# SAP Predictive Analysis and the MLB Post Season

Since September is drawing to a close and October is rapidly approaching, I decided to hunt down some baseball data and see if we can draw any insights on MLB statistics and post-season performance. I'm definitely not an expert in sports statistics, but I pulled the well-known Lahman baseball database and calculated some summary statistics to evaluate team performance. I used SAP Predictive Analysis (with SAP Lumira visualization components) to visualize the data and perform some predictive analytics for the 2013 post season.

## Metrics and Data

I pulled just a few metrics to summarize batting and fielding performance of each team. The metrics I pulled were:

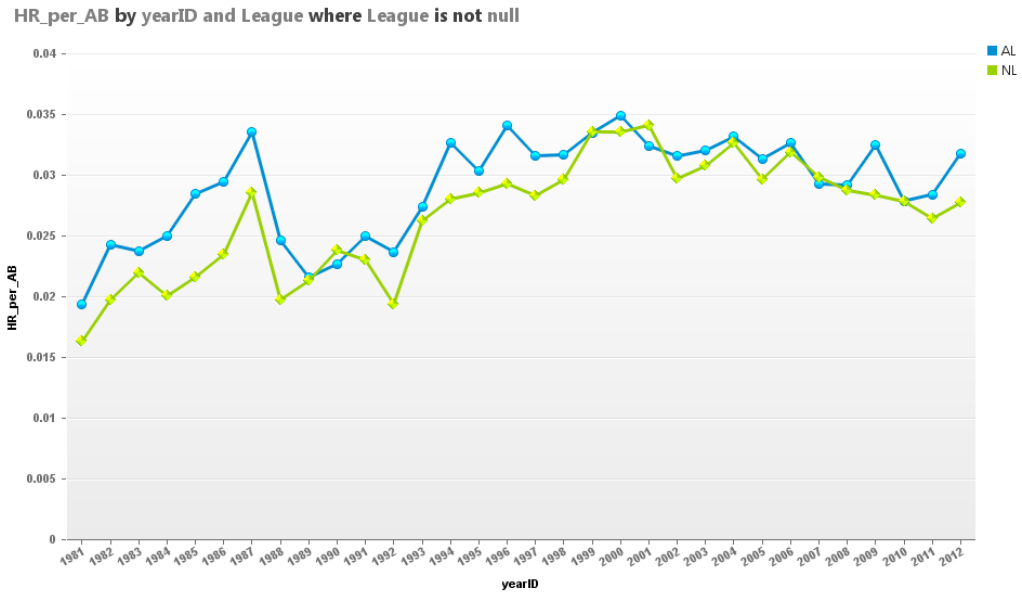
- Put Outs and Errors per inning out for each of the main positions (1B, 2B, 3B, C, CF, RF, LF, SS)
- HR, H, R, 2B, 3B, SO, SB, CS, SF, SH per At Bat

For all teams from 1981 – 2012 during the regular season only.

