Swing Quality Factor

Andrew Chen (陳昱宏)[1], Tony Liao (廖世祺)[2], Charles Chen[3], Mason Chen[4]
[1]University of San Francisco, 184 Ewing Terrace, San Francisco, CA 94118
[2]Executive Coach, 7911 Dry Creek Drive, Dublin, CA 94568
[3]MBB STEM Advisor, 2191 Morrill Ave, San Jose, CA 95132
[4]Stanford Online High School, Palo Alto, CA 94035

Abstract
This paper introduces a new Baseball Statistical Model, Swing Quality Factors (SQF), we analyze the game from the batter's perspective which will be a new concept to help simplifying the decision making process for the pitchers. Using pitch count, pitch location, pitch type and velocity to calculate the correlation between each pitch.

The biggest breakthrough of this project comparing to the previous research is that after a great amount of data have been collected, we will be able to predict the probability of swing quality factor by using JMP 13 statistic software. With this model, various applications are considered which will be the new insights that could benefit both pitchers and hitters.

Keywords: Swing Quality Factor (SQF), Baseball Prediction, JMP 13

1. Introduction

In these days, most of the baseball analytics and statistics are using numbers that are too general to determine whether the player would be helpful and benefits the team. However, we thought that those number are not specific enough to represent a player because those numbers do not show the actual scenario between each pitch. What we have experienced as a pitcher and coaches is that choosing the right type of pitch and location is always difficult, to me, especially in the critical situations like [different counts, numbers of outs, runners or no runner on bases, scores, or regular season or playoff games].

Players are relying on their experience or the scout report to choose their pitch type and location instead of scientific base. Swartz, Grosskopf, Bingham and Swartz (2016) estimate pitch quality through the use of random forest methodology to accommodate the inherent complexity of the relationship between pitch quality and the associated covariates. However, only swing quality factor is looking at the game from batter's perspective which we thought it would be more accurate.

1.1 Pitch Qualification Research

From the research of pitch qualification that has done by Wilson (2015). He used experts to assess the quality of pitches QOP on a scale of -10 to 10. A fitted linear regression equation was obtained to estimate QOP for future pitches based on observed covariates, and used QOP to assess pitchers and compared the approach with conventional pitching measures. However, the correlation between pitches and the quality of the pitch could not be determined clearly since the total number of fastball will always be higher than other pitches such as curve ball, change up, or slider.

From Figure 1 that is showing below we can see the average of quality of pitch on sinker is above 5.0 almost every year started from 2008, but the total number of pitches is always below 100,000 pitches. On the other hand, four-seam fastball is the most common pitch for every pitcher, although the average of quality of pitch is below 5.0 every year, but the total number of pitches is always between 200,000 to 300,000. When only look at numbers that are too general there will be some misunderstanding because the standard of each pitch is not the same. In order to get first strike and take advantages of batters, pitchers will always try to throw the most comfortable and confident pitch which often is a fastball, so the quantity between pitches is so distinct.

NP = Number of Pitches
CH = Change Up, CU = Curveball
FC = Cutter Fastball, FF = Four Seam Fastball
FS = Split Finger Fastball, FT = Two Seam Fastball
KC = Knuckle Curve, KN = Knuckleball
SI = Sinker, SL = Slider

