Intro | Features | Offense | Defense | Pitching | WE/RE/LI | Principles | WAR | Business

Park Factors – 5 Year Regressed

by David Appelman

May 16, 2016

*This article was originally published on Buckeyes and Sabermetrics by Brandon Heipp. The method shown here is how FanGraphs calculates its park factors. For a general overview of park factors without the math, click **here**.*

Getting involved with park factors can be dangerous. There is a lot of confusion/disagreement out there about the purpose of park factors, how they should be calculated, how they should be applied, what they really mean, etc. Hopefully this short word on them will clarify my views and give some sense of the philosophy behind the park factors presented here.

Why Do We Need Park Factors?

Park factors are needed because the thirty major league ballparks all affect the game of baseball in thirty different ways. Sometimes these effects are drastic(Coors Field or Dodger Stadium) and sometimes they're almost negligible. But if you want to fairly value ballplayers, you need to take context into account. PFs are one way to do that.

An argument against PFs, usually made by sabermetric opponents, is something along the lines of “Life is not fair; some players have the fortune to play in a park conducive to their talents, others don’t. Deal with it.” I believe that there is a lot of truth in this argument; for instance, I don’t seek to right all the perceived injustices of the world in doing sabermetrics(like, “what if Edgar Martinez played before the DH” or, on the flip side of the same debate, “what if Edgar Martinez wasn’t wasted by Seattle for three or four years when he was a good player”). But the argument ignores the cardinal rule of sabermetrics: that it is wins and losses(or as building blocks, runs and outs) that matter. A Dante Bichette who has the good fortune to play at Coors Field will put up gaudy numbers–but even in that context, they do not translate to wins. Park Factors are a step along the way of determining the value(or the theoretical value under other circumstances) or player performance. Martinez may have been helped by the DH rule–but it was a reality and using it, he produced real wins for the Mariners. Bichette was helped by Coors Field–but in doing so, it didn’t help the Rockies win baseball games.

How Should Park Factors be Constructed and Calculated?

To answer this question, you first must ask yourself the all-important sabermetric question, “What are you trying to measure”(See the “Ability v. Value” article on this site for more details on the terms I will be throwing around). If you want to measure ability, then certain assumptions and methods may be appropriate–but if you are seeking a value measurement, then those same assumptions can be woefully inappropriate. I will discuss a number of various ways to do PFs and when they should be used:

Runs or Componenets?

Most published PFs are run park factors(e.g. they measure the park’s effect on run scoring). This could be for a variety of reasons: it is easier to use one PF then six or seven, the data is more available(especially pre-Retrosheet), or that is the appropriate choice for the question at hand. However, other people publish event specific park factor, such as the Wrigley Field park factor for doubles.

Component park factors are, at least in my opinion, possibly appropriate for performance measures and absolutely appropriate for ability measures. A park may increase run scoring by 10%, but it does not effect all offensive events by the same factor. So, if you want to know what a player would do in a truly neutral context, you need to adjust his singles, doubles, walks, etc. separately.

But if you are measuring value, that is not at all what you want to do. A player's actual value is almost solely a function of his runs and outs. The park factor allows us to convert the runs and outs to wins, but that is all they should do. In a value method, the park factor's only purpose is to state the true value of the player's runs. I am struggling a bit with how to put this in writing clearly, but it is clear in my mind.

A related question is whether or not to use separate factors for left-handed and right-handed batters. The answer: sure, you can do it if you want ability, but if you want value, absolutely not. The reasoning here is exactly the same as the reasoning for runs v. components.

Adjustment for Team

The title is vague, but that's because I'm tying two ideas under it. The first is Bill James work in the Historical Baseball Abstract. Instead of using park factors and adjusting the player's performance(or the other option, adjusting the league so that they played in the same park as the player), he simply eschewed the league altogether and evaluated the player's performance in the context of the runs scored and allowed by his team. This is an implicit park factor–the RPG in Rockies games is higher then the RPG in Padres games. James also indicated that he liked it better, saying that the player ultimately performs in that context. A Dodger playing in San Francisco's real value to his team's winning does not depend on what a Brave is doing in Chicago. This approach is inherently for a value measure, and personally, I feel that it has a lot of merit.

