Journal of Quantitative Analysis in Sports

Volume 4, *Issue* 2 2008 *Article* 4

Improving Major League Baseball Park Factor Estimates

Rohit A. Acharya, *Harvard University*
Alexander J. Ahmed, *Harvard University*
Alexander N. D'Amour, *Harvard University*
Haibo Lu, *Harvard University*
Carl N. Morris, *Harvard University*
Bradley D. Oglevee, *Harvard University*
Andrew W. Peterson, *Harvard University*
Robert N. Swift, *Harvard University*

Recommended Citation:
Acharya, Rohit A.; Ahmed, Alexander J.; D'Amour, Alexander N.; Lu, Haibo; Morris, Carl N.; Oglevee, Bradley D.; Peterson, Andrew W.; and Swift, Robert N. (2008) "Improving Major League Baseball Park Factor Estimates," *Journal of Quantitative Analysis in Sports*: Vol. 4: Iss. 2, Article 4.

DOI: 10.2202/1559-0410.1108

Improving Major League Baseball Park Factor Estimates

Rohit A. Acharya, Alexander J. Ahmed, Alexander N. D'Amour, Haibo Lu, Carl N. Morris, Bradley D. Oglevee, Andrew W. Peterson, and Robert N. Swift

Abstract

The study of Park Factors (PF) is essential to the correct evaluation of player performance in Major League Baseball. We have identified two important problems with the commonly used formula which has been popularized by ESPN: it produces variable results due to unbalanced scheduling, and it has an inherent inflationary bias. To address these problems, we develop a new estimator for Park Factors using an ANOVA weighted fixed-effects model for run generation. Using simulated data, in addition to run data from 2000 through 2006, we show that this new estimator does not have the biases of the old estimator. From a strategic viewpoint, accurate PF values are needed to properly evaluate free agents and trade proposals, as well as to compare players for postseason awards. We develop a method to adjust statistics using Park Factors called a Neutral Park Adjustment (NPA), which takes into account the Park Factors of the entire schedule of a player, not simply their home park.

KEYWORDS: park factor, baseball, hitter's park, pitcher's park, neutral park adjustment

Author Notes: All of the authors are members of the Harvard Sports Analysis Collective and all are undergraduates in Harvard College, except for Carl N. Morris, who is Professor of Statistics and HSAC faculty advisor. He can be reached at morris@ stat.harvard.edu. Rohit A. Acharya is the HSAC President and can be reached at racharya@fas.harvard.edu. For information about HSAC and their research please visit http://www.hcs.harvard.edu/~hsac/Blog.

Introduction

Not all baseball parks are created equal. Unlike other professional sports, which stipulate exact dimensions for their respective playing fields, Major League Baseball is unique in that individual ballparks have irregularities such that no two ballparks are identical. For example, some parks favor batters more than others due to discrepancies in the size of field, height of fences, size of foul areas, wind speed and direction, hitters' field of vision, among other characteristics. Each stadium biases a player's statistics in a positive or negative direction depending on whether it is a "hitter's park" or a "pitcher's park." This variation among ballparks presents problems for effectively evaluating and comparing the performance of baseball players.

To accurately assess the contribution a particular ballpark has on offensive or defensive production, a metric known as Park Factors (PF) is employed. Equation 1 illustrates the most frequently used formula for deriving PF.

$$PF = \frac{\left(\frac{RS_{home}+RA_{home}}{Games_{home}}\right)}{\left(\frac{RS_{road}+RA_{road}}{Games_{road}}\right)} \tag{1}$$

This formula uses three familiar statistics to determine the estimator: runs scored (RS), runs allowed (RA), and number of games. The intuition behind this formula is that one can isolate the effect of a specific ballpark on the performance of a team by dividing the total runs scored at its home field by the total runs scored while it is on the road. By keeping the team constant, the ratio produced by the formula therefore indicates that the home park is a hitter's park if the value is greater than one, a pitcher's park if it is less than one, or perfectly neutral when compared to the rest of the league if it is exactly equal to one.

This is a beautifully simple, and fairly suggestive, formula. However, even if teams were to play balanced schedules, as they did until 1960, it suffers from inflationary biases. That is, the Park Factors are overestimated for hitter's parks, and underestimated for pitcher's parks. This would be true even if teams were able to play a nearly infinite number of games against each other, so that nearly exact results could be computed for every park and team combination. The law of large numbers does not hold in the correct way for this estimator, or more technically, the estimate is inconsistent and asymptotically biased.

To illustrate this point, consider a simple two-team league. Team A plays their games in a park with $PF = 2.0$ and Team B plays their games in a park with $PF = 0.5$. The teams are equally balanced in every other way, so that both teams would be expected to score four runs per game in a neutral park ($PF = 1.0$). When they play one game in team A's park, the expected score is $8 - 8$, and when they play one game in team B's park, the expected score is $2 - 2$. If we use the ESPN model to recalculate the park factors, the Park Factor for park $A = 4.0$, and the Park Factor for park $B = 0.25$. The actual Park Factor values are thus biased further from neutral, and the effects of each park appear much more severe than they actually are.

