13TH ANNUAL
MIT SLOAN
SPORTS ANALYTICS CONFERENCE
MARCH 1 - 2, 2019 | BOSTON, MA
presented by ESPN

Predicting Major League Baseball Strikeout Rates from Differences in Velocity and Movement Among Player Pitch Types

Baseball Track
Eric P. Martin[1]

1. Introduction

Baseball enthusiasts frequently focus on how individual pitch characteristics, like fastball velocity and curveball movement, impact pitching results. Studying these elements in isolation, however, fails to account for relationships among a pitcher's different pitches. For example, a pitcher's curveball is more effective if its movement is dramatically different than his fastball. The effectiveness of these pitches is dependent on each other. The relationships among all pitches in a pitcher's arsenal—rather than the characteristics of a single pitch—are a larger determinate of success [1]–[4].

Using PITCHf/x information for over 2.5 million MLB pitches thrown by 402 pitchers from 2012 through 2017, this paper predicts pitcher strikeout percentages by examining differences in velocity, movement, and release points among each of their pitch types. The best performing model has a mean absolute error of 2.94 percentage points from a pitcher's actual strikeout percentage. Velocity attributes and differences in vertical movement among pitches have the most significant impact on a pitcher's strikeout percentage. Understanding these performance drivers enables players and coaches to both target the pitch elements most likely to increase strikeouts and to identify promising young pitchers for development.

2. Background

In this analysis, performance is evaluated by pitcher strikeout percentage. Compared to other rate measures, strikeout percentage is one of the most reliable and consistent pitching statistics and the statistic least likely to be affected by chance or a team's defensive ability [5], [6].

In 2002, Gray [7] conducted research to understand how differences among pitches induce strikeouts. His study examined college baseball players swinging at computer-generated images of baseballs having varying speeds and movement. Gray found that hitting was "nearly impossible in a situation where pitch speed is random and in which no auxiliary cues (e.g., pitcher's arm motion or pitch count) are available to the batter" [7, p. 1140]. Batters suffered (and pitchers benefited) as

[1] The E15 Group, Chicago, IL 60091. Email: ermartin@e15group.com or epm145@gmail.com
This research was the thesis subject for a master's degree in predictive analytics from Northwestern University and is available at https://github.com/epm145/MLB-K-Predictions.git

the number of different pitch speeds increased. The study noted the importance for pitchers to learn to throw at least three different pitch types [7], [8].

Subsequent analyses using MLB PITCHf/x data examined the interrelationship between velocity and pitch movement and its effect on swing-and-miss rates [9], swinging strikes [10], batter contact rates [11], and expected runs [12]. Other attempts have been made to quantify how a pitcher's pitch repertoire induces outs. Different "arsenal scores" have been developed to evaluate the effectiveness of a pitcher's collection of pitches [13]–[15]. While their methods varied, each calculation aggregated the impacts of individual pitch types to create a single measure of effectiveness. Unlike the present analysis, however, these studies either did not specifically examine the combined interrelation among a pitcher's entire arsenal of pitches or explore how pitch characteristics such as velocity or movement affect strikeout rates.

Healy, Zhao, and Brooks [16] developed a pitch sequencing model examining a pitcher's strikeout rate as a function of pitch velocity and movement. They discovered pitcher strikeout rates increased when both fastball velocity and vertical movement increased [16]. The study's conclusion suggests "a more detailed model could include information about the number, frequency, and physical properties of a pitcher's off-speed pitches and how well these pitches complement each other and the pitcher's fastball" [16, p. 101]. This proposition was the catalyst for this paper.

3. Methods

This paper sources PITCHf/x information from 2012 through 2017. Pitcher statistics from each season are collected from Baseball-Reference.com [17]. Model-based clustering techniques are used to group pitches from each pitcher based on velocity, horizontal movement, and vertical movement. Several statistical models, trained using data from the 2012–2016 seasons, employ supervised and unsupervised machine learning techniques to predict pitcher strikeout rates. The 2017 season is used as the test set to evaluate each model's accuracy.

