Model Strike Probability for MLB Pitches

David Teuscher

December 11, 2021

Introduction

In Major League Baseball (MLB), there is a defined strike zone based on the rule book. In reality, the strike zone is an abstract concept that is determined and influenced by human judgement. Umpires have nothing to reference except their own sight when calling balls and strikes. As a result, there is variation that results in whether a pitch is called a strike or not. For many pitches, the human element is not a factor since pitches in the middle of the strike zone are essentially always called a strike and pitches more than a couple inches out of the strike zone are always called a ball. When pitches are located around the edge of the strike zone, there is more variation on whether or not the pitch is called a strike and these pitches can play an important role in the game. For example, if a batter is up and the count is 1 ball and 1 strike and the next pitch is taken on the outside corner, the at-bat will greatly be influenced by the call. The approach of the batter and pitcher will change a lot depending on whether the count is 1-2 or 2-1. As a result, it would be useful to estimate the probability of a pitch being called a strike based on location.

Since there is variation in calling strikes, catchers often will attempt to frame a pitch, meaning they try to catch the ball in a way that makes the pitch appear to be a strike even when it isn't. As mentioned previously, it is beneficial for the team if catchers are able to frame pitches well and get more called strikes because at-bats turn in favor of the pitcher if they are ahead of the batter. Also, umpires are likely to have different strike zones, with some umpires having a large strike zone and calling many of the borderline pitches as strikes, while others will have a small strike zone and call many of the borderline pitches as balls. From these observations, it appears that there are three important aspects in determining whether a pitch is a strike or not. First, the location of the pitch is the most important since regardless of umpire or catcher, a pitch down the middle of the strike zone will basically always be called a strike and a pitch 2 feet outside of the strike zone will always be a ball. When the location of a pitch is around the edge of the strike, the catcher and umpire are likely to both have an effect on whether the pitch is called a strike or not as well. In order to account for these three aspects, the goal of this analysis is to fit a model that will determine the probability of a strike based on the location, as well as accounting for the individual impact of catchers and umpires to determine the catchers who frame the best and worst and the umpires who have the largest and smallest strike zones.

The remainder of the paper is outlined as follows. The data and source used for the analysis will be introduced and explained. Then the models that were fit to accomplish the goals of the analysis will be outlined mathematically and justified by diagnostics and model comparison. Finally, the results of the models will be shown and the strengths and weaknesses of the selected model in accomplishing the goals will be outlined in the conclusion.

Data and Exploratory Data Analysis

The data for this analysis was taken from the 2021 MLB season and was obtained from Baseball Savant, which is a site that pulls in data from each MLB from MLB Advanced Media. The data here is the publicly available information about each pitch that MLB releases. For the 2021 season, there was over 700,000 pitches and there were around 300,000 pitches that were not swung at by the batter, which are used for this analysis.

There are a number of variables in the data set, but the ones that are used for this analysis are the horizontal location of a pitch, the vertical location of a pitch, whether a pitcher throws right or left handed, whether a batter hits right or left handed, the catcher, pitcher, umpire, and whether the pitch was called a strike or ball. These variables are provided for every pitch.