Worse yet, since teams no longer play balanced schedules, the results are not only asymptotically biased, but the mean of this estimator for a particular park's Park Factor will depend on the team's schedule. To give an extreme example, suppose hypothetically that Yankee Stadium were taken down and a park exactly like Kansas City's were put in its place in the Bronx, and simultaneously, the stadium in Kansas City were rebuilt exactly like Yankee Stadium. If we now replay the AL schedule for thousands of seasons, with all the players staying in the same cities, then the estimates of the two Park Factors (now moved) would change. This is due to the different schedules the Yankees and Royals would play, although the physical characteristics of the parks would remain constant. Estimates would be biased, but the biases would change.

Our goal in this paper is to improve estimates of Park Factors by reducing the inflationary biases (most importantly the large sample bias), eliminate the bias caused by schedule differences, and reduce the variability of estimates. To address these problems, we develop a new estimator for Park Factors using an ANOVA weighted fixed-effects model for run generation. Using simulated data, in addition to run data from 2000 through 2006, we show that this new estimator does not include the biases of the old estimator.

Improved estimates for Park Factors benefit all members of the baseball community, from fans to general managers and award voters, by creating more accurate formulae for adjusting hitters' and pitchers' performances to neutral fields. Additionally, these factors will allow for more accurate predictions of how a player's performance will change if he moves to a new park due to a trade or free agency.

Literature Review

Despite revolutionary work by sabermetricians in the field of baseball analysis, the mainstream sports media has resisted embracing non-traditional statistics such as Value Over Replacement Player (VORP) and Equivalent Average (EqA). That baseball is played on the field by players, not by computers, is a common sentiment among baseball journalists. Even so, several statistical creations such as Walks plus Hits per Inning Pitched (WHIP), On-base Percentage (OBP), and Park Factors have become generally accepted as useful and credible statistics. "Here is a fundamental fact of baseball: the playing environment affects the game on the field," wrote Dayn Perry [6], a prominent baseball writer for FOX Sports. While the importance of Park Factors may no longer be disputable, the magnitude and direction of these effects is still a subject of some confusion.

ESPN deserves much of the credit for popularizing the use of the Park Factor estimate given in Equation 1 through its website. The widespread acceptance of this model, and also the several references in analytical and topical papers to ESPN's calculated Park Factor estimates, initially provided the impetus for us to analyze the effectiveness of this specific model. As a result, we will now refer to the model in Equation 1 as the ESPN model in the remainder of this paper for the sake of brevity and clarity; however, it should be noted that ESPN did not devise the equation itself.

Deciding to use ESPN's estimates as our basis of comparison, we noticed a discrepancy in the actual Park Factors, produced by running data through ESPN's model, and the published versions on its website. In *each* year from 2001 to 2003, ESPN's calculated Park Factors had eight mistakes, this out of only 30 teams. In fact, five of these erroneous Park Factors were off by over 45%, and ten were off by more than 35%. Although they seem to have calculated the values more accurately in 2004 and 2006 where we found no errors, in 2005 they seem to have stopped short of the end of the season as 27 of their 30 figures are incorrect (though by margins less than 1%).

Although it is difficult to fully escape mathematical error, the frequency and the obviousness of the errors leads one to question how much thought ESPN has put into their methodology as well. Their figures have been so widely referenced by sportswriters and statisticians alike that it is astounding such errors have persisted. To give an example for how severe these errors are, especially with the intrinsic biases of the model, ESPN still shows that Oakland's park factor was 1.357 in 2001, 0.703 in 2002, and 0.515 in 2003, despite there having been no alterations to McAfee Coliseum in this century. Such numbers are totally unrealistic, and we have corrected their errors for the comparisons in this paper as the focus here is to compare estimation formulae and not calculator input errors.

However, moving on from issues with data entry and looking more at methodology, the literature on Park Factors is rich in explanations of the possible causes of variability between parks, but it lacks many sophisticated models for quantifying these effects. Some studies [3] have shown, for instance, the effect of temperature on velocity. Temperature and altitude effects have a significant effect on runs scored, Colorado's Coors Field being the most prominent example, yet these effects are difficult to quantify. The important question, when analyzing park effects, is not why they affect the game, but by how much. By adding variables for park dimensions to the analysis, Click [2] was unable to improve on his estimates for next year's Park Factors, indicating not only the biases in the model, but also the significant effects of weather, temperature, wall heights, and other factors on runs scored.

Several sabermetricians have acknowledged the flaws in the standard model for calculating Park Factors and have tried to correct for these biases. Sheehan [9] for instance, describes the scheduling bias and its effect on Park Factors. Sheehan asserts that his colleague, Clay Davenport, uses a method for calculating Park Factors that is "actually weighted by games played in each individual park." Sheehan, unfortunately, does not disclose Davenport's methodology.