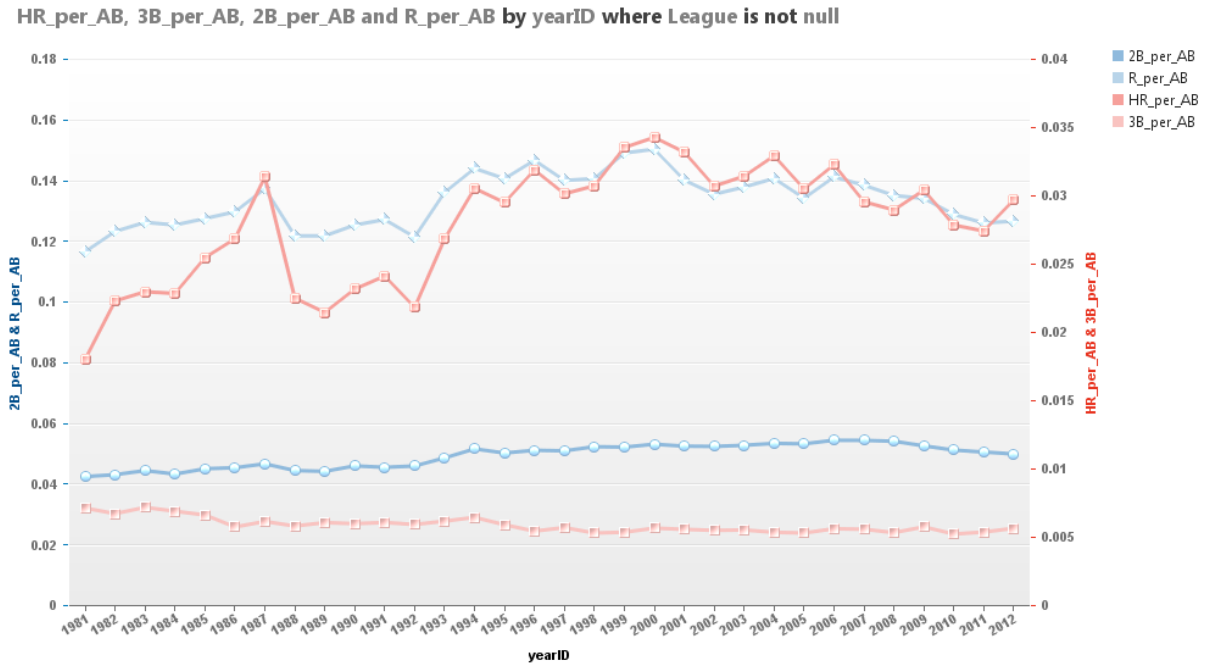
## Visualizing Changes Over Time

Not being familiar with the ins and outs (get it?) of baseball, I decided to look for trends over time—would these metrics be consistent, or have strategies changed over the years?

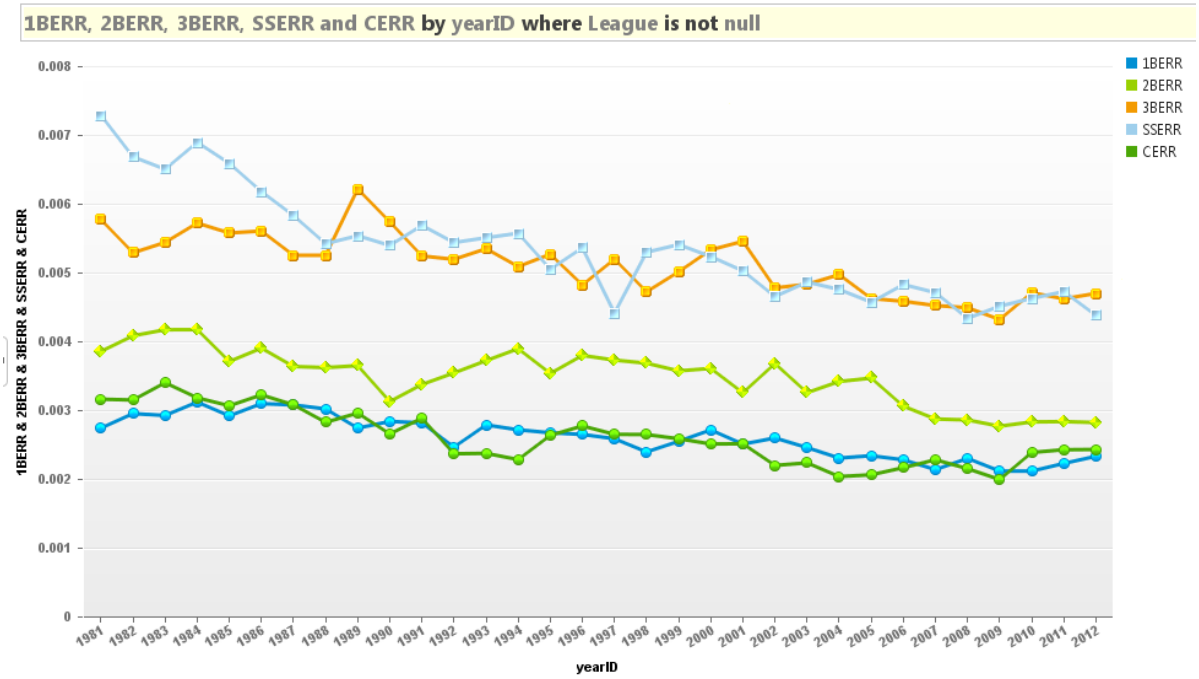
I can look at these trends over time by league (HR increased through the early 2000s and have since been decreasing):