Year	Pitch	All	CH	CU	FC	FF	FS	FT	KC	KN	SI	SL
2017	qop Max	9.91	9.35	8.91	8.90	9.91	8.84	9.71	8.66	7.66	9.67	9.13
	qop Avg	4.56	4.35	4.36	4.34	4.70	4.35	5.03	4.41	4.02	4.97	4.11
	NP	729,396	73,202	61,708	35,863	260,069	10,842	102,405	18,213	2,678	43,931	119,895
2016	qop Max	9.99	9.11	9.08	9.21	9.66	8.45	9.91	8.83	8.89	9.99	8.87
	qop Avg	4.59	4.37	4.40	4.50	4.82	4.37	5.08	4.50	4.14	5.08	4.24
	NP	715,245	73,372	62,305	34,705	258,726	10,363	95,847	15,121	4,256	48,133	108,807
2015	qop Max	9.90	8.80	9.13	9.50	9.76	8.85	9.66	8.85	8.26	9.90	8.90
	qop Avg	4.58	4.32	4.36	4.53	4.81	4.43	5.10	4.46	4.15	5.05	4.22
	NP	712,273	75,688	54,158	39,408	255,565	11,227	92,489	15,480	3,850	57,211	103,159
2014	qop Max	9.75	8.83	9.30	9.46	9.71	8.73	9.75	8.88	8.15	9.75	9.09
	qop Avg	4.57	4.34	4.52	4.53	4.75	4.15	5.08	4.46	4.15	5.10	4.21
	NP	708,663	73,207	58,169	41,055	243,028	10,867	94,649	14,735	3,176	63,655	101,178
2013	qop Max	10.00	8.84	9.35	9.59	10.00	8.65	9.63	8.75	8.02	9.69	9.13
	qop Avg	4.57	4.34	4.61	4.59	4.74	4.14	5.08	4.33	4.10	5.08	4.24
	NP	720,217	72,968	62,216	38,854	253,062	9,290	96,194	7,809	3,125	60,891	110,982
2012	qop Max	10.03	9.36	9.02	9.49	9.99	8.56	9.73	8.85	8.36	10.03	8.99
	qop Avg	4.57	4.29	4.65	4.51	4.73	4.15	5.09	4.62	4.16	5.10	4.26
	NP	723,185	73,427	65,565	40,656	245,513	9,723	90,358	6,609	2,859	73,545	110,055
2011	qop Max	10.21	9.00	9.15	8.98	9.62	8.71	9.76	8.99	8.52	10.21	8.89
	qop Avg	4.47	4.24	4.64	4.49	4.61	4.16	5.01	4.52	4.41	5.01	4.21
	NP	717,060	73,924	59,068	42,315	238,933	8,980	83,149	5,335	4,450	83,855	111,262
2010	qop Max	10.31	8.83	9.24	9.34	9.50	8.94	9.76	9.09	8.33	10.31	9.10
	qop Avg	4.46	4.16	4.52	4.44	4.63	4.11	4.99	4.43	4.38	5.03	4.12
	NP	737,143	78,678	60,453	40,480	241,957	9,506	85,912	4,482	4,385	97,644	107,287
2009	qop Max	9.98	9.49	9.01	9.20	9.71	9.48	9.84	8.65	8.43	9.98	9.08
	qop Avg	4.51	4.25	4.58	4.41	4.66	4.20	5.02	4.58	4.12	5.10	4.17
	NP	711,945	69,474	58,376	37,643	243,332	8,276	80,601	4,055	2,778	94,594	106,801
2008	qop Max	10.07	9.11	9.00	9.21	9.78	8.69	10.07	9.00	8.34	9.84	9.15
	qop Avg	4.47	4.26	4.60	4.40	4.62	4.33	5.00	4.61	4.20	5.00	4.20
	NP	702,619	69,244	56,390	36,687	238,225	9,200	73,995	2,956	3,606	102,170	104,175

Figure 1: The Measurement of QOP from 2008 to 2017

1.2 The Best Fastball

Moore (2009) claimed that Jeremy Greenhouse presented a new method for evaluating who throws best pitches in baseball which the principle is to evaluate pitches based on their outcomes. For example, if two pitchers throw exactly the same pitch, one of the pitcher may get an out, and the other pitcher gives up a hit. The outcome-based method would give exactly the same credit to both pitchers, so the researcher used regression methods to looked specifically at fastballs thrown by right-handed pitchers to right-handed batters, and to investigate the quality of various types of fastballs thrown by various pitchers. As the result, this model may be biased to pitchers with "intangibles" because every pitchers' deliveries are not identical.