Many sabermetricians have noted the variability of the results, but none have proposed an adequate solution for solving the problem. Several authors have correctly noted that by adding more years to the calculation and increasing the sample size that statistical noise is reduced, resulting in a more unbiased estimate of the true Park Factor. Click [2] correctly explains this, but also notes that even by adding several years to the calculations, much of the between-year variability remains. He asserts that Park Factors are not as constant as they should be, but does not propose any methodological improvements to correct for the inherent biases.

There have been attempts to improve the ESPN model in the sabermetric community. Thorn et al [10] make a number of useful modifications to the popular Park Factor methodology. In order to determine the Park Factor for year n, they average the calculated factors for years $n-1$, n, and $n+1$, dependent on whether the values for the respective year is available. For example, the final Park Factor for a ballpark for the most recent year is given in Equation 2.

$$PF_{final}(n) = \frac{(PF\,(n-1) + PF\,(n))}{2}. \tag{2}$$

The second modification regards an innings pitched corrector (IPC), which is given in Equation 3.

$$IPC = \frac{18.5 - \left(\frac{Wins_{home}}{Games_{home}}\right)}{18.5 - \left(\frac{Wins_{road}}{Games_{road}}\right)} \tag{3}$$

IPC is an extra factor that accounts for the fact that when the home team is winning

DOI: 10.2202/1559-0410.1108

going into the bottom of the ninth inning, the game will end a half-inning earlier. However, this factor does not achieve the correction that it is intended for since in the ideal situation of a team winning all of their games, this factor would be $\frac{17.5}{18.5}$ (with 18.5 being the average number of half-innings per game if the home team always bats in the ninth). This therefore assumes that there are no extra innings in the home wins, and that there is an average of one half of an extra inning in away wins. Although the extra half inning in away wins is plausible, the 17.5 innings in home wins does not allow the home team to ever go into extra innings or stage a ninth inning comeback. This IPC factor in Equation 3 may be well-intentioned; however, it is not formulated correctly.

Thorn et al [10] also attempt to institute an “Other Parks Corrector” (OPC). This is intended to “make corrections for the fact that the other road parks’ total difference from the league average is offset by the park rating of the club that is being rated.” In layman’s terms, this means that the parks on the road may be skewed toward being either hitter’s or pitcher’s parks. If all the ballparks are hitter’s parks, it would make the park that is under evaluation appear to be more of a pitcher’s park than it really is. Their formula for OPC is given in Equation 4.

$$OPC = \frac{No.Teams}{(No.Teams - 1) + RF} \tag{4}$$

Run Factor (RF) indicates the teams traditional Park Factor, as derived in Equation 1, divided by the IPC. Therefore, the OPC does not take the other Park Factors directly into consideration, which is certainly a weak point when this offsetting is supposed to be concerned with just that.

Methodology

To improve upon the ESPN estimator, we have imposed two requirements: it must be Fisher consistent (i.e. it must give the correct answer when applied to the expected runs scored), and it must be robust to schedule changes by accounting properly for how many games each team plays in each park. To meet these constraints, we chose to use an ANOVA weighted fixed-effects model to estimate the Park Factors. Assuming our model for run generation is realistic, the fixed effects estimator for the treatment effect (in this case, the Park Factor) will be consistent, and nearly unbiased, and differences in weighting (in this case, to handle scheduling imbalances) reduce the estimators variance while keeping the asymptotic mean consistent.

Because there are relatively few interleague games, and because the designated hitter rule differs in the National League (NL) and the American League (AL), but will be confounded with Park Factors, we estimated the AL and NL separately.

We assume that the runs a team scores in a game are generated by a linear process. Controlling for home field advantage and the offensive and defensive strength of the home and away teams, the Park Factor adds or subtracts runs from the number that would have been scored in a neutral park. Using $R^h_{h,v}$ to represent the average number of runs scored by team h when h was the home team against team v, and $R^v_{h,v}$ to represent the average number of runs scored by v when playing h at home, we present the model in Equations 5 and 6.

$$R^h_{h,v} = \mu + O_h - D_v + PF_h \tag{5}$$

$$R^v_{h,v} = \mu + O_v - D_h + PF_h + \tau \tag{6}$$

For example, if we were encoding a matrix for fitting one year of AL games, the model includes a constant term μ, a constant fitted for being the visiting team τ, thirteen dummy indicators to estimate each teams offensive strength, another thirteen dummy indicators of defensive/pitching strength, and thirteen Park Factor constants for the other thirteen teams in the American League. The fourteenth team is omitted to avoid multicolinearity, and its value is recovered by scaling the coefficients so that their arithmetic mean is zero. Thus, in the AL case, the dependent variable vector $\vec{R}$ contains $364 = 14*13*2$ rows corresponding to each possible combination of teams in each park, and the regression matrix is a 364 by 41 matrix, with a column for each of the 41 estimated coefficients. Likewise the regression matrix in the NL case, with $480 = 16*15*2$ rows, would produce a matrix of dimensions 480 by 47.

Using the average number of runs per game between two teams introduces heteroskedasticity due to different sample sizes, ranging from three to ten games played against the same opponent in its park. We have accounted for that by weighting each observation in inverse proportion to its variance (i.e. in proportion to the number of games played). Since averages have $\frac{1}{n}$ times the variance of the underlying random variable, we weight each entry by the number of games that were played between those two teams in the given park. As mentioned above, this weighting allows us to correct for the uneven statistical uncertainty introduced by unbalanced schedules without biasing our estimator in the limit.

