SAFE2: A Hierarchical Model of Pitch Framing

Sameer K. Deshpande and Abraham J. Wyner

The Wharton School
University of Pennsylvania

4 April 2017

Abstract

Since the advent of high-resolution pitch tracking data (PITCHf/x), many in the sabermetrics community have attempted to quantify a Major League Baseball catcher's ability to "frame" a pitch (i.e. increase the chance that a pitch is called as a strike). Especially in the last three years, there has been an explosion of interest in the "art of pitch framing" in the popular press as well as signs that teams are considering framing when making roster decisions.

We introduce a Bayesian hierarchical model to estimate each umpire's probability of calling a strike, adjusting for pitch participants, pitch location, and contextual information like the count. Using our model, we can estimate each catcher's effect on an umpire's chance of calling a strike. We are then able to translate these estimated effects into average runs saved across a season. We also introduce a new metric, analogous to Jensen, Shirley, and Wyner's Spatially Aggregate Fielding Evaluation metric, which provides a more honest assessment of the impact of framing.

1 Introduction

The New York Yankees and Houston Astros played each other in the 2015 American League Wild Card game in October 2015, with the winner continuing to the next round of the Major League Baseball playoffs. During and immediately after the game, several Yankees fans took to social media expressing frustration that home plate umpire, Eric Cooper, was not calling pitches as balls or strikes consistently for both teams, thereby putting the Yankees at a marked disadvantage. Even players in the game took exception to Cooper's decision making: after striking out Yankees catcher, Brian McCann, argued with Cooper that he was calling strikes on similar pitches when the Astros were pitching but balls when the Yankees were pitching. Figure 1 shows two pitches thrown during the game, one by Astros pitcher, Dallas Keuchel, and the other by Yankees pitcher, Masahiro Tanaka.

Figure 1: Both pitches missed the strike zone (outlined in red) and by rule, should have been called balls. Keuchel's pitch (left) was called a strike while Tanaka's pitch (right) was called a ball. Screenshot source: http://www.fangraphs.com/blogs/how-the-astros-wound-up-with-a-bigger-zone/

Both pitches were thrown in roughly similar locations, near the bottom-left corner of the *strike zone*, the rectangular region of home plate shown in the figure. According to the official rules, since neither pitch passed through the strike zone, the umpire ought to have called both pitches a ball. Cooper, however, called the pitch from Keuchel a strike but the pitch from Tanaka a ball. That Cooper did not adhere strictly to the official rules is hardly surprising; Green and Daniels (2014) observe that umpires are influenced by extraneous factors beyond pitch location, such as their previous ball/strike decisions and the state of the at-bat. During the television broadcast, the announcers speculated that the difference in Cooper's strike zone enforcement was the ability of Astros catcher Jason Castro to "frame" pitches, catching them in such a way to increases Coopers' chances of calling a strike (Sullivan, 2015).

Though pitch framing has received attention from the sabermetrics community since 2008, it has generated tremendous interest in the popular press (see, e.g., Lindbergh, 2013; Pavlidis, 2014; Sanchez, 2015) and among team officials (see, e.g., Drellich, 2014; Holt, 2014) in the last three or four years, due to its apparently large impact on team success. According to Woodrum (2014), most studies of framing, including the most recent by Judge et al. (2015) for the website Baseball Prospectus, estimate that a good framer can save his team as many as 25 runs, on average, over the course of the season. By the traditional heuristic of 10 runs per win (Cameron, 2008), these results suggest that the way a good framer catches a few pitches per game may be worth as many as 2 to 3 wins for his team, on par with the offensive production of the game's elite batters. The ostensibly large impact that framing may have on team success has, until recently, been overlooked and undervalued. Sanchez (2015) highlights the catcher Jonathan Lucroy, whose framing accounted for about 2 wins in the 2014 would have been worth about \$14M, writing that "the most impactful player in baseball today is the game's 17^{th} highest-paid catcher."

Returning to the two pitches in Figure 1, Cooper may have been more likely to call the Keuchel pitch a strike because of Castro's framing. However, looking carefully at Figure 1, we see that the two pitches are quite different, making it difficult to immediately attribute the difference in calls to Castro. First, Keuchel's pitch is much closer to the displayed strike zone than Tanaka's and it was thrown in a 1–0 count while Tanaka's was thrown in a 1–1 count. We also note that the batters, catchers, and pitchers involved in each pitch are, necessarily, different. In particular, Keuchel is a left-handed pitcher and Tanaka is a right-handed pitcher. Any of these factors, then, may have contributed to Cooper being more likely to call Keuchel's pitch a strike. Of course, it could also be the case that Cooper was equally likely to call both pitches a strike and the different calls are simply due to noise. This raises the questions: what effect did Castro have on Cooper's called strike probability, over and above factors like the pitch location, count, and the other pitch participants? And what impact does such an effect have on his team's success?

Existing attempts to answer these questions fall broadly into two categories: those that do not fit statistical models of the called strike probability and those that do. The first systematic study of framing, Turkenkopf (2008) falls into the first category. For each catcher, he counted the number of strikes called on pitches throw outside Walsh (2007)'s approximate strike zone. Walsh (2007) approximated the strike zone with a rectangular box over home plate and set the horizontal (resp. vertical) boundaries by first estimating the proportion of called strikes as a function of the pitches horizontal (resp. vertical) coordinates and then identifying the 50% point. Turkenkopf (2008) then took the counts of "extra strikes" received by each catcher and converted them into a measure of runs saved using estimate of 0.161 runs saved per strike. Missing from this analysis, however, is any consideration of the other players and the umpire involved in the pitch, as well as the context in which the pitch was thrown (e.g. count, run differential, inning, etc.). This omission could overstate the apparent impact of framing since it is not immediately clear that a catcher deserves all of the credit for an extra strike. More recently, Rosales and Spratt (2015) proposed an iterative method to distribute credit for a called strike among the batter, catcher, pitcher, and umpire. Unfortunately, this procedure is rather ad hoc and the statistical properties are unknown, making it very difficult to ascertain the uncertainty and validity of the the resulting estimates of framing's impact on the game.

The second broad category of framing studies begin by fitting a statistical model of the called strike probability that accounts for the above factors. Armed with such a model, one then can estimate the predicted called strike probability with and without the catcher. The difference in these probability reflects the catcher's apparent framing effect on that pitch. By weighting these effects by an appropriate estimate of the value of "stealing a strike," and summing over all pitches caught by a catcher, one can arrive at an estimate of the value of framing. Marchi (2011) fit a mixed-effects logistic regression model, expressing the log-odds of a called strike as a function of the identities of the pitch participants and interactions between them. This model, however, does not systematically incorporate pitch location, meaning that the resulting estimates of framing effects are confounded by the location just like Turkenkopf (2008)'s. To our knowledge, the most systematic attempt to study framing is Judge et al. (2015). They introduced a mixed-effects probit regression model built in two stages: first, they used a proprietary model to estimate the probability of called strike as a function of its location, the count, the handedness of the batter, and the ballpark. They then fit a probit regression model using the participants of the pitch and the proprietary probability estimate from their first step as covariates. They incorporated random effects for each of the pitch participants as well as for the interaction between the catcher and the initial called strike probability estimate. Curiously, they model the probit transformation of the called strike probability with an existing, proprietary estimate of the same probability, in essence linearizing the probit function. This modeling choice lacks statistical justification and potential consequences on the validity of and uncertainty in the final estimates of framing impact are unknown and unexplored. Further, it is not clear whether this existing estimate was computed with the same data used to fit the final probit model, further complicating the uncertainty quantification for their final framing estimates.

Both Judge et al. (2015)'s and Marchi (2011)'s models assume that the only difference between umpires is some base-rate of calling strikes (i.e. they do not interact the umpire with any other covariate). It is arguably more realistic to believe that factors like the pitch location and count do not exert a constant effect across umpires. More relevant to framing is the preclusion of umpire bias towards certain players. In light of this, denoting the probability that the i^{th} pitch called by umpire u is called a strike by p_i^u, we will consider models of the following general form:

$$\log\left(\frac{p_i^u}{1-p_i^u}\right) = \theta_0^u + \theta_{b_i}^{u,B} + \theta_{c_i}^{u,C} + \theta_{p_i}^{u,P} + \theta_{co_i}^{u,CO} + f^u(x_i, z_i) \quad (1)$$

where $\theta^{u,B}$, $\theta^{u,C}$, and $\theta^{u,P}$ are vectors of the *partial effects* of the batters, catchers, and pitchers on umpire u, respectively, where b_i, c_i, and p_i are indices corresponding to which players are involved in that pitch, x_i and z_i are the horizontal and vertical coordinates of the pitch, and f^u is an umpire-specific function of pitch location that we will specify later.

The model in Equation 1 is quite flexible as it allows for pitch location, count, and individual batters, catchers, and pitchers to affect umpires differently. We pause for a moment to appreciate the sheer number of parameters implied by Equation 1. In our 2014 dataset (described in more detail in Section 2), there are 93 different umpires, 1010 batters, 101 catchers, and 719 pitchers. Even without accounting for the count and the semi-parametric location term (specified in Section 2.2), the model in Equation 1 includes over 160,000 parameters, which must be estimated from a total of 308,388 called pitches. This vast over-parametrization immediately raises concern about overfitting. In the context of framing, overfitting called strike probabilities in a single year could skew retrospective estimates of the impact framing has. More importantly, it diminishes our ability to forecast the impact a catcher's framing has into the future, a key consideration for teams making roster decisions.