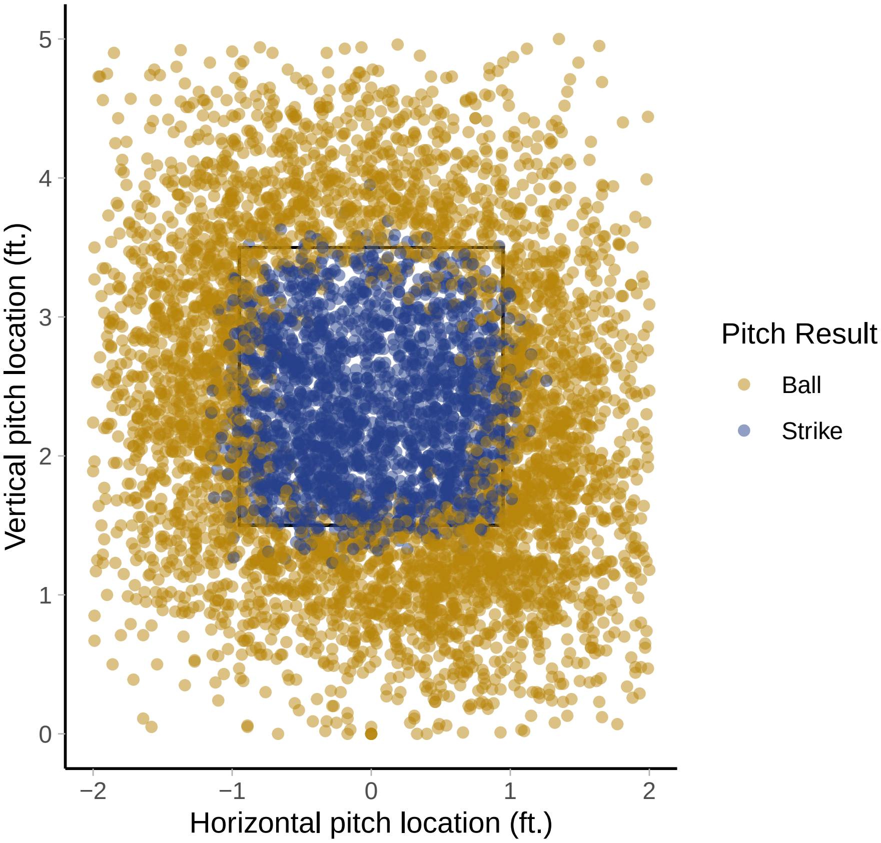

Figure 1: Location of 10,000 pitches from the 2021 MLB season showing the result of the pitch

A random sample of 10,000 pitches was plotted in Figure 1 to show the general relationship between location and the result of the pitch. The width of the strike zone is the exact width based on the rule book definition, but the height of strike zone is an average of all MLB players since the height of the strike zone is from a player's knees up to their chest. Pitches that are in the center of the strike zone or are far outside of the strike zone are always called strikes or balls, respectively, but around the edges of the strike zone there is a mixture of called balls and strikes, which shows that there is more variation in called balls and strikes around the edge of the strike zone.

Models

There were two models fit for this analysis. The first model is a generalized additive model (GAM) with smoothing components and coefficients. The second model is a generalized linear mixed model. The models are defined below:

Generalized Additive Model

$$y_i \sim Bernoulli(\pi_i)$$

$$g(\mu_i) = log(\frac{\pi_i}{1-\pi_i}) = \eta_i$$

$$\eta_i = \beta_0 + I(Throws_i = R)\beta_1 + I(Stands_i = R)\beta_2 + s(H_i, V_i)$$

where y_i is whether a pitch is called a strike or not, the intercept β_0 is for a left-handed pitcher and a left-handed batter, with coefficients for right-handed pitchers, β_1, and right-handed batters, β_2. A smoothing function, $s()$, is used for H_i and V_i, which is the horizontal and vertical locations of the ith pitch respectively and $s()$ is a thin plate regression spline.

Generalized Linear Mixed Model

$$Y_{ij} \sim Bernoulli(y_{ij}|\mu_{ij})$$

$$\mu_{ij} = E(y_{ij}|\mu_j)$$

$$g(\mu_{ij}) = \mathbf{x_i'}\beta + \mathbf{z_i}\mathbf{u}$$

where y_i is whether or not a pitch was called a strike and there is a fixed effect, β for the probability of the pitch being called a strike, which is the estimated probability of a pitch being a strike from the GAM previously defined. Random effects are included for the catcher, umpire, and the pitcher in the model.

Model Justification

There are a number of justifications that need to be made for the previous model. Inference was the goal of this analysis since it is not very useful to predict whether a certain pitch will be a strike. As a result, the predictive abilities of the models considered were not explored.

