	0.33	3.6	21.2	0.17
Na Sung-bum (C1)	8.4	25.3	0.33	1.0	25.8	0.04

With the adjustment, Kim and Na see their BB% and K% decrease. Kim's projection falls to an 9.2% BB% (above MLB average) and a 19.9% K% (better than the MLB average), for a BB/K of 0.46. That's a BB/K that is slightly above the MLB average and players at that mark had success in 2020; Corey Seager posted a 151 wRC+ with a 0.46 BB/K while Andrew McCutchen posted a 103 wRC+. There's a path to success with a 0.46 BB/K, and this is a reasonable first-year projection for Kim.

How does this Kim Ha-seong adjusted projection compare to other popular projection systems?

System	BB%	K%	BB/K
Un-adjusted	11.4%	14.5%	0.79
Adjusted	9.2%	19.9%	0.46
ZiPS	9.1%	17.2%	0.53
Depth Charts	9.2%	23.4%	0.39
Davenport Translations	10.3%	8.9%	1.15

Turns out it looks pretty similar! That's a good sign that this model is on the right track and producing realistic numbers, especially with the cluster adjustment. The Davenport Translations included at the bottom are Clay Davenport's translation of what Kim's 2020 KBO stats are worth in MLB in 2020 rather than a projection, one reason they're more optimistic. The ZiPS projections by Szymborski are some of the best in the industry and give Kim a similar outlook to our model.

Na Sung-bum is projected for a 1% BB% and a 26% K%, coming out to a 0.04 BB/K ratio. That's historically bad; the lowest BB% since 2000 was 1.5% by Dee Gordon in 2018 (76 wRC+), and the lowest BB/K was 0.08 in 2017 by Tim Anderson (79 wRC+). When Park Byung-ho came to MLB, he posted an 8.6% BB% and a 32.8% K%, for a 0.26 BB/K which resulted in a 79 wRC+. This is not a favorable projection for Na Sung-bum.

Table 11: MLB Career Numbers vs Projected

Hitter	MLB BB%	MLB K%	MLB BB/K	Pred BB%	Pred K%	Pred BB/K
Preston Tucker	6.3	23.2	0.27	9.2	14.3	0.64
Kim Hyun-soo	9.9	16.6	0.60	10.6	14.1	0.75
Jose Miguel Fernandez	4.9	12.2	0.40	12.2	10.0	1.22
Dixon Machado	6.7	18.0	0.37	9.5	12.1	0.78
Aaron Altherr	9.5	28.7	0.33	12.3	21.7	0.57

Since I have to wait for the 2021 season to play out before judging how my Kim Ha-seong and Na Sung-bum projections fared, the best way to "test" this model is on the players who have made the transition from MLB to KBO. This is not an ideal method because I'm testing a future projection against past performance.

On the whole, the projected metrics outpace the true career values for these players. However, I have a few theories as to why these models consistently overshot. First, these are using these players' un-adjusted KBO swing metrics rather than applying an adjustment or using their MLB swing metrics. Second, hitters change, especially when they get more playing time and a chance to refine their approach with consistent at-bats. Aaron Altherr may be a slightly different hitter now than he was in MLB, hence a reduction in his projected K%.

Third, I'm comparing a future projection against each of these player's career values, not their most recent MLB performance. Using career numbers doesn't always give us an honest representation of a player.

- In his first season in MLB, Kim Hyun-soo posted a 0.71 BB/K with a 10.1% BB% and a 14.7% K% before a hamstring injury derailed his MLB stint
- In 2018, Aaron Altherr saw 285 PAs for the Phillies and posted a 0.40 BB/K. In 2019, he received just 66 PAs with three different teams and posted a 0.12 BB/K
- Jose Miguel Fernandez posted a 0.40 BB/K in 123 MLB PAs in 2018. With the Angels AAA team, he posted a 0.97 BB/K in 2018

Using past performance to evaluate a future projection is not the best method of evaluation. However, I am limited by a small, biased sample, which requires using the statistics that are available. If I had more year-over-year KBO swing data, I'd be able to more rigorously test this model and its results on a larger sample of players, one that included both MLB to KBO and KBO to MLB transfers.