In this paper, we leverage high-resolution pitch tracking data from the PITCHf/x system, which is briefly described in Section 2.1, to estimate how much a catcher influences umpires' chances of calling strikes and how large an impact such effects have on his team's success. In Section 2.3, we introduce several simplifications of the model in Equation 1 that still elicit umpire-to-umpire heterogeneity but vastly reduces the number of parameters. All of these models are fit in a hierarchical Bayesian framework, which provides very natural uncertainty quantification for our framing estimates. Such quantification, notably absent in previous framing studies, is vital, considering the fact that several teams are making framing-based roster decisions (Drellich, 2014; Holt, 2014; Sanchez, 2015). We compare the predictive performances of these models in Section 3.1 and assess the extent to which incorporating umpire-specific count and player effects lead to overfitting. We then translate our estimates of catcher effects from the log-odds scale to the more conventional scale of average runs saved. We introduce two metrics in Section 4 to estimate the impact framing has on team success. We conclude with a discussion and outline several potential extensions of our modeling efforts.

2 Data and Model

We begin this section with a brief overview, adapted primarily from Fast (2010) and Sidhu and Caffo (2014), of our pitch tracking dataset before introducing the hierarchical logistic regression model used to estimate each umpire's called strike probability.

2.1 PITCHf/x Data

Starting in 2006, the TV broadcast company Sportvision began installing system of cameras in major league ballparks to estimate the trajectory of each pitch thrown, known as PITCHf/x. During the flight of each pitch, these cameras take 27 images of the baseball and the PITCHf/x software fits a quadratic polynomial to the 27 locations to estimate its trajectory (Sidhu and Caffo, 2014). This data is transmitted to the MLB Gameday application, which allows fans to follow the game online (Fast, 2010). In addition to collecting pitch trajectory data, an MLB Advanced Media employee records game-state information during each pitch. For instance, he or she records the pitch participants (batter, catcher, pitcher, and umpire) as well as the outcome of the pitch (e.g, ball, swinging strike, hit), the outcome of the at-bat (e.g. strikeout, single, home run), and any other game action (e.g. substitutions, baserunners stealing bases). The PITCHf/x system also reports the approximate vertical boundaries of the strike zone for each pitch thrown. Taken together, the pitch tracking data and game-state data provide a high-resolution pitch-by-pitch summary of the game, available through the MLB Gameday API.

Though our main interest in this paper is to study framing effects in the 2014 season, we collected all PITCHf/x data from the 2011 to 2015 regular season. In Section 2.2, we use the data from the 2011–2013 season to select the function of pitch location $f^u(x, z)$ from Equation 1. We then fit our model using the 2014 data and in Section 3.1, we assess our model's predictive performance using data from 2015. In the 2014 season, there were a total of 701,490 pitches, of which 355,293 (50.65%) were *taken* (i.e. not swung at) and of these taken pitches, 124,642 (35.08%) were called strikes. We approximated the rule book strike zone by averaging the vertical boundaries recorded by the PITCHf/x system and aligning the horizontal boundaries with the edges of home plate. We restricted our attention to the $N = 308,388$ taken pitches which were located within one foot of this approximate strike zone. In all, there were a total of $n_U = 93$ umpires, $n_B = 1010$ batters, $n_C = 101$ catchers, and $n_P = 719$ pitchers.

2.2 Adjusting for Pitch Location

Intuitively, pitch location is the main driver of called strike probability. The simplest way to incorporate pitch location into our model would be to include the horizontal and vertical coordinates recorded by the PITCHf/x system as linear predictors so that $f^u(x, z) = \theta^u_x x + \theta^u_z z$, where θ^u_x and θ^u_z are parameters to be estimated. While simple, this forces an unrealistic left-to-right and top-to-bottom monotonicity in the called strike probability surface. Another simple approach would be to use a polar coordinate representation, with the origin taken to be the center of the approximate rule book strike zone. While this avoids any horizontal or vertical monotonicity, it assumes that, all else being equal, the probability of a called strike is symmetric around this origin.

Such symmetry is not observed empirically, as seen in Figure 2, which divides the plane above home plate into 1" squares whose color corresponds to the proportion of pitches thrown in the three year window 2011 – 2013 that pass through the square that are called strikes. Also shown in Figure 2 is the approximate strike zone, demarcated with the dashed line, whose vertical boundaries are the average of the top and bottom boundaries recorded by the PITCHf/x system. If the center of the pitch passes through the region bound by the solid line, then some part of the pitch passes through the approximate strike zone. This heat map is drawn from the umpire's perspective so right handed batters stand to the left of home plate (i.e. negative PX values) and left-handed batters stand to the right (i.e. positive PX values). We note that the bottom edge of the figure stops 6 inches off of the ground and the left and right edges end 12 inches away from the edges of home plate. Typically, batters stand an additional 12 inches to the right or left of the region displayed.

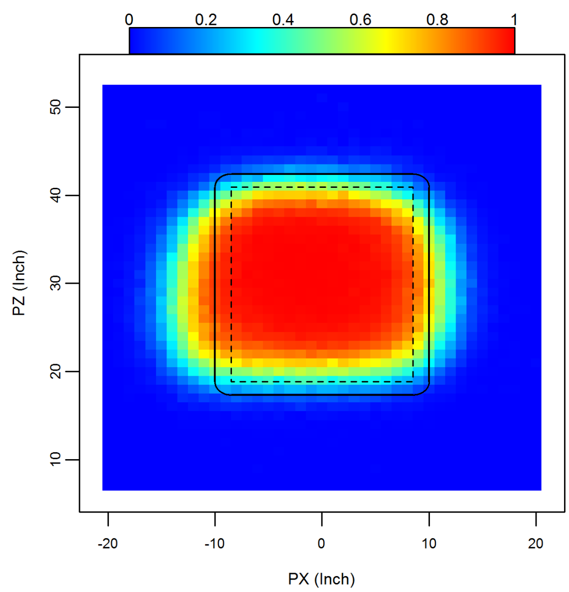
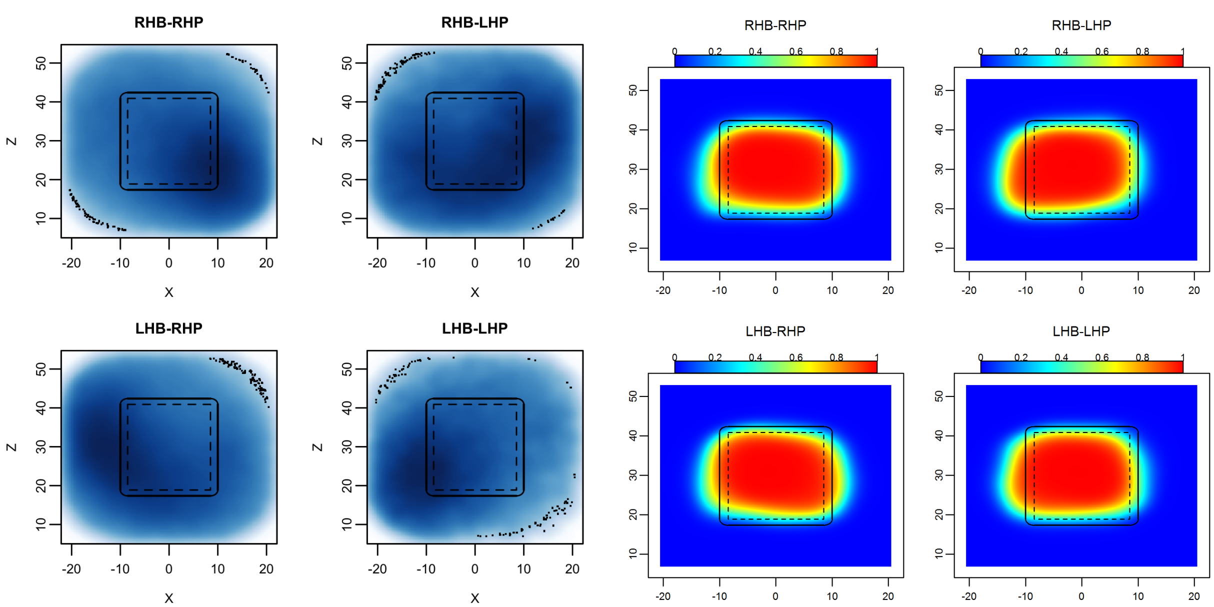


Figure 2: Heat map of empirical called strike probabilities, aggregate over the three-year window 2011 – 2013. The boundary of the approximate 2014 rule book strike zone is shown in dashed line. If the center of the pitch passes through the region bounded by the solid line, some part of the pitch passes through the approximate strike zone.

Rather than specifying a explicit parametrization in terms of the horizontal and vertical coordinates, we propose using a smoothed estimate of the historical log-odds of a called strike as an implicit parametrization of pitch location. This is very similar to the approach taken in Judge et al. (2015), who included the estimated called strike probability as a covariate in their probit model.

Following the example of Mills (2014), we fit generalized additive models with a logistic link to the data aggregated from 2011 – 2013. As seen in Figure 3a, however, pitch location can vary considerably based on the handedness of the batter and pitcher. Because of this, we actually fit four separate GAM's, hereafter referred to as the "hGAMs" or "historical GAMs", that express the log-odds of a called strike as a smooth function of the pitch location, one each for the four combinations of pitcher and batter handedness. Figure 3b shows the four estimated called strike probability surfaces. Recall that these figures are drawn from the umpires' perspective and that right-handed batters stand about one foot further to the left of the plotted region while left-handed batters stand one foot further to the right of the right edge of the plotted region.



(a) Pitch frequency

(b) hGAM called strike surfaces

Figure 3: (a)Smoothed density of pitch locations based on cross-handedness. (b) hGAM estimated called strike probability based on handedness.