Then there is what has traditionally been done in the Pete Palmer PFs which have appeared in The Hidden Game of Baseball and Total Baseball. In addition to quantifying the park's effect on run scoring, Palmer publishes a separate PF for the hitters and pitchers. In the BPF(Batter's Park Factor) for instance, there is an adjustment made for the fact that the batters did not face the team's pitchers. In order to facilitate a truly equal comparison with the other batters in the league, the quality of pitchers faced should be equal.

This ties into the James approach, because my take on it(at least value-wise) is, "Why does it matter?" Sure, it's easier if you don't have to face your teammates who lead the league in ERA. But the wins and losses you have produced were done solely against your competitor on the field. If the Yankees have good offense and pitching, they will benefit from not having it square off–and they will win real games as a consequence. Value must be grounded in real wins and losses.

But if you are going for ability, the idea of adjusting for quality of opponents is a good one. However, I think it is best left as a separate adjustment, because lumping it in with the park factor obscures which part of the adjustment is for the park and which part is for the not facing teammates factor. But it's Pete's book and if that's how we wants to do it, more power to him. This is a style issue, not a methodology one.

In an Ideal World

Here I will start to discuss the choices I have made in calculating the PFs published here. I wanted to save this until the end, but I need to clarify my terms before writing the next section. Usually, we take a PF and apply it to a player's complete batting line. But what if the player has played more on the road then at home? Isn't he unfairly docked by applying a PF that assumes equal play at home and on the road to his road-heavy stats? Yes. And it's not just "unfair", it also goes against value logic. If the performance actually occurred on the road, it should be evaluated in that context.

The Big Bad Baseball Annual followed this approach in their later editions. They calculated home and road park factors for each team(the road factors for each team are not 1.00 as they are for the league because the team does not play on the road in its home park, while all the other teams do. They should be around 1, though) and adjusted the players home and road stats separately, then adding them back up. This is a great approach, if you have the data, time, and sanity. Heck, if you have the player's stats by each specific park rather then just home/road, go for it. But you can see that this would get very complicated very quickly.

And when you actually calculate the PF, most people will just weigh the RPG at home against the RPG on the road. Well, the more advanced approaches will, instead of comparing home context to road context, compare home context to league context. If there are T teams in a league, the aggregate league context is based 1/T% on each park. To sum for a team, (T-1)*Road+Home((T-1)/T% is road, 1/T% home). In a league with 14 teams, 13/14 of the league context is based on a given team's "road" park; 1/14 on its home park.

But then, don't teams play unbalanced schedules? If the Orioles play 10% of their road games at Yankee Stadium and just 4% at Safeco Field, shouldn't their road context be based 10% on Yankee Stadium and 4% on Safeco rather then 1/13=8% on both? Sure. But it's just more complexity with a limited effect on the ultimate answers the sabermetric methods will yield.

Another problem with these kind of adjustments(although more so left/right adjustments) is the sample size. And so the question also arises, how many years data should you use? Some people use just one year, others three, others five. There are arguments, and good ones, to be made for all viewpoints. Pros for one year include taking into account conditions that change from year to year such as the weather. Also, while the "road" context used in figuring the PF is assumed to be constant, it actually varies based on the other parks in the league. When Coors Field opened, teams scored more runs on the road then the previous year because they were now playing 8% of their road games at a launching pad. A one year PF keeps the other parks constant.

On the other hand, multiple year PFs give you more data and an increased sample size. Personally, I side with multiple year. With regression thrown in.

Any time we observe something, it is just a sample and not the true frequency or proportion the event should occur at. By taking our observed values and regressing them towards the mean, we can increase the reliability and lessen the chance that we "overadjust". On the other hand, sometimes we might lose a real effect because it did not occur due to chance. But on the whole, we will be better off. The degree of regression that is appropriate is a totally different subject and one that deserves more study and real statisticians(unlike myself) working on it.

