

1 Introduction

“The STRIKE ZONE is that area over home plate the upper limit of which is a horizontal line at the midpoint between the top of the shoulders and the top of the uniform pants, and the lower level is a line at the hollow beneath the kneecap. The Strike Zone shall be determined from the batter’s stance as the batter is prepared to swing at a pitched ball.”mlb (2016) Those two sentences define the strike zone in the rulebook for Major League Baseball (MLB). Calling balls and strikes is easier said than done, however, as pitches cross home plate at speeds of up to 100mph and frequently move in different directions as they cross the plate. MLB umpires have faced increased scrutiny in recent years as video technology enables every player, coach, fan, and league official to be a critic and review every call an umpire makes for accuracy.

We are interested not only in an umpire’s accuracy – the percentage of calls that matches classification according to the strike zone defined in the rule book – but to be able to learn a representation of a strike zone that can define a probability of a pitch being called a ball or strike based on its location, and, later on, game situation. We use nonlinear classification methods to learn the strike zones, then reduce their dimension using kernel PCA, enabling us to perform inference on the principal components for each strike zone. This methodology makes it possible to determine how umpires’ strike zones change in numerous game situations. Our work is useful for players seeking to gain an edge on the field, league officials determining which umpires should be promoted, and for umpires who are looking to improve their skills.

References

(2016): “The official rules of major league baseball,” .