Pitcher ELO ratings

Once the simulations for the MLB season were complete (see previous article on how the predictions are made http://simodds.com/Static/BaseballResearch.pdf), I started looking for ways to improve the predictive performance for individual games. In aggregate the ELO prediction performance is reasonable. However, since baseball is only partially a "team" sport, the aggregate approach breaks down for any individual game. Why is baseball not a true team sport? The starting pitcher has a large effect on the outcome of any individual game. Since ELO ratings act over all games it will predict too many losses for an elite pitcher, and too many wins for a complete bum. This doesn't really matter for a season long prediction simulation but it could matter for predicting the outcome of a playoff series.

How do we determine the performance of a pitcher in the context of an ELO rating? This is trickier than it sounds. Since ELO is purely based on a team's aggregate performances (wins or losses and margin of victory or loss), using a pitcher's individual statistics is not particularly helpful. To account for the importance of a starting pitcher's stats I'd have to find a way to map them to ELO. That would likely mean that team statistics would need to be included as well and at that point I might as well go to a full player level baseball simulation. As I've always said, a player level simulation is likely to be more accurate than an ELO based system. However, the goal of these sims is simplicity first, followed by the most accuracy possible using a team aggregate rating.

So, since ELO is based on the team's results, the pitcher ELO will also be determined by the team's results. This means that we will examine the game outcomes for any games in which a pitcher is the starter. Clearly this rating will be noisy. A pitcher with less stamina has less control over the outcome of a game. Also, a good pitcher on a poor hitting team cedes more control to the bullpen, since the score is likely close. However, the hypothesis going into the analysis is that using game outcome per starting pitcher appearance will provide a useful signal amidst the noise.

On we go.

Research

The data used in this analysis comes from retrosheet. I have all baseball event data since 2008 in a database thanks to the Chadwick tools for parsing raw retrosheet game files. I could have used data prior to 2008 as well but decided to keep the sample size smaller to start. If the analysis indicates areas of interest for followup then the database will likely be expanded.

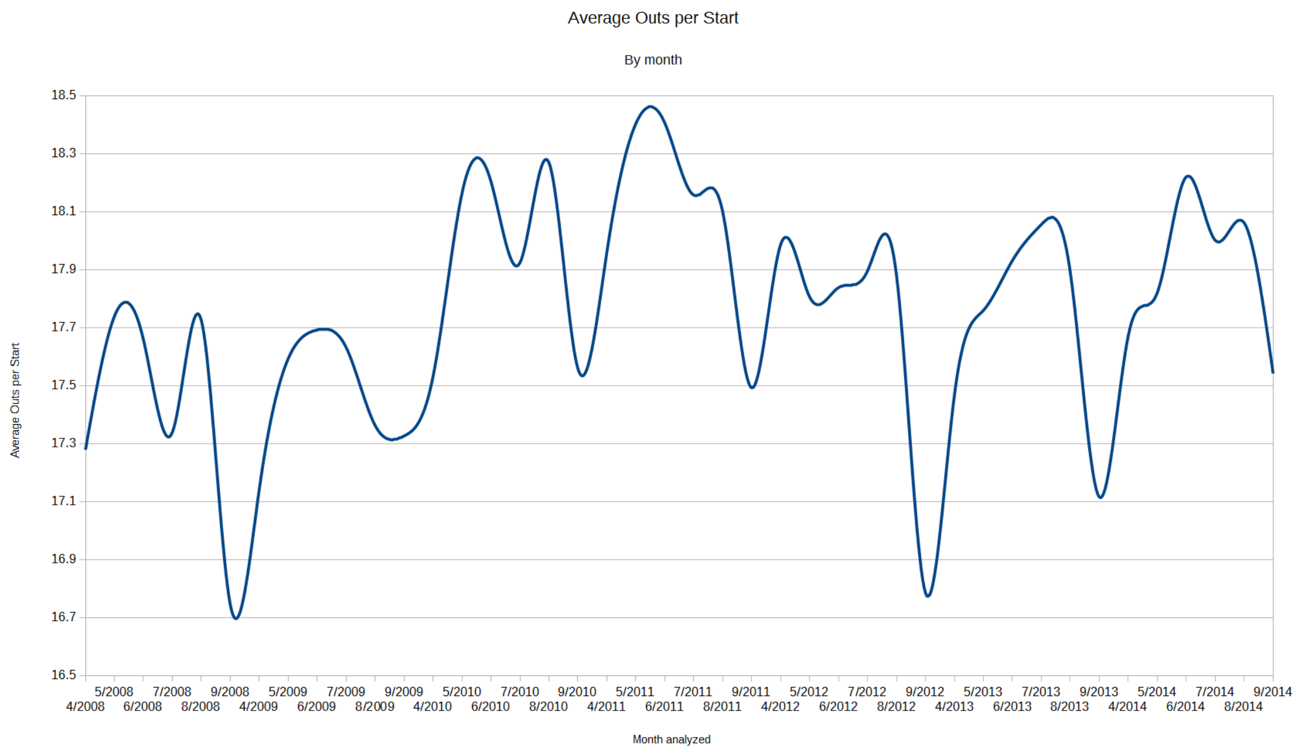
This section is divided up by metrics. First, the number of outs per starting pitching performance will be tabulated for later use. Second, the overall game outcome for each starting appearance will be compared to the expected outcome of the game based on the team's respective ELO ratings at the beginning of the game. Third, variables that can influence the predictive power of a pitcher's ELO rating are examined.