Limitations

This project is limited by two major factors. The lack of publicly available swing data, even if hand charted, from the KBO makes this type of analysis very difficult. While I am confident that MLB swing decisions are stable year-over-year, I am not able to confirm a similar trend in the KBO, though I suspect swing decisions remain stable in the KBO as well.

The other major limitation is the small sample of players who have not only played in both the KBO and MLB, but played in those leagues at a time when swing-type data was being collected. For example, the best comp for Kim Ha-seong is Kim Hyun-soo while Na Sung-bum bears a striking resemblance to Park Byung-ho. While I can examine how their BB% and K% changed from the KBO to MLB and attempt to apply that to Kim and Na, I cannot see how their underlying swing decisions changed.

The players that I have swing data on in MLB and the KBO represent a biased sample. They've transitioned from MLB to the KBO because they were unable to find MLB playing time and likely at or below MLB replacement level. However, I used them as a test case for how swing decisions transfer between the two leagues because they are all I have. A future project could use mean reversion techniques to account for the league average player.

To improve on this project, I need more swing data from the KBO. Pairing that data with the players who have transferred between the two leagues would give us more of an idea of how swing decisions transfer, improving our clustering algorithm and adjusting player swing attributes.

References

Alex Chamberlain (2016). Yasmany Tomas' Plate Discipline Makes Me Nervous.
https://fantasy.fangraphs.com/yasmany-tomas-plate-discipline-makes-me-nervous/

Ben Howell (2020). KBO Wizard.
https://benhowell71.shinyapps.io/KBO_Wizard/

Clay Davenport. DT Report for the 2020 Korean Baseball Organization.
http://claydavenport.com/stats/webpages/2020/2020pageKBOyearALL.shtml

Dan Kurtz. myKBOstats.net.
https://mykbostats.com/

Dan Szymborski (October 2020). Ha-seong Kim Has Big MLB Aspirations and Projections.
https://blogs.fangraphs.com/ha-seong-kim-has-big-mlb-aspirations-and-projections/

Driveline Baseball (2020, June 17). An Introduction to College WAR - cWAR.
https://www.drivelinebaseball.com/2020/06/an-introduction-to-cwar/

Jim Alpert. Called Strike Package.
https://bayesball.github.io/Intro_to_CalledStrike_Package.html

Justin Choi (2020, December 28). wOBA against 95+ MPH pitches.
https://twitter.com/justinochoi/status/1343582622702682112?s=20

Justin Perline (2015, April 1). Projecting the NPB to MLB Transition for Batters.
https://www.beyondtheboxscore.com/2015/4/1/8322487/projecting-npb-to-mlb-transition-batters-japan-plate-discipline-iso

Justin Perline (2015, April 16). Projecting the NPB to MLB Transition Part 2.
https://www.beyondtheboxscore.com/2015/4/16/8415791/projecting-the-npb-to-mlb-transition-part-2

Russell Carleton, Pizza Cutter (2007, November 14). 525,600 minutes: How do you measure a player in a year.
http://web.archive.org/web/20080102094412/http:/mvn.com/mlb-stats/2007/11/14/525600-minutes-how-do-you-measure-a-player-in-a-year/

Sung Min Kim (2018, March 18). How Korean Baseball Briefly Shortened Time of Game. *https://blogs.fangraphs.com/how-korean-baseball-briefly-shortened-time-of-game/*

Sung Min Kim (2019, April 23). The KBO Appears to be De-Juicing Its Baseballs. *https://blogs.fangraphs.com/the-kbo-appears-to-be-de-juicing-its-baseballs/*

Sung Min Kim (2019, August 16). Let's Check in on the KBO's De-Juiced Baseballs. *https://blogs.fangraphs.com/lets-check-in-on-the-kbos-de-juiced-baseballs/*