We see that between 2011 and 2013, there are clear differences in how umpires called strikes, depending on combination of pitcher and batter handedness. For each 2014 called pitch set in our dataset, we estimate the log-odds of that pitch being called a strike with the approach historical GAM. We use these historical estimates as continuous predictors in our model, so that potential player effects and count effects may be viewed as adjustments to these historical baselines.

2.3 Bayesian Logistic Regression Models

As alluded to in Section 1, we consider several models of increasing complexity to predict umpires' calls in 2014. Before fully specifying these models, we label the 93 umpires $u_1, \ldots, u_{93}$ and we further suppose that umpire u calls a total of n^u total pitches. Consider the i^{th} pitch called by umpire u. Let $y_i^u = 1$ if it is called a strike and let $y_i^u = 0$ if it is called a ball. Let h_i^u be a vector of length four, encoding the combination of batter and pitcher handedness on this pitch and let $\mathbf{LO}_i^u$ be a vector of length four, containing three zeros and the estimated log-odds of a strike from the appropriate historic GAM. If x_i^u and z_i^u denote the PITCHf/x coordinates of this pitch, we take $f^u(x_i, z_i) = \mathbf{LO}_i^{u\top}\Theta_{LO}^u$. Note that for each combination of batter and pitcher handedness, we first standardized the corresponding historical GAM estimates to have standard deviation 1. Finally let $\mathbf{CO}_i^u, \mathbf{CA}_i^u, \mathbf{P}_i^u$ and $\mathbf{B}_i^u$ be vectors encoding the count, catcher, pitcher, and batter involved with this pitch. We can re-write the model from Equation 1 as

$$\log\left(\frac{P(y_i^u = 1)}{P(y_i^u = 0)}\right) = h_i^{u'}\Theta_0^u + \mathbf{LO}_i^{u'}\Theta_{LO}^u + \mathbf{CO}_i^{u'}\Theta_{CO}^u + \mathbf{CA}_i^{u'}\Theta_{CA}^u + \mathbf{P}_i^{u'}\Theta_P^u + \mathbf{B}_i^{u'}\Theta_B^u$$

For identifiability, we specify a single catcher, Brayan Pena, and count, 0–0, as baseline values.

We are now ready to present several simplifications of the general model from Equation 1 of gradually increasing complexity. We begin first by assuming that the players and count have no effect on the log-odds of a called strike (i.e. that $\Theta_{CO}^u, \Theta_{CA}^u, \Theta_P^u$, and Θ_B^u are all equal to the zero vector for each umpire). This model, hereafter referred to as Model 1, assumes that the only relevant predictor of an umpire's ball/strike decision is the pitch location but allows for umpire-to-umpire heterogeneity. We model, *a priori*,

$$\Theta_0^{u_1}, \ldots, \Theta_0^{u_{93}} | \Theta_0 \sim N\left(\Theta_0, \tau_0^2 I_4\right)$$

$$\Theta_{LO}^{u_1}, \ldots, \Theta_{LO}^{u_{93}} | \Theta_{LO} \sim N\left(\Theta_{LO}, \tau_{LO}^2 I_4\right)$$

$$\Theta_0 | \sigma_0^2 \sim N\left(0_4, \sigma_0^2 I_4\right)$$

$$\Theta_{LO} | \sigma_{LO}^2 \sim N\left(\mu_{LO}, \sigma_{LO}^2 I_4\right)$$

The parameters τ_0^2 and τ_{LO}^2 capture the umpire-to-umpire variability in the intercept and location effects and we may view Θ_0 and Θ_{LO} as the mean intercept and location effects averaged over all umpires. By placing a further level of prior hierarchy on Θ_0 and Θ_{LO}, we are able to "borrow strength" between umpires. Further priors on the hyper-parameters σ_0^2 and σ_{LO}^2 induce a dependence between the the four combinations of handedness, allowing us to "borrow strength" between them as well.

We now consider four successive expansions of Model 1. In Model 2, we introduce count effects, assuming that the Θ_{CO}^u's are all equal to some common value Θ_{CO}. Model 3 elaborates on Model 2 by including catcher effects, under a similar assumption that a catcher exerts the same effect on all umpires (i.e. $\Theta_{CA}^u = \Theta_{CA}$ for all u and some Θ_{CA}). Models 4 and 5 further expand on Model 3, with Model 4 adding a constant pitcher effect to Model 3 and Model 5 adding a constant batter effect to Model 4. *A priori*, we model

$$\Theta_{CO} | \sigma_{CO}^2 \sim N\left(0_{11}, \sigma_{CO}^2 I_{11}\right)$$

and consider similar, zero-mean spherically-symmetric Gaussian priors for Θ_{CA}, Θ_P and Θ_B.

Models 2 – 5 represent vast simplifications to the general model in Equation 1 as they assume that there is no umpire-to-umpire variability in the count or player effects. This assumption is arguably simplistic, which leads us to consider Model 6, which builds on Model 3 by allowing umpire-specific count effects and asserts

$$\log\left(\frac{P(y_i^u = 1)}{P(y_i^u = 0)}\right) = h_i^{u'}\Theta_0^u + \mathbf{LO}_i^{u'}\Theta_{LO}^u + \mathbf{CO}_i^{u'}\Theta_{CO}^u$$

We retain the same prior specification for Θ_0^u and Θ_{LO}^u as in Model 3 but now model

$$\begin{aligned}\Theta_{CO}^{u_1}, \cdots, \Theta_{CO}^{u_{93}} | \Theta_{CO}, \tau_{CO}^2 &\sim N\left(\Theta_{CO}, \tau_{CO}^2 I_{11}\right) \\ \Theta_{CO}^u | \sigma_{CO}^2 &\sim N\left(0_{11}, \sigma_{CO}^2 I_{11}\right).\end{aligned}$$

We finally consider Model 7, which is identical to Model 6 but now allows for umpire-specific catcher effects. We adopt a very similar prior specification, modeling

$$\begin{aligned}\Theta_{CA}^{u_1}, \cdots, \Theta_{CA}^{u_{93}} | \Theta_{CA}, \tau_{CA}^2 &\sim N\left(\Theta_{CA}, \tau_{CA}^2 I_{100}\right) \\ \Theta_{CA}^u | \sigma_{CA}^2 &\sim N\left(0_{100}, \sigma_{CO}^2 I_{100}\right).\end{aligned}$$

Throughout, we place independent Inverse Gamma(3,3) hyper-priors on the top-level variance parameters $\sigma_0^2, \sigma_{LO}^2, \sigma_{CO}^2, \sigma_{CA}^2, \sigma_P^2$ and σ_B^2. It remains to specify the hyper-parameters $\tau_0^2, \tau_{LO}^2, \tau_{CO}^2$ and τ_{CA}^2, which capture the umpire-to-umpire variability in the intercept, location, count, and catcher effects, respectively. For simplicity, we fix these hyper-parameters to be equal to 0.25 in the appropriate models. To motivate this choice, consider how two umpires would call a pitch thrown at a location where the historical GAM forecasts a 50% called strike probability. According to Model 7, the difference in the two umpires' log-odds of a called strike follows a $N\left(0, 2(\tau_0^2 + \tau_{CO}^2 + \tau_{CA}^2)\right)$ distribution, *a priori*. Taking $\tau_0^2 = \tau_{CO}^2 = \tau_{CA}^2 = 0.25$ reflects a prior belief that there is less than 10% chance that one umpire would call a strike 75% of the time while the other calls it a strike only 25% of the time. For simplicity, we take $\tau_{LO}^2 = 0.25$ as well.

Of course, we could easily continue this systematic enumeration of nested models until we reach the full model of Equation 1. As we will see in Section 3.1, the increased complexity of such models is accompanied by less generalizability.

3 Model Performance and Comparison

3.1 Predictive Performance

We fit each model in Stan (Carpenter et al., 2016) and ran two Markov Chain Monte Caro chains for each model. After burning-in the first 2000 iterations, the Gelman-Rubin $\hat{R}$ statistic for each parameters was less than 1.1, suggesting convergence. We continued to run the chains, after this burn-in, until each parameter's effective sample size exceeded 1000. For Models 1 – 4, running the sampler for 4000 total iterations was sufficient but for the more complex models, we needed 6000 iterations. Unsurprisingly, the more parameters, the slower the mixing.

Using the simulated posterior draws from each model, we can approximate the mean of the posterior predictive distribution of the called strike probability for each pitch in our 2014 dataset. Table 1 shows the misclassification rate and mean square error for models 1 – 7 over all 2014 pitches and over two separate regions, as well as the error for the historical GAM forecasts. Region 1 consists of all pitches that were thrown within 1.45 inches of the average 2014 strike zone and Region 2 consists of pitches that were thrown between 1.45 and 2.9 inches away from the average 2014 strike zone boundary. Since the radius of the ball is about 1.45 inches, Region 1 consists of pitches that were "borderline" calls, according to the official rule book, and Region 2 consists of pitches that missed this strike zone by at least a ball's width.

Table 1: In-sample misclassification rate (MISS) and mean square error (MSE) for several models ($\times 100$)

		Model 1	Model2	Model 3	Model 4	Model 5	Model 6	Model 7	hGAM
Overall	MISS	10.326	10.121	10.017	9.894	9.852	9.929	**9.642**	10.496
	MSE	7.287	7.150	7.059	6.988	6.932	7.002	**6.817**	7.426
Region 1	MISS	24.833	24.133	23.590	23.241	23.184	23.341	**22.524**	25.797
	MSE	16.274	15.899	15.584	15.424	15.304	15.439	**14.989**	16.796
Region 2	MISS	21.421	21.195	20.916	20.654	20.541	20.763	**20.272**	21.527
	MSE	15.333	15.129	14.912	14.726	14.589	14.804	**14.400**	15.579

We see that Models 1– 7 outperform the historical GAMs overall and in both Regions 1 and 2. This is hardly surprising, given that the hGAMs were trained on data from 2011–2013 and the other models were trained on the 2014 data. Recall that Model 1 only accounted for pitch location. As we successively incorporate count (Model 2), catcher (Model 3), pitcher (Model 4), and batter (Model 5) effects, we find that the overall error drops. We see that the overall performance of Models 4 and 5 are quire similar, though the incorporation of batter effects seems to help most in Regions 1 and 2. We finally see that Model 7 has the best performance across the board, followed by Model 5. This is hardly surprising, as Model 7 included over 11,000 parameters, thanks to the umpire-specific catcher and count effects, while Model 5 has the second highest number of parameters, 2582.