Rank	Player	Value	Type	Control	Velocity	Movement
1	Zack Greinke	-0.0313	FT	1.13	2.68	0.90
2	Roy Halladay	-0.0304	FT	0.73	2.19	0.72
3	Ronald Belisario	-0.0181	FB	0.38	2.05	0.33
4	Ubaldo Jimenez	-0.0166	FB	0.33	2.14	0.25
5	Jonathan Broxton	-0.0164	FB	0.18	1.87	0.09
6	Felix Hernandez	-0.0155	FB	0.38	1.94	0.25
7	Roy Halladay	-0.0150	FC	0.50	0.96	0.55
8	Heath Bell	-0.0149	FB	0.41	1.39	0.28
9	Mariano Rivera	-0.0130	FC	0.19	1.32	0.32
10	Bobby Jenks	-0.0123	FB	0.31	1.15	0.07
11	Daniel Bard	-0.0122	FB	0.13	1.42	-0.01
12	Brandon Morrow	-0.0118	FB	0.11	1.20	0.21
13	Joel Zumaya	-0.0112	FB	-0.08	1.95	-0.14
14	Vin Mazzaro	-0.0106	FB	0.34	1.04	0.26
15	Andrew Bailey	-0.0101	FC	0.38	0.51	0.37
16	J.J. Putz	-0.0095	FB	0.20	0.96	0.14
17	Joe Nathan	-0.0092	FB	0.19	0.63	0.26
18	Freddy Dolsi	-0.0090	FB	0.12	1.47	0.12
19	Chris Carpenter	-0.0090	FB	0.24	1.34	0.26
20	Kevin Jepsen	-0.0090	FB	0.18	0.99	0.13

Figure 2: Ranking the Best Fastball Based on Control, Velocity, and Movement

The researcher claimed that Zack Greinke has the best fastball in Major League in 2009, but as a former pitcher and coaches we thought the ranking standard need to be more specific since the intentions of two-seam fastball and four-seam fastball are different. If we look at Figure 2 above, the ratio between four-seam fastball and two-seam fastball is 15: 2, but how can two different type of pitches have comparison? We all know that under normal condition, the average velocity of four-seam fastball should always be higher then two-seam fastball. Also, from location (control) to movement then to velocity is actually the right order for a scout to evaluate a pitcher. Therefore, we recreated a graph which shows the right order of evaluating a pitcher and only based on fastball to see the difference.
From figure 3 we can see that Heath Bell's fastball ranked was at 8th and now he is at 1st because of his ability of locating a four-seam fastball which we consider the most import factor that a pitcher must has.

1		Location	Movement	Velocity
2	H. Bell	0.41	0.28	1.39
3	R. Belisario	0.38	0.33	2.05
4	F. Hernandez	0.38	0.25	1.94
5	V. Mazzaro	0.34	0.26	1.04
6	U. Jimenez	0.33	0.25	2.14
7	B. Jenks	0.31	0.07	1.15
8	C. Carpenter	0.24	0.26	1.34
9	J.J. Putz	0.2	0.14	0.96
10	J. Nathan	0.19	0.63	0.26
11	J. Broxton	0.18	0.09	1.87
12	K. Jepsen	0.18	0.99	0.13
13	D. Bard	0.13	-0.01	1.42
14	F. Dolsi	0.12	0.12	1.47
15	B. Morrow	0.11	0.21	1.2
16	J. Zumaya	-0.08	-0.14	1.95

Figure 3: New Ranking Based Only on Four-Seam Fastball and Location

1.3 Prediction of Hot and Cold Zones for Hitters

As we know, the strike zone is usually divided into a 3x3 grid. However, Fast (2011a) concluded that 3x3 grids are not as sufficient as 2x2 grids to whether a player can hit better in certain zones. The researcher used regression to compare performance in a given zone in one half of the sample along with performance in the other eight zones in the same half of the sample to predict performance in that zone in the other half of the sample.

