Northfield Information Services

Coronado, California

December 9, 2001

Conference Paper

"The Advantages of Using *As First Reported* Data With *Current* Compustat Data For Historical Research"

Presented by:

Marcus C. Bogue, III, Ph.D. and Morris E. Bailey

Charter Oak Investment Systems, Inc.

Wellesley Hills, MA

Introduction

Let's talk about tools: hammers and screwdrivers, for instance. Both the hammer and the screwdriver are useful and powerful tools. Each is indispensable in its own way for the purpose it was intended. Building a case, for instance, can require both a screw-driver and a hammer. But picking-up the wrong tool at the wrong time (pounding a screw with a hammer; or trying to bang on a nail with a screwdriver) can be risky business.

What's the point? There is power in using the right tool for a given task at the right time. And there is risk, strange outcomes, danger in not understanding the consequences of using a tool for other than its intended purposes.

Standard and Poors' Compustat is, in our estimation, the premier fundamental database available for serious equities research. For more than 30 years Compustat has been the preeminent vehicle for corporate modeling and investment research. The Compustat dataset has length, breadth, depth, coverage, consistency, backed by a corporate commitment to quality enjoyed by few, if any, others. By and large, there's not much of anybody else to talk about. The Compustat dataset speaks for itself.

Over the years, S&P has made significant efforts to enlarge and evolve the dataset and change what's offered and available with the times. Compustat was originally focused on currently available industrial companies (since that's what the US economy was focused on in the '60's). In the early '70's, historical data on dead companies was made available to counteract concerns over survivor bias. In the early 1990's Finance companies were added to the standard dataset.

When talking about the flagship Compustat product (which we will call "*Current Compustat*") certain database construction principals are important to keep in mind.

1. When a company is added to the dataset, that company's back history is provided if it is available.
2. The dataset is "current" and is based on the data available in a company's latest SEC filings. The latest 10-Q and 10-K data is entered into the current monthly publication of the Compustat data.

The design of the *Current Compustat* dataset not only took into account the SEC and latest corporate information available, it was also oriented toward the analytic technology and modes of the times. Back in the late 1960's when Compustat was in its formative stages, analytical styles and quantitative tools were different than they are today. In the 1960's:

1. State-of-the-art analytic technology consisted of (mostly green) 13 column pads of paper and (mostly yellow) pencils.
2. "High tech" was a *sharpened* #2 pencil used by someone who could touch-add. A Monroe calculator was a luxury. (Whirr…whirr…whirr... clunk: I can still hear it.)

And thanks in part to leading texts of the times (Graham and Dodd, for instance) the analytic game was to take the most recent and up-to-date information available on each company, one-by-one, and use the best available and most current data to *forecast* that company's future performance – mind you, that is future defined almost exclusively as NOW- (real time-on-the-date-I'm-doing-the-analysis)-Forward

This, then, is the task for which the *Current Compustat* data was designed. Many, probably most, still use the *Current Compustat* data that way for exactly that purpose (PC's and spreadsheets, not 13-column pads, notwithstanding). And there is nothing risky and everything right in using *Current Compustat* data to take either a company or a portfolio of companies' most recent performance and model its future. In fact, we would argue Compustat remains the premier vehicle for doing just exactly that sort of analysis. So if you want to *forecast* use *Current Compustat* data.

So is there an issue? Where's the risk?

Tools and technologies have changed. Analytic techniques have evolved. "Quant" analysis now often means more than a company-by-company, next quarter's earnings forecast based on last-and-a-year-ago's numbers. Accordingly, S&P has again evolved Compustat's data offerings.

Out of concern for the quality of the dataset as sometimes used for quantitative research for backtesting purposes, Compustat and Charter Oak collaborated starting seven years ago to structure another look at the Compustat data. This new look is specifically designed to be used *retrospectively* – that is not *forecasting* but *back-casting*. This second look of the Compustat data is called Compustat *As-First-Reported* (or "Compustat AFR") data. It is available as an additional dataset, a companion to *Current Compustat*.

Compustat AFR was specifically designed and implemented to assist the user in retrospective, historical, cross-sectional and time series analyses of the Compustat data.

The purpose of this paper is to define some of the salient attributes and advantages of each of the two Compustat datasets (*Current Compustat* and *Compustat AFR*) and to discuss: (1) each datasets' strengths and uses, as well as (2) the possible consequences of using one dataset for the purposes for which the other dataset was intended.

Specifically in this paper we will address two attributes of *Current Compustat* (namely, Backfilled Data and Data Restatements, each of which is an important asset when properly employed). And we will enumerate how use of the *Compustat AFR* data in conjunction with *Current Compustat* data addresses analytic and historical research *backcasting* needs.