Short relief pitchers are excluded from the analysis. Their tendency to enter games in favorable match-ups often biases their statistics. This analysis therefore only includes pitchers who both faced an average of 10 batters per appearance and pitched at least 1,000 pitches in a season. The resulting training set contains 894 observations for 359 different pitchers from 2012–2016 while the 2017 test set includes 170 pitchers. Each observation in the dataset represents pitching information from a full MLB season in an effort to normalize both opposing batter talent and potential measurement discrepancies across stadiums due to systematic errors in stadium PITCHf/x and Statcast measurement systems [18]–[24].

While this analysis focuses on the relationship among each pitcher's pitch types, it ignores other influential pitching effects that induce outs. For example, pitch location and sequencing are not evaluated here, although they undoubtedly affect pitcher performance. (See [7], [25]–[27]). Future extensions of this analysis may seek to incorporate these elements.

3.1. Identifying Pitch Types – Clustering

Identifying different pitch types can be difficult. For example, a sinker from one pitcher may move the same as another pitcher's four-seam fastball [25]. PITCHf/x pitch type designations are sometimes inconsistent and can label the same ball movement differently among pitchers [18]. Nevertheless, whether a pitch is labeled a sinker or a four-seam fastball is irrelevant to a hitter. Hitters are simply concerned with velocity and movement. What matters is a pitcher's ability to deceive a hitter by changing the velocity, movement, and location of each pitch.

This analysis determines each pitcher's pitch repertoire by using the model-based clustering algorithm introduced by Pane et al. [28] instead of PITCHf/x pitch classifications. Pane et al. [28] determined that model-based clustering more accurately identifies MLB pitch types than either k-means or neural network clustering. Model-based clustering is also better able to identify pitches with differing variances and often reduces the number of small clusters.

In this analysis, pitches are assigned to clusters according to velocity, horizontal movement, and vertical movement using an agglomerative hierarchical clustering method based on maximum likelihood criteria for parameterized Gaussian mixture models [29]–[31].[2] Each pitcher's pitches are separated into nine alternative cluster configurations: all pitches are assigned to either one, two, three, etc., pitch clusters. As set forth by Pane et al. [28], Bayesian Information Criterion (BIC) values for each alternative cluster configuration are adjusted using penalties based on the number of clusters and high intra-cluster correlation coefficients. This reduces the number of small clusters and provides a more reliable representation of each pitcher's pitch types. Accordingly, the present analysis uses the cluster configuration for each pitcher with the lowest adjusted BIC value.

Identifying outlier pitches in a model-based clustering method is especially important since the number of clusters and their variance structures are unknown [28]. This analysis identifies outliers by first determining each cluster's mean using the minimum covariance determinant method developed by Rousseeuw and Van Driessen [34]. The Mahalanobis distances of the points from each cluster's mean are used to create 97.5% Gaussian confidence ellipsoids [34]–[36]. Pitches more than two standard deviations from their cluster's mean are removed as outliers. Between 3–6% of each player's pitches are therefore removed ($\mu = 4.5\%$, $\sigma = 0.4\%$). Pitch clusters are deleted if they contain fewer than either 10 pitches in a season or one pitch per game appearance. Finally, three pitchers in the dataset have only two pitch clusters and, due to their low incidence, are excluded from this analysis.

Table 1. Number of Pitchers with 3–9 Pitch Clusters in the Training and Test Sets

	Number of pitch clusters						
Data	3	4	5	6	7	8	9
Training	105	287	235	114	79	42	32
Test	32	72	39	18	5	2	2
Total	137	359	274	132	84	44	34

[2] To adjust for changes PITCHf/x made reporting velocity in 2017, velocities in the data set for 2017 are adjusted to indicate values 50 feet from home plate [32], [33].