Of course, comparing in-sample predictive performance provides little insight into the degree to which these models have overfit the data, our primary motivation for considering such model elaborations. One way to assess this is to look at how well our fitted models predict umpires' calls in 2015. Table 2 compares such out-of-sample predictive performance, by considering pitches from the 2015 season for which the associated batter, catcher, pitcher, and umpire all appeared in our 2014 dataset.

Table 2: Out of sample misclassification rate (MISS) and mean square error (MSE) for several models ($\times 100$)

		Model 1	Model2	Model 3	Model 4	Model 5	Model 6	Model 7	hGAM
Overall	MISS	10.705	10.571	10.498	10.470	**10.459**	10.551	10.624	10.851
	MSE	7.540	7.422	7.405	7.383	**7.380**	7.414	7.454	7.630
Region 1	MISS	25.566	24.931	24.532	24.476	**24.391**	24.583	24.754	26.736
	MSE	16.720	16.357	16.178	16.157	**16.138**	16.181	16.256	17.247
Region 2	MISS	23.593	23.373	23.218	**23.093**	**23.093**	23.283	23.308	23.732
	MSE	16.865	16.638	16.590	16.533	**16.502**	16.602	16.614	16.997

Now we see that Model 5 has the best performance overall. The fact that Models 7 has worse out-of-sample performance despite having far-and-away the best in-sample performance is a clear indication of overfit. Somewhat surprisingly, it appears that Model 6 also slightly overfits the data despite have fewer parameters than Model 5.

Comparing predictive performance on 2015 data is arguably not the best means of diagnosing overfitting. Starting in 2009, Major League Baseball began reviewing and grading umpires' ball/strike decisions and Roegelle (2014) and Mills (2014) have documented year-to-year changes in umpires' strike zone enforcement ever since. As a result, umpire tendencies are non-stationary and we cannot reasonably expect a model that explicitly tries to identify such tendencies in one season to forecast future umpire decisions particularly well. A potentially more appropriate way to diagnose overfitting issues would be to holdout a random subset of our 2014 data, say 10%, re-fit each model on the remaining 90% of the data, and assess the predictive performance on the held-out 10%. Table 3 shows the average misclassification and mean square error over 10 such holdout sets.

Table 3: Hold-out misclassification rate (MISS) and mean square error (MSE) for several models ($\times 100$)

		Model 1	Model2	Model 3	Model 4	Model 5	Model 6	Model 7	hGAM
Overall	MISS	10.396	10.170	10.089	10.056	**10.054**	10.127	10.169	10.501
	MSE	7.331	7.202	7.118	7.090	**7.073**	7.139	7.161	7.421
Region 1	MISS	25.112	24.395	23.939	**23.890**	23.968	24.022	24.158	25.89
	MSE	16.432	16.081	15.796	15.738	15.719	15.837	15.876	**15.52**
Region 2	MISS	21.263	21.047	20.801	20.757	**20.631**	20.896	20.835	21.432
	MSE	15.372	15.161	14.964	14.856	**14.802**	14.998	15.009	15.52

The results of Tables 2 and 3 indicate that Models 6 and 7 really are over-fitting the data. While we did not extend the model construction in Section 2.3 to include the general model of Equation 1, the results of Tables 2 and 3 suggest that the general model would over-fit the data even more than Models 6 and 7. It is interesting to note that Models 4 and 5 behave quite similarly on the holdout sets, indicating that the incorporation of batter effects does not substantially improve Model 4's predictions.

3.2 Full Posterior Analysis

We now examine the posterior samples from Model 5 more carefully. Figure 4 shows box plots of the posterior distributions of the catcher effects.

We see that there are some catchers, like Hank Conger and Buster Posey, whose posterior distributions are entirely supported on the positive axis, indicating that, all else being equal, umpires are more likely to call strikes when they are caught by these catchers as opposed to the baseline catcher, Brayan Pena. On the other extreme, there are some catchers like Tomas Tellis with distinctly negative effects. As we would expect, catchers who appeared very infrequently in our dataset have very wide posterior distributions. For instance, Austin Romine caught only 61 called pitches and his effect has the largest posterior variance among all catchers. It is interesting to see that all of

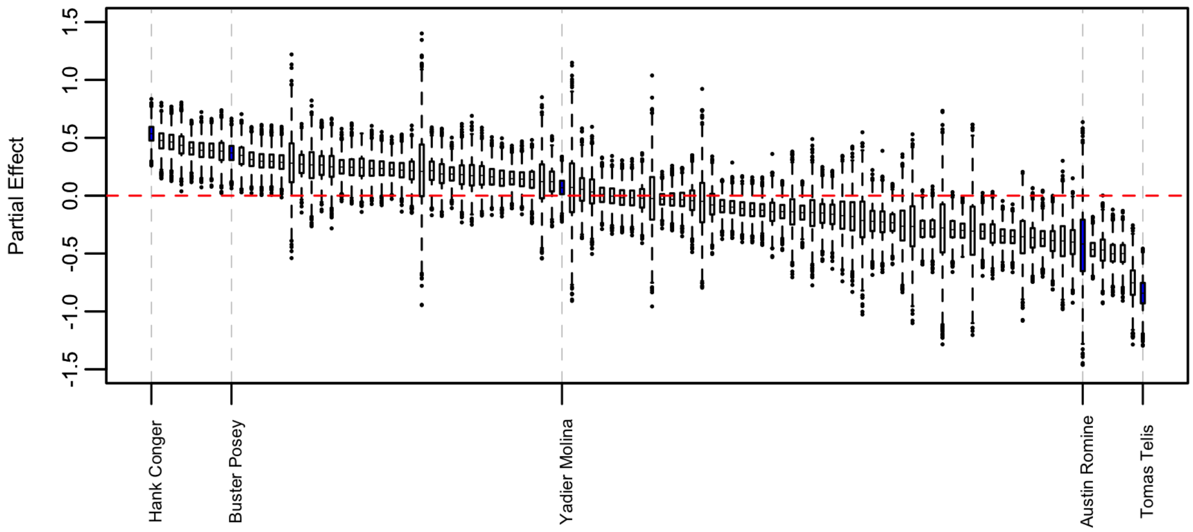


Figure 4: Comparative box plots of catcher effects sorted by posterior mean

the catcher effects are contained in the interval [-1.5,1.5], despite the prior placing nearly 20% of its probability outside this interval.

Figure 5 shows the approximate posterior densities of the count effects. Recall that these are the partial effects relative to the baseline count of 0–0. As we might expect, umpires are more likely to call strikes in 3–0 and 2–0 counts than in 0–0 counts and much less likely to call strikes in 0–2 and 1–2 counts, all else being equal. Somewhat interestingly, we find that umpires are slightly less likely to call strikes in a 3–1 count than they are in a 1–0 count.

We also see that the posterior distribution for the effects of a 3–0 counts are considerably wider than those for a 0–1 count, indicating that we are much more uncertain about the effect of the former two counts than the latter two. This is due the rather large disparity in the numbers of pitches taken in these counts: in our dataset, there were more than five times called pitches thrown into a 0–1 count than into a 3–0 count (37,513 versus 6,162).

All of the count effects are contained in the interval $[-1.5, 1.5]$ suggesting that the overall effect of count and catcher of a somewhat similar magnitude. For instance, suppose that Madison Bumgarner is pitching to Yasiel Puig in an 0–0 count. Further suppose that the pitch is thrown in a location

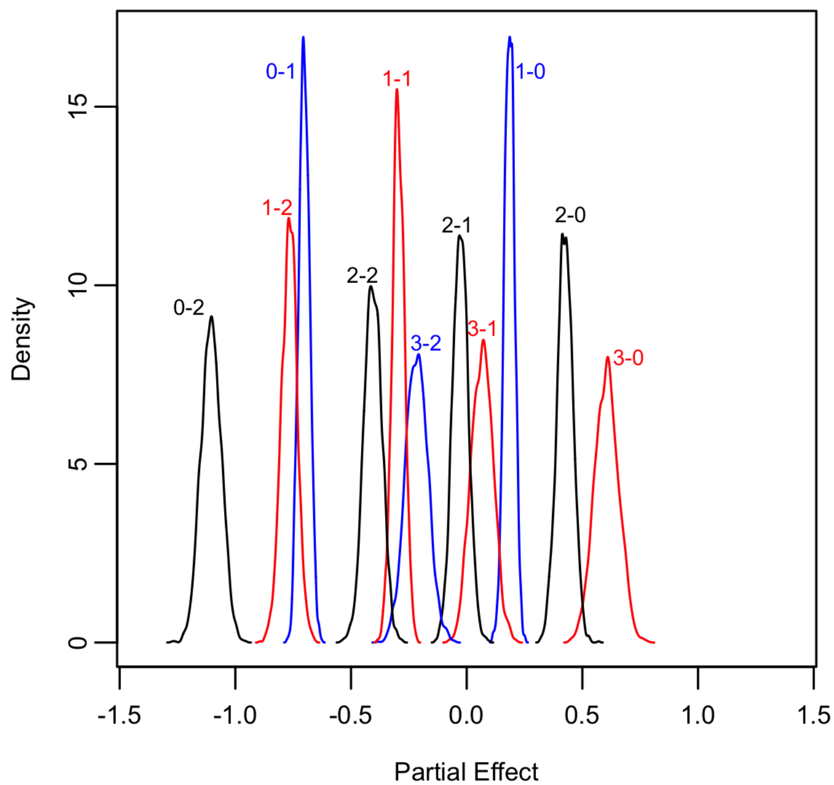


Figure 5: Posterior densities of the partial effect of count. Densities computed using a standard kernel estimator

where the hGAM called strike probability forecast is exactly 50%. According to our model, if this pitch were hypothetically caught by the baseline catcher, Brayan Pena, it would be called a strike only 54% of the time. In contrast, if the same pitch had been caught by Buster Posey, our model estimates the called strike probability to be 64%. If Posey had caught the same pitch but the count were 2–2 instead of 0–0, the forecast would be 55%. In this way, at least for this pitch, the effect of swapping the baseline catcher with Posey on an 0–0 pitch is the same as changing the count from 0–0 to 2–2 with Posey catching.