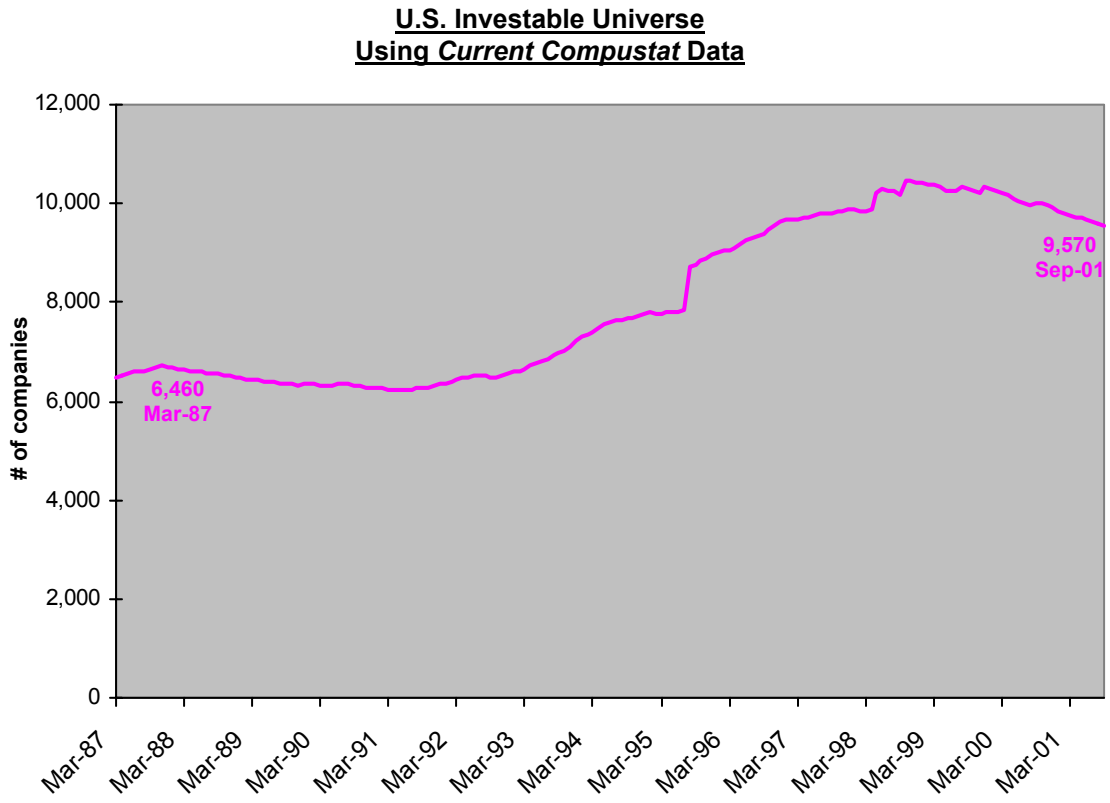
Backfilled Data and Survivorship Bias

The Compustat dataset increases in size as new companies come into being and as existing companies grow to recognizable size. Over the years Compustat has selected and added companies to the dataset routinely on a continuing basis. At certain times large increments have been added as database size and composition enhancement projects have been undertaken.

To measure the growth in coverage of the Compustat data, we defined a set of companies using *Current Compustat* data that could reasonably be called a US domestic Investable Universe. The definition of that set of companies is as follows: Companies in the *Current Compustat* dataset excluding mutual funds and Canadian companies with a stock price at the beginning of each calendar month.

Between March 1987, and September 2001, the Investable Universe grew by 3,110 companies or 48%, as shown in the Figure 1 below.

Figure 1



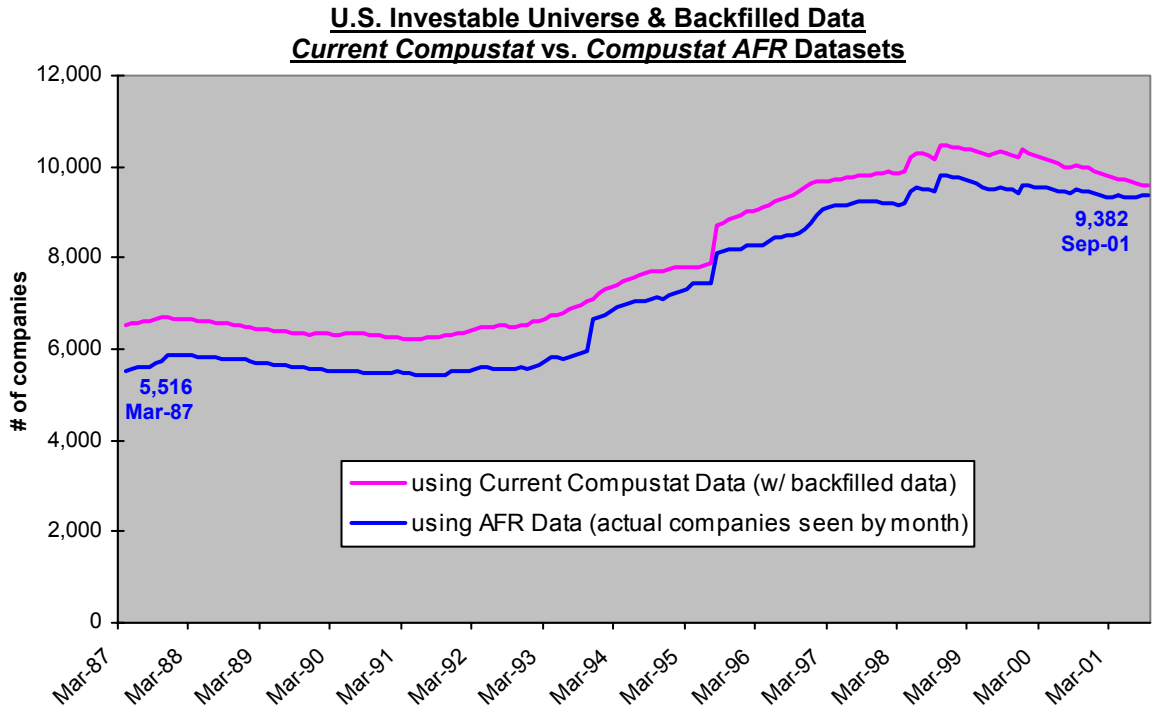
However, the *Current Compustat* dataset contains backfilled data. (Backfilled data = data added to the dataset for financial reporting periods prior to the date on which a company first has made its appearance in the dataset.) As a result of backfilled data, the analysis of the Investable Universe shown in Figure 1 significantly *overstates* the actual size of the Investable Universe as you could and would have actually seen it each month. And the Figure 1 analysis significantly *understates* the contribution of Compustat in adding company coverage to the dataset.

