Working with Data, Keeping the Mop Handy

Paula Mikrut, CFA
March, 2015

As a quantitative analyst at OakBrook Investments, I wish I could tell you that I spend my days reading research papers on investments, brainstorming new ideas to take advantage of inefficiencies in the markets with the rest of the research team, and developing and running simulations to test our ideas. These are some of the things I do, and they are the most fun and challenging part of my job, but they aren't where I spend most of my time. If you're a client of ours, you should actually be happy about that.

I actually spend most of my time testing our investment ideas, and I spend most of that time testing the data we use as input into those fun and challenging activities I just mentioned.

About fifteen years ago, I was hired by a company that aggregated and sold market data to traders. I was brought in to start and run a business unit that would add value to this data by detecting problems and cleansing it before delivering it to clients. When I was first contacted about the position, I was skeptical about whether we could add enough value to justify the expense. As I learned more about market data, I was appalled at what I saw. As we received it from our vendors, the incoming data was rife with errors as obvious as negative stock prices, prices that were off by orders of magnitude due to missing decimal points, and data for securities that didn't exist.

I learned from my clients that these errors could cost them millions of dollars and that some had departments whose sole job was to scrub the data they got from us before passing it on to their analysts and traders. I also learned that it was a problem in particular for people who were making trades based on the results of analytical programs, where a small number of bad data points could throw off calculations in ways that weren't obvious if people weren't looking at the raw data.

When I looked into how such obvious problems could exist, I found that there were steps in the trading process where opportunities for human error still existed. For example, a floor trader might make an error when writing down an order, or the clerk who was responsible for entering the order might type something incorrectly. And, even fifteen years ago, the exchanges produced such a flood of information that quality checking was a gargantuan task. The question for my business unit was not whether we could add enough value to justify our existence, but how we could ever hope to stay in front of the problems?

In 2010, I started at my current position at OakBrook. I knew that financial market data would be critical to my job. I assumed, though, that since most trades were now electronic, and since financial reports were filed electronically, the problem of human error would have been solved, and data quality would not be a major concern for me.

I was wrong.

In some ways, the data I work with now is much better than it was fifteen years ago. I have not, for example, run into a negative stock price since I started at OakBrook. In other ways, though, it is just as problematic. Today's issues almost always revolve around corporate actions or the interactions between individual data vendors and the integrator that delivers it to us.

OakBrook has been in business since 1998. We have built a database that contains information about stocks that have belonged to all of the indices that we track. It is largely a point-in-time database, which means that the information was collected in real-time and reflects what was known about a stock on a specific date in history. We combine information from this database with data that we query from our market data vendors and use it to test investment ideas.

In the rare case, this is pretty straightforward. AutoZone, Inc., for example, has been in business continuously since at least the beginning of 1998. In that time, the company has not changed its name, its ticker, or its CUSIP (the identifier that we use to track stocks). It has undergone no splits. If I query my market data vendors for information on AutoZone, I can be pretty sure what I'm getting. Unfortunately, AutoZone really is the rare case. When I went looking, it was the only stock I could find in our database that has existed since the company's inception and hasn't been involved in any corporate actions that might, if we aren't careful, distort the results of a back test.

Splits and dividends are usually easy to deal with, but other corporate actions are trickier. For a simple example, consider Aluminum Company of America, which changed its name to Alcoa Inc. in January, 1999. Along with the name, the company also changed its CUSIP, but nothing else about the company changed. It continued to operate the same business in the same industry. When I run investment simulations, I would consider this to be the same stock after the name change as it was before, and I need to be able to build a seamless data history for the purpose of testing and performance measurement. However, my data vendors disagree. I currently have seven ways to access market data through three different vendors. Four of those access methods tell me the company represented by the old CUSIP doesn't exist. If I wasn't careful, the stock could drop out of my tests before 1999 and skew my results.

