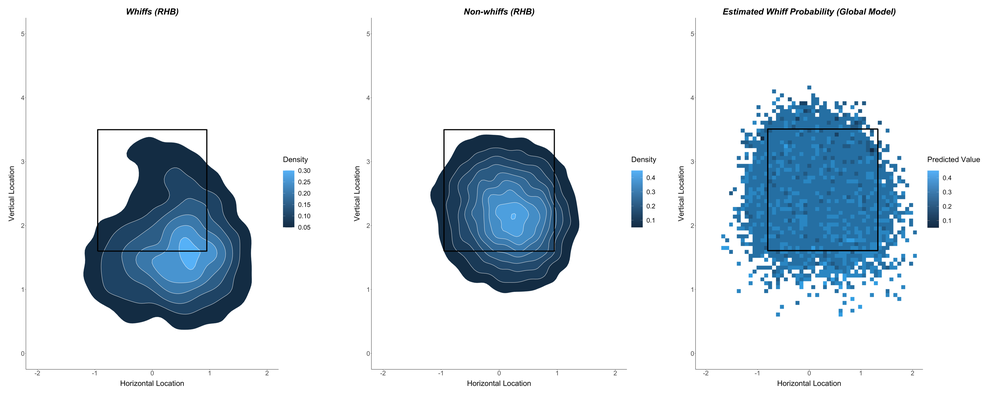
Assessing Spatial Heterogeneity in Whiff Rate Using Geographically Weighted Regression (GWR)

Background

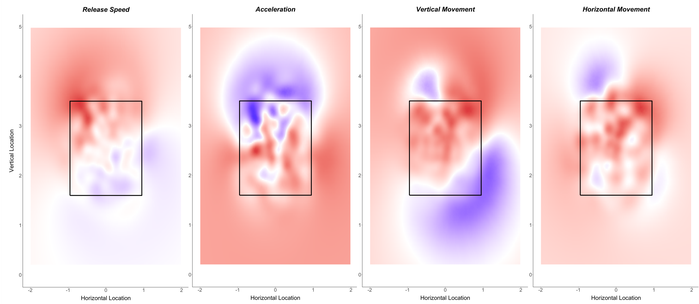
While the use of **spatial statistical analysis** has become increasingly ubiquitous in academic fields such as geography and ecology, its adoption in sports analytics is far less extensive—introducing an exciting opportunity for professional sports organizations to leverage a novel and under-utilized family of statistical approaches for a variety of complex data problems. Spatial statistics has particularly germane applications in the sport of baseball, where play-by-play data is nearly always accompanied by each observation's vertical and horizontal locations in space—whether it be the coordinates of a pitch as it crosses home plate or the hit location of a batted ball. In these cases, the analysis of baseball data can benefit from the use of spatial statistical techniques.

One such spatial approach is **Geographically Weighted Regression (GWR)**. Whereas the fitted coefficient values of a typical parametric model may fail to capture spatial variations in the data, GWR models the **local relationships** between predictors and the response variable. In order to demonstrate the capacity of GWR to create accurate and insightful statistical inferences from spatial data, I create a model of **whiff probability** as it varies in coordinate space. For this analysis, whiffs are defined as events where a batter swings at a pitch and fails to make contact. The non-uniform distribution of whiffs in space—evinced below—suggests an underlying spatial component to the data generating process. As a result, a global parametric model produces confounded coefficient estimates and fails to account for pitch location. GWR produces a more reliable estimation of the relationships between different pitch quality parameters and whiff probability.

