Swinging Strike Probabilities

Pitches that start high but end up near the strike zone result in the most swinging strikes

Pitch f/x data

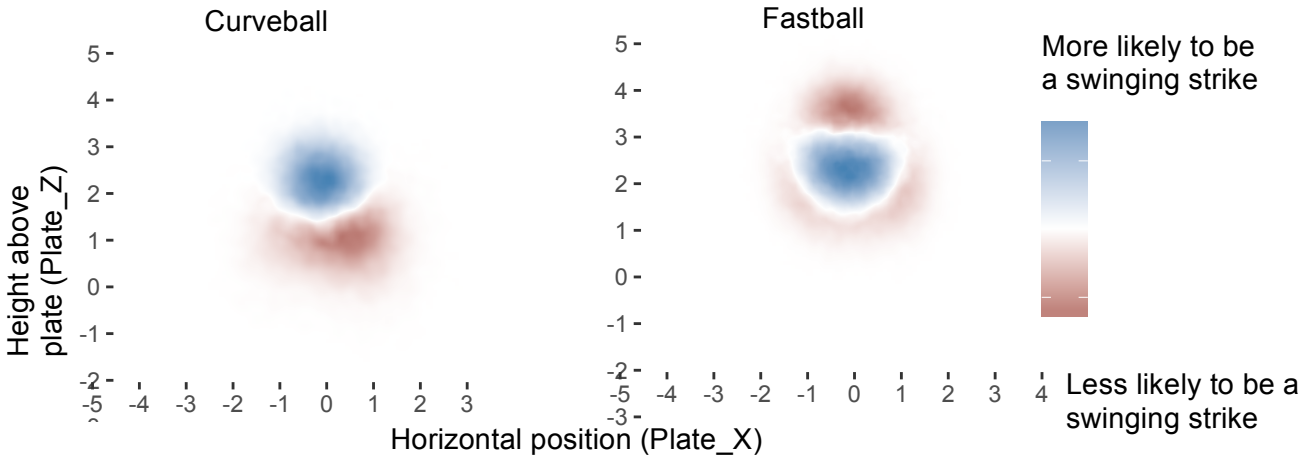
1.2 Million pitches from 2011-2015 seasons were analysed, from 12,143 games, 1757 batters and 1300 pitchers.

Although fastballs make up the majority of pitches thrown (60%) they are the least likely to result in a swinging strike. 15% of all fastballs result in a swinging strike, half the rate of most other pitch types.

Pitch Type	CB	CH	FB	KN	SC	SL
How common are swinging strikes? *(% Swinging Strike)*	29%	29%	15%	22%	18%	31%
How common is this pitch type? *(% of total pitches)*	8%	13%	61%	<1	<1	16%

Visualizing the Results: Plate Location

Swinging strike probability depends highly on the final location of the pitch, since the pitch must be close enough to generate a swing. Locations that are more likely to result in a swinging strike (blue) are in the same position regardless of pitch type. Curveballs that result in swinging strikes tend to cross the plate relatively high, especially compared to other curveballs, while fastballs cross the plate relatively low compared to other fastballs.

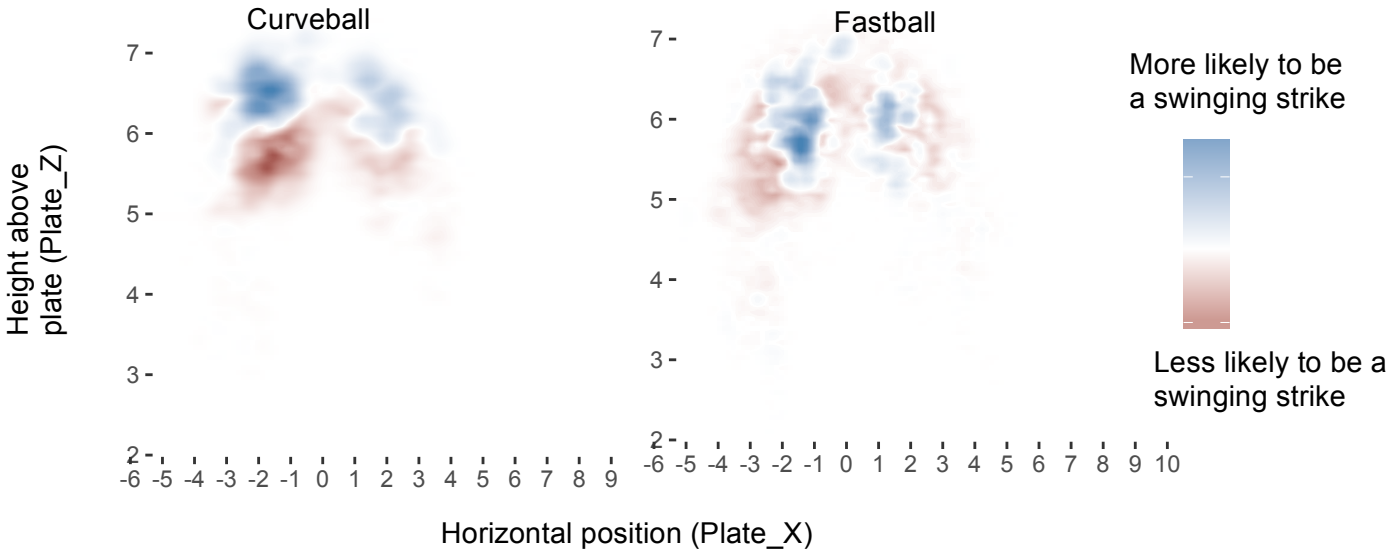


This analysis was produced by calculating the two-dimensional density of swinging strike versus non-swinging strike pitches and subtracting the two densities. This resulted in a "net density" where positive values (blue) are associated with a higher probability of a pitch resulting in a swinging strike compared to negative values (red) which are less likely to result in a swinging strike. Similar density plots were made to visualize the results of all variables.

Statistical Modelling

A logistic regression model was used to estimate the probability of a pitch resulting in a swinging strike. Model 1 compared all features of interested (pitch type, velocities, accelerations etc.) while Model 2 included all two-way interactions between each pitch type and other quantitative variables. This

model was fit on the entire training set, and resulted in a 68% accuracy of prediction for Model 1 and 71% accuracy for Model 2, based on AUC.[i]
The most significant predictors, based on standardized coefficients are:
*** 1. Fastballs with high initial start position (Init_pos_Z)
* 2. Curveballs, Sliders and Knuckleballs, with high initial start position (Init_pos_Z), although knuckleballs in general were less likely to result in swinging strikes



Next steps

1. Incorporate additional data such as pitch count, inning etc.
2. Create pitch-specific models (fastball, curveball etc.)
3. Engineer features of interest, such as difference between starting and finishing position.
4. Evaluate more complex interactions between variables, such as velocity and position. The significance of parameters in this more complex model can be best assessed using a LASSO-regression. This is relatively easy to implement in R provided that adequate computational power is available.

Analysis by E. Nitsch 28 February 2016

[i] Accuracy was based on re-fitting to the training sample. In practice the test-sample would have been used to assess the accuracy of fit, or a holdout on the training sample. Since actual values were not available for the test sample, the accuracy of prediction cannot be assessed on a real trial at this stage. I expect the second model may be in danger of overfitting the data, which must be monitored when assessing the accuracy of fit on a true holdout sample.