In fact, on average over the period March 1987 to September 2001, 11% of this Investable Universe was not there at any month you might have looked; prior to September 1993, over 15% of that Investable Universe wasn't there to find.

Here in Figure 2, using *Compustat AFR* data, is the actual size of the Investable Universe as it was seen in each month-end cut of the then-current Compustat dataset from March 1987 through September 2001. The Compustat Investable Universe actually increased in

size by over 70%, 3,866 companies, (not by 48%) between March 1987 (5,516 companies) and September 2001 (9,382 companies).

Figure 2

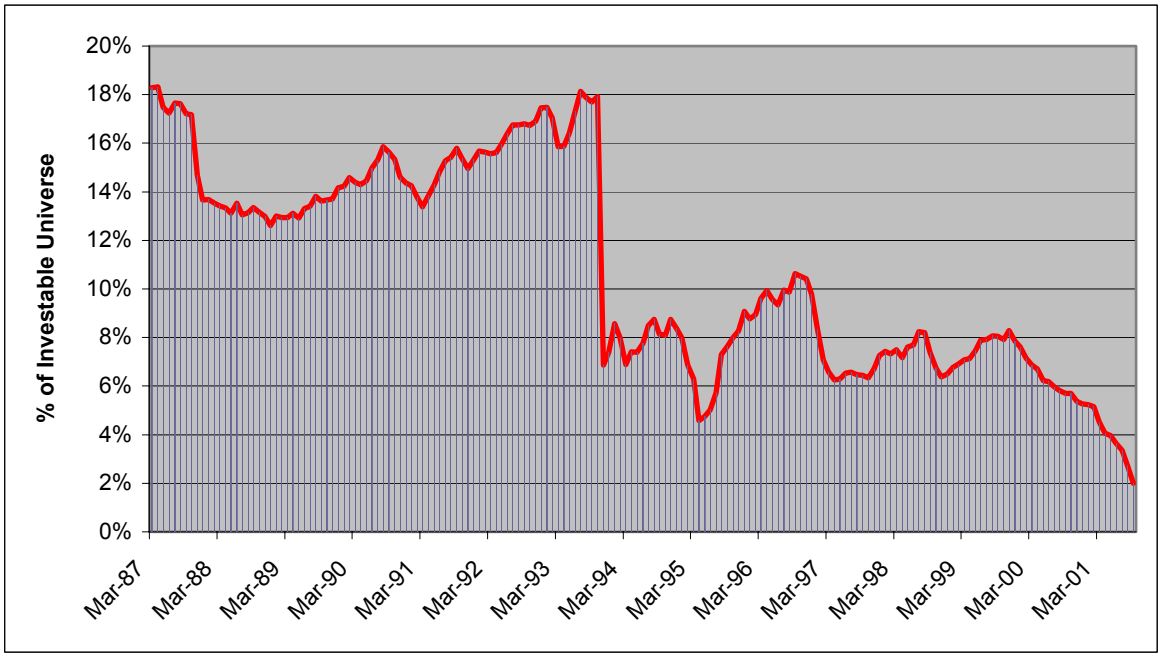


The difference between the two lines is the number of companies in each time period with backfilled data. The maximum number of companies with backfilled data in any month was 1,069 companies in October 1993. The minimum number of companies with backfilled data in any month was 188 companies in September 2001. As is observable in Figure 2, you would expect the number of backfilled companies to decrease the closer you get to "now," to the current date, since in the immediate months preceding "now" there's not enough "back" yet to "fill." If we give ourselves an additional year's worth of "back" to allow for "fill," one sees that there was a minimum of 342 companies with backfilled data (in April 1995) in any month up through September 2000. During the entire time period March 1987 to September 2000 the number of companies with backfilled data averaged 748 per month.

Figure 3 shows the percentage of companies with backfilled data in each month, using the Investable Universe as defined with *Compustat AFR* data as a base. The larger percentage of backfilled company data prior to 1993 is a result of the projects to increase the size, and broaden the types, of companies in the Compustat dataset undertaken in the early 1990's.

Figure 3

Percent of Companies with Backfilled Data: March 1987 to September 2001



From March 1987 through October 1993 the average number of companies with backfilled data was 15.1% of the total companies in the then-current, *Compustat AFR* dataset; while from October 1993 through September 2000 the companies with backfilled data averaged 7.6% of the *AFR* Investable Universe. During the time period from March 1987 to September 2001 the number of companies with backfilled data averaged 10.8%. If we truncate the analysis a year earlier (September 2000) the number of companies with backfilled data averaged 11.2%.

So backfilled data exists. Can there be anything wrong with that?

Answer: Whether backfilled data is good, bad or indifferent depends on the analytical task on which one has embarked, the composition of the dataset one is examining and the time period over which observations for those companies are chosen.