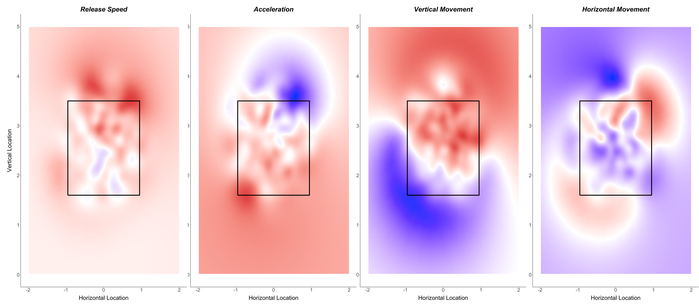


Modeling Results

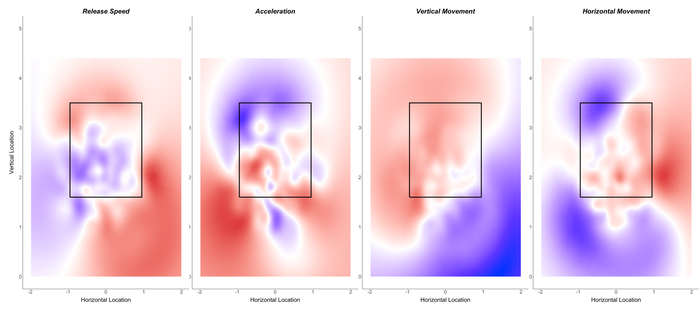
Fastball Against LHB



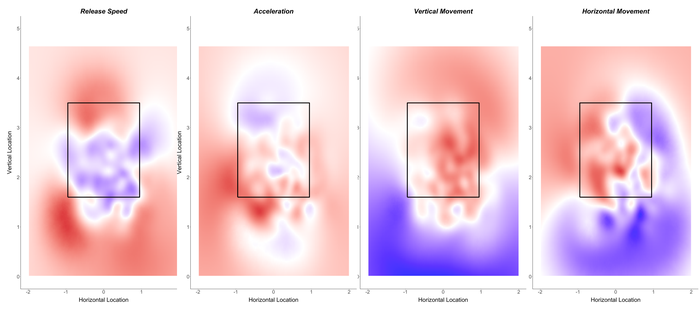
Fastball Against RHB



Off-speed Against LHB



Off-speed Against RHB



Negative Coefficient Estimate — *Positive Coefficient Estimate*

Methodology

The first step in any spatial approach is empirically confirming whether the data exhibits spatial heterogeneity. The degree of **spatial autocorrelation** can be determined statistically by calculating the Moran's I measure of the response variable's spatial distribution. For whiffs in the pitch-by-pitch data, the calculated Moran's I statistic of 0.1418 has a p-value smaller than the $\alpha = 0.01$ significance level. In this case, the p-value represents the probability of observing the calculated statistic from a non-spatially heterogeneous distribution of Moran's I, as approximated by a Monte-Carlo simulation.

With sufficient statistical evidence of spatial autocorrelation, I begin constructing the GWR. I start by specifying a global parametric model (i.e., logistic regression) and performing exploratory data analysis to identify relevant **model features**. Because GWR is highly computationally intensive, it is best to exclude redundant predictors to avoid unnecessary training time. This is especially true when the goal of the statistical analysis is inference as opposed to prediction—here, I am particularly interested in the coefficient estimates of just a small number of pitch quality parameters. Next, I use leave-one-out cross validation to select the optimal bandwidth of the Gaussian kernel—a bell-shaped function that weights observations based on their proximity to the location being modeled. The **kernel bandwidth** determines the neighborhood distance used for each local regression equation, and thus dictates the local variation of parameter estimates. Finally, I run the GWR with the selected predictors and hyperparameters, training the model on a random sample of 25,000 pitches. The resulting **coefficient estimates** are rasterized and plotted at each spatial point.

Interpretation

The above plots illustrate the **spatial variation** in the estimated relationships between each of the selected **parameters** and whiff probability for four different GWR models. Separate models were created for fastballs and off-speed pitches to account for differences in the distributions of both the response and predictors with respect to pitch type. Each plot is rendered from the **catcher's perspective**.

Red shaded regions represent locations in coordinate space where the given parameter has a positive estimated relationship with whiff rate. Blue shaded regions represent locations in coordinate space where the given parameter has a negative estimated relationship with whiff rate*. For example, the release speed of a fastball is, on average, positively associated with whiff rate in the upper left corner of the strike zone against left-handed batters.

*One parameter where the interpretation is slightly more ambiguous is vertical movement, since observations can take on both positive (upward) and negative (downward) values. In this case, blue shaded regions represent locations where an increase in downward break is positively associated with whiff rate and an increase in "upward" break (or rise) is negatively associated with whiff rate. Red shaded regions represent locations where an increase in downward movement is negatively associated with whiff rate and an increase in upward break is positively associated with whiff rate.

Extensions and Sources

Having established a framework for assessing spatial heterogeneity in baseball pitch-by-pitch data, it is possible to extend the outlined modeling approach to new variables of interest, including continuous responses (e.g., exit velocity, BA, xwOBA). There are also several improvements to be made in the GWR construction. This includes standardizing parameter values for more meaningful and consistent interpretations of coefficient estimates. Additionally, because over 500,000 individual pitches are observed each season, it is possible to train the model on larger datasets to obtain more accurate results.

All data was sourced from MLB **Statcast**. Statistical analysis was performed in **R** using the *Tidyverse* and *spgwr* packages.

Riley Leonard, Cornell University (April 2023)
Presented at the *Cornell Sports Research Conference*