This weighted least squares regression produces estimates for the linear Park Factor coefficients. These coefficients represent how many runs per game played at a particular park can be attributed to the park itself. To make these estimates

DOI: 10.2202/1559-0410.1108

comparable to the ones derived from the ESPN model, we have transformed them into multiplicative constants. To do so, we divide each factor by the mean number of runs scored by a team per game and add it to one. Upon rescaling these values so their geometric mean is one, we obtain our new estimators for the Park Factor.

For simplicity, we have focused on runs scored in this section. However, this methodology is as easily applied to other statistics such as hits, doubles, triples, and home runs, or any other cumulative statistic, simply by switching the dependent variable in the regression.

Results

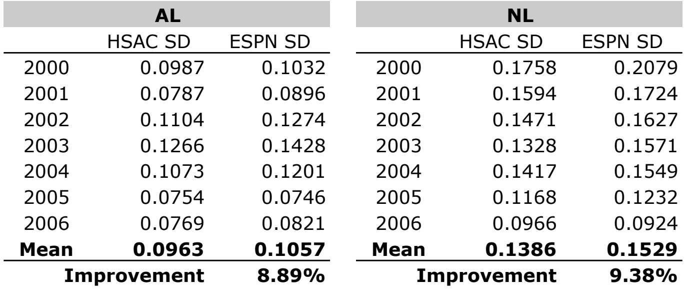
We tested our estimator (henceforth the HSAC estimator) against the ESPN model estimator on both real and simulated run data. In order for the test to run successfully, the standard estimators have been made comparable to our results by restricting the dataset to a single league, and by scaling them to have a geometric mean of one.

Real Data

For a first-order assessment on the effectiveness of the HSAC estimator, we tested it on seven years of season run data from 2000 through 2006. This was only a first-order assessment because the small sample size (at most seven years) kept us from drawing any robust conclusions about the sampling distribution for either one of the estimators. The work done here was mainly performed to generate our own estimates for the 2000 through 2006 seasons, and to lay the groundwork for the simulations to follow. Our data were gathered from game logs at www.retrosheet.org[1]. We ran a separate regression for each year, and assessed the estimators on two criteria: first, on within year standard deviation to check the standard estimator's inflationary bias, and second, on between year standard deviation to see which estimator produced more consistent results. The seven-year averages for each of the estimators and the relevant standard deviations are shown in Tables 1, 2, and 3.

First we considered the within year standard deviation as a heuristic for the amount by which our estimator corrected the inflationary bias of the ESPN estimator. We found that in both the AL and NL, the spread of the Park Factors

[1] The information used here was obtained free of charge from, and is copyrighted by, Retrosheet. Interested parties may visit Retrosheet at http://www.retrosheet.org.



AL			NL		
	HSAC SD	ESPN SD		HSAC SD	ESPN SD
2000	0.0987	0.1032	2000	0.1758	0.2079
2001	0.0787	0.0896	2001	0.1594	0.1724
2002	0.1104	0.1274	2002	0.1471	0.1627
2003	0.1266	0.1428	2003	0.1328	0.1571
2004	0.1073	0.1201	2004	0.1417	0.1549
2005	0.0754	0.0746	2005	0.1168	0.1232
2006	0.0769	0.0821	2006	0.0966	0.0924
Mean	**0.0963**	**0.1057**	**Mean**	**0.1386**	**0.1529**
	Improvement	**8.89%**		**Improvement**	**9.38%**

Table 1: Within Year Standard Deviations for the ESPN and HSAC estimators. The HSAC estimator created a smaller spread within in year, suggesting bias correction by reducing the inflationary biases inherent in the ESPN model.

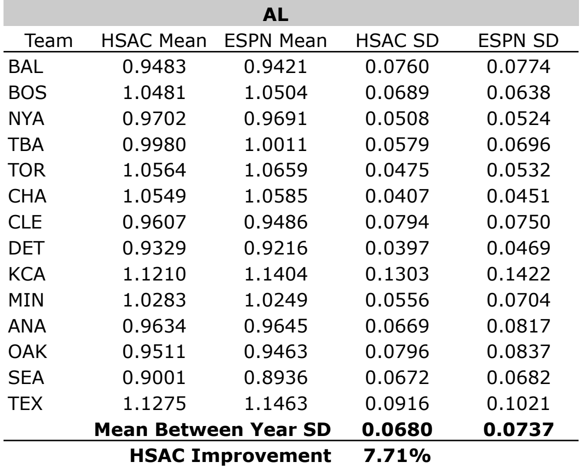
within a given year was on average lower, with 8.89% and 9.38% improvements respectively. This was, of course, only a first-order assessment because the actual size of the ESPN bias is unknown. However, the consistent “tucking in” of the estimators suggested that our estimator is providing bias correction by removing the inflationary tendencies mentioned earlier.