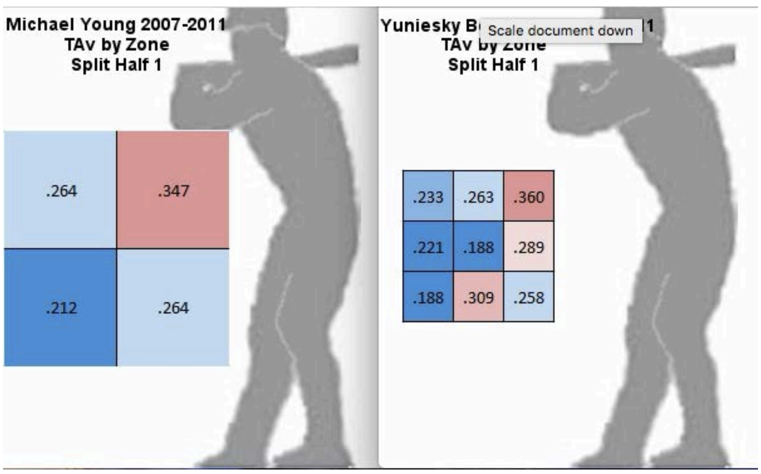


Figure 4: Comparison of Cold and Hot Zones between 2x2 grids and 3x3 grids

If we look at Figure 4 above**,** Michael Young's hot zone in 2x2 grids is inside high and the cold zone is outside low. When we look at the 3x3 grids, he has 2 hot zones including inside high and middle low. People might think the strike zone of 3x3 grids conducts more details than 2x2 grids. However, the sample size which the researcher considers as plate appearance is too small which causing the inaccuracy according to 3x3 grids.

Let's say if there are total of 20 pitches that had thrown to the strike zone, but only 3 out of 20 were thrown to the middle low box. If Michael Young had executed those middle low pitches relatively good, then the result will not be accurate because only a few amount of pitches was located there. On the other hand, if 10 out of 20 pitches had thrown to the outside low corner and Michael Young did not executed as well as pitches that were thrown to the middle low, then the hitting average of outside low corner will be relatively low compared to the hitting average of middle low pitches.

1-4 Strike Zone Pitch Quality

This research that has done by Roegele (2013a) targeted specifically for pitches inside the strike zone and used wOBA to determine what made pitches successful, and used the 3x3 grids custom strike zone formulas (Figure 5) which created by Mike Fast (2011b) to calculated this modified form of wOBA for each of these sub-zones for each pitch type.

RHB zone: -1.03 < px < 1.00 and (0.92 + batter_height*0.136) < pz < (2.60 + batter_height*0.136)

LHB zone: -1.20 < px < 0.81 and (0.35 + batter_height*0.229) < pz < (2.00 + batter_height*0.229)

Figure 5: px (the horizontal location of the pitch crossing the front of home plate, in feet), pz (the height of the pitch at the front of home plate)

At that point, they were able to answer the question of "How will are these pitchers locating their pitches within the strike zone?" The approach of his was very similar to what Swing Quality Factor is trying to do and this research have provided a good start for us. However, how to properly and accurately measure the px and pz is the biggest disadvantage of the equations since there might have some deviation because feet and height are not "solid" numbers.

According to the second research, Roegele (2013b) added velocity as another consideration. He measures closers' velocity and location using wOBA and this time a strong positive correlation has shown on the graph (Figure 6)

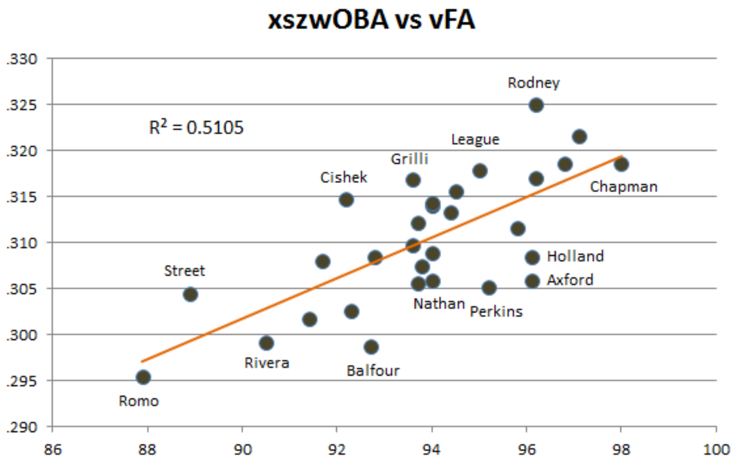


Figure 6: x-axis (velocity), y-axis (strike zone location quality)

As this model has shown potentials on evaluating pitches using wOBA, our model of SQF has a even bigger advance since it can combines all three factors including location, movement, velocity into our measurement. We collect numbers during every plays that had happened instead of using "equation" to calculated baseball.