Table 1 lists the distribution of pitch clusters in the training and test sets. Almost 60% (59.5%) of the pitchers have either four or five pitch types. Notably, a relatively large proportion of pitchers (15.2%) have pitches grouped into seven or more clusters. This is unusually high, especially since the PITCHf/x classification system identifies only 6.1% of observations having seven or more pitch types. The high number of pitch clusters in this analysis is likely due to PITCHf/x measurement differences across stadiums—especially in earlier seasons when differences were more pronounced [18]–[24], [37]. Specifically, 72.8% of the observations with 7–9 pitch clusters occurred in the 2012 or 2013 seasons compared to only 27.2% in the four seasons from 2014–2017. Moreover, players with the Detroit Tigers (11) and St. Louis Cardinals (9) had the most pitchers with seven or more pitch clusters. In 2013, these two teams' stadiums had the largest PITCHf/x measurement discrepancies relative to the rest of the league [37]. Future iterations of this analysis may seek to control for these PITCHf/x park effects.

3.2. Distance Between Pitch Clusters

Measuring the distance between pitch clusters is one way to quantify differences between pitch types. Any measure of the difference between pitches must account for natural correlations between measurement axes—in this case velocity, horizontal movement, and vertical movement. Healey, et al. [38] demonstrated why the Mahalanobis distance is best suited to measure differences in velocity and movement between pitches. Gravity's effect on slower moving objects causes them to drop more than their higher velocity analogues. Velocity and vertical movement are therefore highly correlated (r = 0.68). The Mahalanobis distance accounts for this by dividing the standardized version of each value by the covariance matrix [39], [40]. The result is a unitless measure of the Mahalanobis distance (D) for each point (x) defined as:

$$D = (x\text{-}m)^{\mathrm{T}} C^{-1} (x\text{-}m) \quad (1)$$

where m is the vector of mean variable values and C^{-1} is the inverse covariance matrix of the variables [28].

When calculating Mahalanobis distances, separate correlation matrixes are used for left-handed and right-handed pitchers. The opposite arm angles from these pitchers result in natural differences in horizontal movement. Pitches by left-handed pitchers move left horizontally (from the catcher's perspective) while pitches from right-handed pitchers move in the opposite direction, with the effects more pronounced at lower velocities (Table 2).

Table 2. Velocity and Movement Correlation Matrixes for Left- and Right-Handed Pitchers

Left-Handed Pitchers			
	Velocity	Horizontal	Vertical
Velocity	1.000	0.523	0.687
Horizontal	0.523	1.000	0.523
Vertical	0.687	0.523	1.000

Right-Handed Pitchers			
	Velocity	Horizontal	Vertical
Velocity	1.000	-0.574	0.685
Horizontal	-0.574	1.000	-0.519
Vertical	0.685	-0.519	1.000

3.3. Independent Variables

The mean velocities, horizontal movements, vertical movements, and pitch release points for each pitch cluster are calculated across a pitcher's group of pitch types. For each of these measures,

variables are created to indicate the minimum, maximum, and range of the cluster means for every pitcher. A pitcher's maximum velocity is therefore the highest mean cluster velocity among a pitcher's pitch types rather than the highest velocity single pitch thrown that season. A related variance measure also calculates the size of the interquartile range (IQR) between the 25th and 75th percentiles for all pitches thrown by a pitcher for each of the above measures—regardless of pitch type.

Variables identifying pitch release points are included based on pitch tunneling research to account for visual cues signaling the velocity or movement of an incoming pitch [41]. As this analysis examines the relationship between pitch movements rather than the ordinal direction of movement, measurements are not separated based on either batter or pitcher handedness.

The maximum and average Mahalanobis distances between all pitch clusters are also measured. A weighted measure of the Mahalanobis distances among all pitch types is also created to aggregate differences between pitch types into a single variable. The cluster with the most pitches (the top cluster) serves as the anchor from which all pitches are measured. The Mahalanobis distances of each pitch cluster from the center of the top cluster are weighted by pitch frequency and added together (Weighted Mahalanobis Distance or WMD):

$$\text{WMD} = \sum_{2}^{k}\left[\text{D}_k\left(\frac{\text{n}_k}{\text{n}_{total}}\right)\right] \quad (2)$$

Where *D* = Mahalanobis distance from the center of the cluster of pitches thrown most frequently, *k* = number of clusters, and *n* = number of pitches. There is a small correlation between strikeout percentage and WMD (r = 0.16) suggesting that the most frequently thrown pitch serves to set-up the remaining pitches.