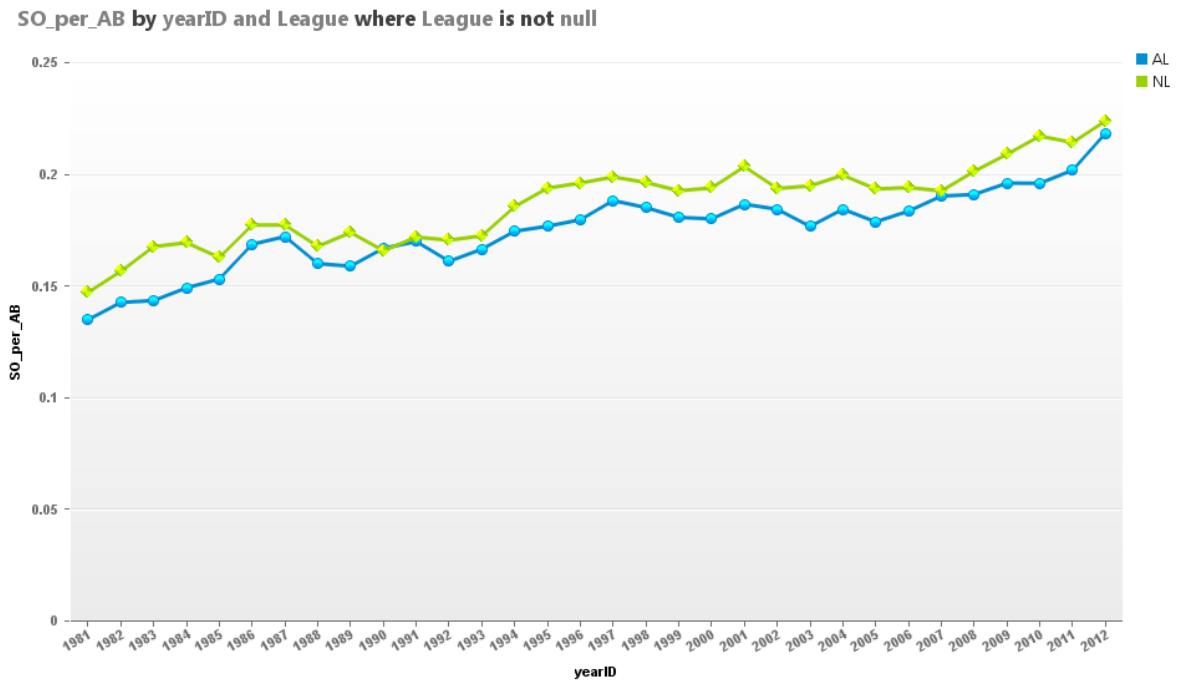


And by metric, it looks like the frequency of 2B hits has been increasing, while 3B hits have been steadily decreasing. Though interestingly, the number of runs per AB has been relatively steady since 1993.

