Comments on Thesis

Ben Goodrich

July 1, 2022

I am giving you an A- on your thesis. You have a good question that is germane enough to the QMSS but one where you can use your knowledge of physics that most QMSS students or sabermetricians do not have. You have a lot of relevant data and an obvious missingness problem to tackle. All of those are good ingredients for a QMSS thesis, even though a lot of QMSS theses are lacking in one or more of these areas. Your R code is kind of a bunch of chunks without a lot of explanation, but it runs and does much more than you could have done when you started the QMSS program. I personally feel as if I understood your description of the problem, the physics of seam shifted wake (SSW), and the recent literature on it, but that was more because I had already read those papers. I kind of doubt that a random QMSS person or a random general manager of a baseball team would be able to follow the details, even though they might be interested in your conclusion. But the main drawback of your thesis is that it does not quite reach a conclusion as to how much SSW matters.

The only real mistake in your thesis is that a unit vector is defined as a vector whose *squared* elements sum to 1, rather than a simplex whose elements are non-negative and sum to 1. Healey's equation (11) is

$$\boldsymbol{\omega}^\top = \left[\mp\sqrt{1-\omega_y^2}\cos\left(\theta\right) \quad \omega_y \quad \mp\sqrt{1-\omega_y^2}\sin\left(\theta\right)\right]$$

where $\theta = \tan^{-1}\left(\frac{\mp\sqrt{1-\omega_y^2}\sin(\theta)}{\mp\sqrt{1-\omega_y^2}\cos(\theta)}\right) = \tan^{-1}\left(\tan\left(\theta\right)\right)$ is what Statcast refers to as the "spin_axis" (but mapped to radians). Healey's formula yields a unit vector whenever $-1 \leq \omega_y \leq 1$ since,

$$\begin{aligned}\boldsymbol{\omega}^\top\boldsymbol{\omega} &= \left(1-\omega_y^2\right)\cos^2\left(\theta\right) + \omega_y^2 + \left(1-\omega_y^2\right)\sin^2\left(\theta\right) \\ &= \left(1-\omega_y^2\right)\left(\cos^2\left(\theta\right) + \sin^2\left(\theta\right)\right) + \omega_y^2 \\ &= 1-\omega_y^2+\omega_y^2 = 1.\end{aligned}$$

I could have saved you from that mistake if I had gotten to see a draft earlier, although it would just allow you to omit some of the plots.

So, the question becomes what is this spin vector for a pitch or pitcher's pitch type? If we make the (perhaps strong) assumption that the spin vector is distributed according to the Fisher-von Mises distribution, then the aggregated data underlying the Statcast Leaderboard provides sufficient statistics for maximum likelihood estimates (MLEs), which means that you get the same MLEs from the aggregated data that you would get if you had access to the disaggregated data

$$\left[\widehat{\mu}_x \quad \widehat{\mu}_y \quad \widehat{\mu}_z\right] = \frac{1}{\widehat{R}}\left[\overline{\omega}_x \quad \overline{\omega}_y \quad \overline{\omega}_z\right]$$

where $\widehat{R} = \sqrt{\overline{\omega}_x^2 + \overline{\omega}_y^2 + \overline{\omega}_z^2}$ and $\overline{\omega}_x = \frac{1}{N}\sum_{n=1}^{N} \omega_{x,n}$ seems to be what Statcast calls "image_spin_x". As it turns out, $\widehat{R}$ is pretty close to 1 for most pitches in the aggregated data, so the concentration parameter of the Fisher-von Mises distribution is pretty high since $\widehat{\kappa} \approx \frac{\widehat{R}\left(3-\widehat{R}\right)}{1-\widehat{R}^2}$ and the circular standard deviation, $\widehat{\sigma} = \sqrt{-2\ln\left(\widehat{R}\right)}$, is pretty low. That suggests either that the assumption of the Fisher-von Mises distribution is too restrictive or else that there is not a lot of variability in the spin vector for pitches of a particular type by the same pitcher. On the other hand, it seems as if a pitcher needs to release the pitch very consistently in order to generate substantial movement due to SSW.

