Applied Mathematics at the Ballpark

The Life of One Sabermetrician

Ben Baumer

I am not quite sure how I became a New York Mets fan growing up in western Massachusetts—Red Sox territory—but it happened. By the time I was nine years old, I had a 1986 NLCS Mets pennant on my wall, a Dwight Gooden T-shirt on my back, and the familiar blue-and-orange baseball cap on my head.

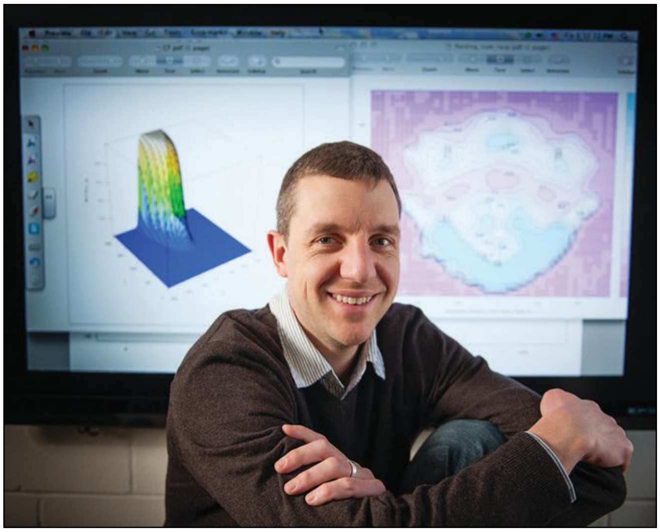
Like any Little Leaguer, I dreamed often of *playing* for my favorite team—but never working for it. But the summer of 2003 brought me a tantalizing proposition. I had completed a master's degree in applied mathematics at the University of California, San Diego. I was back home to plan my next move. Out of the blue, a close family friend asked if I would be interested in doing statistical analysis for the Mets.

"Huh?" was all I could muster.

He persisted. "You should think about it. Have you read *Moneyball*?"

I hadn't. But by then, Michael Lewis's account of the Oakland A's 2002 season was atop the best-seller list. In the book, Lewis argues that the management's extensive use of statistical analysis fueled its unlikely playoff berth. The notion that statistics could help teams make better decisions about baseball players was sweeping through front offices. According to my connection, the Mets were in the market for a statistical analyst.

After several months of interviews, phone calls, emails, and desperate living in Brooklyn with no job, I became that statistical analyst. I am convinced that what earned me the offer was not my ability to perform cutting-edge statistical research, but my ability to integrate statistics with a good-enough understanding of baseball that I could present quantitative findings in a digestible fashion. Also, like all entry-level employees in sports, I was willing to work for peanuts.



Jim Gipe

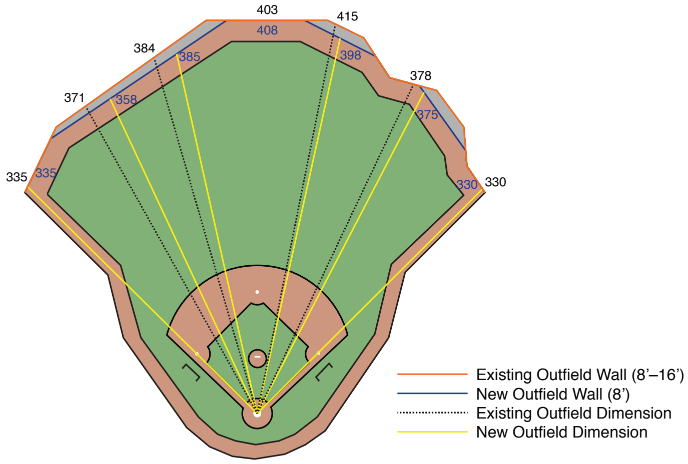
Ben Baumer.

My Time with the Mets

During the eight and a half years that I worked for the Mets, I saw dramatic changes in the way the team processed statistical information. One of the first things I was asked to do was to evaluate a group of free agent relief pitchers. Which ones were most promising, and why? Immediately, I recognized that this task was representative of a larger class of assignments that I would be given again and again.

I built a database of player statistics and a web-based front end that automated the generation of reports. Moving forward, I could update the database as new info came in. My colleagues could look up players on their own and have our internal metrics automatically computed for them.