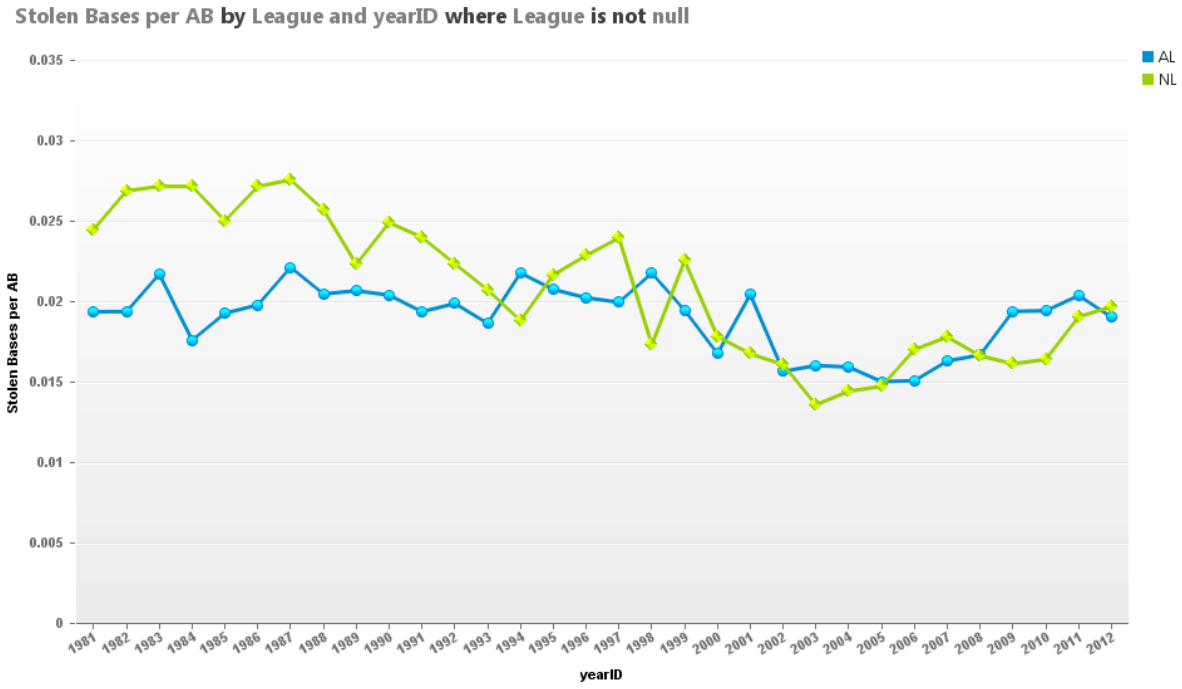



Perhaps these decreasing trends in scoring are driven by an improvement in fielding or pitching? Steady decreases in errors per inning out for all infield positions and increases in strikeouts per at bat suggest this could be the case.



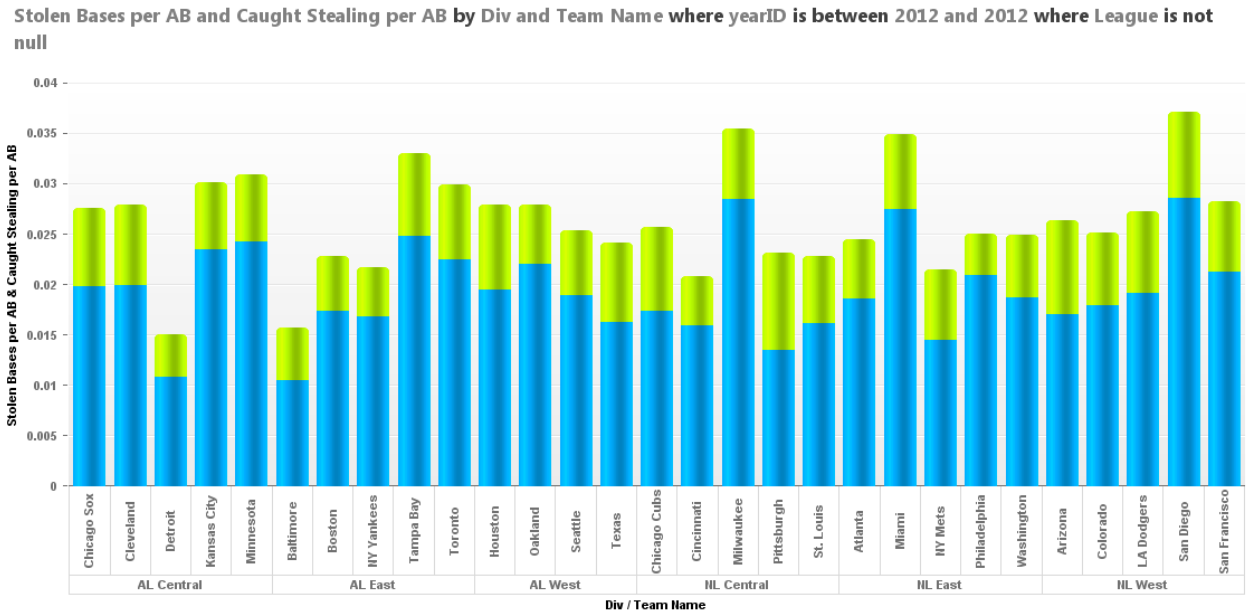



Stolen bases have always fascinated me in baseball, so I wanted to look at the prevalence of stealing over time. Interestingly, this is the one metric that showed significant differences between the leagues with NL teams stealing much more frequently than AL teams through the early 90s, though over time they have tracked much more closely and now show similar trends.