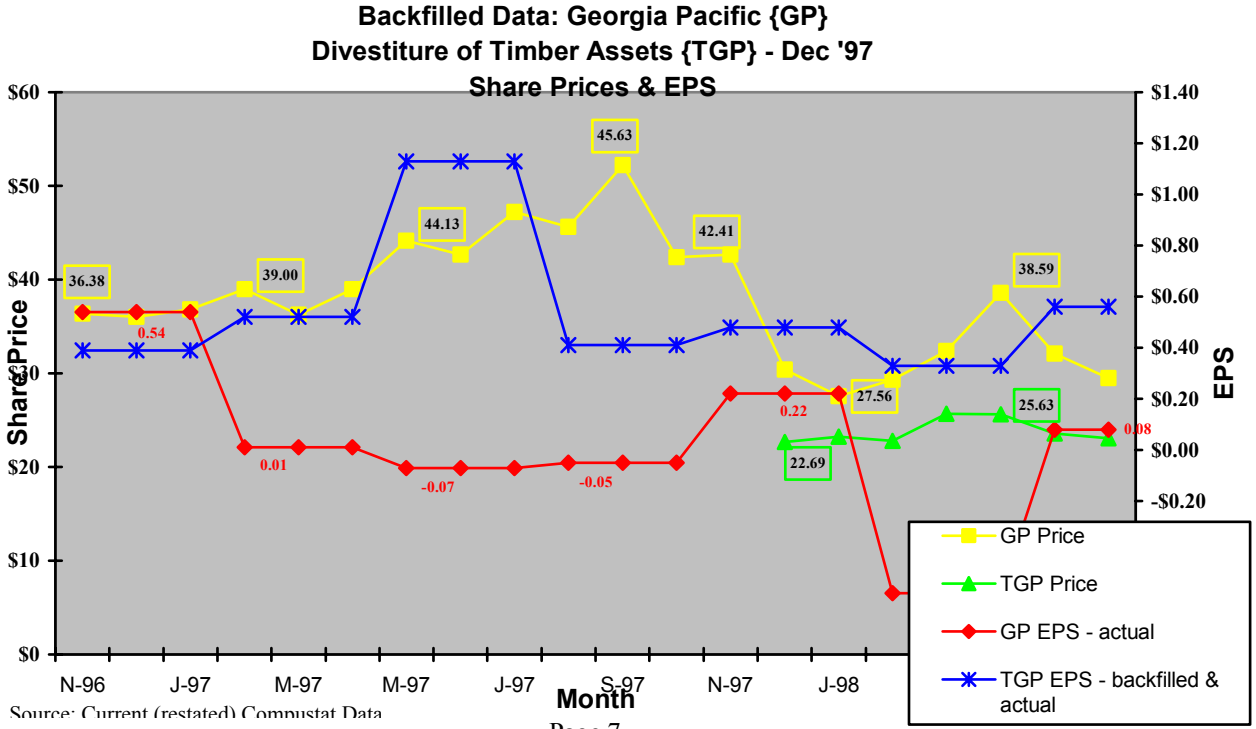
Compustat backfills data with good and honorable intent and for obvious reason: to provide the user a basis for analyzing the added company's performance; to provide a base from which to forecast. If bottoms-up, company-by-company, now-forward forecasting is what you are about, Compustat policy of providing backfilled data when a company is added to the dataset is an essential and unmitigated asset, and one of the essential competitive advantages of the Compustat data. Typically, two to three years

worth of data for the period prior to a company's Compustat database debut is included. However, there have been times (particularly when major database coverage enhancement projects have occurred) when up to ten years and more worth of backfilled data is included. If you are forecasting a company's performance, backfilled data is as important and valid an asset today as it was 30 years ago. If *forecasting* is the task, *Current Compustat* is the dataset.

On the other hand if what you are doing is *backcasting* -- looking back historically overtime and across companies in the dataset, or, for instance, if you are using the dataset to apply historical portfolio selection criteria, backfilled data can pose certain risks, certain dangers. For instance, unless an *as reported, as-was, where-was* dataset such as *Compustat AFR* is used…

1. There is the risk that you will include data for a company, or that your model will select a company, that literally didn't exist. This can happen with data for a spin-off backfilled prior to the spin-off date, or when a company goes public and data from the red herring is included in the dataset prior to the issue date of the stock. As an example, Figure 4 shows the inclusion of earnings for both Georgia Pacific (GP) parent actual and the Timber Group (TGP) backfilled prior to the Timber Group spin-off in December 1997.

Figure 4



(For those who want a clearly defined universe on which to consistently base their research, the corollary argument that you are skewing your research by selecting companies that weren't seen in the Compustat dataset even though they may have been real back-then will also resonate as a valid point.)

2. There is the risk that in a backcasted cross sectional analysis you will double count a spun-off company's results – the results being included once in the mother company's actual results and a second time in the spun-off company's backfilled results prior to the spin-off date. (See also GP and TGP EPS prior to December 1997 in Figure 4.)
3. There is the risk that including backfilled data will skew calculated returns as well as biasing cross-sector analyses. This is the Survivorship Bias issue: that by and large it is only the more successful companies that succeed in becoming noticed, meaningful and in getting added to the dataset. Depending on the analytic point-in-time you choose, returns calculated for prior periods will wander over time based on the publication date of the data you use. Therefore your calculated returns will reflect the companies that happen to have been added since, and the backfilled data included in, the publication date of the dataset you used. (At this point it is interesting to note that many of the studies that we now all view as pillars of finance employed a single publication of the then-current *Current Compustat* dataset. Therefore, the results and conclusions obtained from those studies may have been in degree, if not in whole, affected by the inclusion of backfilled data.)

If you knew ***now*** what you're going to know ***later*****…..or….. few factors beat next month's price**

Clairvoyance (the handmaiden of survivorship bias) is an asset, as demonstrated by the differing returns for the S&P 500 based on which S&P 500 one chooses and over what time period one looks.