Given the shape of the strike zone, the relationship between the horizontal and vertical location and the probability of a pitch being a strike is a complex relationship, so the relationship is better represented by the smooth function instead of a linear component of the log odds of the probability of being a strike. The other variables, whether the batter hits left or right handed and whether the pitcher is right and left handed were included because it seems the the strike zone varies some based on where the pitcher is throwing from and where the batter is standing. Both of these variables were significant and when testing the change in deviance. The AIC for the model decreased by including these variables. As a result, it seemed appropriate to include these variables.

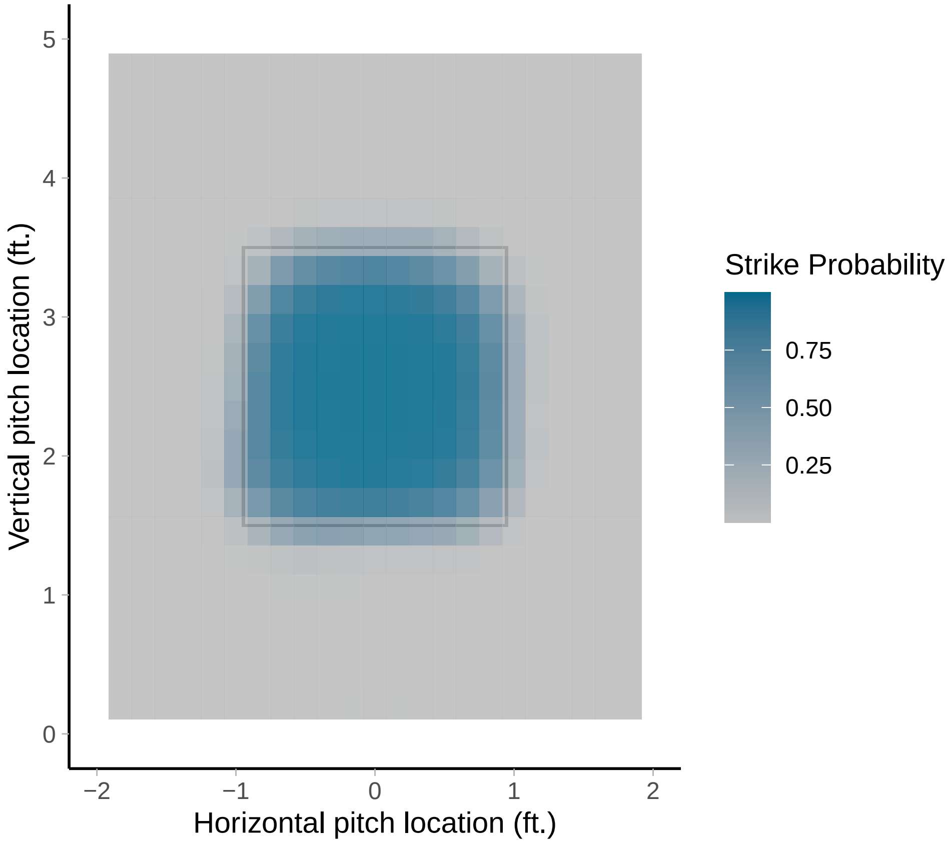

Figure 2: Estimated strike probability from the GAM

There are other variables that could have been included, but the model selected appeared to be a good model for modeling the probability of a strike as illustrated by Figure 2 and Figure 3 across all possible batter-pitcher combinations. Figure 2 shows that there is a high probability of a strike in the center of the strike zone and that probability is high for most of the strike zone and then decreases around the edges and quickly becomes almost zero not too far outside of the strike zone. Figure 3 shows the standard errors for the model, which show that there is a lot more uncertainty around the edges of the strike zone, which shows the variation that occurs in those areas that have been mentioned previously.