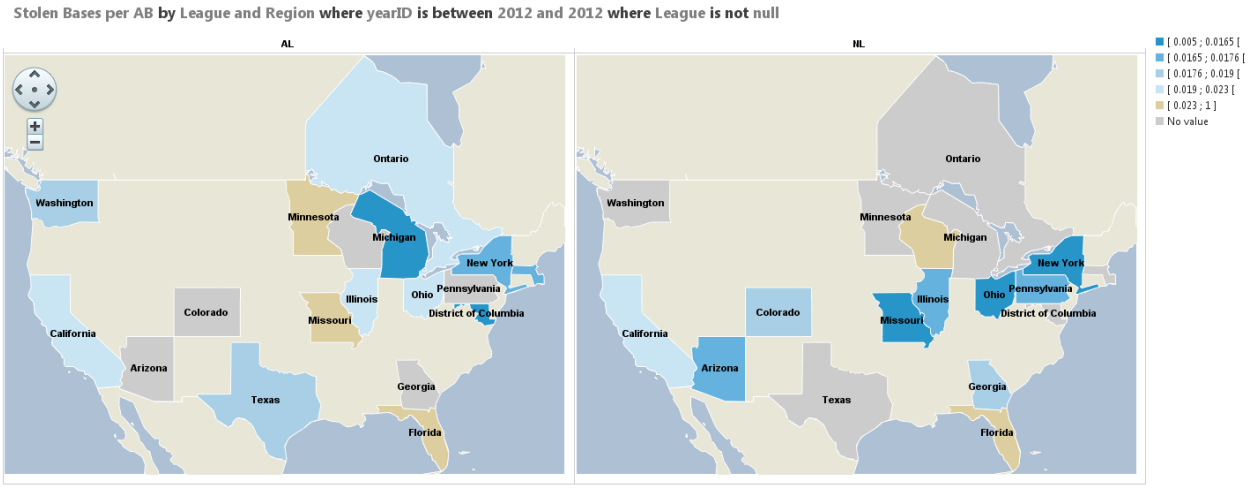


## Visualizing Differences by Team

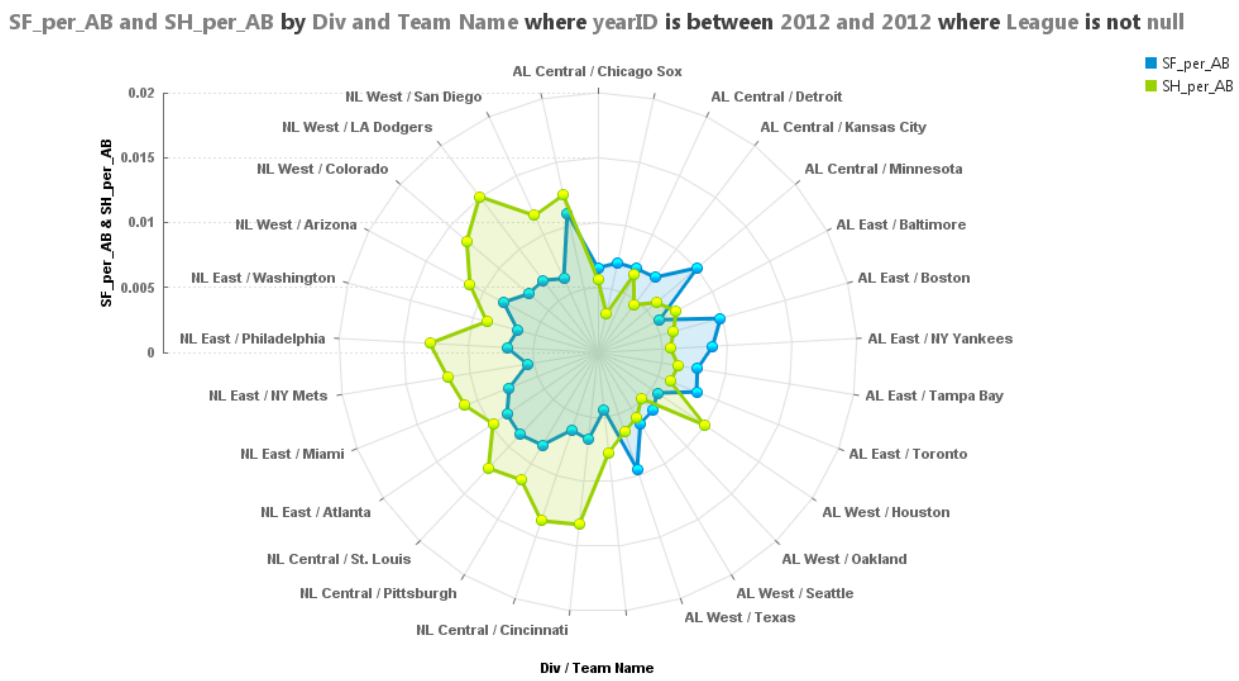
Still on stolen bases, which teams are most effective at stealing? For the 2012 season, Milwaukee, Miami, and San Diego had the highest frequency of stealing, with Milwaukee, Miami, Oakland, Minnesota, and Kansas City stealing most effectively (fewest CS per SB). Pittsburgh, Arizona, and Baltimore are the least effectiveat stealing bases, with Pittsburgh successfully stealing only 58.4% of the time.