An entropy variable is created to measure uncertainty caused by both the number of pitch types and the frequency with which each is thrown. The equation introduced by Shannon [42][43] is used to calculate each pitcher's entropy value:

$$\text{Entropy} = \sum_{i=1}^{k}\frac{n_i}{n_{total}}\left(log_2\frac{n_i}{n_{total}}\right) \quad (3)$$

Where *n* = number of pitches, *k* = number of clusters, and *i* is the cluster number.

Since strikeout percentage is highly correlated with strike rate (r = 0.30), pitchers with the same pitch characteristics may have different strikeout percentages simply because one pitcher throws more strikes than the other. As this analysis seeks to isolate the impact of differences in velocity and movement on strikeout percentages, accuracy differences are considered by including a variable identifying pitcher strike rates. Similarly, an indicator variable identifies whether pitchers played in the National League, American League, or both each season since National League pitchers generally have higher strikeout percentages from facing opposing pitchers in the lineup.

Finally, cross-validation within the training data reveals that logarithmically transforming the independent variables produces the lowest in-sample MAE. Natural logarithms of the independent variables are therefore used to improve normality and increase prediction accuracy.

3.4. Models Examined

Thirteen types of models are evaluated using strikeout percentage as the dependent variable: multiple linear regression without (MLR) and with (MLR+) interaction terms, lasso and ridge shrinkage regression methods, principal components (PCR) and partial least squares (PLS) regressions, polynomial regression, regression splines, generalized additive models (GAM), random forests, boosted tree-based models, neural networks, and support vector machines (SVM). As the dependent variable is a proportion, predictions below zero or above one are capped to ensure predictions remain within an acceptable range. None of the predictions in the test set fell outside this range.

The polynomial, spline, and GAM models use prediction errors from k-fold cross-validation to identify the best iteration of each independent variable to model. For example, each variable is individually tested against strikeout percentage in a simple linear regression model using one to five polynomial orders. The order of each variable with the lowest in-sample prediction error is used in the polynomial model. The spline model uses the same technique to identify the optimal number of degrees of freedom for either natural or smoothing spline variables. The GAM model uses either the polynomial, natural spline, smoothing spline, or local regression (LOESS) version of each variable with the lowest in-sample error. Stepwise AIC selection is used in the MLR, MLR+, polynomial regression, regression spline, and GAM models to identify independent variables. Variables with high multicollinearity are removed from each model to ensure all variables have variance inflation factor values less than 10. Finally, several models use k-fold cross-validation to identify optimal hyperparameters including: the best lambda tuning values for the ridge and lasso regression models; the number of trees in the random forest model; the variable interaction depths and number of trees in the boosted model; the number of components used in the PCR and PLS models; the number of hidden layers to use in the neural net model; and the optimal cost and gamma values for the SVM model.

4. Results

4.1. Model Results

Many statisticians agree that fastball velocity and strike rates affect strikeout percentages. A linear regression model with these two independent variables is used as a standard/control with which to measure the success of the present analysis over established metrics. Using the training data to fit the model, and strike percentage as the dependent variable, the control model has an adjusted r-squared value of 0.2297 and a MAE of 0.0331 against the test set.

The models in this analysis have MAE values between 0.0294 and 0.0324 against the test set (Table 3). The random forest model performs the best with a MAE of 0.0294—or 2.94%—which is 0.0037 points (0.37%) less than the control model.