And then there is the problem of how to express the PF when we are done. Let's say that by whatever approach, we determine a home context RPG of 11 and a road context RPG of 10. A quick PF would be 11/10=1.10. The park increased offense by 10%. But that is only for performances that occurred in the park. The road numbers should not be adjusted at all. So we can't just adjust a player's RC down by 10%

(unless it is only his home RC). To apply it to the whole stat line, we need to water it down by 50%(since 50% of the games are played at home). So to apply it, we need (1.10+1)/2=1.05, or analogously, (1.10-1)/2+1=1.05.

What Should The Denominator Be?

In this section, I will focus solely on run park factors, since that is what I use and I have not really given deep thought to proper denominators for other events. Some people do park factors for singles and other hits in play with a denominator of balls in play. Others do them with plate appearances. Voros McCracken uses all sorts of different denominators. I'm not going there, right now.

So, for a run PF, the denominator should be…outs. Or innings or games, which are essentially equivalent to outs. Why outs? Well, let's take a player who creates 100 runs and makes 350 outs in 600 plate appearances, but plays in a park with a 1.05 PF(increases offense by 5%). Let's say that the 1.05 is based on using R/PA as the rate stat for the teams in calculating the PF. What a 1.05 PF means, then, is that his R/PA has been inflated by 5%. So his 100/600=.167 R/PA should be adjusted to .167/1.05=.159. But then we are going to use R/O in the valuation method(at least in my stats here). So now we have adjusted the runs, but the out rate is also affected by the park(the OBA at Coors is higher then the OBA at Dodger Stadium). So now we need an outs park factor too. We need to change the number of outs he's made since that effect was not considered in the park factor.

But wait a second, aren't outs constant? If you play at Coors Field, you still have 27 outs in a game and you use them all. Outs are constant across parks. At first, this seems like it is really going to complicate things, but luckily, the math works out well. Say the 1.05 was actually a PF based on team R/G(which is essentially equivalent to R/O). The 1.05 PF based on R/G means the player's R/O needs to be divided by 1.05. So it goes from 100/350=.286 to .286/1.05=.272. Now, how many runs has he created? .272 runs/out*350 outs =95.24 runs. And he still has 350 outs. But 95.24 is also equal to 100/1.05. In other words, you can apply the R/O PF directly to a player's RC total and still get the correct rate stat answer.

How I Calculate PF

I lifted the method here essentially from Craig Wright in The Diamond Appraised, with regression factors published by MGL thrown in. It is fairly straightforward and most of the concepts involved have been touched on above.

First, I use up to five years of data(so this year, it is 2000-2004). If the park has not been used for all five years, or has had major renovations, I use the appropriate smaller time length. One departure is the Montreal/San Juan Expos. Although the Expos have played at Olympic Stadium for many years, playing twenty games or whatever in Puerto Rico has really thrown a monkey wrench into things. One approach(and perhaps the best) would be to use five years data for Olympic Stadium and the two years data for San Juan, and weight them by the percentage of games played at each venue. However, I have chosen to just look at the two seasons separately and with the stadiums combined(I'm pretty sure they played the same number of games in San Juan in 03 and 04 so we don't have to worry about the percentages for each season being different. Next year we won't have to worry about this, but we will have to worry about a one year park factor in Washington).

I have applied regression as well, but we'll touch on that last. I have taken RPG at home and RPG on the road. Call these H and R respectively. Then, the raw PF is H*T/((T-1)*R+H). (NOTE 11/2005: This formula previously lacked the "*T" in the numerator before a reader pointed out the error) All this does is make the "road context" 13/14 of the opposing stadium and 1/14 of your stadium, as it is for the league(assuming 14 teams here; T is # of teams in the league). Then, take the raw PF, add 1 to it, and divide by 2. This just makes it applicable to composite stats rather then just home stats. Call this intermediate PF iPF.