Although it took me much longer to complete this first assignment than it could have, it saved me thousands of hours in the long run. By the time I left the Mets, multiple dedicated servers powered numerous



internal websites, upon which most members of the baseball operations department (including front office, field staff, and scouts) depended every day.

This type of innovation was happening simultaneously all over the world, in every industry. Today, most baseball teams have a similar infrastructure. Teams like the Tampa Bay Rays are heavily invested in statistical research and technology, and dedicate as many as eight people to using statistics and analytics to help the general manager make better decisions. Only a handful of teams do not have any interest in such things. In these cases, the objection is likely philosophical.

The work that I did for the Mets is typical of what is now being called data science, and for such jobs, a solid foundation in mathematics, statistics, and computer science, along with a working knowledge of the application domain—baseball in my case—is essential. But one of the most challenging aspects of my job was that none of my colleagues was trained in the more technical fields, nor was anyone working on the same type of problems. This created multiple obstacles.

First, I had nowhere to turn at the office when I got stuck (and I did!). Compounding the problem was that the work needed to be kept private, so I couldn't share ideas with colleagues at other teams or like-minded individuals outside of the organization. Thus, when it came to something I didn't know, I had to learn it on my own. I remember teaching myself Bayesian statistics, because I wanted to understand a fielding model that was published in the *Annals of Applied Statistics*.

Second, any statistical analyst will ultimately have to convince her boss that what she is arguing is true. If your argument is based on a complicated statistical model, then it becomes especially challenging to persuade a nontechnical audience. To give up and beg, "trust me, this works," doesn't cut it. You must be able to think creatively about graphical displays, simplified models, analogies, and examples as part of your presentation.

More subtly, learning to speak the *lingua franca* of your field helps others to see you as one who belongs and helps you talk in a way colleagues will understand. For example, for batters, strikeouts are bad, and walks are good. Therefore, a commonly cited metric is a batter's strikeout-to-walk ratio—lower is better. When discussing a player in a roomful of colleagues, you might be tempted to say: "He has a high strikeout-to-walk ratio." But that's not the way they talk. You may do better saying something less precise, but more understandable, like: "He doesn't control the strike zone" or even "He's a free-swinger."

Citi Field Redesign

One of the more interesting projects on which I worked was the redesign of the outfield wall at the Mets's home ballpark of Citi Field, which took place during the summer of 2011.

Citi Field was built in 2009. Although Jody Gerut of the Padres hit the third pitch in Citi Field history for a home run, the park saw precious few homers. By midseason in 2011, just 1.9 percent of all plate appearances resulted in a home run, the lowest rate among all 30 major league parks.

Moreover, the home run park factor, which attempts to control for the quality of the home team, was the fifth lowest in baseball. The situation was so dire that no Mets left-handed batter had ever hit an opposite-field home run at Citi Field!

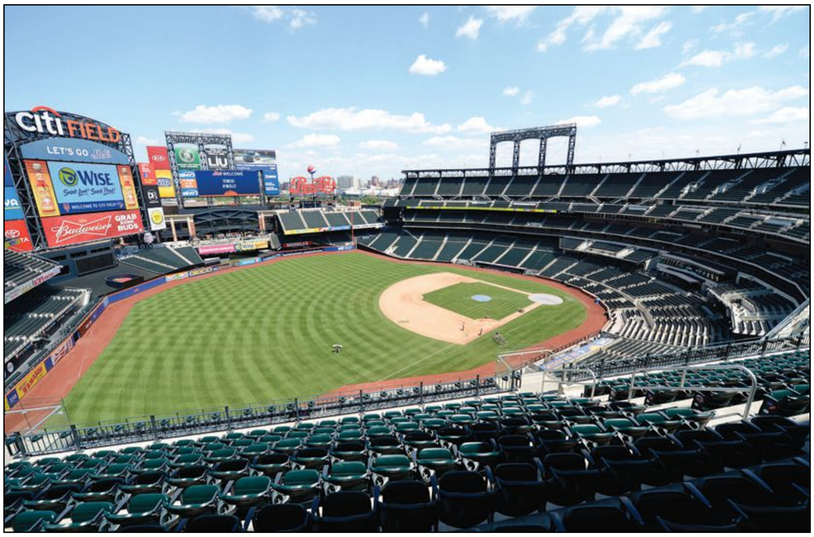
Publicly, several prominent right-handed hitters complained about the imposing 16-foot left field wall, which was not terribly deep, but was very tall (outfield walls are most often eight feet tall). Sandy Alderson, the team's general manager, spoke about making the park more "fair." Management also worried that the difficulty of hitting a homer at Citi Field was becoming a psychological burden that kept the Mets batters from hitting home runs in games on the road as well.