2. Difference Between Today's Training and SQF

In today's training, the strategies are usually based on one game, or maybe per batter, instead of per pitch. The simple reason was that there is no model designed for this process and also the model will require very specific statistical tools. There is no solution right now because the most common statistical results including Earning Run Average (ERA), Walks Plus Hits Per Inning Pitched (WHIP), Strikeout per 9 Innings, or Number of Strikeouts per walk (K/BB). All of these data which were provided by the baseball community are still per batter or per game based, and they do not record the true effectiveness of each pitch. Not only baseball professionals but also fans are trying to simplify the game.

Graphs | All Graphs | Season Stats | Splits | Game Log | Play Log | Compare | PITCHf/x | Spray Charts | Heatmaps

Batting | Pitching | Standard | Advanced | Batted Ball | Win Probability | Pitch Type | Plate Discipline | Fielding | Value

Dashboard | Customize

Postseason | Projections | Minor Leagues | Regular Season

Season	Team	W	L	SV	G	GS	IP	K/9	BB/9	HR/9	BABIP	LOB%	GB%	HR/FB	ERA	FIP	xFIP	WAR
2006	Devil Rays	6	8	0	21	21	124.2	7.51	2.74	1.30	.328	72.4 %	42.7 %	13.7 %	4.84	4.39	3.99	1.7
2007	Devil Rays	12	8	0	31	31	215.0	7.70	1.51	1.17	.282	71.8 %	43.4 %	11.1 %	3.85	3.86	3.65	4.3
2008	Rays	14	8	0	33	33	215.0	6.70	1.67	1.00	.287	73.3 %	46.3 %	9.8 %	3.56	3.82	3.87	3.9
2009	Rays	11	12	0	33	33	219.2	6.84	2.13	1.19	.308	71.2 %	42.3 %	11.2 %	4.14	4.02	3.85	3.5
2010	Rays	13	15	0	34	33	203.1	8.28	2.26	1.50	.341	68.4 %	41.3 %	13.8 %	5.18	4.24	3.55	1.7
2011	Rays	16	12	0	33	33	249.1	8.12	2.35	0.94	.258	79.6 %	46.2 %	11.1 %	2.82	3.42	3.25	4.5
2012	Rays	15	10	0	33	33	227.2	8.82	2.29	0.99	.292	71.9 %	52.3 %	13.4 %	3.52	3.47	3.24	3.9
2013	Royals	13	9	0	34	34	228.2	7.71	2.68	0.79	.298	79.5 %	41.6 %	8.6 %	3.15	3.47	3.72	4.5
2014	Royals	6	3	0	11	11	73.1	7.73	1.84	0.86	.298	73.0 %	47.0 %	10.1 %	2.95	3.43	3.42	1.5
2014	ZiPS (R)	8	7		22	22	147.0	8.18	2.38	0.92	.315	73.1 %			3.65	3.51		2.6
2014	ZiPS (U)	14	10		33	33	221.0	8.03	2.20	0.90	.313	74.8 %			3.41	3.48		4.2
2014	Steamer (R)	9	7	0	21	21	131.0	7.57	2.42	0.92	.291	71.9 %			3.68	3.63		2.1
2014	Steamer (U)	15	10	0	32	32	205.0	7.63	2.21	0.90	.294	71.6 %			3.42	3.56		3.7
Total	- - -	106	85	0	263	262	1756.2	7.73	2.17	1.08	.298	73.6 %	44.7 %	11.4 %	3.76	3.77	3.60	29.5

Figure 7: Present Day Baseball Statistical Analysis

The concept of trying to simplify baseball is the fundamental issue that is causing many pitchers are lack of guidance while they choosing the right type of pitch and location. Also, most of the baseball professionals think that simplifying the game will help pitchers and fans to understand the game better. For example, Neil Weinberg (2014) said that “Advanced stats might seem like a complicated foreign language, but becoming conversational is actually very easy.” It probably seems reasonable to people why they want to simplify baseball, but pitchers are not gaining much help from those approaches.