Out per starting pitcher appearance

The more outs a starting pitcher records, the more influence that pitcher has on the eventual game outcome. It is expected that the average number of outs recorded by a starting pitcher will be a useful variable for adjusting the pitcher's ELO.

Average Outs Per Start	17.735

How does the average number of outs change over time?



Not much of a pattern, although it seems pretty clear that the average number of outs per start declines in September (larger rosters?). To confirm that impression, here is the data per month over all years.

April	17.58
May	17.90
June	18.00
July	17.86
August	17.90
September	17.22

Among pitchers with at least 100 starts in those years, here are the top five best per start.

Pitcher	Outs per start
Cliff Lee	21.38
Roy Halladay	21.24
CC Sabathia	20.87
Felix Hernandez	20.75
Adam Wainwright	20.51

ELO change in starting appearance

For all games from 2008 on, the difference between the expected outcome and the actual outcome was calculated. This ELO rating difference was then mapped to a winning percentage. Summing all of the differences in winning percentage over all games for each starting pitcher gives a total influence of the pitcher compared to expected outcomes. This accounts for strength of the pitcher's team, strength of the opposing team, and the home field advantage. It does not account for influence from the bullpen or offense, except to the degree those contribute to the team's overall ELO rating. Top 5 pitchers since 2008 with more than 100 appearances. NOTE: This does not mean the pitchers below are the “best”. Being a good pitcher on a poor team will almost certainly show as better than a very good pitcher on a great team.

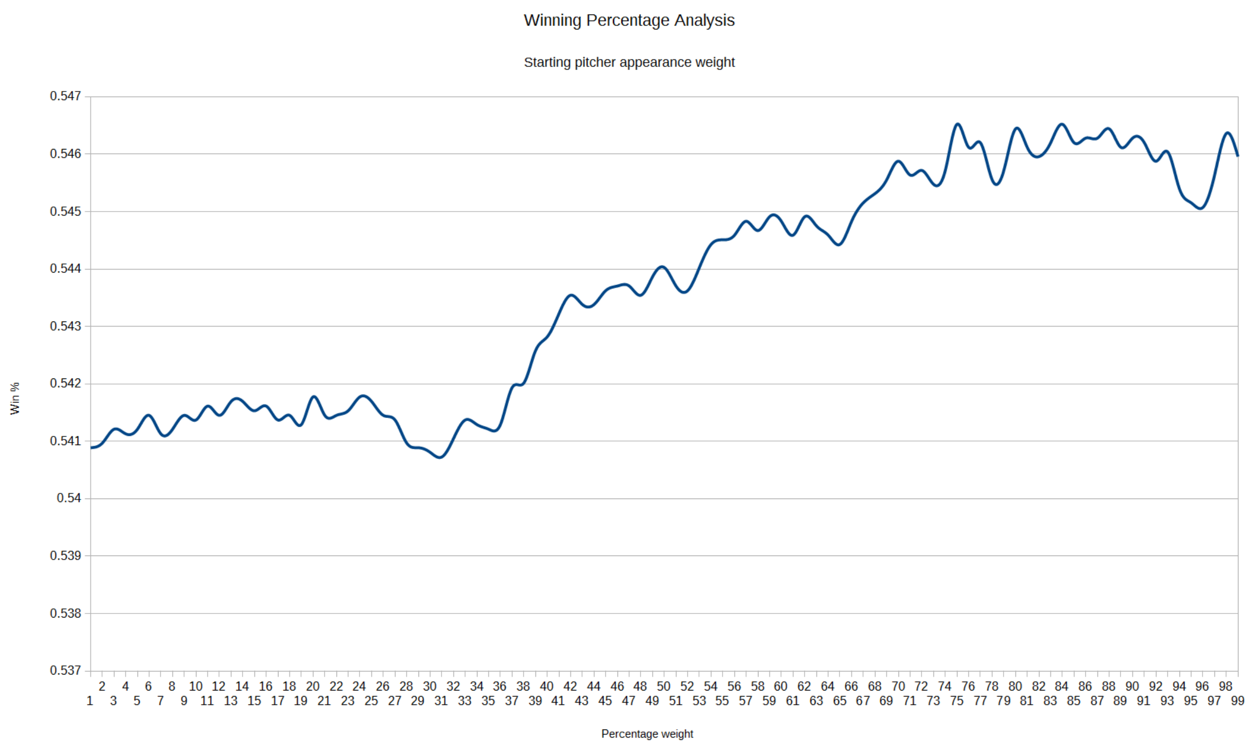
	Avg Win Percentage Improvement
Adam Wainwright	0.106
Jorge de la Rosa	0.104
Clayton Kershaw	0.103
Chris Tillman	0.100
Zack Greinke	0.097