Regardless of the veracity of these claims, three of my colleagues and I were asked to estimate the impact of several redesigns of the outfield Citi Field wall under consideration.

I suspected that with a problem this complex, visualization was going to play an important role, so I started by plotting some data. Every major league team gets a data feed that contains the (x,y)-coordinates of every batted ball that lands in play. It was easy to extract only fly balls that were hit at least 300 feet at Citi Field.

I plotted these and overlaid other informative elements: the current outfield configuration, the "average" configuration of a major league park, and each of the proposed configurations. After color-coding the different outcomes of each batted ball (e.g., double, out, home run), it was easy to see which balls were likely to be affected by the proposed changes.

We quickly discovered that our data was not perfectly accurate (is it ever?). For starters, it was collected by human beings who watched video and then marked the eyeballed (x,y) location on a sheet of paper—not by GPS or some automatic video capture. The locations were reasonably accurate, but occasionally illogical things would appear in the plots, like a double being hit over the wall. This did not inspire confidence.



Courtesy of the New York Mets

The new Citi Field.

More problematic was the fact that we were using a two-dimensional data set to address a three-dimensional problem. The distance of the left field wall was only part of the problem—just as significant was the fact that the wall was twice as high as a normal outfield wall. We lacked data on how high a ball had landed on the wall (the z-coordinate), and so we couldn't distinguish a ball that smacked the base of the wall from one that bounced off the top. Under the redesign, the former would likely not be a home run, but the latter certainly would be. Of course, answering this question definitively also depended on the trajectory of each batted ball, about which we also had no data.

In the end, we based our analysis on a combination of high-tech and low-tech thinking. First, we used the database to identify a set of a few hundred balls in play that we thought, based on geometry and data, would likely be affected by the proposed change. Then we enlisted a very capable intern (who now has a full-time job doing analysis for another club) to watch the video of each of these plays. They were categorized as "yes," "no," or "maybe" in terms of their likelihood of becoming a home run. The yeses gave us our point estimate, and the maybes provided our margin of error.

During the ensuing off-season, the left field fence was brought down to a normal height of eight feet and moved in to create room for a raised seating area. In right center field, the wall was brought in considerably, and a patio area was created in right field by bringing in that fence.

How did our estimate perform in the wake of these changes? From 2009 to 2011, the home run rate at Citi Field was 18.6 per 1,000 plate appearances. Since the start of the 2012 season, the home run rate at Citi Field has been 25.0 per 1,000 plate appearances, a figure that falls safely within our estimated confidence interval.

It's important to recognize that the success of our work came from an interaction of both technical and nontechnical labor. It would not have been feasible to watch video of the thousands of balls that had been put into play in Citi Field without first identifying the relevant few. By the same token, data analysis alone would not have yielded accurate estimates or inspired the confidence one would like to have in a multimillion-dollar renovation. In this case, it required both technical ability and creative thinking to find an optimal solution.

Conclusion

Working in baseball was a special chapter in my life that I can never forget or trivialize. Although I was exceptionally lucky to be the right person in the right place at the right time in getting the job, the opportunity was available to me only because of my applied mathematics background. Moreover, my ability to succeed as a statistical analyst was predicated on a firm grasp of statistics, the ability to compute with data, and a working knowledge of my field of application. I'm sure these skills are universal for any occupation in applied mathematics, but I was fortunate enough to begin my career working for my favorite team. ■

Ben Baumer teaches in the Mathematics and Statistics Department at Smith College and is a coauthor of The Sabermetric Revolution. *From 2004 to 2012, he was the statistical analyst for the front office of the New York Mets.*

Email: bbaumer@smith.edu

http://dx.doi.org/10.4169/mathhorizons/22.1.18