We then tested to see whether our estimator was more consistent from year to year given that the parks in question had not changed. We were able to estimate this for all of the AL parks since none has changed since 2000, and found that the HSAC estimator had 7.71% smaller variance across years. In the NL case, we were limited to the eight parks that had not changed since 2000, but still found that the HSAC estimator was 2.62% more consistent.

Simulated Data

Building off of the first-order assessments using real data, we generated a set of simulated data that we used to rigorously test the quality of the two estimators.

We used the 2006 season as our “seed” season to generate a set of fictitious seasons as follows. We used our estimated coefficients for offensive strength, defensive strength, home field advantage, and intercept from our 2006 regression, and the ESPN Park Factors from 2006 to generate an expected runs vector analogous to the one we used as the dependent variable in our regressions above. We then used the actual 2006 season schedule, and treated each game as a draw from a normal distribution centered at the expected run value, and standard deviation given by $\left(\frac{11}{8}E\left[R\right]\right)^{0.6}$ as estimated by Morris and Christiansen [4] using MLB



		AL		
Team	HSAC Mean	ESPN Mean	HSAC SD	ESPN SD
BAL	0.9483	0.9421	0.0760	0.0774
BOS	1.0481	1.0504	0.0689	0.0638
NYA	0.9702	0.9691	0.0508	0.0524
TBA	0.9980	1.0011	0.0579	0.0696
TOR	1.0564	1.0659	0.0475	0.0532
CHA	1.0549	1.0585	0.0407	0.0451
CLE	0.9607	0.9486	0.0794	0.0750
DET	0.9329	0.9216	0.0397	0.0469
KCA	1.1210	1.1404	0.1303	0.1422
MIN	1.0283	1.0249	0.0556	0.0704
ANA	0.9634	0.9645	0.0669	0.0817
OAK	0.9511	0.9463	0.0796	0.0837
SEA	0.9001	0.8936	0.0672	0.0682
TEX	1.1275	1.1463	0.0916	0.1021
	Mean Between Year SD		**0.0680**	**0.0737**
	HSAC Improvement		**7.71%**	

Table 2: 7 Year Means and Standard Deviations for the ESPN and HSAC estimators. The HSAC estimator had smaller variability between years.

data at the time of their publication.

We generated 5000 seasons in this manner and for each produced an HSAC and ESPN model estimate for each team's Park Factor. This allowed us to construct a probability density for each team, and calculate a mean-squared error for each estimator, allowing us to assess the relative efficiency of the Park Factor estimators.

In these tests, the HSAC estimator dominated the ESPN estimator for each team in the AL and the NL. The HSAC estimator had an average relative efficiency of 3.85 for the AL and 2.16 for the NL, with the relative efficiency as high as 14 for one team (KCA) in the AL. The density plots in Figures 1 and 2 show that the HSAC estimator is nearly unbiased, in contrast to the ESPN estimator and is much more precise. The relative efficiency of the HSAC estimator for each team is printed in the corresponding density plots.

Interpretation

Both real data and simulated data experiments showed that the HSAC estimator dominates the ESPN estimator in terms of mean-squared error. It corrects the biases that the ESPN estimator suffers from and reduces the variability of the estimate, establishing that the HSAC estimator generally provides noticeably better estimates of Park Factors than the ESPN method.