Table 4 elaborates on this example and shows the estimated called strike probability for the same pitch as a function of count and catcher. That is, we forecasted the called strike probability, averaged across umpires, on a pitch thrown by Bumgarner to Puig at a location where the hGAM forecast was 50% for many combinations of catcher and count. To highlight the relative size of the catcher and count effects, we have subtracted a baseline 54%, the called strike probability when the catcher is Pena and the count is 0–0, from all of these probabilities.

Table 4: Difference in forecasts of Bumgarner's pitch to Puig, relative to the baseline called strike probability, for various combinations of catcher and count

	0–0	0–1	0–2	1–0	1–1	1–2	2–0	2–1	2–2	3–0	3–1	3–2
Hank Conger	0.12	-0.04	-0.14	0.16	0.06	-0.06	0.21	0.12	0.03	0.24	0.14	0.07
Buster Posey	0.09	-0.08	-0.18	0.13	0.02	-0.10	0.18	0.08	-0.01	0.21	0.10	0.04
Brayan Pena	0.00	-0.17	-0.26	0.04	-0.07	-0.18	0.10	-0.01	-0.10	0.14	0.02	-0.05
Tomas Telis	-0.20	-0.34	-0.39	-0.16	-0.26	-0.35	-0.10	-0.20	-0.28	-0.06	-0.18	-0.25

Recall that the baseline called strike probability is 54% on such a pitch. According to our model, the effect of changing the count from 0–0 to 0–2 when Posey is receiving the pitch is about the same as changing the count from 0–0 to 1–2 with the baseline catcher receiving the pitch. We note that the called strike probability forecasts for Tomas Tellis are much lower than for Hank Conger, Posey, and the baseline catcher. For instance, our model estimates that umpires on average would call this pitch a strike only 15% of the time if it were thrown in a 0–2 count and Tellis was receiving, in contrast to 40% for Conger, 36% for Posey, and 28% for Pena.

Figure 6 shows the 50% and 90% contours of the called strike probability when Bumgarner is pitching to Puig and being caught by Posey for two umpires, Angel Hernandez and Mike DiMuro, and the average umpire in a 2–0 and 0–2 count. If the center of a pitch passes within the region bound by the dashed gray line, then some part of the ball passes through the approximate rule book strike zone, shown in gray. Puig is a right-handed batter, meaning that he stands on the left-hand side of the approximate rule book strike zone, from the umpire's perspective.

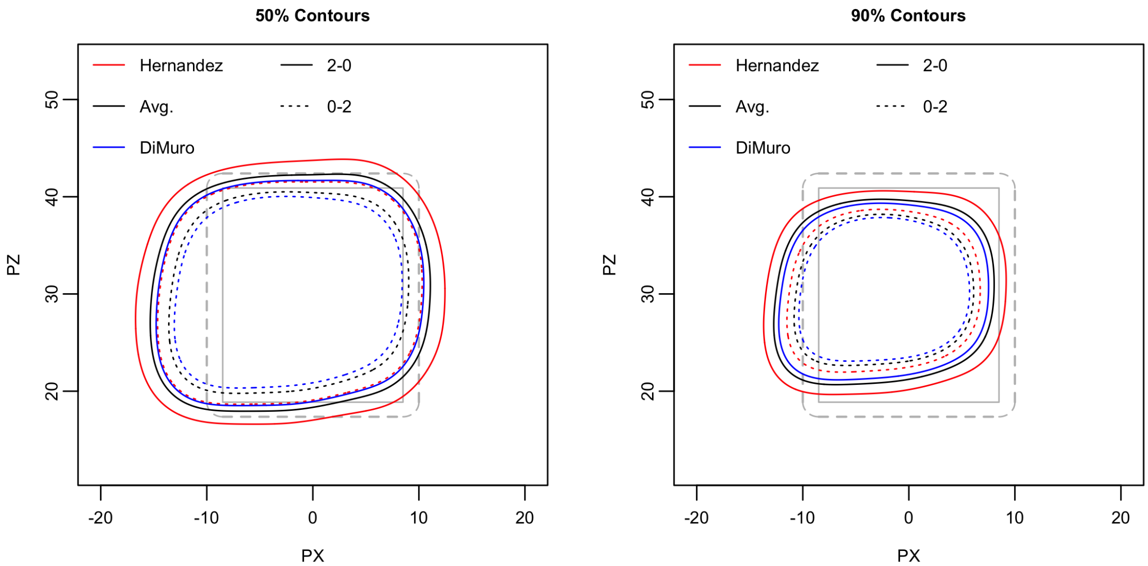


Figure 6: 50% and 90% contours for called strike probability for the Bumgarner-Puig-Posey matchup with different umpires

Across the board, Hernandez's contours enclose more area than the average umpire's contours and DiMuro's contours enclose less area than the average umpire's. For instance, on a 2–0 count, Hernandez's 50% contour covers 4.37 sq. ft., the average umpire's covers 3.87 sq. ft., and DiMuro's covers 3.53 sq. ft. Additionally, as anticipated by Figure 5, the 2–0 contours cover more area than the 0–2 contours. Each of the 50% contours extend several inches to the left, or *inside*, edge of the approximate rule book strike zone. At the same time, the same contours do not extend nearly as far beyond the right, or *outside*, edge of the strike zone. This means that, Hernandez, DiMuro, and the average umpires are more likely to call strikes on pitches that miss the inside edge of the strike zone than they are on pitches that miss the outside edge by the same amount. We moreover find that even the 90% contour on a 2–0 contours extend a few inches past the inside edge of the strike zone, implying that Hernandez, DiMuro, and the average umpire will almost always call a strike that misses the inside edge of a the strike zone so long as it is not too high or low. Interestingly, we see that the leftmost extents of DiMuro's and the average umpire's 90% contours on a 0–2 pitch nearly align with the dashed boundary on the inside edge. A pitch thrown at this location will barely overlap the rule book strike zone, indicating that at least on the inside edge, DiMuro and umpires on average tend to follow the rule book prescription, calling strikes over 90% of the time.

Perhaps unsurprisingly at this point, we find that the same is not true at the top, bottom, or outside edge. Indeed, we find that the rightmost extent of average umpires' 90% contour on a 0–2 pitch lies several inches within the outside edge of the strike zone. We also see that the rightmost extent of the average umpires' 50% contour on the 0–2 pitch So in the space of about four and a half inches, the average umpires' called strike probability drops dramatically from 90% to 50%. It is worth noting that, according to the rule book, all of these pitches ought to be called a strike.

4 Impact of framing

In keeping with the spirit of most studies of pitch framing, we turn our attention now to estimating the impact these effects could have on the game.

Formally, let S be a random variable counting the number of runs the pitching team gives up after the current pitch to the end of the half-inning. Using similar notation as in Section 2.3, let $\mathbf{h}$ encode the handedness of the batter and pitcher and let $\mathbf{lo}$ be the estimated log-odds of a called strike from the appropriate historical GAM. Let $\mathbf{co}, \mathbf{c}$, and $\mathbf{u}$ denote the count, catcher, and umpire involved in the pitch. Finally, let $\mathbf{b}$ and $\mathbf{p}$ be the batter and pitcher involved in the pitch. Denote the baseline catcher Brayan Pena by c_0.