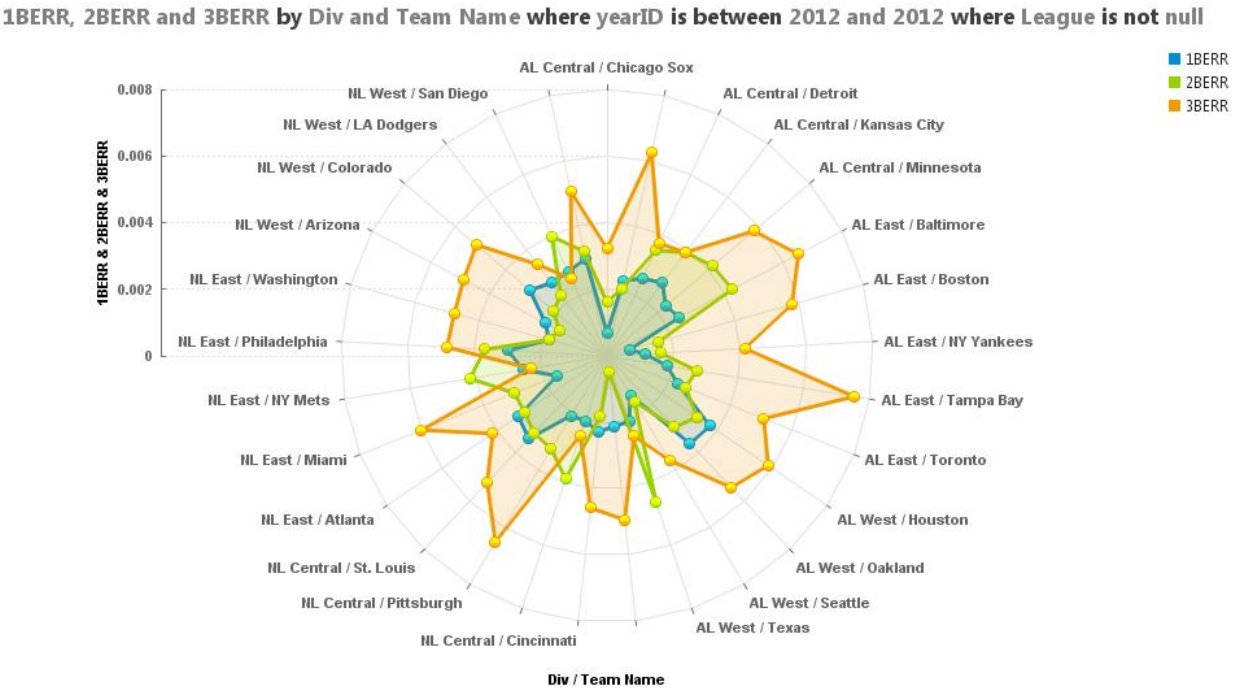

We can also visualize the base stealing geographically by state, with darker blue states stealing less than lighter blue states. Generally, base stealing seems to be less popular in the AL and the East.