Then we apply regression. I have used weights suggested by MGL in a post on baseballboards.com in 2000. These weights seem arbitrary, but maybe he had a good study to base them on. Anyway, I think their reasonable and that's why I use them.
The Final PF is:
1-(1-iPF)*X
Where X=.6 for 1 year, .7 for 2 years, .8 for 3 years, and .9 for 4+ years.

To give you an idea of the effect of this, say there's a park with an observed iPF of 1.25(this is a Coors-type effect). If that was observed over 1 year, we'd use a PF of 1.15. For 2 years, 1.175, for 3 years, 1.20, and for 4 years, 1.225.

UPDATE: 10/05…For the 2005 park factors, I am adding a home run park factor. It is calculated in exactly the same way as the run park factor. This means that it is based on HR/G, which as I explained above, is roughly proportional to HR/Out. While R/Out is a proper measure, HR/Out is not. HR/PA would be a much better measure of HR ability. But when you consider how this specific park factor is intended to be applied, it is not problematic. First, it is based on the team data, so the kind of problems you would have with HR/Out applied to Barry Bonds do not exist. But more importantly, they are more valuable if you want to figure out how many homers a player would hit in a different context. If a park reduces outs and increases homers(hello, Coors), the HR PF may overstae the increase in home runs caused by the park. But a player who moves to Coors will still hit more homers both due to the true HR factor AND the fact that less outs for the team equals more PAs for him. So I think that the PF based on HR/G is actually more useful for a fantasy player. Also, the differences between the HR/G and HR/PA PFs would be very small–most parks are close to 1.00 of course anyway.

I have also added a spreadsheet with five year PFs for all teams, 1901-2004. The guiding philosophy was to try to include as much data as possible. If there are five possible years of data to be used for a park, they will all be used, even if four of the seasons were in the past or in the future. The source of the raw data was KJOK's excellent park database.

I treat a park as new if there are major changes to the dimensions, but I did not by any means do a complete historical survey to find out when those changes have taken place, so some that probably should have been treated differently are not. If you have specific data on when a change should have (or shouldn't have) been made, feel free to leave a comment and I will try to incorporate these changes when I update the chart some time in the future.

Additionally, when a team moves, and a new team immediately moves in (for example, the Senators of '60 and '61), this is treated as a new team. Also, in cases in which teams have played a significant (which I defined as around ten or more) number of games in a different stadium in the same year, those years are treated as being a new park (an example is the Dodgers playing games in New Jersey the two years before they moved from Brooklyn). Whenever a "new park" of this sort is established, when the old order is restarted it is treated as another new park.

The reason the park factors are only shown through 2004 is that my ideal data set is two previous years, the year in question, and two future years. For most of the parks active in 2005, we will after 2007 be able to fill this dataset, and so I don't want to publish a park factor now and change it later. However, there are a few parks where the 2004 or 2003 factors are not yet settled because they are new and there are not yet five years of data available. In these cases, I have listed a PF but marked it as one that will change in the future.

Now I will give an example of how I chose the years to be considered in figuring the PF. Suppose we look at the Diamondbacks, who have played in Bank One Ballpark since 1998. In 1998, we have no previous data, but we do have four future years of data we can use, so the sample is 1998-2002. For 1999, we have one previous year, so we take three future years, and get 1998-2002. For 2000, we have two previous years and two future years, so we use 1998-2000. In 2001, we use the two previous years (1999 and 2000), and two future years (2002 and 2003), making the total sample 1999-2003.

Let's also consider the end of the Braves' tenure in Fulton-County Stadium. The last season there was 1996. For 1994, we have two previous years (92 and 93) as well as two future years (95 and 96), so we use 1992-1996. For 1995, we have just one future year, so we use three previous years, and also use 1992-1996, and the same for 1996.

Facebook Twitter Reddit Email

‹ Game Score Player Graphs ›

David Appelman is the creator of **FanGraphs**.

Comments are closed.

X

X