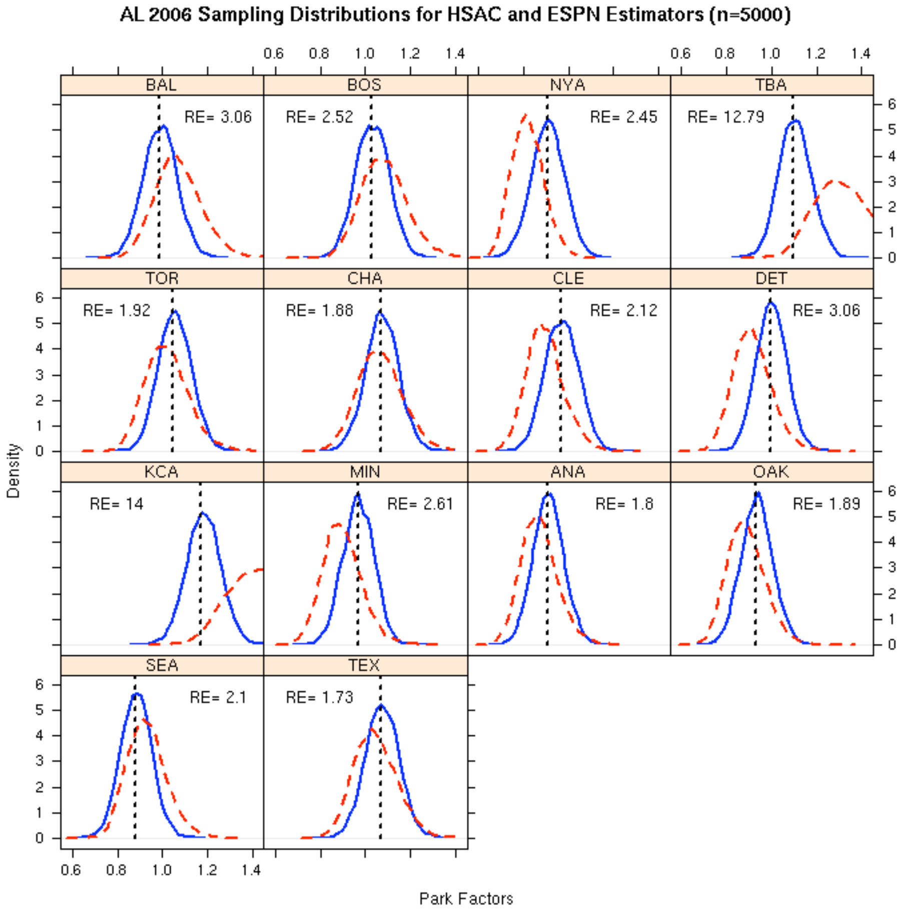


Figure 1: Sampling distributions for the HSAC (solid curve) and ESPN (dotted curve) estimators on simulated data based on the AL 2006 season with relative efficiencies. The HSAC estimator had a mean relative efficiency of 3.85, and ran as high as 14 for KCA.

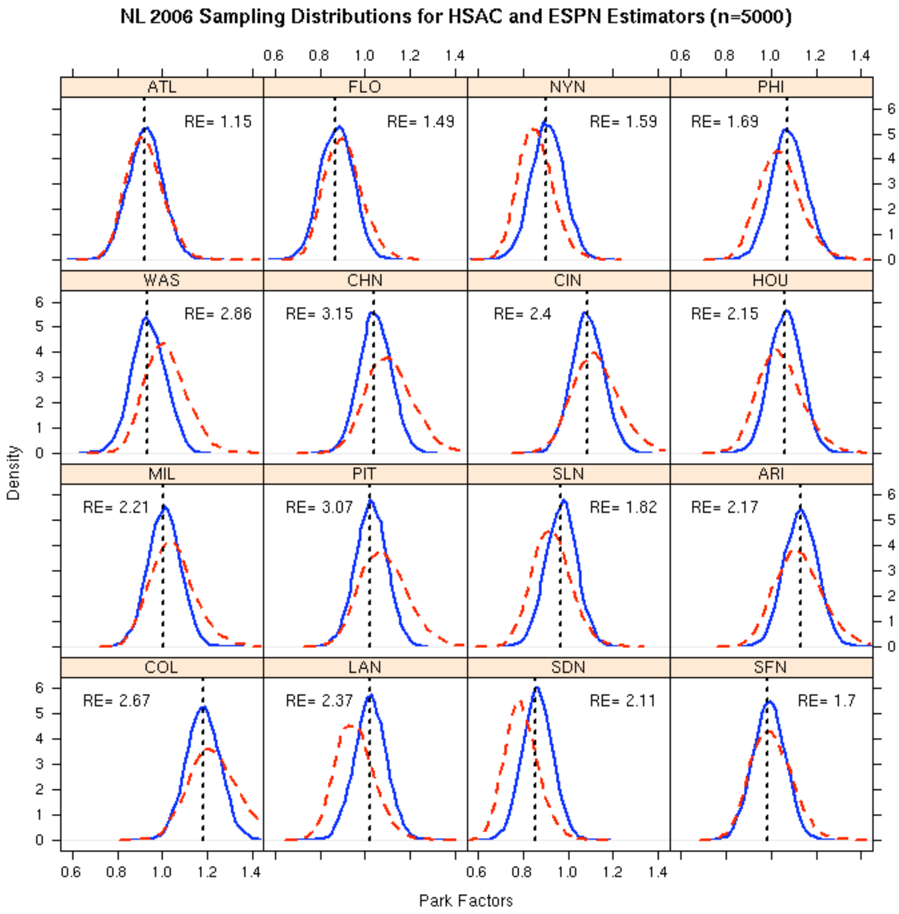
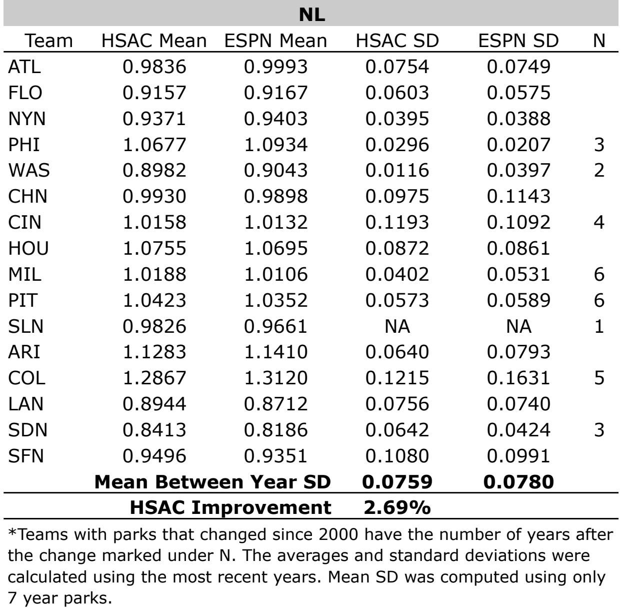


Figure 2: Sampling distributions for the HSAC (solid curve) and ESPN (dotted curve) estimators on simulated data based on the NL 2006 season with relative efficiencies. The HSAC estimator had a mean relative efficiency of 2.16.