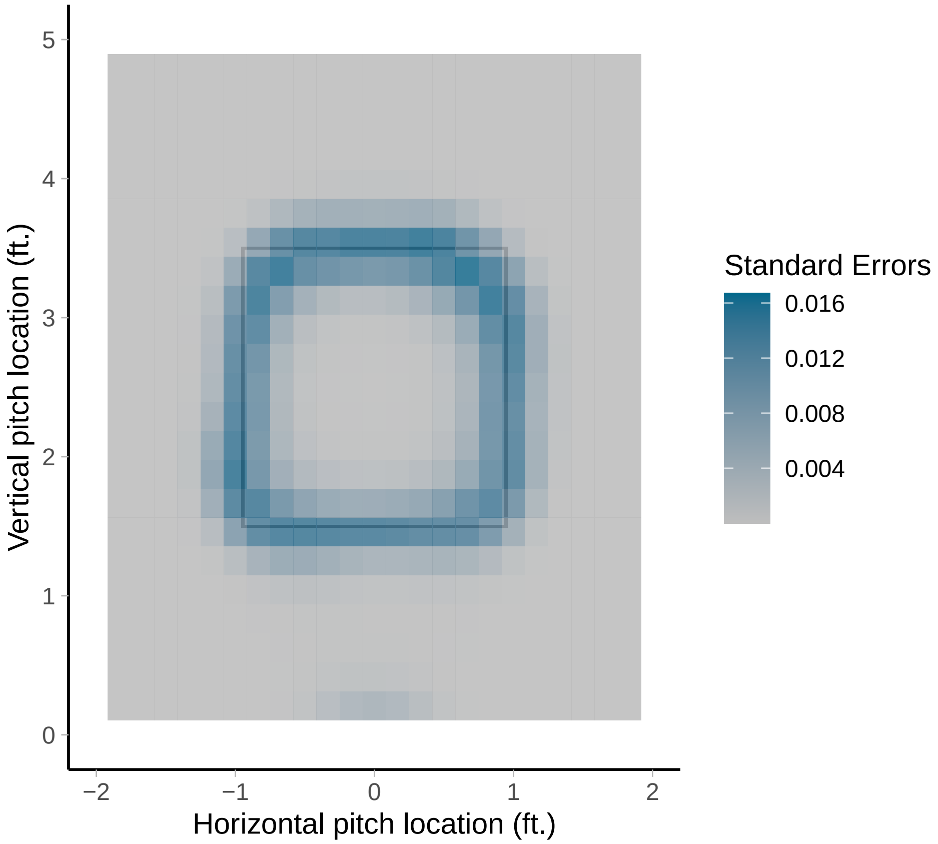

Figure 3: Standard errors of the estimates from the GAM

There were other link functions that were explored for both the GAM and GLMM, but there was minimal difference in the AIC of each model when using different link functions, so the logit link was used for this analysis simply because it is common and people are often familiar with the logit link. Any other link could be used and the same results given from the logit link used here should be expected.

After using the GAM to model the complex relationship between the location of a pitch and the probability that it is a strike, a generalized linear mixed model (GLMM) was used to estimate the individual effects of catchers and umpires. Often random effects are used because they are not of interest to the analyst, but in this scenario, the random effects are desired because they can show the individual difference from league average. As a result, the mixed model provides the capabilities of understanding the influence of both catchers and umpires on the probability of a pitch being called a strike.

The mixed model used the estimated probability of a pitch being a strike as a fixed effect in the model and then had random effects for the pitcher, umpire, and catcher. The reason for this was because there was difficulty including the smoothing term for location as part of the GLMM, but this approach has been taken in research as shown by Brooks. Consequently, there were no concerns about this approach of using the estimated probability from the GAM as a fixed effect in the GLMM

Results

The interesting results of the GAM model are shown in Figure 2. There is no reasonable interpretation of the smoothing parameter and the other parameters are not especially important since the main goal of the GAM is to estimate the probability of a strike based on the location of the pitch. The results from the GLMM are the more interesting results because they provide information about the best catchers for framing and umpires with large and small strike zones.