For compactness, let $\xi = (\mathbf{u}, \mathbf{co}, \mathbf{lo}, \mathbf{b}, \mathbf{p}, \mathbf{h})$ and observe that every pitch in our dataset can be identified by the combination $(\mathbf{c}, \xi)$. For each catcher c, let $\mathcal{P}_c$ be the set of all called pitches caught by catcher c:

$$\mathcal{P}_c = \{(\mathbf{c}, \xi) : \mathbf{c} = c\}.$$

Finally, let $TAKEN$ be an indicator for the event that the current pitch was taken and let $CALL \in \{Ball, Strike\}$ be the umpire's ultimate call. We will be interested in the expected value of S, conditioned on $(\mathbf{c}, \xi)$, the fact that pitch was taken, and the umpire's call. Assuming that, conditional on these factors, S is independent of the pitch participants and location, we have

$$E[S|\mathbf{c}, \xi, TAKEN] = \sum_{CALL} E[S|COUNT, TAKEN, CALL] P\left(CALL|\mathbf{c}, \xi, TAKEN\right)$$

To determine the expected number of runs given up that can be attributed to a catcher c, we may consider the counter-factual scenario in which the catcher is replaced by the baseline catcher, Brayan Pena, with all other factors remaining the same. In this scenario, the expected number of runs the fielding team gives up in the remaining of the half-inning is $E\left[S|\mathbf{c} = c_0, \xi, TAKEN, CALL\right]$. We may interpret the difference

$$E\left[S|\mathbf{c} = c, \xi, TAKEN, CALL\right] - E\left[S|\mathbf{c} = c_0, \xi, TAKEN, CALL\right]$$

as the average number of runs saved (i.e. negative of runs given up) by catcher c's framing, relative to the baseline. A straightforward calculation shows that this difference is exactly equal to

$$f(c, \xi) = (P(Strike|\mathbf{c} = c, \xi, TAKEN) - P(Strike|\mathbf{c} = c_0, \xi, TAKEN)) \times \rho(COUNT),$$

where

$$\rho(COUNT) = E[S|COUNT, TAKEN, Ball] - E[S|COUNT, TAKEN, Strike].$$

We can interpret the difference in called strike probabilities above as catcher c's *framing effect*: it is precisely how much more the catcher adds to the umpires' called strike probability than the baseline catcher, over and above the other pitch participants, pitch location, and count. We can easily simulate approximate draws from the posterior distribution of this difference using the posterior samples from Model 5. We interpret ρ as the value of a called strike in a give count: it measures how many more runs a team is expected to give up if a taken pitch is called a ball as opposed to a strike. Since the true value of ρ is not known, we estimate the two terms in the difference using historical data from 2011 to 2014.

In a seminal 1963 paper, George Lindsey introduced the idea of tracking the distribution of the number of runs given up in a half-inning as a function of the game state to evaluate different strategies in baseball. He found, for example, that at the beginning of an inning, with no outs and no baserunners, the fielding team will give up, on average 0.46 runs in the remainder of the half-inning but when there are 2 outs and three baserunners, they will give up 0.82 runs, on average, in the remainder of the half-inning (Lindsey, 1963). This number is known as run expectancy and as Albert (2015) notes, it has formed the bedrock of many analyses of baseball data ever since Lindsey's original paper. While most uses of run expectancy focus on the average number of runs scored as a function of the configuration of the baserunners and number of outs, to study the impact of framing, we instead control only for the count. To estimate the value of a called strike in an 0 –1 count, for instance, we observe first that over the fours reasons from 2011 to 2014, 182,405 0 – 1 pitches were taken (140,667 balls, 41,738 called strikes). The fielding team gave up an average of 0.322 runs following a ball on a taken 0–1 pitch, while they only gave up an average of 0.265 runs following a called strike on a taken 0–1 pitch. So conditional on an $0-1$ pitch being taken, a called strike saves the fielding team 0.057 runs, on average. Table 5 shows the number of runs scored after a called ball or a called strike for each count, as well as an estimate of ρ. Also shown is the relative proportion of each count among our dataset of taken pitches from 2011 to 2014. We see, for instance, that a called strike is most valuable on a 3–2 pitch but only 2.1% of the taken pitches in our dataset occurred in a 3–2 count. The weighed average run value of a called strike is 0.11 runs, slightly smaller than value of 0.14 used by Judge et al. (2015). The discrepancy is due primarily to the fact that we estimate the run value based only on taken pitches while other attempts at valuing strikes include swinging strikes and strikes called off of foul balls. For the purposes of framing, we argue that our estimates are more natural.

With our posterior samples and estimates of ρ in hand, we can simulate draws from the posterior distribution of $f(c, \xi)$ for each pitch in our dataset. An intuitive measure of the impact of catcher c's framing, which we denote RS for "runs saved" is

$$RS(c) = \sum_{(\mathbf{c},\xi)\in\mathcal{P}_c} f(c, \xi).$$

Table 5: Empirical estimates of run expectancy and run value, with standard errors in parentheses

Count	Ball	Strike	Value of Called Strike ρ	Proportion
0-0	0.367 (0.002)	0.305 (0.002)	0.062 (0.002)	36.2%
0-1	0.322 (0.002)	0.265 (0.004)	0.057 (0.004)	12.5%
0-2	0.276 (0.003)	0.178 (0.007)	0.098 (0.008)	5.5 %
1-0	0.427 (0.003)	0.324 (0.003)	0.103 (0.005)	11.5%
1-1	0.364 (0.003)	0.280 (0.004)	0.084 (0.005)	8.8%
1-2	0.302 (0.003)	0.162 (0.006)	0.140 (0.006)	6.9%
2-0	0.571 (0.007)	0.370 (0.006)	0.201 (0.009)	3.9%
2-1	0.468 (0.005)	0.309 (0.006)	0.159 (0.008)	4.0%
2-2	0.383 (0.004)	0.165 (0.006)	0.218 (0.007)	4.8%
3-0	0.786 (0.013)	0.481 (0.008)	0.305 (0.015)	1.9%
3-1	0.730 (0.010)	0.403 (0.009)	0.327 (0.014)	1.8%
3-2	0.706 (0.008)	0.166 (0.008)	0.540 (0.011)	2.1%

The calculation of $RS(c)$ is very similar to the one used by Judge et al. (2015) to estimate the impact framing has on the game. Rather than using fixed baseline catcher, Judge et al. (2015) reports the difference in expected runs saved relative to a hypothetical average catcher. According to their model, Brayan Pena, our baseline catcher, was no different than this average catcher, so our estimates of RS may be compared to the results of Judge et al. (2015).

Table 6: Top and Bottom10 catchers according to the posterior mean of RS.

Rank	Catcher	Runs Saved (SD)	95% Interval	N	BP
1	Miguel Montero	25.71 (5.03)	[15.61, 35.09]	8086	11.2 (8172)
2	Mike Zunino	22.72 (5.17)	[12.56, 32.31]	7615	20.4 (7457)
3	Jonathan Lucroy	19.56 (5.69)	[8.16, 30.49]	8398	16.4 (8241)
4	Hank Conger	19.34 (3.24)	[12.93, 25.65]	4743	23.8 (4768)
5	Rene Rivera	18.81 (3.69)	[11.63, 25.89]	5091	22.5 (5182)
6	Buster Posey	17.01 (4.14)	[8.79, 25.01]	6385	23.6 (6190)
7	Russell Martin	14.35 (4.41)	[5.85, 22.77]	6388	14.9 (6502)
8	Brian McCann	14.01 (3.95)	[6.18, 21.66]	6335	9.7 (6471)
9	Yasmani Grandal	12.88 (2.98)	[7.18, 18.69]	4248	14.5 (4363)
10	Jason Castro	12.61 (4.43)	[3.80, 21.08]	7065	11.5 (7261)
92	Josmil Pinto	-6.49 (1.41)	[-9.32, -3.76]	1748	-6.9 (1721)
93	Welington Castilo	-6.70 (4.28)	[-15.19, 1.78]	6667	-15.6 (6661)
94	Chris Iannetta	-7.50 (4.46)	[-16.18, 1.08]	6493	-7.3 (6527)
95	John Jaso	-7.76 (2.41)	[-12.50, -3.07]	3172	-11.3 (2879)
96	Anthony Recker	-8.37 (2.33)	[-13.29, -3.93]	2935	-13 (3102)
97	Gerald Laird	-8.68 (1.87)	[-12.29, -4.99]	2378	-9.6 (2616)
98	A. J. Ellis	-12.90 (3.79)	[-20.10, -5.38]	5476	-12.3 (5345)
99	Kurt Suzuki	-17.67 (4.25)	-[26.07, -9.35]	6811	-19.5 (7110)
100	Dioner Navarro	-18.81 (4.68)	[-28.00, -9.40]	6659	-19.8 (6877)
101	Jarrod Saltalamacchia	-23.98 (4.35)	[-32.76, -15.87]	6498	-34 (6764)

According to our model, there is little posterior uncertainty that the framing effects of the top 10 catchers shown in Table 6 had a positive impact for their teams, relative to the baseline catcher. Similarly, with the exception of Welington Castilo and Chris Iannetta, we are rather certain that the bottom 10 catchers' framing had an overall negative impact, relative to the baseline. We estimate that Miguel Montero's framing saved his team 25.71 runs on average, relative to the baseline. That is, had he been replaced by the baseline catcher on each of the 8,086 called pitches he received, his team would have given up an additional 25.71 runs, on average. Unsurprisingly, our estimates of framing impact differ from those of Judge et al. (2015)'s model. This is largely due to differences in the model construction, valuation of a called strike, and collection of pitches analyzed. Indeed, in some cases, (e.g. Montero and Rene Rivera), they used more pitches to arrive at their estimates of runs saved while in others, we used more pitches (e.g. Mike Zunino and Jonathan Lucroy). Nevertheless, our estimate are not totally incompatible with theirs; if we re-scale their estimate to the same number of pitches we consider, we find overwhelmingly that these re-scaled estimates fall within our 95% posterior credible intervals.

4.1 Spatially Aggregate Framing Effect

Looking at Table 6, it is tempting to say that Miguel Montero is the best framer. After all, he is estimated to have saved the most expected runs relative to the baseline catcher. We observe, however, that Montero received 8086 called pitches while Conger received only 4743. How much of the difference in the estimated number of runs saved is due to their framing ability and how much to the disparity in the called pitches they received?

A naive solution is to re-scale the RS estimates and compare the average number of runs saved on a per-pitch basis. While this solution would account for the differences in number of pitches received, it does not overcome the fact that Montero appeared with different players than Vazquez and that the spatial distribution of pitches he received is not identical to that of Conger. In other words, even if we convert the results of Table 6 to a per-pitch basis, the results would still be confounded by pitch location, count, and pitch participants.