NL					
Team	HSAC Mean	ESPN Mean	HSAC SD	ESPN SD	N
ATL	0.9836	0.9993	0.0754	0.0749	
FLO	0.9157	0.9167	0.0603	0.0575	
NYN	0.9371	0.9403	0.0395	0.0388	
PHI	1.0677	1.0934	0.0296	0.0207	3
WAS	0.8982	0.9043	0.0116	0.0397	2
CHN	0.9930	0.9898	0.0975	0.1143	
CIN	1.0158	1.0132	0.1193	0.1092	4
HOU	1.0755	1.0695	0.0872	0.0861	
MIL	1.0188	1.0106	0.0402	0.0531	6
PIT	1.0423	1.0352	0.0573	0.0589	6
SLN	0.9826	0.9661	NA	NA	1
ARI	1.1283	1.1410	0.0640	0.0793	
COL	1.2867	1.3120	0.1215	0.1631	5
LAN	0.8944	0.8712	0.0756	0.0740	
SDN	0.8413	0.8186	0.0642	0.0424	3
SFN	0.9496	0.9351	0.1080	0.0991	
	Mean Between Year SD		**0.0759**	**0.0780**	
	HSAC Improvement		**2.69%**		

*Teams with parks that changed since 2000 have the number of years after the change marked under N. The averages and standard deviations were calculated using the most recent years. Mean SD was computed using only 7 year parks.

Table 3: 7 Year* Means and Standard Deviations for the ESPN and HSAC estimators. For the limited number of teams with unchanged stadiums over 7 years, the HSAC estimator had a slightly smaller variability.

Applications of Park Factors

Once the Park Factors are determined for each team in the American League, they can provide very useful insights into a players statistics. Using our Park Factor adjustments, we can strip away variables that might bias comparisons between players. If a player plays predominantly in hitter's parks, then his statistics should not be considered equally to a player who spends most of his season in pitcher's parks. This becomes particularly important in several areas of baseball analysis.

Park Factors could be calculated for any cumulative statistic, such as home runs and doubles, and can also be done with average statistics, such as batting average, ERA, and OPS, but as the results we have here are for runs scored, the following applications of Park Factors will be based on the sabermetric "Runs Created" (RC). The Runs Created formula, originally developed by Bill James, is given in Equation 7.

$$RC = \frac{(Hits + Walks) * Total\ Bases}{At\ Bats + Walks} \tag{7}$$

Other versions of this formula have been developed; however, this is the simplest runs created calculation, so it is the one we used to apply the Park Factors for runs. To adjust a player's RC for a season, one must take into account all of the ballparks in which they played during that season. Our calculations do not account for games played in stadiums in the other league (interleague road games) because the Park Factors do not necessarily translate between leagues due to rule differences, particularly the existence of a designated hitter in American League lineups. Once we have information on a player's RC for each ballpark, we divide each total by the park's Park Factor, and then sum them to get a seasonal adjusted RC total. This allows us to compare players on different teams in a more equitable manner than ever before.

This is extremely important for MVP, Cy Young, and Rookie of the Year consideration, and it can even be used with career statistics to assess a player's candidacy for the Hall of Fame. We have found that the Park Factors can often affect, either positively or negatively, a players RC total by as much as 10% over a season.

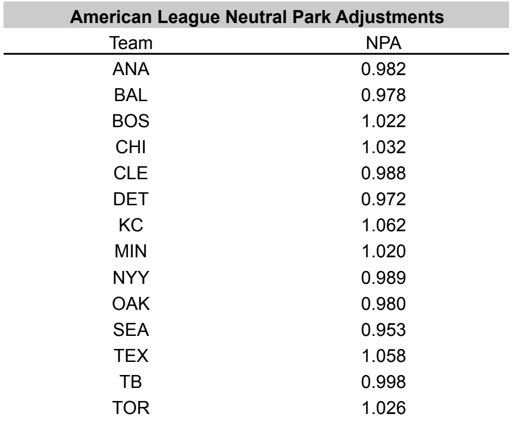
The second application of Park Factors allows general managers and front office personnel to determine how players would perform differently when playing their home games in different parks. It is not as simple as the first type of Park Factor application, though, because when general managers consider possible trades, they must realize that the portion of games played in each park and against each team will change. We have developed a method to adjust statistics for the change in schedule when a free agent is signed by a different team or when players switch teams in a trade.