As shown in Figure 5, for the 10 years ended in December 2000 the annualized return on the set of companies that constituted the S&P 500 at the end of the period (with the return weighted by their market caps at the end of December 2000) was 32.8%. The annualized 10 year return over the same time period for the set of companies that constituted the S&P 500 at the beginning of each month (with the returns weighted by their market caps at the beginning of each month and rebalanced monthly) was 17.9%, or 14.9% lower than the return on the ending composite.

Sliding this 10-year return analysis back five years changes the magnitude of the difference, but not the phenomenon. As shown in Figure 6, the annualized 10-year return on the set of companies that constituted the S&P 500 at December 1995 (with the return weighted by their market caps at the end of December 1995) was 21.0%. The annualized 10 year return from 1985 to 1995 for the set of companies that constituted the S&P 500 at the beginning of each month (with the returns weighted by their market caps at the

beginning of each month and rebalanced monthly) was 14.9%, 6.1% lower than the return on the ending composite.

Figure 5

Return on the S&P 500: 1990 - 2000 *

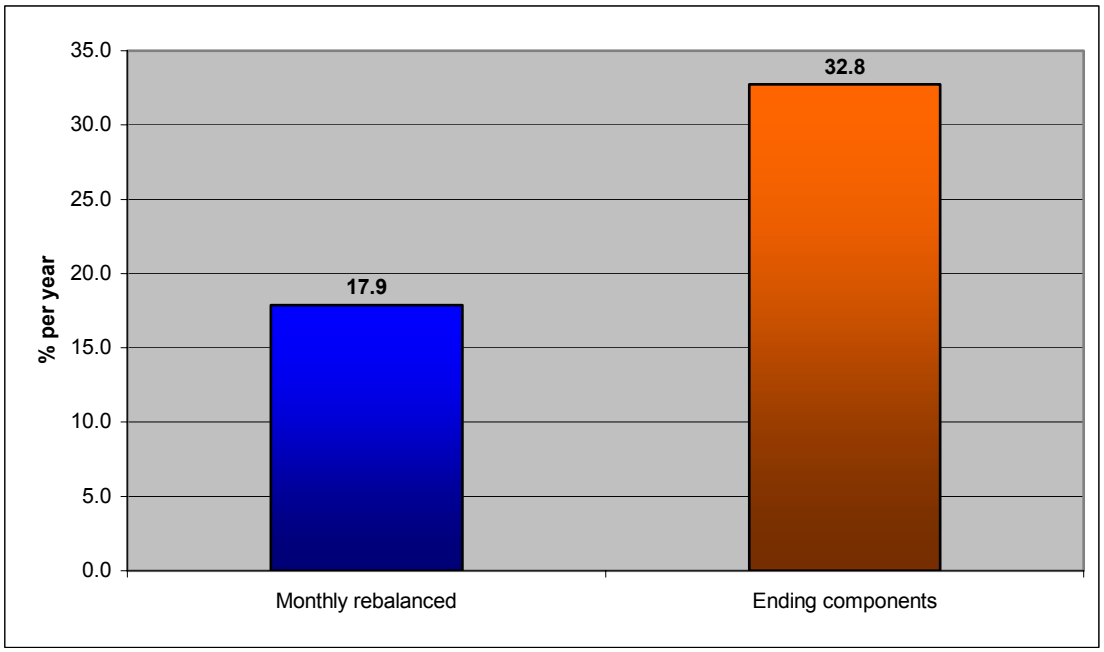
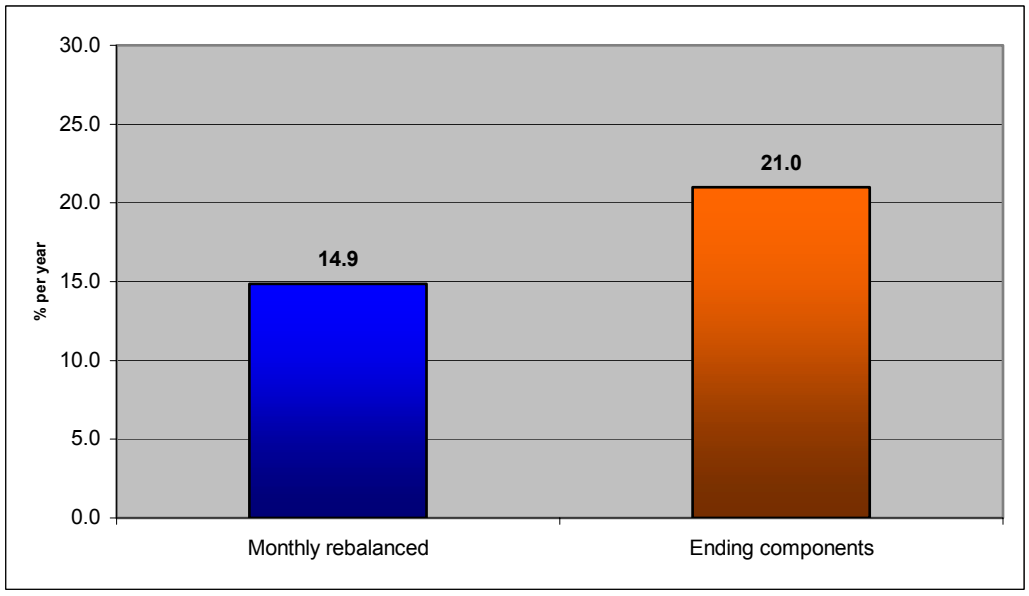