Table 3. Test Set Errors for All Models

Model	MAE
Random Forest	0.0294
Multiple Linear Regression	0.0298
Principal Components Regression	0.0299
Boosted	0.0300
Ridge Regression	0.0300
Lasso Regression	0.0301
Regression Spline	0.0301
Multiple Linear Regression with Interaction Terms	0.0304
Partial Least Squares Regression	0.0305
Polynomial Regression	0.0308
Neural Network	0.0309
Generalized Additive Models	0.0310
Support Vector Machine	0.0324
Control Model: Strike Rate + Maximum Velocity	0.0331

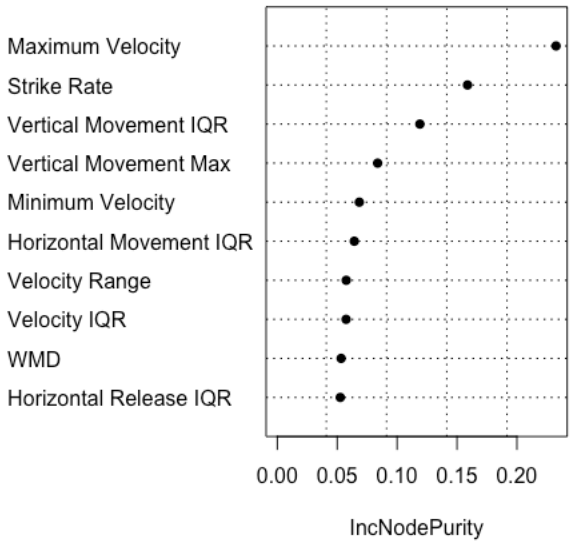

Figure 1. Random forest model variable importance plot (top ten variables).

Maximum velocity, strike rate, and vertical movement IQR have the greatest impact on strikeout percentage (Figure 1). The influence of high pitch velocity and the ability to throw strikes on a pitcher's strikeout percentage is consistent with both intuition and previous research [44], [45]. Somewhat surprising, however, is the importance of vertical movement relative to other pitch characteristics. Vertical movement IQR and maximum vertical movement are more important predictors of strikeout percentage than the ability to change speeds (i.e., velocity range and velocity IQR). In fact, the correlation with strikeout percentage is almost as high for vertical movement IQR ($r = 0.27$) as it is for a pitcher's strike rate ($r = 0.29$).

Conversely, there is no correlation between the number of pitch types and a pitcher's strikeout percentage (r = -0.01). What matters more is the maximum velocity, strike rate, and vertical movement of a pitcher's pitches.

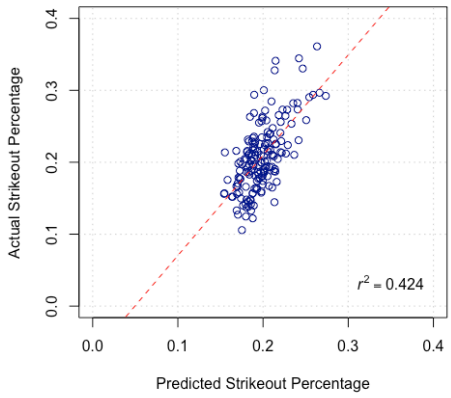

Figure 2. Predicted strikeout percentages against actual test set values

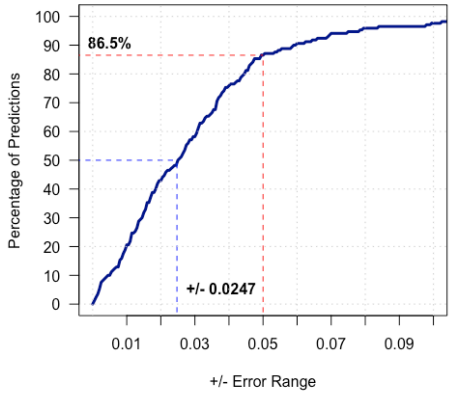

Figure 3. Percentage of model predictions within error ranges.

Figure 2 shows predicted verses actual strikeout percentages in the test set with an adjusted r-squared value of 0.424. A large number of predictions are within five percentage points of their actual values (Figure 3). Of the 170 predictions, 147 (86.5%) are within five points of the actual strikeout percentages with a median of 2.47 percentage points. Table 4 lists the 10 predictions with the lowest errors against the test set with each having an absolute prediction error of 0.21 percentage points or less.