The bottom five pitchers with more than 100 appearances are:

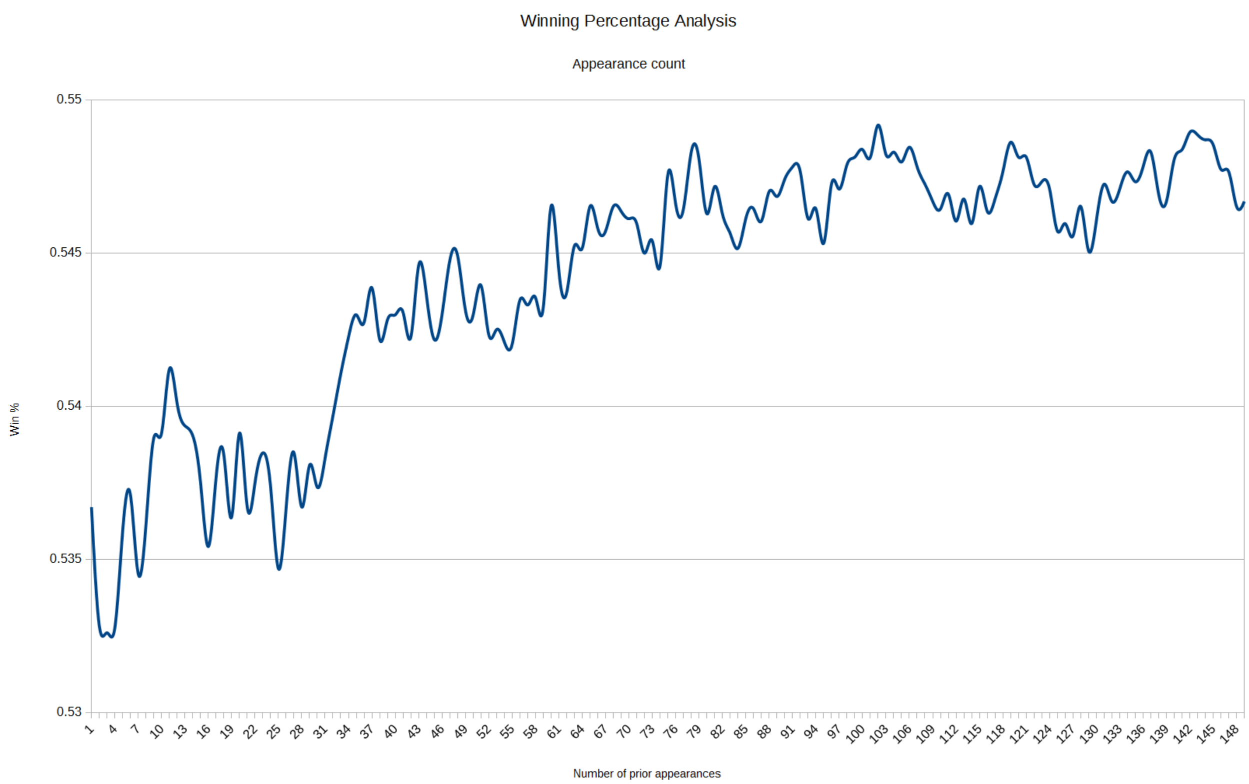
	Avg Win Percentage Improvement
Jeff Francis	-0.079
Roberto Hernandez	-0.071
Charlie Morton	-0.070
Jhoulys Chacon	-0.065
Brandon McCarthy	-0.062