Figure 6

Return on the S&P 500: 1985 - 1995 *



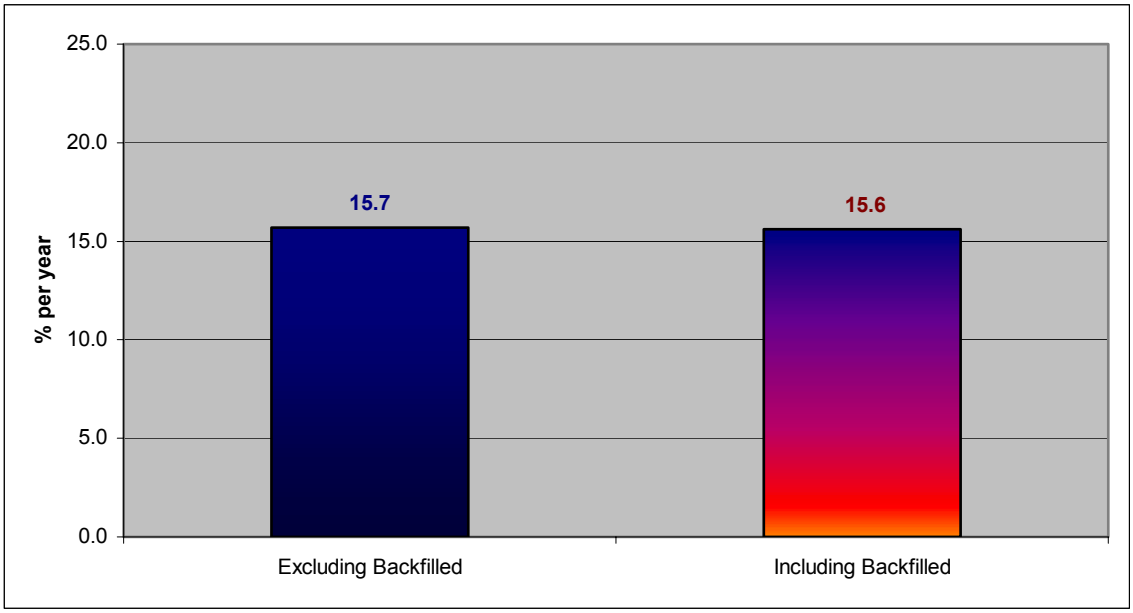
* Definition of returns: For the "Monthly Rebalanced" portfolios, the return is the annualized ten year return for a portfolio constructed from the S&P components at the beginning of each month weighted by their relative market values at the beginning of each month. For the S&P "Ending Components" (S&P 500 as it was at the end of the period) portfolios, the return is the annualized ten year return for a portfolio constructed from the S&P components at the end of the ten year period that were tradable each month weighted by their relative market values at the end of the ten year period.

You are undoubtedly by now stifling a yawn and well into an irritated "no kidding" or "so what?" This "look-ahead" phenomenon (Survivorship Bias writ forward not back), is second-nature conceptually to all of us for an index like the S&P 500.

The point is that fundamentally the same notion can apply to the Compustat dataset. Which Compustat dataset one looks at and over what period of time can yield varying results. For instance, as shown in Figure 7, for the 10-year period ended December 2000, the 10 year annualized returns on the entire Compustat Investable Universe rebalanced monthly for companies that both existed at the beginning of the month with returns weighted by their market cap at the beginning of the month were virtually identical both including backfilled companies' data (15.6%) or excluded backfilled companies' data (15.7%).

Figure 7

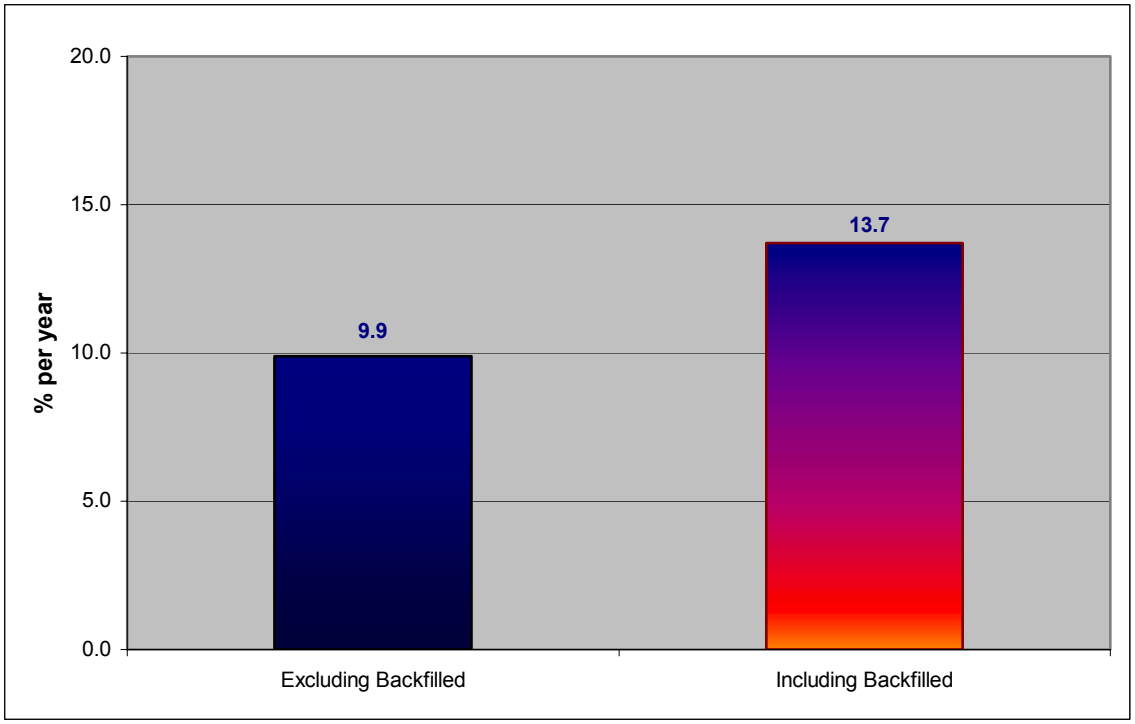
Return on the Compustat US Investable Universe: 1990 - 2000 **