Table 4. Top 10 Random Forest Model Predictions

	2017 Strikeout Rates (%)		
Player	Prediction	Actual	Absolute Error
Chad Kuhl	21.05	21.07	0.02
Cole Hamels	17.02	17.07	0.05
Jhoulys Chacin	20.20	20.13	0.07
Chris Stratton	19.84	19.92	0.08
Trevor Williams	18.40	18.28	0.12
Chad Bell	19.45	19.59	0.14
Zack Wheeler	21.16	20.98	0.18
Sam Gaviglio	15.42	15.61	0.18
Daniel Norris	18.64	18.82	0.18
Mike Leake	16.87	16.67	0.21

5. Conclusions

The factors having the greatest impact on a pitcher's strikeout rate are maximum velocity, strike rate, and differences in vertical movement among pitches. Vertical movement differences are more influential than differences in velocity, horizontal movement, or pitch release points among pitches. Moreover, the number of pitch types does not impact a pitcher's strikeout percentage. This suggests a pitcher looking to increase his strikeout percentage should prioritize maximizing the break of his curveball or increasing the rise on his fastball before looking to add a cutter or slider to his arsenal. Similarly, all other things being equal, pitchers with a plus fastball and plus curveball may be more effective at striking out batters than those with plus pitches having less vertical movement.

In addition to identifying the pitch characteristics most likely to increase strikeouts, the insights in this analysis may be used to identify potentially high strikeout pitchers with otherwise unremarkable statistics. Within a game, this analysis may also help managers determine when to remove a pitcher. As a pitcher increases his pitch count, managers can monitor the variables most likely to affect that pitcher's strikeout rate.

Going forward, several additional analyses could build upon these findings. Do the factors influencing strikeout percentage change based on a batter's time through the batting order? Which elements are most indicative of a relief pitcher's strikeout percentage? To what extent does a batter's contact rate impact the efficacy of these pitch characteristics? Understanding these effects would further the understanding of how relationships among a pitcher's pitch types impact his ability to strikeout batters. These findings could also be used to identify the best situational pitchers and adjust pitching strategies based on the match-up.

References

[1] Branch, J. (2015, Oct. 4). Baseball talk, and all that stuff. *The New York Times* (New York print ed.), A1.

[2] Sarris, E. (2018, Jan. 23). What Jack Flaherty has in common with Clayton Kershaw. *Fangraphs*. Retrieved from https://www.fangraphs.com/blogs/what-jack-flaherty-has-in-common-with-clayton-kershaw/.

[3] Davis, E. (2016, Aug. 11). Greg Maddux was a power pitcher despite the low velocity. *SB Nation Beyond the Boxscore*. Retrieved from https://www.beyondtheboxscore.com/2016/8/11/12423936/greg-maddux-velocity-finesse-power-pitcher-no-hope-for-batters.

[4] Rescan, A. (2017, Oct. 12). Kyle Hendricks's greatness is about more than control and command: His velocity-less success depends on the movement, too. *Beyond the Boxscore*. Retrieved from https://www.beyondtheboxscore.com/2017/10/12/16464244/kyle-hendricks-cubs-game-five-nlds-velocity-control-command-movement.

[5] Woolner, K. & Perry, D. (2006). Why are pitchers so unpredictable? [In] *Baseball Between the Numbers, Why everything you know about the game is wrong* (pp. 48-57). Keri, J. (ed.). Basic Books:New York, NY.

[6] Albert, J. (2006). Pitching statistics, talent and luck, and the best strikeout seasons of all-time. *Journal of Quantitative Analysis in Sports*, 2(1), Article 2. https://doi.org/10.2202/1559-0410.1014.

[7] Gray, R. (2002). Behavior of college baseball players in a virtual batting task. *Journal of Experimental Psychology: Human Perception and Performance*. 28(5), 1131-1148. http://doi.org/10.1037//0096-1523.28.5.1131.

[8] Ryan, N., & House, T. (1991). *Nolan Ryan's pitcher's bible: The ultimate guide to power, precision, and long-term performance*. New York, NY: Simon & Schuster.

[9] Hale, J. (2013, Oct. 30). Baseball ProGUESTus: Is speed enough?: A PITCHf/x look at the effect of fastball velocity and movement. Baseball Prospectus. Retrieved from https://www.baseballprospectus.com/news/article/22139/baseball-proguestus-is-speed-enough-a-pitchfx-look-at-the-effect-of-fastball-velocity-and-movement/.