I'm always looking for companies that have undergone interesting corporate actions that I can use to test my data, and my current favorite is Google. In April, 2014, Google issued a new share class, Class C. These shares were issued to existing Class A and Class B shareholders on a one-for-one basis. One of the interesting things about this transaction is that our data vendors each treated it differently: one as a split, one as a stock dividend, and one as a spin-off. If I wasn't careful, a transaction like this could cause problems if I mixed data from different vendors in a test. In fact, we have to be careful

whenever we mix vendors' data, since they often disagree about the effective dates of simple corporate actions or even whether a particular action occurred.

Then there is my all-time favorite stock to test with: Liberty Media. Here is my best description

of Liberty Media's corporate actions history.

Liberty Media has ownership stakes in several other media, communications, and entertainment businesses. It's a company that always seems to be in the midst of a spin-off or an acquisition, a name or ticker change, or the issuance of a new share class. Its subsidiaries often have similar names, and it can be challenging to figure out the corporate pedigree and history of one of its stocks.

So, why is this important? I'm glad you asked (it means that you're still reading).

The biases that researchers can introduce into their testing are well-known and widely discussed (e.g., survivorship bias and look-ahead bias). Data integrity may be like cyber-security was until recently, where two common reactions were, "Sure it's important, but I'm sure that my bank (data integrator) has solved that problem. Isn't that their job?" and "I'm sure there are problems, but what can I do?"

The data integrators I work with claim to track corporate actions and changes to securities for us. I'm skeptical. It's in my nature. Whenever I start a project that involves data we haven't used before,

I spend a lot of time wandering around in the data, trying to imagine the ways that errors may have been introduced, and testing to see whether the data is good enough to use for a research project.

For example, the research team recently agreed on a project that involved the use of credit quality. I looked and found that the data was available from one of our integrators, in a database provided by a large, well-known data provider. I looked at data across the range of dates we planned to use for the project. I found that, as I went back in time, more and more of the stocks I looked at had no data of the particular type I was querying. So I called our data integrator. In this, or any of the similar conversations I have with our data integrators, their side of the conversation goes something like, "It can't be our fault. It's not our fault. There is no way it's our fault." My side involves supplying more and more evidence and examples of bad or missing data, until they finally see the light and say, "Huh, I guess we have a problem."

The same thing happened on another recent project where I ran some simulations using a familiar data set in a new way. The simulations looked promising, so I began testing the data to make sure that the results were real and not due to data errors. When I looked, I found what I thought was too much missing data and took the issue to my data integrator. Once we got past the "It can't be our fault," phase of the conversation, we moved into, "This isn't that big of a deal. You're not talking about a lot of missing data. Most of our data is right!"

This was an interesting statement. "Most of our data is right," is both correct and irrelevant.

I explained to them that most of the information in a data set is in the stocks that have undergone some kind of corporate action, and that the stocks most likely to be missing were exactly the stocks that were most likely to have an outsized impact on my test results. I shared with them the fact that I had found another source for this data and estimated that one-half of the excess return that showed up in my tests was a result of their missing data. I was able to convince them to investigate the

issue, which resulted in the decision to completely re-load one data provider's data from scratch, a project that has now been going on for ten months. Sigh.

I take comfort in a recent article in the New York Times[1] which estimated that data scientists spend 50-80% of their time doing "data janitor work". According to Monica Rogati, vice president for data science at Jawbone, "At times, it feels like everything we do." It does. But I'm glad that I'm not alone.

It would be nice, and my job would be more fun, if I could spend most of my time finding and testing new ideas. But the whole point of my job is to produce the best results we can for our clients. Our first rule has to be, "Do No Harm."

When we run simulations to make decisions about our investment process, we analyze the results to ask whether it's reasonable to extrapolate the past into the future. To understand this, we need to analyze whether we are accurately modeling the past. For the sake of our clients, those of us who do investment research need to keep our mops and buckets handy at all times.

[1] Lohr, Steve. "For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights". *New York Times*. 17, Aug. 2014