To overcome the dependence of our estimates of framing's impact on these confounding factors, we propose to *integrate* $f(c, \xi)$ over all ξ rather than summing $f(c, \xi)$ over $\mathcal{P}_c$. Such a calculation is analogous to the spatially aggregate fielding evaluation (SAFE) of Jensen et al. (2009). In that paper, they first estimated the average number of runs saved by a player successfully fielding a ball put in play as a function of the location and velocity of the ball. They then integrated this against the estimated density of the location and velocity of balls put in play. The result was an overall fielding metric un-confounded by the disparities in players' fielding opportunities. Unlike their setting, however, our confounding factors, ξ, are high-dimensional and, with the exception of the term related to pitch location, discrete. This makes it difficult to construct a kernel density estimate of the distribution of ξ in the same way that Jensen et al. (2009) did. Instead, we propose to integrate $f(c, \xi)$ against the empirical distribution of ξ and define catcher c's spatially aggregate framing effect, denoted SAFE2, to be

$$SAFE2(c) = 4000 \times N^{-1} \sum_{\xi} f(c, \xi), \tag{2}$$

The sum in Equation 2 may be viewed as the number of expected runs catcher c saves, relative to the baseline, if he participated in every pitch in our dataset. We then re-scale this quantity to reflect the impact of his framing on 4000 "average" pitches. We opted to re-scale by SAFE2 by 4000 as the average number of called pitches received by catchers who appeared in more than 25 games was just over 3,992. Of course, we could have easily re-scaled by a different amount.

Once again, we can use our simulated posterior samples of the Θ^u's to simulate draws from the posterior distribution of $SAFE2$. Table 7 shows the top and bottom 10 catchers ranked according to the posterior mean of their $SAFE2$ value, along with the posterior standard deviation, and 95% credible interval for their $SAFE2$ value. Also shown is the a 95% interval of each catchers marginal rank according to $SAFE2$.

We see that several of the catchers from Table 6 also appear in Table 7 The new additions to the top ten, Christian Vazquez, Martin Maldonado, Chris Stewart, and Francisco Cervelli were ranked $13^{th}, 17^{th}, 18^{th}$ and 19^{th} according to the RS metric. The fact that they rose so much in the rankings when we integrated over all ξ indicates that their original rankings were driven primarily

Table 7: Top and Bottom10 catchers according to the posterior mean of $SAFE2$.

Rank	Catcher	Mean (SD)	95% Interval	95% Rank Interval
1.	Hank Conger	16.20 (2.72)	[10.84, 21.50]	[1, 11]
2.	Christian Vazquez	14.33 (2.94)	[8.26, 20.03]	[1, 19]
3.	Rene Rivera	14.04 (2.76)	[8.75, 19.31]	[1, 18]
4.	Martin Maldonado	13.24 (3.33)	[6.73, 19.68]	[1, 24]
5.	Miguel Montero	12.36 (2.42)	[7.50, 16.90]	[2, 22]
6.	Yasmani Grandal	11.90 (2.76)	[6.56, 17.29]	[2, 27]
7.	Mike Zunino	11.78 (2.69)	[6.51, 16.74]	[2, 26]
8.	Chris Stewart	11.63 (3.28)	[5.21, 18.03]	[1, 30]
9.	Buster Posey	11.16 (2.73)	[5.74, 16.51]	[2, 30]
10.	Francisco Cervelli	10.45 (3.21)	[4.06, 16.72]	[2, 36]
92.	Jordan Pacheco	-11.73 (3.80)	[-19.26, -4.30]	[68, 98]
93.	Koyie Hill	-11.79 (5.67)	[-22.48, -0.68]	[53, 100]
94.	Josh Phegley	-12.05 (4.66)	[-21.40, -3.20]	[64, 99]
95.	Austin Romine	-12.76 (9.78)	[-32.14, 5.81]	[30, 101]
96.	Jarrod Saltalamacchia	-14.00 (2.53)	[-19.11, -9.26]	[82, 99]
97.	Brett Hayes	-14.04 (4.06)	[-21.51, -5.93]	[73, 100]
98.	Gerald Laird	-14.96 (3.21)	[-21.17, -8.69]	[81, 99]
99.	Josmil Pinto	-15.04 (3.27)	[-21.57, -8.78]	[82, 100]
100.	Carlos Santana	-22.48 (4.63)	[-31.63, -13.26]	[93, 101]
101.	Tomas Telis	-25.06 (3.85)	[-32.41, -17.27]	[98, 101]

by the fact that they all received considerably fewer pitches in the 2014 season than the top 10 catchers in Table 6. In particular, Vazquez received 3198 called pitches, Cervelli received 2424, Stewart received 2370, and Maldonado received only 1861.

Interestingly, we see that now Hank Conger ranks ahead of Miguel Montero according to the posterior mean SAFE2, indicating that the relative rankings in Table 6 was driven at least partially by disparities in the pitches the two received than by differences in their framing effects. Though Conger emerges as a slightly better framer than Montero in terms of SAFE2, the difference between the two is small, as evidenced by the considerable overlap in their 95% posterior credible intervals.

We find that in 95% of the posterior samples, Conger had anywhere between the largest and 11^{th} largest SAFE2. In contrast, we see that in 95% of our posterior samples, Tomas Tellis's SAFE2 was among the bottom 3 SAFE2 values. Interestingly, we find much wider credible intervals for the marginal ranks among the bottom 10 catchers. Some catchers like Koyie Hill and Austin Romine appeared very infrequently in our dataset. To wit, Hill received only 409 called pitches and Romine received only 61. As we might expect, there is considerable uncertainty in our estimate about their framing impact, as indicated by the rather wide credible intervals of their marginal rank.

4.2 Year-to-year reliability of SAFE2

We now consider how consistent SAFE2 is over multiple seasons. We re-fit our model using data from the 2012 to 2015 seasons. For each season, we restrict attention to those pitches within one foot of the approximate rule book strike zone from that season. We also use the log-odds from the GAM models trained on all previous seasons so that the model fit to the 2012 data uses GAM forecasts trained only on data from 2011 while the model fit to the 2015 data uses GAM forecasts trained only on data from 2011 to 2014. When computing the values of SAFE2, we use the run values given in Table 5 for each season. There were a total of 56 catchers who appeared in all four of these seasons. Table 8 shows the correlation between their SAFE2 values over time.

Table 8: Correlation of SAFE2 across multiple seasons.

	2012	2013	2014	2015
2012	1.000	0.700	0.564	0.406
2013	0.700	1.000	0.714	0.613
2014	0.564	0.714	1.000	0.581
2015	0.406	0.613	0.581	1.000

In light of the non-stationarity in strike zone enforcement across seasons, it is encouraging to find moderate to high correlation between a player's SAFE2 in one season and the next. Interestingly, the correlations between 2012 SAFE2 and 2013 SAFE2 and the correlation between 2013 SAFE2 and 2014 SAFE2 are greater than 0.7, but the correlation between 2014 SAFE2 and 2015 SAFE2 is somewhat lower, 0.581. While this could just be an artifact of noise, we do note that there was a marked uptick in awareness of framing between the 2014 and 2015 seasons, especially among fans and in the popular press. One possible reason for the drop in correlation might be umpires responding to certain catcher's reputations as elite pitch framers by calling stricter strike zones, a possibility suggested by Sullivan (2016).

5 Discussion

We systematically fit models of increasingly complexity to estimate the effect a catcher has on an umpire's likelihood of calling a strike over and above factors like the count, pitch location, and other pitch participants. We found evidence that some catchers do exert a substantially positive or negative effect on the umpires but that the magnitude of these effects are about as large as the count effects. Using the model that best balanced fit and generalization, we were able to simulate draws from the posterior predictive distribution of the called strike probability probability of each taken pitch in 2014. For each pitch, we estimated the apparent framing effect of the catcher involved and, following a procedure similar to that of Judge et al. (2015), we derived an estimate of the impact framing has on the game, RS. Our RS metric is largely consistent with previously reported estimates of the impact of catcher framing, but a distinct advantage is our natural quantification of the estimation uncertainty. We find that there is considerable posterior uncertainty in this metric, making it difficult to estimate precisely the impact a particular catcher's framing had on his team's success.

While the construction of RS is intuitive, we argue that it does not facilitate reasonable comparisons of catchers' framing since, by construction, the metric is confounded by the other factors in our model. We propose a new metric, $SAFE2$, that integrates out the dependence of RS on factors like pitch location, count, and other pitch participants. In some sense, $SAFE2$ compares catchers by computing the impact each catcher's framing would have had had he received every pitch in our dataset. Like RS, there is a considerable uncertainty in our $SAFE2$ estimates. While we are able to separate the posterior distributions of $SAFE2$ of good framers from bad framers, there is considerable overlap in the posterior distributions of $SAFE2$ within these groups, making it difficult to distinguish between the good framers or between the bad framers. Despite this, we find rather high year-to-year correlation in $SAFE2$, though there is a marked drop-off between 2014 and 2015. This coincides with the increased attention on framing in the sports media and sabermetrics community following the 2014 season. One potential explanation for this drop-off is that umpires adjusted their strike zone enforcement when calling pitches caught by catchers with reputations as good framers.