When I was a pitcher, our coaches always encourage us to keep our mind simple, at the best is just think about only one thing when we pitch. Before pitching, we should think about things like where should I throw the ball to, with what type of pitch to throw. Then just execute it with the correct mechanics. By doing all the things that coaches had mentioned, it is much more likely to accomplish the team’s goal, to get outs in order to win the game.

3. Methods

So how does Swing Quality Factor help for pitching efficiency? Instead of spending time looking into the massive amount of general stats that we have mentioned, we are using Swing Quality Factors to make our decision making process easier. All the pitchers want to know include whether the batter will swing at the pitch, making contact or not, and how hard the contact will be.

In the mathematical form, a set of Swing Quality Factors are introduced, where as using N to represent “No Swing”, M for “Swing and Miss”, S for “Strong Contact”, F for “Fair Contact”, and W for “Weak Contact”. Data Collected with SQF Information including pitch count, ball or strikes, Swing Quality Factors, and location. Even more specific factors are velocity, types of pitch, result, intended target, and on/off the target. Some people might not be very familiar with what is intended target and on/off target because the present baseball statistics do not have them on their record. Intended target means the original location where the catcher wants the pitcher to throw the ball, and on/off target means whether the pitcher hit the target or not.

Batter #	Pitch #	target	swing factor	B/S	location	velo	pitch	result	on/ off
L1	1	9	N	B	9	96	FF		OFF
L1	2	9	N	B	7	87	CU		OFF
L1	3	9	N	B	3	94	FF		OFF
L1	4	6	N	B	3	93	FF	WALK	OFF
R2	5	7	N	S	7	93	FF		ON
R2	6	7	N	B	3	94	FF		OFF
R2	7	4	N	B	8	87	CU		OFF
R2	8	4	CH	S	2	94	FF	E	OFF
L3	9	9	N	B	8	87	CU		OFF
L3	10	9	N	S	3	94	FF		OFF
L3	11	9	N	S	6	87	CU		OFF
L3	12	9	CH	S	4	94	FF	FOUL	OFF
L3	13	9	CW	S	9	88	CU	GO	ON
R4	14	9	N	B	9	93	FF		OFF
R4	15	7	N	S	6	93	FF		OFF
R4	16	8	M	S	9	88	CU		OFF
R4	17	7	N	B	1	94	FF		OFF
R4	18	9	N	B	6	90	CU		OFF
R4	19	8	CM	S	9	87	CU	FOUL	OFF
R4	20	9	M	S	1	93	FF	SO	OFF
L5	21	9	N	B	3	94	FF		OFF
L5	22	9	N	S	9	89	CU		ON
L5	23	9	N	B	8	89	CU		OFF
L5	24	9	CM	S	6	94	FF	SINGLE	OFF

Figure 8: Swing Quality Factor Analysis

In order to analyze pitching types and Swing Quality Factors, we use JMP 13 statistics software. There are three kinds of platform that we are using as our main tool. The first one is called Distribution Platform, which is used to study the interaction and conditional probability among the pitching and hitting variables to identify the correlations relevant to the swing contact quality. Then we combined the second and the third platform known as the Fit Model and Profiler Platform.

The Fit models are able to build the Transfer Function (Regression Predictive Model), and by analyzing all the collected data such as Hitting Side (Right or Left Handed), Ball/Strike, Pitching Location, Pitching Velocity, and Pitching Target, we can use the Profiler Platform to predict the probability of swing quality factor.