Anyway, the aggregated data allows you to make the same kind of conclusions that Healy sought without having to make a lot of his assumptions in order to estimate an average spin vector from the pitch-by-pitch data. It seems that what Statcast calls "movement_inferred" is the result of one of Nathan's formulas applied to the "9Ps" data, which assumes a quadratic flight path of the pitch estimated from noisy points. That was as good as anyone could do before Hawkeye was introduced, but if you knew the spin vector and the velocity vector, then the Magnus effect should take the ball in the direction of their cross-product,

$$\boldsymbol{\omega} \times \mathbf{v} \quad = \quad \begin{bmatrix} \omega_y v_z - \omega_z v_y \\ \omega_z v_x - \omega_x v_z \\ \omega_x v_y - \omega_y v_x \end{bmatrix},$$

And since the velocity of a pitch in the y-direction is so much larger in magnitude than in the x and z directions,

$$-\tan^{-1}\left(\frac{\omega_y v_z - \omega_z v_y}{\omega_x v_y - \omega_y v_x}\right) \quad \approx \quad -\tan^{-1}\left(\frac{-\omega_z y_y}{\omega_x v_y}\right) = \tan^{-1}\left(\frac{\omega_z}{\omega_x}\right) = \theta.$$

These calculations are confusing because it depends on whether you download the data from the pitcher or the batter's perspective, but if you download from the batter's perspective, then the "hawkeye_measured" variable is usually approximately the same as `atan2(mu_z, mu_x) * 180 / pi + ifelse(mu_z < 0, 360, 0)`. I think the couple outliers that are visible off the 45-degree line in the corners of a plot of these two quantities are due to neglecting v_x and v_z in the Magnus direction calculation, which is occasionally sufficient to flip the sign in the denominator. In any event, if there were no seam-shifted wake (SSW), then all the points would be on the 45-degree line so the non-outlier deviations from the 45-degree line suggest SSW. Your figure 10 contrasts "hawkeye_measured" with "movement_inferred" rather than `atan2(mu_z, mu_x) * 180 / pi + ifelse(mu_z < 0, 360, 0)`, so some of the deviation from the 45-degree line is just due to noise in their "movement_inferred" calculation.

What can be said about ω_y on a particular pitch in the absence of the Hawkeye measurement of it? Here, I think your supervised learning process was fine, but as it turns out, ω_y is hard to predict well with methods like random forrests and MARS. Or more specifically, it is decent for pitches like four-seam fastballs and curveballs where there is thought to be little SSW and worse for the pitches that might have substantial SSW. Even though the total "spin_rate" is publicly reported, since the proportion of "active_spin" (a.k.a. spin efficiency) is approximately $\sqrt{1-\omega_y^2}$, without a good (and independent of the 9Ps) estimate of ω_y, it is hard to determine what the flight of the pitch would be if there were only drag and Magnus effects.

Since "spin_axis" is rounded to the nearest degree in the pitch-level data anyway, you could impute a spin vector via rejection sampling. For each pitch, draw a proposed spin vector from a Fisher-von Mises distribution with parameters $\widehat{\boldsymbol{\mu}}$ and $\widehat{\kappa}$ (using `Rfast::rvmf` in R) and keep that proposal if and only if its $\theta = \tan^{-1}\left(\frac{\omega_z}{\omega_x}\right) \times \frac{180}{\pi} + (\omega_z < 0\,?\,360 : 0)$ rounds to the observed "spin_axis" for the pitch. That is hard to do with dplyr because you might

have to draw multiple proposals for a pitch before you get one that is accepted, but it can be done with a `for` loop.

That would still leave you with the question of how much SSW was there on a particular pitch? With an imputed spin vector, I suppose you could do a Healey / Nathan thing, where you decompose acceleration (after removing drag and gravity) into lift and side forces and see whether the latter is associated with success on the pitch (whiff, weak contact, etc.). But one way or another, you need to be able to produce an answer to the question of how much does SSW matter?