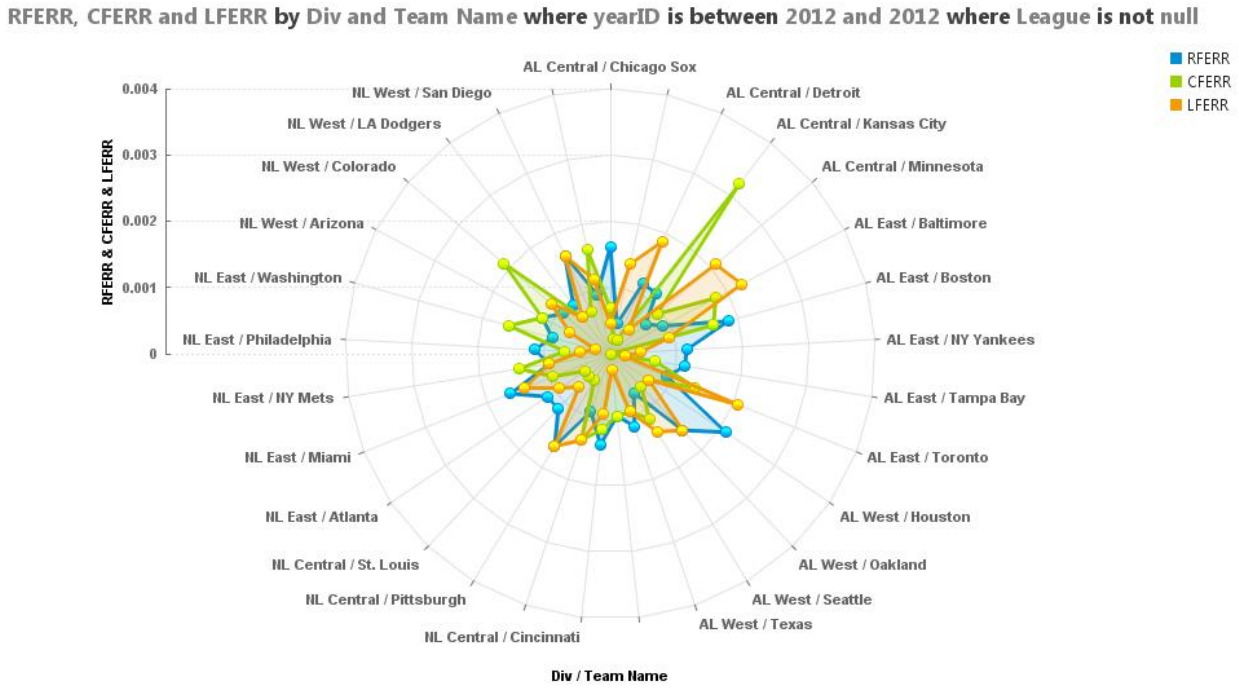

Shifting over towards batting strategy, we can compare the frequency of sacrifice flies and hits by team. Sacrifice hits seem much more common in NL teams, and even sacrifice flies are relatively uncommon in the AL except for a few teams (Minnesota, Texas, Toronto, Tampa, the Yankees, and Boston).




Let's look at fielding performance for infielders by team in 2012. Interestingly, it appears that the NL teams have lower frequency of errors across all positions, but especially for 3B errors.



Outfield fielding performance seems to differ less between leagues, but the NL still seems to have slightly better fielding. Maybe Kansas City should look at firing their center fielder (though it looks like their main center fielder was injured in April).




## Predicting Post-Season Performance

Enough with understating the stats. I'm interested in being able to figure out how my precious Boston Red Sox will fare in the post season this year, since they're at the top of the leader board at the end of the season. I built a modeling dataset that uses the regular season team statistics discussed above to predict the outcome of a post-season matchup. This model uses all post-season games from 1981-2012 to train the model, and I have scored it on every possible (and impossible) matchup for the 2013 post season. (2013 regular season data was pulled from ESPN and a variety of other sources since it is not yet included in the Lahman database).

The post season information in the Lahman database is at the series-level, so I used this to create simulated records for each game in a series (ex. if there was a 4-3 series between Team 1 and Team 2, there are 4 records with Team 1 winning and 3 with Team 2 winning), so this model predicts the outcome of a series between teams. It also only takes into account position-level statistics, not statistics of any particular player, so if the team lineup changes significantly in the post season due to injury, it will likely not be as accurate. In order to predict the likelihood of winning, I used my Custom R Logistic Regression algorithm, but similar analysis could also be done with a decision tree model.

In addition to the team metrics discussed previously, I calculated some comparison metrics for each potential game—for example, the 1BERR_diff is the difference between the error rates for 1st basemen

for the two teams. In addition to these difference variables, I also included the series level (DIV, WS, WC, CS) and league type (AL, NL, or IL for interleague play). This helps gauge the actual matchup between teams rather than simply relying on a single team's stats to predict their likelihood to win in the post season. Adding these difference metrics increase the accuracy of prediction (AUC) from approximately 0.69 to 0.722. Which variables were most predictive in the model?

| | | | | |
|---|---|---|---|---|
| 1BERR_diff | SO_per_AB | CERR | DFERR | SF_per_AB |
| SSERR_diff | 2BERR_diff | 2BPO_diff | H_per_AB | R_per_AB |

Enough statistics. Let's get to the ~~betting~~ baseball. While my Red Sox have clinched the divisional title, I'm not sure exactly who they will be playing in the AL Division Series. However, my model can evaluate the performance of every team against every other team in the division at the division level. The model actually returns 2 probabilities: the probability of BOS vs. another team and another team vs. BOS. Comparing the probability of BOS to win for each of these scenarios gives us an idea of the outcome of the matchup. In the chart below, BOS wins against all teams except BAL, where BAL's probability of winning is 56% compared to BOS's probability of winning (52%), as well as Detroit and Texas. So, I will thank my lucky stars that Boston probably does not have to play Baltimore in the Division Series, and keep my fingers crossed for the final matchup (hopefully Detroit over Texas).