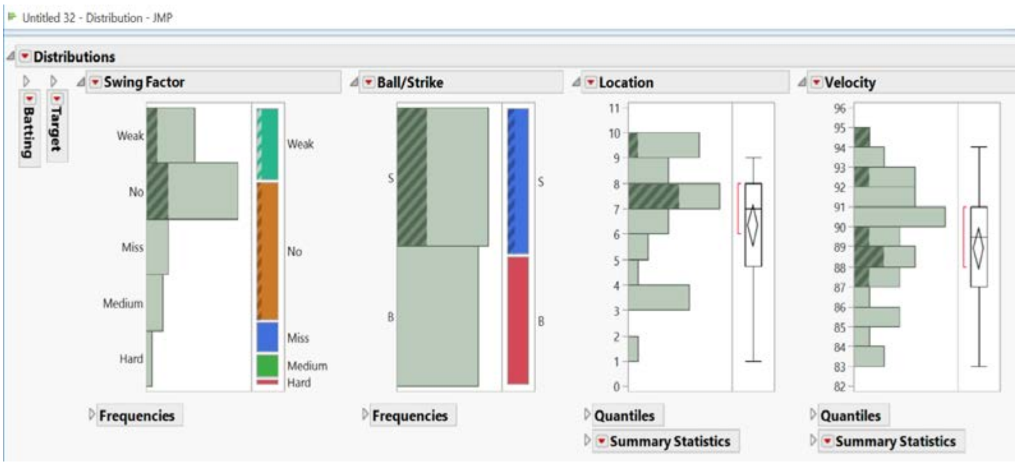


Figure 9-1: Interactive Distributions #1: Pitch Locations

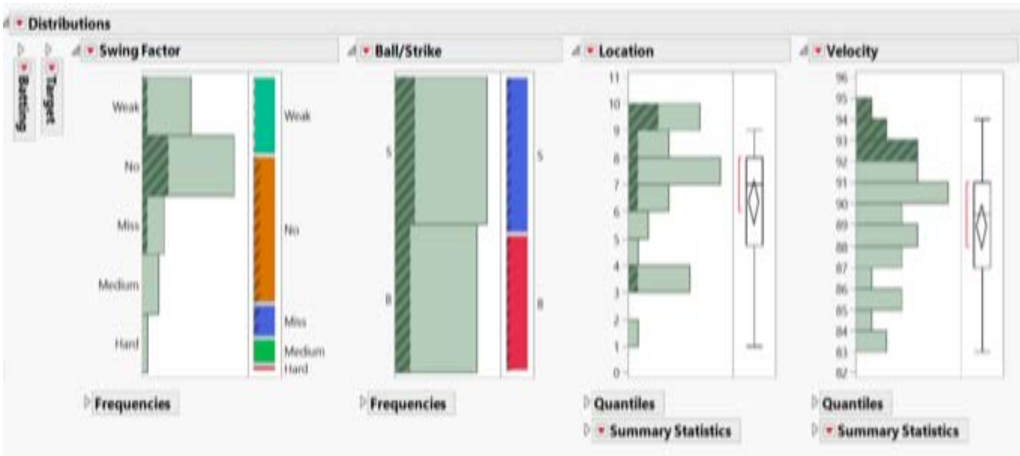


Figure 9-2: Interactive Distributions #2: Pitch Velocity

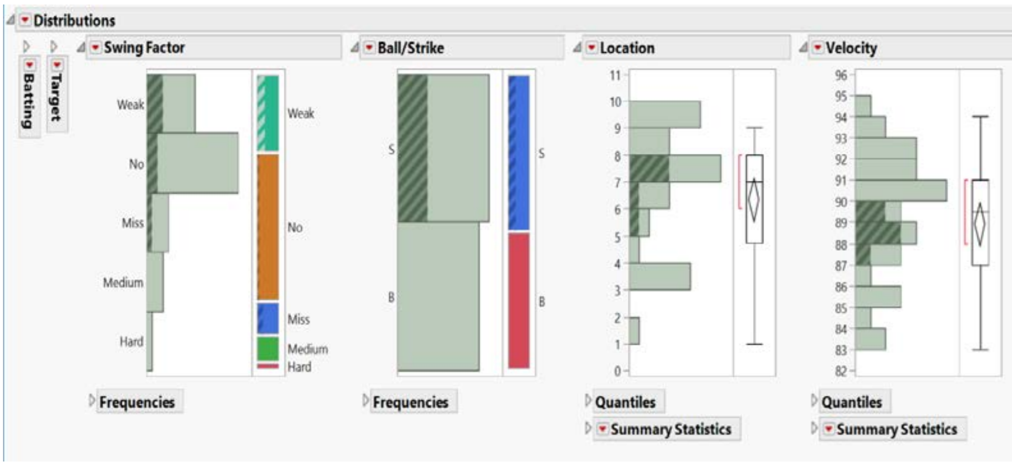
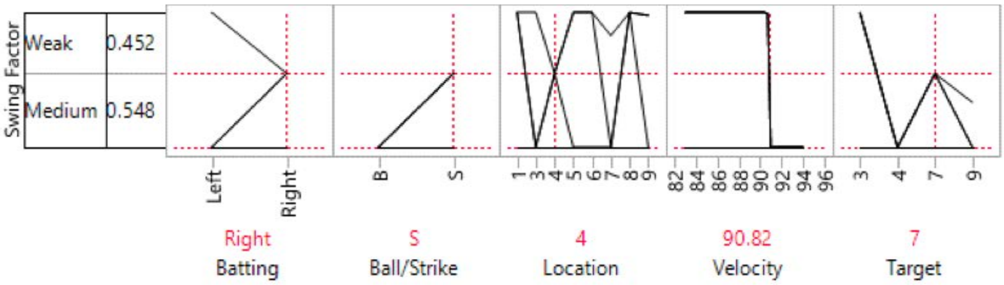