Table 1: Ten Highest Random Effects for Umpires and Catchers

Umpire	Random Effect	Catcher	Random Effect
Doug Eddings	0.7401611	Max Stassi	0.3806580
Lance Barrett	0.4010628	Jose Trevino	0.3054926
Bill Miller	0.3966738	Kyle Higashioka	0.2946369
Phil Cuzzi	0.3861889	Cam Gallagher	0.2825110
Ron Kulpa	0.3332829	Jonah Heim	0.2698605
Brian O'Nora	0.3253105	Sean Murphy	0.2429557
Roberto Ortiz	0.3005747	Omar Narvaez	0.2367468
Brian Gorman	0.2991598	Austin Hedges	0.2295280
Vic Carapazza	0.2953889	Reese McGuire	0.2085689
Jeremy Riggs	0.2408721	Jorge Alfaro	0.1789057

Table 2: Ten Lowest Random Effects for Umpires and Catchers

Umpire	Random Effect	Catcher	Random Effect
Edwin Moscoso	-0.4356161	Riley Adams	-0.4360781
Carlos Torres	-0.3601212	Drew Butera	-0.4015985
Adrian Johnson	-0.3094090	Zack Collins	-0.3997225
Kyle McCrady	-0.2801225	Chance Sisco	-0.3913650
Mark Wegner	-0.2658314	Rafael Marchan	-0.3147549
Ryan Wills	-0.2387429	Bryan Holaday	-0.2873769
Alfonso Marquez	-0.2329446	Austin Wynns	-0.2809902
Larry Vanover	-0.2269151	Nick Fortes	-0.2436609
David Rackley	-0.2207834	Luis Torrens	-0.2374558
Lew Williams	-0.2041942	Wilson Ramos	-0.2262182

Table 1 provides the ten highest random effects for catchers and umpires during the 2021 season. There is a specific interpretation for these parameters, but the best way to think about these parameters in this context is that the larger the random effect, the more pitchers the catcher got called a strike over the course of the season or the umpire called more pitches a strike during the season. As a result, a pitcher would want to pitch with one of these catchers catching him or one of these umpires behind the plate for the game, since they are likely to call more strikes.

Table 2 provides the ten lowest random effects for catchers and umpires during the 2021 season. Again the best way to understand these parameters, rather than a formal interpretation, is to know that the more negative the random effect, the fewer strike calls a catcher got or the umpire called fewer strikes. As a result, a pitcher would not want to pitch to one of these catchers or one of these umpires behind the plate for the game because the catchers don't get extra strikes called and these umpires have a smaller strike zone.

These results show that the catcher and umpire can influence the probability of a pitch being called a strike, after accounting for the location of a pitch. These results also make sense when comparing to other sources. For example, Baseball Savant lists the catchers for the 2021 season and provides their strike rate for each catcher. The top catchers from the model are at or near the top of the list according to MLB and the same thing happens for the worst catchers as well. When comparing umpires, another analysis that was written about in the Washington Post was done from the start of the 2021 season to mid August and looked at the expected number of strikes against the number of called strikes for umpires. These results also reflect the results based on the random effects for umpires in this analysis. Based on both of these comparisons, the model appears to effectively accomplish the goals of estimating the probability of a strike based on location and the additionally determining the influence of catchers and umpires on that probability after accounting for location.

Conclusions

The goals of this analysis were to model the probability of a pitch being called a strike based on location as well as determining the influence of umpires and catchers of that probability. As shown in the previous section, the models proposed effectively estimate the probability of a strike based on the location and provide estimates for how catchers and umpires influence the strike probability. There is evidence that catcher framing is an important skill for catchers and that catchers are able to get strike calls better than other catchers, which is important to teams.

There are a number of potential factors that could be furthered studied to better understand this relationship. First, for baseball there is an expected number of runs until the end of an inning based on the current state of play. When pitches are called a strike, but are actually a ball or are called a ball when it is actually a strike, the run expectancy for the inning changes. As a result, the change in run expectancy could be calculated for each of these catchers and an idea of how many runs a catcher prevented or gave up could be calculated, which makes it easier to measure the impact of the catcher on winning. Second, the GAM and GLMM that were used in this analysis were very simple and could likely be improved as well to get better umpire and catcher effects.

It seems to be a fair conclusion to say the proposed model in this analysis is a good baseline and provides a good idea of how location, umpires, and catchers influence whether a pitch is called a strike or not, but further improvements could certainly be made to this model in the future as well.