[10] Roegele, J. (2014, Nov. 24). The effects of pitch sequencing. *The Hardball Times*. Retrieved from https://www.fangraphs.com/tht/the-effects-of-pitch-sequencing/.

[11] Carleton, R. A. (2015, Feb. 3). Baseball therapy: The power of changing speeds. *Baseball Prospectus*. Retrieved from https://www.baseballprospectus.com/news/article/25494/baseball-therapy-the-power-of-changing-speeds/.

[12] Bonney, P. (2015, Mar. 6). Defining the Pitch Sequencing Question. *The Hardball Times*. Retrieved from https://www.fangraphs.com/tht/defining-the-pitch-sequencing-question/.

[13] Sarris, E. (2014, Dec. 16). Toward a pitching arsenal score statistic. *Rotographs*. Retrieved from https://www.fangraphs.com/fantasy/toward-a-pitch-arsenal-score-ranking-statistic/.

[14] Schwartz, D. (2014, Dec. 19). Pitch arsenal score part deux. *Rotographs*. Retrieved from https://www.fangraphs.com/fantasy/pitch-arsenal-score-part-deux/.

[15] Jackman, S. (2015). Pitch arsenal scores. *The Hardball Times*. Retrieved from https://www.fangraphs.com/tht/pitch-arsenal-scores/.

[16] Healey, G. & Zhao, S. (2017c). Using PITCHf/x to model the dependence of strikeout rate on the predictability of pitch sequences. *Journal of Sports Analytics*, 3, 93-101. http://doi.org/10.3233/JSA-170103.

[17] Sports Reference LLC. (2018). *Baseball-Reference.com - Major League Statistics and Information*. Retrieved from https://www.baseball-reference.com/.

[18] Fast, M. (2010b, June 17). The Internet cried a little when you wrote that on it. *The Hardball Times*. Retrieved from https://www.fangraphs.com/tht/the-internet-cried-a-little-when-you-wrote-that-on-it/.

[19] Fast, M. (2011, Mar. 2). Spinning yarn: How accurate is PitchTrax? *Baseball Prospectus*. Retrieved from https://www.baseballprospectus.com/news/article/13109/spinning-yarn-how-accurate-is-pitchtrax/.

[20] Garik. (2011, Feb. 10). Being cautious with using Pitchf/x data to evaluate stuff: The case of Kyle Drabek. *SB Nation Beyond the Boxscore*. Retrieved from https://www.beyondtheboxscore.com/2011/2/10/1982529/being-cautious-with-using-pitchf-x-data-to-evaluate-stuff-the-case-of.

[21] Marchi, M. (2011, Feb. 25). Fine tuning PITCHf/x location data. *The Hardball Times*. Retrieved from https://www.fangraphs.com/tht/fine-tuning-pitchf-x-location-data/.

[22] Arthur, R. (2017, Apr. 28). Baseball's new pitch-tracking system is just a bit outside: As MLB switches from PITCHf/x to Statcast, the new tool is going through growing pains. *FiveThirtyEight*. Retrieved from https://fivethirtyeight.com/features/baseballs-new-pitch-tracking-system-is-just-a-bit-outside/.

[23] Kagan, D. (2018, Jan. 23). The physics of RoboUmp. *The Hardball Times*. Retrieved from https://www.fangraphs.com/tht/the-physics-of-roboump/.

[24] Boyle, W., O'Rourke, S., Long, J. & Pavlidis, H. (2018, Jan. 29). Robo strike zone: It's not as simple as you think. *Baseball Prospectus*. Retrieved from https://www.baseballprospectus.com/news/article/37347/robo-strike-zone-not-simple-think/.

[25] Healey, G., Zhao, S. & Brooks, D. (2017a, July 10). Measuring pitcher similarity. *Baseball Prospectus*. Retrieved from https://www.baseballprospectus.com/news/article/32199/prospectus-feature-measuring-pitcher-similarity/.