Figure 9-3: Interactive Distributions #3: Pitch Type

The way of Regression Predictive Model works is put in all the collected data then the model will show the probability of next SQF at the cross-section at any particular interest scenario. For example, if we only want to focus on Pitching Location as the Interactive Distributions. The model will tell us how effective the next pitch would be for any particular pitchers. The model can also focus on the effectiveness of velocity and pitching types, and show us the percentage of probability of SQF (Weak Contact, No Swing, Swing and Miss, Medium contact, or Hard Contact.)



Fit Model & Profiler: Swing Contact Quality

4. Conclusions and Future Work

By introducing Swing Quality Factors, we make it possible to analyze the effectiveness of each pitch by studying the interactions between pitchers and hitters in sequential at each bat. SQF is also capable of identifying the most effective pitch choice in critical situation. Our current work shows the correlations between SQF and location, ball/strikes, type of pitches and velocity. Last but not least, the JMP Interactive Distribution Platform can help to discover the batting pattern in a favor to the pitchers (pitch location, pitch velocity, pitch type).

The most important breakthrough of this project is that the JMP Fit Model and Profiler Platform can help predict the probability of the Swing Contact Quality at any particular pitching scenario to help building pitchers' confidence by knowing their winning pitches. SQF even has future potential which baseball professionals can further utilize JMP Analytics capability to analyze Baseball Analytics using the conditional probability to optimize the Pitching Sequence.

Acknowledgement

I would like to thank my statistic advisor Dr. Charles Chen, baseball coach Tony Liao, and my working partner Mason Chen for helping me and giving me advices throughout this project.

References

Fast, M. (2011a). Spinning yarn: can we predict hot and cold zones for hitters? In Baseball Prospectus, http://www.baseballprospectus.com/article.php?articleid=15363.\

Fast, M. (2011b). Spinning Yarn: A Zone of Their Own. Retrieved from https://www.baseballprospectus.com/news/article/14572/spinning-yarn-a-zone-of-their-own/

Moore, C. (2009). Best fastballs in baseball. In Baseball Analysts: Rich Lederer & Friends, http://baseballanalysts.com/archives/2009/09/.

P. Swartz, M. Grosskopf, D. Bingham & Tim B. Swartz (2017) The

Quality of Pitches in Major League Baseball, The American Statistician, 71:2, 148-154, DOI: 10.1080/00031305.2016.1264313

Roegele, J. (2013a). Strike zone pitch quality, part I: location. In SB Nation: Beyond the Box Score, http://www.beyondtheboxscore.com/2013/3/29/4153440/strike-zone-pitch-qualitypart-i-location-pitchfx-sabermetrics.

Roegele, J. (2013b). 2013 Closers: pitch location vs velocity. In SB Nation: Beyond the Box Score, http://www.beyondtheboxscore.com/2013/4/5/4171286/2013-closers-pitch-location-vsvelocity-strike-zone-quality-pitchfx-sabermetrics.

Wilson, J. (2015). Pitch quantification part 1: between-pitcher comparisons of QOP with conventional statistics, http://www.qopbaseball.com.