However, as shown in Figure 8, sliding the include-vs-exclude backfilled data 10-year return analysis back five years yields a significantly different result. The annualized return on the entire Compustat Investable Universe rebalanced monthly for companies that both existed at the beginning of the month with returns weighted by their market cap at the beginning of the month for the 10 year period 1985-1995 was 13.7% using "ending composite" *Current Compustat* data (i.e., with backfilled data included). Excluding backfilled companies' data using the *Compustat AFR* dataset the return on the as-seen, "monthly rebalanced" dataset with the same beginning-of-month inclusion and market weight criteria was 9.9%, more than 1/4th lower, for the same time period.

Figure 8

Return on the Compustat US Investable Universe
1985 - 1995 **



** Definition of returns: Annualized returns for companies in the Compustat Investable Universe rebalanced monthly, weighted by the relative market values at the beginning of the month: (1) for the companies currently included each month ("Including Backfilled") using the Current Compustat dataset; and (2) for the companies actually there each month ("Excluding Backfilled") using the Compustat AFR dataset.

Backfilled Data - Conclusion

Whether labeled "Survivorship Bias," "Look-ahead Bias," or by its 50-cent, academically preferred moniker "Retrospective Inclusion Bias," backfilled data is important. It can be either a cornerstone asset or unnecessary noise (read "risk"), depending on what analytical task is undertaken. Including or excluding backfilled data will make a difference in the results you get. User control of backfilled data requires the use of both the *Current Compustat* dataset and *Compustat AFR* dataset.

Data Restatements

Data restatements occur when a company issues a current report to the SEC in which changes to data values for the company's previously reported periods are made. In Compustat, data restatements affect quarterly data, not normal annual data (although a company's 10-year annual summary does get restated). Restatements are mandated by SEC guidelines and can be triggered by such events as:

1. FASB rule changes.
2. Significant changes in corporate accounting presentation.
3. Certain corporate actions, such as mergers, acquisitions, divestitures.

While we know of no general coding of what triggers a given data restatement, anecdotal evidence and spot checking reveal that far and away the majority of data restatements occur as a result of corporate actions related to changes in corporate structure.

In the *Current Compustat* dataset, Compustat's policy is to faithfully represent the most current view of a company as reported in the latest quarterly report to the SEC. If that latest 10-K or 10-Q contains restatements of a previously reported quarter's data, Compustat over-writes these latest data restatements into the prior period's data records. In *Current Compustat* the original, as-they-were-first-reported data values are lost.

Future (now-forward) projections for a company should obviously be based on the current structure of the forecasted company and the latest and most accurate presentation of the company's data. Having previous and year-ago quarterly comparisons on the same restatement basis as current data is an enormous asset in projecting a company's future results. If you are *forecasting*, *Current Compustat* with its restated quarterly data is a uniquely powerful tool. We know of no dataset other than *Current Compustat* that offers the latest historical quarterly data restatements from which to forecast a company's performance.

If, on the other hand, you are performing an historical analysis back through time, *backcasting* either for a single company or across companies, or if you are building

historically constructed factors for portfolio selection, quarterly data restatements skew results. Not all companies have restated quarterly data. Furthermore, not all data restatements are of a significant enough magnitude to disrupt your analysis. However, where restatements do occur and are significant, the risks, the dangers associated with the *backcasted* use of quarterly data restatements can be categorized as follows:

1. Longitudinal data disruptions. A time series analysis of, for instance, a measure of growth or momentum will be thrown off by the timing and magnitude of the restated data values. If a company has frequent quarterly data restatements (and some seem to restate prior quarters' data virtually every quarter), the resulting changes in the restated time series of quarterly data for the company can result in a virtually random set of data values reconstituted ex post facto from various points in time.
2. Cross-sectional data differences. Data restatements are idiosyncratic corporate events based on the "look-ahead" bias inflicted by each company's unfolding corporate reality. The magnitude of the difference between a company's restated versus its *as-first-reported* data values vary totally unpredictably from trivial to tectonic-plate shifting. Cross-sectional analyses which include both companies who have never had data restatements and others who frequently restate data risk drawing conclusions based on random relations between data.
3. Factor Friction. Not all data items can be restated. Fundamental data items are. Market related data can't be. (The stock price was the stock price; the shares outstanding were the shares outstanding.) Factors which use only fundamental data items will be off by the magnitude of the difference in their restated to *as-first-reported* component items. The resulting factor changes using restated (*Current Compustat*) versus the as-reported (*Compustat AFR*) data may or may not yield differences. But factors that employ a combination of fundamental and market data (P/E, Market/Book, Sales/Price, etc.) will always show differences using restated data.
4. Wrong Ranking; Partial Portfolios. Using restated data values in any *backcast*, historical analysis will change both the rank order of the companies tested against a factor and can change the composition of the resulting portfolio.