Variables that can influence predictive power

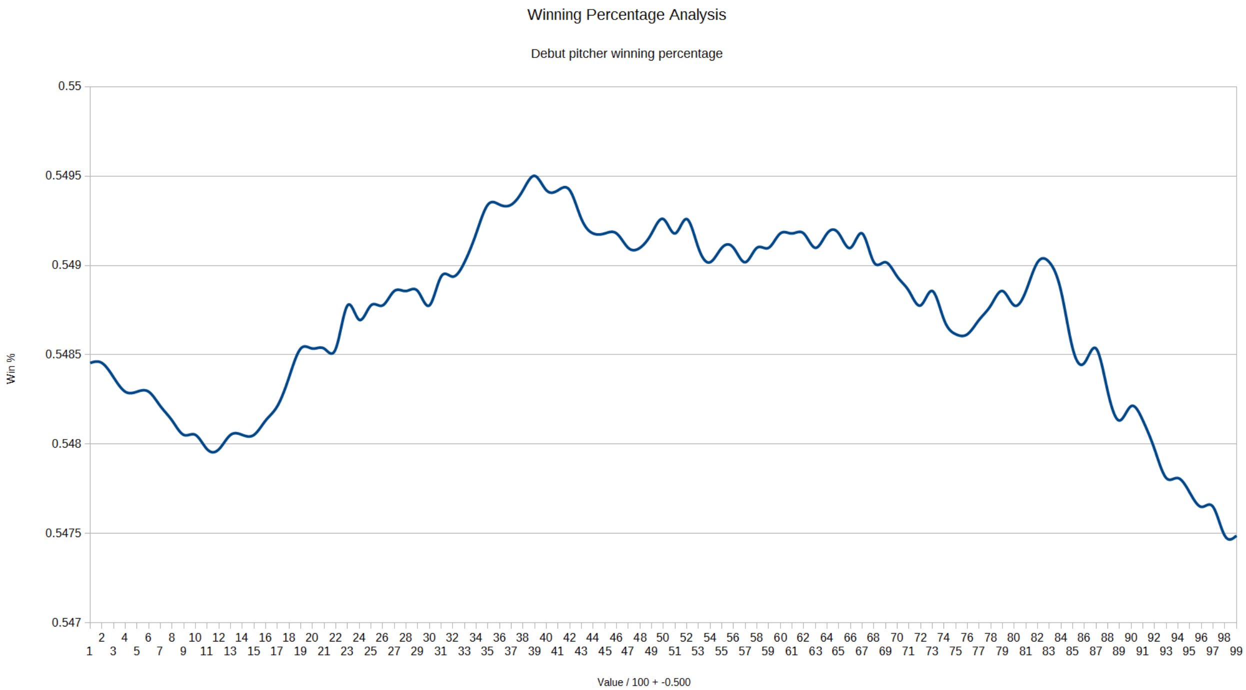
1. Appearance Weight – A number between zero and one (a percentage) that multiplies the effect of the number of prior starting pitcher appearances for a pitcher. The appearances are measured as prior to the current game being analyzed. The idea is that the more appearances a pitcher has, the more likely their current pitcher ELO rating is a robust value that has predictive merit. As you can see in the chart below, the best weight is about 84%. A weight of zero would dismiss the number of appearances as a consideration altogether.



2. Appearance Count – The number of prior appearances as a starting pitcher that are included in the ELO calculation pool. There is likely a balance between the desire to have a large number of appearances that can determine an accurate ability metric, and the wish to avoid a too long timeframe that will miss the affects of aging or injury. Any pitcher that has less than the maximum desired appearances will have their pitcher ELO contribution scaled back in a linear fashion. For example, a pitcher with only 25% of the desired appearance count will only have 25% of their performance contributed to the team total ELO for the game matchup, good or bad. In this analysis, an appearance count of 122 was best. The x-axis on the graph below is value + 20 because of a mistake. Although 122 was best, notice that the graph falls off only very slowly, if at all. It implies that further inclusion is not better or worse, so I chose the earliest, highest value.



3. Debut pitcher winning percentage expectation – What do we do with a pitcher who has no prior experience? This value is the default winning percentage expectation for a debut pitcher. It is expected that it will be negative, on the assumption that even great pitchers often start out a bit shaky. The graph below shows that the best option is (-0.110). A substantial performance hit!



4. Inning Weight – Within the pitcher's prior starts they threw a certain number of average outs per start. Since we know the league average we can further weight the pitcher's participation by how long they throw on average. Intuitively, the longer a pitcher pitches, on average, the more robust the predictive ability for that pitcher.

Conclusion

Using a variety of variables the accuracy of a pitcher based ELO was improved. Further improvement could likely be gained by adding additional data and running it through a genetic algorithm but that is overkill for now.

The baseline prediction ability for team ELO by itself is 55.9%. This is a reasonable success rate but certainly nothing special. Gambling with this value by itself against Vegas odds will quickly cause you to lose all your money! For perspective, if you always bet for the home team the success rate (not against the odds) is 53.7%.

NOTE: The winning percentages on the graphs are lower than the final values because of changes made after the graphs were produced. However, the shapes of the graphs are accurate.

With the best combination of pitcher values included, the success rate rises to 56.5%. This seems small (it is small!) but it also narrows the difference between the Vegas odds and these predictions. Certainly they are no longer a guaranteed loss. Considering the overall simplicity of this approach, compared to maintaining rosters and having a full baseball simulation that needs to be run multiple thousands of times, this seems like good bang for the buck. All that is required to make a reasonable prediction is the game matchup and team ELOs at the start of play, and the starting pitchers.