Each American League team has a Neutral Park Adjustment (NPA), which is a weighted average of all the AL parks in which a team plays in a season.

$$NPA_x = \frac{(PF_1 * Games_1) + (PF_2 * Games_2) + \ldots + (PF_n * Games_n)}{Total\, Games\, Played} \quad (8)$$

This calculation in Equation 8 is much more accurate for analyzing statistics than purely looking at a team's home Park Factor. For example, the Texas Rangers, whose Park Factor is one of the highest in the league at 1.127, happen to have the only hitters park in the AL West. When all the games played in Seattle, Anaheim, and Oakland are considered, as well as the rest of the Rangers' schedule, the Rangers' NPA drops to 1.058.

These statistics can not only standardize RC stats for players in an MVP race, they can be used to predict what a player's stats would have been like had he played that season for another team. If Ichiro Suzuki of the Seattle Mariners,



American League Neutral Park Adjustments

Team	NPA
ANA	0.982
BAL	0.978
BOS	1.022
CHI	1.032
CLE	0.988
DET	0.972
KC	1.062
MIN	1.020
NYY	0.989
OAK	0.980
SEA	0.953
TEX	1.058
TB	0.998
TOR	1.026

Table 4: Neutral Park Adjustments for American league teams using 7 year Park Factors and the 2006 schedule.

who plays in one of the AL's best pitcher's parks, had played the 2006 season for Kansas City, his numbers would have been significantly different.

$$RC_{new\ team} = RC_{old\ team} * \frac{NPA_{new\ team}}{NPA_{old\ team}} \tag{9}$$

Using Equation 9, Ichiro's RC would jump from 102.8 to 114.45. While Seattle and Kansas City have the biggest difference in NPA in the American League, not all manipulated RC statistic changes would appear so drastic.

Similarly, we can look at the case of Justin Verlander, 2006 rookie of the year and starting pitcher for Detroit. If he had played the 2006 season for the Texas Rangers instead of the Tigers, his ERA would jump from 3.69 to 3.79 according to Equation 10.

$$ERA_{new\ team} = ERA_{old\ team} * \frac{NPA_{new\ team}}{NPA_{old\ team}} \tag{10}$$

While it is not completely comprehensive, Equation 9 provides an initial prediction of what a player's statistics would be like for various teams. The above formulae are only a brief example of the uses for our Park Factor analysis, and in no way form a comprehensive list. There are undoubtedly others, and we encourage further research to be done in this area.

Conclusion

In this paper, we established a new methodology to estimate Park Factors that does not suffer from the biases inherent to the fairly ubiquitous model popularized by ESPN. Unfortunately, the lack of longer-term data in Major League Baseball, particularly due to the park relocation undergone by eight National League teams, makes it extraordinarily difficult to assess the true contribution of a ballpark to a team's offense or defensive strength. While we openly admit this difficulty, we still feel that the ESPN model for Park Factors is inadequate and requires improvement. Its theoretical errors are too significant for the amount it is currently quoted. One purpose of our paper was to shed light on this astounding fact.

However, the main purpose of our paper was to demonstrate the advances that our model makes to the study of Park Factors. Our results show marked improvement over ESPN's estimates when looking at both real data and simulations. Our methodology is significantly more thorough than that used by ESPN, and as a result our findings are free of bias. Our work, however, is by no means the last step necessary in the study of Park Factors. The statistic is far too important to evaluating player performances, and consistent debate is essential. We have no doubt that further research can and will improve upon our work. We have merely attempted to incite a deeper look into the realm of Park Factors.

References

[1] "Batting Park Factor." Wikipedia. 2007.
http://en.wikipedia.org/wiki/Batting_Park_Factor.

[2] Click, James. "The Only Constant is Change." *Crooked Numbers*. 2005.
http://www.baseballprospectus.com/article.php?articleid=3814.

[3] Fox, Dan. "On Atmosphere, Probability, and Prediction." *Schrodinger's Bat*. 2007. http://www.baseballprospectus.com/article.php?articleid=6816.

[4] Morris, C. N. and C. L. Christiansen. "Hierarchical Models for Ranking and Identifying Extremes," *Bayesian Statistics 5*, eds. J. M. Bernardo, J. O. Berger, A. P. David and A. F. M. Smith, Oxford: Oxford, 1996.

[5] "Park Factor." ESPN: The Worldwide Leader in Sports. 2007.

http://sports.espn.go.com/mlb/stats/parkfactor.

[6] Perry, Dayn. "Putting the Park Back in Park Factors." *Can of Corn*. 2005. http://www.baseballprospectus.com/article.php?articleid=4250.

[7] Perry, Dayn. "Stats 101: Environmental Effects." 2007. http://msn.foxsports.com/mlb/story/7191988.

[8] Rybarczyk, Greg. "Home Run Park Factor - A New Approach." 2007. http://www.hardballtimes.com/main/article/home-run-park-factor-a-new-approach.

[9] Sheehan, Joe. "Park Effects." *The Daily Prospectus*. 2001. http://www.baseballprospectus.com/article.php?articleid=1164.

[10] Thorn, Jim, Pete Palmer, Michael Gershman, Matthew Silverman, Sean Lahman, and Greg Spira. *Total Baseball: The Official Encyclopedia of Major League Baseball*. Total Sports Publishing: New York, 2001.