Our findings may have several implications for Major League Baseball teams. The uncertainty in both RS and $SAFE2$ make it difficult to precisely value pitch framing with any reasonable degree of certainty. For instance, the 95% credible interval of Jonathan Lucroy's RS is [8.16, 30.49]. Using the heuristic of 10 expected runs per win and $7M per win (Cameron, 2014; Pollis, 2013), our model suggests that Lucroy's framing was worth anywhere between $5.7M and $\$21.34M$. In light of the non-stationarity between seasons and the recent drop-off in correlation in SAFE2, it is difficult to forecast the impact that any individual catcher's framing will have into the future. The observed overlaps in the posterior distribution of $SAFE2$ means that with a single season's worth of data, we cannot discriminate between good framers with the same certainty that we can separate good framers from bad framers. As a concrete example, our model indicates that both Miguel Montero and Hank Conger were certainly better framers than Jarrod Saltalamacchia, but it cannot tell which of Montero or Conger had a larger, positive impact. From a team's perspective then, our results suggest that while there are catchers who can significantly influence umpires' decisions, it is difficult to estimate and value the impact this influence has over the course of a season with any certainty.

There are several extensions of and improvements to our model that we now discuss. While we have not done so here, one may derive analogous estimates of RS and SAFE2 for batters and pitchers in a straightforward manner. Our model only considered the count into which a pitch was thrown but there is much more contextual information that we could have included. For instance, Rosales and Spratt (2015) have suggested the distance between where a catcher actually receives the pitch and where he sets up his glove before the pitch is thrown could influence an umpire's ball-strike decision making. Such glove tracking data is proprietary but should it be become publicly available, one could include this distance along with its interaction with the catcher indicator into our model. In addition, one could extend our model to include game-state information that frequently appear in other analyses of baseball data such as the ball park, the number of outs in the half-inning, the configuration of the base-runners, whether or not the home team is batting, and the number of pitches thrown so far in the at-bat. One may argue that umpires tend to call more strikes late in games which are virtually decided (e.g. when the home team leads by 10 runs in the top of the ninth inning) and easily include measures related to the run-differential and time remaining into our model. Expanding our model in these directions may improve the overall predictive performance slightly without dramatically increasing the computational overhead. Somewhat more substantively, it is well-known that pitchers try to throw to different locations based on the count, but we make no attempt to model or exploit this relationship. One possible extension would be to fit a separate historical GAM for each combination of batter handedness, pitcher handedness, and count, though care must be taken to ensure sufficient sample size in each bin.

In principle, it is possible to extend our model fitting strategy to encompass the fully parametrized model of Equation 1. The results presented here suggest that doing so will result in substantial overfitting, without extremely careful prior specification. In particular, the first level prior on the umpire-specific catcher effects would need to be very tightly concentrated around some unknown location to induce enough shrinkage. It may also be the case that there exists a subset of catchers whose effect on a subset of umpires may be identified and it would be interesting to consider the question of pitch framing from a variable selection standpoint.

We incorporated pitch location in a two-step procedure: we started from an already quite good generalized additive model, trained with historical data, and used the forecasted log-odds of a called strike as a predictor in our logistic regression model. Much more elegant would have been to fit a single semi-parametric model by placing, say, a common Gaussian process prior on the umpire-specific functions of pitch location, $f^u(x, z)$ in Equation 1. Designing a computationally efficient MCMC procedure for such a model is a key challenge. Durante and Dunson (2014) derived a Gibbs sampler to fit logistic regression models of dynamic networks with Gaussian process priors using Polson et al. (2013)'s Polya-Gamma data augmentation strategy. A similar augmentation could be used here with suitable modifications made for the hierarchical structure. Approximation techniques like those described in Banerjee et al. (2013) may also mitigate the computational burden of this semi-parametric approach. Choosing an appropriate mean function and covariance kernel for the Gaussian process prior is also highly non-trivial.

Finally, we return to the two pitches from the 2015 American League Wild Card game in Figure 1. Fitting our model to the 2015 data, we find that Eric Cooper was indeed much more likely to call the Keuchel pitch a strike than the Tanaka pitch (81.72% vs 62.59%). Interestingly, the forecasts from the hGAMs underpinning our model were 51.31% and 50.29%, respectively. In other words, at least on these two pitches, Cooper deviated substantially from the historical called strike probability estimates. Looking a bit further, had both catchers been replaced by the baseline catcher, our model estimates a called strike probability of 77.58% for the Keuchel pitch and 61.29% for the Tanaka pitch, indicating that Astros' catcher Jason Castro's apparent framing effect (4.14%) was slightly larger than Yankee's catcher Brian McCann's (1.30%). The rather large discrepancy between the apparent framing effects and the estimated called strike probabilities reveals that we cannot immediately attribute the difference in calls on these pitches solely to differences in the framing abilities of the catchers. Indeed, we note that the two pitches were thrown in different counts: Keuchel's pitch was thrown in a 1–0 count and Tanaka's was thrown in a 1–1 count. Similar to our findings from the 2014 season in Figure 5, in 2015, umpires were much more likely to call strikes in a 1–0 count than they were in a 1–1 count, all else being equal. Interestingly, had the Keuchel and Tanaka pitches been thrown in the same count, our model still estimates that Cooper would be consistently more likely to call the Keuchel pitch a strike, lending some credence to disappointed Yankees' fans' claims that his strike zone enforcement favored the Astros. Ultimately, though, it is not so clear that the differences in calls on the two pitches shown in Figure 1 specifically was driven by catcher framing as much as it was driven by random chance.

Acknowledgements

We wish to thank Wharton Computing's Research and Innovation team for their help in using their high-performance computing cluster for this work.

References

Albert, J. (2015). Beyond Run Expectancy. *Journal of Sports Analytics*, 1(1):3 – 18.

Banerjee, A., Dunson, D. B., and Tokdar, S. T. (2013). Efficient Gaussian Process Regression for Large Datasets. *Biometrika*, 100(1):75–89.

Cameron, D. (2008). Win Values Explained: Part Five. http://www.fangraphs.com/blogs/win-values-explained-part-five/.

Cameron, D. (2014). The Cost of a Win in the 2014 Off-Season. http://www.fangraphs.com/blogs/the-cost-of-a-win-in-the-2014-off-season/.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2016). Stan: A Probabilitistic Programming Language. *Journal of Statistical Software*.

Drellich, E. (2014). What Conger Trade Means for Castro and Astros Rotation. http://blog.chron.com/ultimateastros/2014/11/05/astros-make-trade-with-angels-for-catcher-hank-conger/.

Durante, D. and Dunson, D. B. (2014). Nonparametric Bayes Dynamic Modelling of Relational Data. *Biometrika*, 101(4):883–898.

Fast, M. (2010). *What the Heck is PITCHf/x?*, pages 153–158. The Hardball Times Annual.

Green, E. and Daniels, D. (2014). What Does it Take to Call a Strike? Three Biases in Umpire Decision Making. In *Sloan Sports Analytics Conference 2014*.

Holt, R. (2014). How Important is Pitch Framing? http://www.beyondtheboxscore.com/2014/12/12/7375383/how-important-is-pitch-framing.

Jensen, S. T., Shirley, K. E., and Wyner, A. J. (2009). Bayesball: A Bayesian Hierarchical Model For Evaluating Fielding in Major League Baseball. *Annals of Applied Statistics*, 3(2):491–520.

Judge, J., Pavlidis, H., and Brooks, D. (2015). Moving Beyond WOWY: A Mixed Approach to Measuring Catcher Framing. http://www.baseballprospectus.com/article.php?articleid=25514.

Lindbergh, B. (2013). The Art of Pitch Framing. http://grantland.com/features/studying-art-pitch-framing-catchers-such-francisco-cervelli-chris-stewart-jose-molina-others/.

Lindsey, G. R. (1963). An Investigation of Strategies in Baseball. *Operations Research*, 11(4):477–501.

Marchi, M. (2011). Evaluating Catchers: Quantifying the Framing Pitches Skill. http://www.hardballtimes.com/evaluating-catchers-quantifying-the-framing-pitches-skill/.

Mills, B. M. (2014). Social Pressure at the Plate: Inequality Aversion, Status, and Mere Exposure. *Managerial and Decision Economics*, 35:387–403.

Pavlidis, H. (2014). You Got Framed. http://espn.go.com/espn/feature/story/_/id/11127248/how-catcher-framing-becoming-great-skill-smart-teams-new-york-yankees-espn-magazine.

Pollis, L. (2013). How Much Does a Win Really Cost? http://www.beyondtheboxscore.com/2013/10/15/4818740/how-much-does-a-win-really-cost.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian Inference For Logistic Models Using Polya-Gamma Latent Variables. *Journal of American Statistical Association*, 108:1339–1349.

Roegelle, J. (2014). The Strike Zone Expansion Is Out of Control. http://www.hardballtimes.com/the-strike-zone-expansion-is-out-of-control/.

Rosales, J. and Spratt, S. (2015). Who Is Responsible For a Called Strike? In *Sloan Sports Analytics Conference 2015*.

Sanchez, R. (2015). Jonathan Lucroy Needs a Raise. http://espn.go.com/mlb/story/_/id/12492794/jonathan-lucroy-needs-raise.

Sidhu, G. and Caffo, B. (2014). MONEYBaRL: Exploiting Pitcher Decision-Making Using Reinforcement Learning. *Annals of Applied Statistics*, 8(2):926–955.

Sullivan, J. (2015). How the Astros Wound Up With a Bigger Zone. http://www.fangraphs.com/blogs/how-the-astros-wound-up-with-a-bigger-zone/.

Sullivan, J. (2016). The Beginning of the End for Pitch-Framing? http://www.fangraphs.com/blogs/the-beginning-of-the-end-for-pitch-framing/.

Turkenkopf, D. (2008). Framing the Debate. http://www.beyondtheboxscore.com/2008/4/5/389840/framing-the-debate.

Walsh, J. (2007). Strike Zone: Fact vs. Fiction. http://www.hardballtimes.com/strike-zone-fact-vs-fiction/.

Woodrum, B. (2014). The State and Future of Pitch-Framing Research. http://www.hardballtimes.com/the-state-and-future-of-pitch-framing-research/.