[26] Sidle, G. & Tran, H. (2018). Using multi-class classification methods to predict baseball pitch types. *Journal of Sports Analytics*, 4(1), 85-93.

[27] Trueblood, M. (2018, Feb. 1). Rubbing mud: The Cubs have already mined these tunnels. *Baseball Prospectus*. Retrieved from https://www.baseballprospectus.com/news/article/37461/rubbing-mud-cubs-already-mined-tunnels/.

[28] Pane, M. A., Ventura, S.L., Steorts, R.C., & Thomas, A.C. (2013). Trouble with the curve: Improving MLB pitch classification. *arXiv:1304.1756v1 [stat.AP]*. Retrieved from https://arxiv.org/pdf/1304.1756.pdf.

[29] Fraley, C., Raftery, A.E., Murphy, T.B., & Scrucca, L. (2012). mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. *Technical Report No. 597, Department of Statistics, University of Washington*. Retrieved from https://www.stat.washington.edu/sites/default/files/files/reports/2012/tr597.pdf.

[30] Fraley C., & Raftery A. E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97(458), 611-631.

[31] Evans, K., Love, T., & Thurston, S. W. (2015). Outlier identification in model-based cluster analysis. *Journal of Classification*, 32, 63-84. http://doi.org/10.1007/s00357-015-9171-5.

[32] Cameron, D. (2017, Apr. 4). About all these velocity spikes. *Fangraphs*. Retrieved from https://www.fangraphs.com/blogs/about-all-these-velocity-spikes/.

[33] Nathan, A. & Brooks, D. (2017, Apr. 5). Prospectus Feature: Estimating Release Point Using Gameday's New Start-Speed. *Baseball Prospectus*. Retrieved from https://www.baseballprospectus.com/news/article/31529/prospectus-feature-estimating-release-point-using-gamedays-new-start-speed/.

[34] Rousseeuw, P., & Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41, 212-223.

[35] Hardin, J. & Rocke, D.M. (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. *Computational Statistics & Data Analysis*, 44, 625-638.

[36] Hardin, J. & Rocke, D.M. (2005). The distribution of robust distances. *Journal of Computational and Graphical Statistics*, 14(4), 928-946. http://doi.org/10.1198/106186005X77685.

[37] Roegele, J. (2013, Sept. 13). Basic 2013 PITCHf/x velocity park effects: Calculating basic PITCHf/x velocity park effects and discussing the sources of error that are inherent in the numbers. *SB Nation Beyond the Boxscore*. Retrieved from https://www.beyondtheboxscore.com/2013/9/13/4720852/basic-2013-pitchfx-velocity-park-effects-error-sabermetrics.

[38] Healey, G., Zhao, S. & Brooks, D. (2017b). Measuring pitcher similarity: Technical details. *viXra.org*. Retrieved from http://vixra.org/pdf/1705.0098v1.pdf.

[39] Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*. 2(1), 49-55.

[40] De Maesschalck, R., Jouan-Rimbaud, D., Massart, D.L. (2000). Tutorial: The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50, 1–18.

[41] Pavlidis, H., Judge, J., & Long, J. (2017, Jan. 24). Prospectus feature: Introducing pitch tunnels. *Baseball Prospectus*. Retrieved from https://www.baseballprospectus.com/news/article/31030/prospectus-feature-introducing-pitch-tunnels/.

[42] Shannon, C. (1948a). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(3), 379-423, 27(4), 623-656, July, October, 1948.

[43] Shannon, C. (1948b). A Mathematical Theory of Communication (continued). *The Bell System Technical Journal*, 27(4), 623-656.

[44] Arthur, R. (2014, Feb. 6). Baseball proGUESTus: Entropy and the eephus. *Baseball Prospectus*. Retrieved from https://www.baseballprospectus.com/news/article/22758/baseball-proguestus-entropy-and-the-eephus/.

[45] Cameron, D. (2009, Feb. 17). Velocity and K/9. *Fangraphs*. Retrieved from https://www.fangraphs.com/blogs/velocity-and-k9/.