| | vs. BOS | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **BOS vs.** | **Baltimore** | **Boston** | **Chicago Sox** | **Cleveland** | **Detroit** | **Houston** | **Kansas City** | **LA Angels** | **Minnesota** | **NY Yankees** | **Oakland** | **Seattle** | **Tampa Bay** | **Texas** | **Toronto** |
| Baltimore | | 0.56 | 0.72 | 0.59 | 0.36 | 0.73 | 0.57 | 0.71 | 0.54 | 0.63 | 0.74 | 0.55 | 0.53 | 0.52 | 0.72 |
| Boston | 0.52 | | 0.7 | 0.57 | 0.34 | 0.71 | 0.55 | 0.69 | 0.52 | 0.61 | 0.73 | 0.53 | 0.51 | 0.5 | 0.71 |
| Chicago Sox | 0.28 | 0.29 | | 0.32 | 0.15 | 0.47 | 0.31 | 0.44 | 0.28 | 0.36 | 0.49 | 0.29 | 0.27 | 0.26 | 0.46 |
| Cleveland | 0.5 | 0.52 | 0.69 | | 0.32 | 0.7 | 0.54 | 0.68 | 0.5 | 0.6 | 0.71 | 0.51 | 0.49 | 0.48 | 0.69 |
| Detroit | 0.57 | 0.59 | 0.74 | 0.62 | | 0.75 | 0.6 | 0.73 | 0.57 | 0.66 | 0.76 | 0.58 | 0.56 | 0.55 | 0.75 |
| Houston | 0.27 | 0.29 | 0.44 | 0.32 | 0.15 | | 0.3 | 0.43 | 0.27 | 0.35 | 0.48 | 0.28 | 0.26 | 0.25 | 0.45 |
| Kansas City | 0.43 | 0.44 | 0.61 | 0.48 | 0.26 | 0.63 | | 0.6 | 0.42 | 0.52 | 0.64 | 0.44 | 0.41 | 0.4 | 0.62 |
| LA Angels | 0.23 | 0.24 | 0.39 | 0.27 | 0.12 | 0.4 | 0.25 | | 0.23 | 0.3 | 0.42 | 0.24 | 0.22 | 0.21 | 0.4 |
| Minnesota | 0.45 | 0.47 | 0.64 | 0.5 | 0.28 | 0.65 | 0.48 | 0.63 | | 0.55 | 0.67 | 0.46 | 0.44 | 0.43 | 0.64 |
| NY Yankees | 0.41 | 0.42 | 0.59 | 0.46 | 0.24 | 0.61 | 0.44 | 0.58 | 0.4 | | 0.62 | 0.42 | 0.39 | 0.39 | 0.6 |
| Oakland | 0.44 | 0.45 | 0.63 | 0.49 | 0.27 | 0.64 | 0.47 | 0.61 | 0.43 | 0.53 | | 0.45 | 0.42 | 0.42 | 0.63 |
| Seattle | 0.3 | 0.31 | 0.47 | 0.34 | 0.16 | 0.49 | 0.32 | 0.46 | 0.29 | 0.38 | 0.5 | | 0.28 | 0.28 | 0.48 |
| Tampa Bay | 0.43 | 0.45 | 0.62 | 0.48 | 0.26 | 0.63 | 0.46 | 0.61 | 0.42 | 0.52 | 0.65 | 0.44 | | 0.41 | 0.62 |
| Texas | 0.84 | 0.85 | 0.92 | 0.87 | 0.71 | 0.92 | 0.86 | 0.91 | 0.84 | 0.88 | 0.93 | 0.85 | 0.83 | | 0.92 |
| Toronto | 0.09 | 0.09 | 0.17 | 0.11 | 0.04 | 0.18 | 0.1 | 0.16 | 0.09 | 0.12 | 0.19 | 0.09 | 0.08 | 0.08 | |

Want to follow my predictions for the rest of the season? Interested in matchups for your favorite team? Follow me on Twitter @HillaryBlissDFT to find out the results! Go SOX!

Hillary Bliss, Analytics Practice Lead
Decision First Technologies
Hillary.bliss@decisionfirst.com
twitter @HillaryBlissDFT

Hillary Bliss is the Analytics Practice Lead at Decision First Technologies, and specializes in data warehouse design, ETL development, statistical analysis, and predictive modeling. She works with clients and vendors to integrate business analysis and predictive modeling solutions into the organizational data warehouse and business intelligence environments based on their specific operational and strategic business needs. She has a master's degree in statistics and an MBA from Georgia Tech.