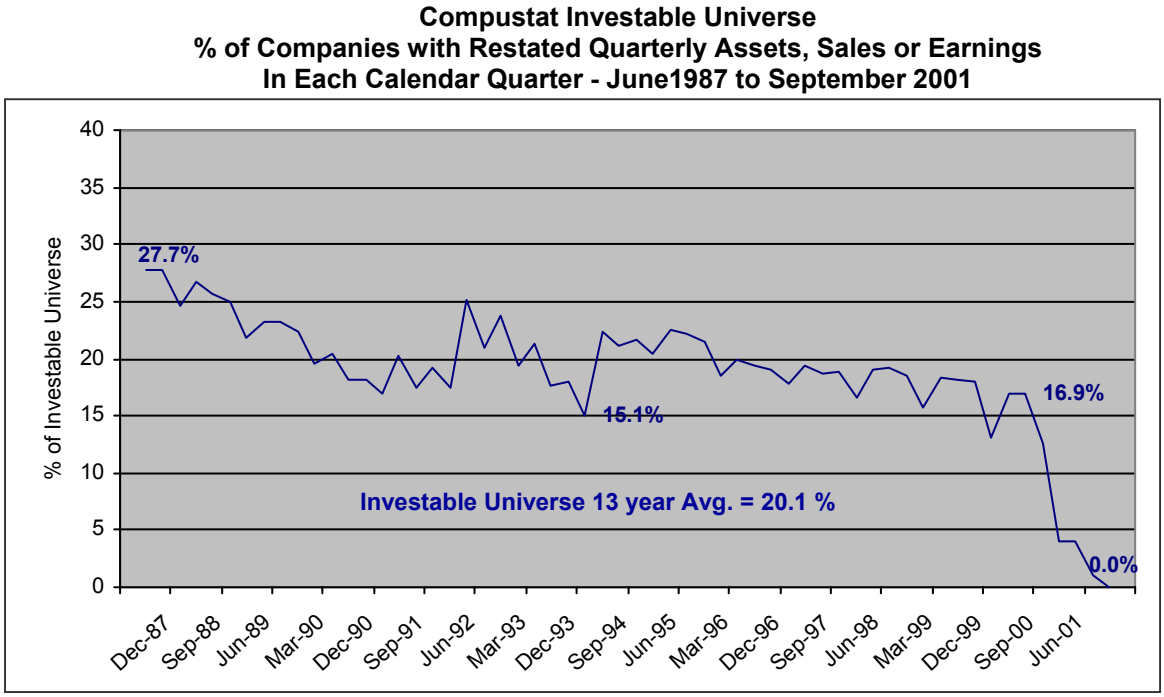
Let's examine the impact of quarterly data restatements by asking two questions using the *Current (restated) Compustat* and the *Compustat AFR* datasets:

1. How prevalent are data restatements?
2. Do they matter? What is the magnitude of the difference between various values calculated using the restated (*Current Compustat*) versus as-first-reported (*Compustat AFR*) data?

Prevalence of Data Restatements

Since 1987, on average one-in-five companies in the Investable Universe in each calendar quarter have restated quarterly data. As shown in Figure 9, the percentage of companies with quarterly data restatements varied from a low of 15.1% of the Investable Universe companies in December 1993 to a high of 27.7% of the Investable Universe companies in March 1987.

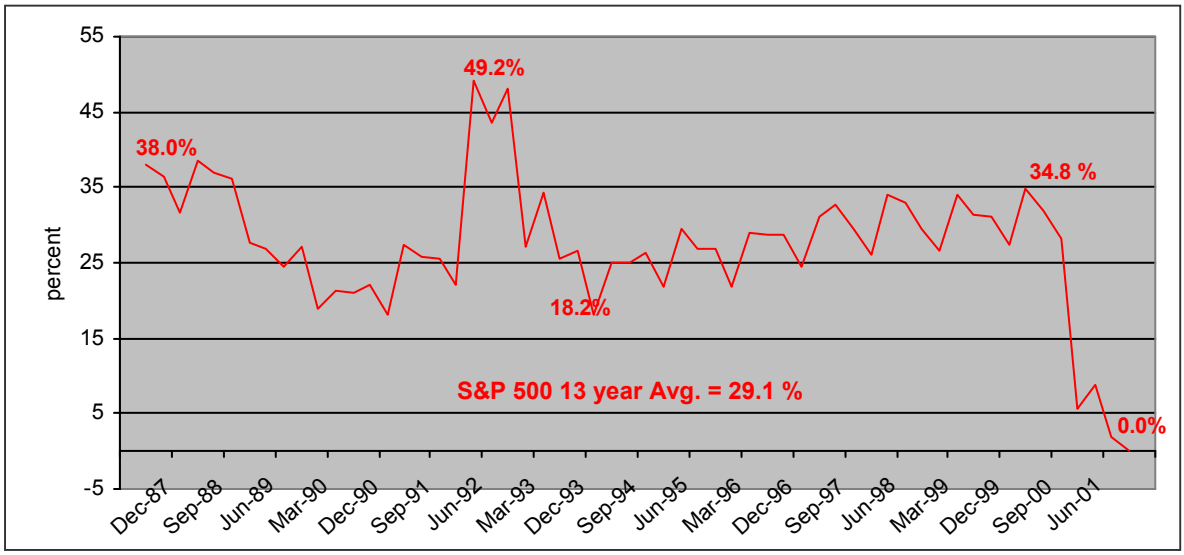
Figure 9



Data restatements are disproportionately a large company phenomenon. As shown in Figure 10, the percent of the S&P 500 companies with data restatements since 1987 was nearly half-again larger than for the Investable Universe as a whole. Over the calendar quarters since 1987 on average 29.1% of the S&P 500 companies had restated data. This percentage ranged from a low of 18.2% in December 1993 to a high of 49.2% in March 1992.

Figure 10

S&P 500
% of Companies with Restated Quarterly Assets, Sales or Earnings
In Each Calendar Quarter - June1987 to September 2001

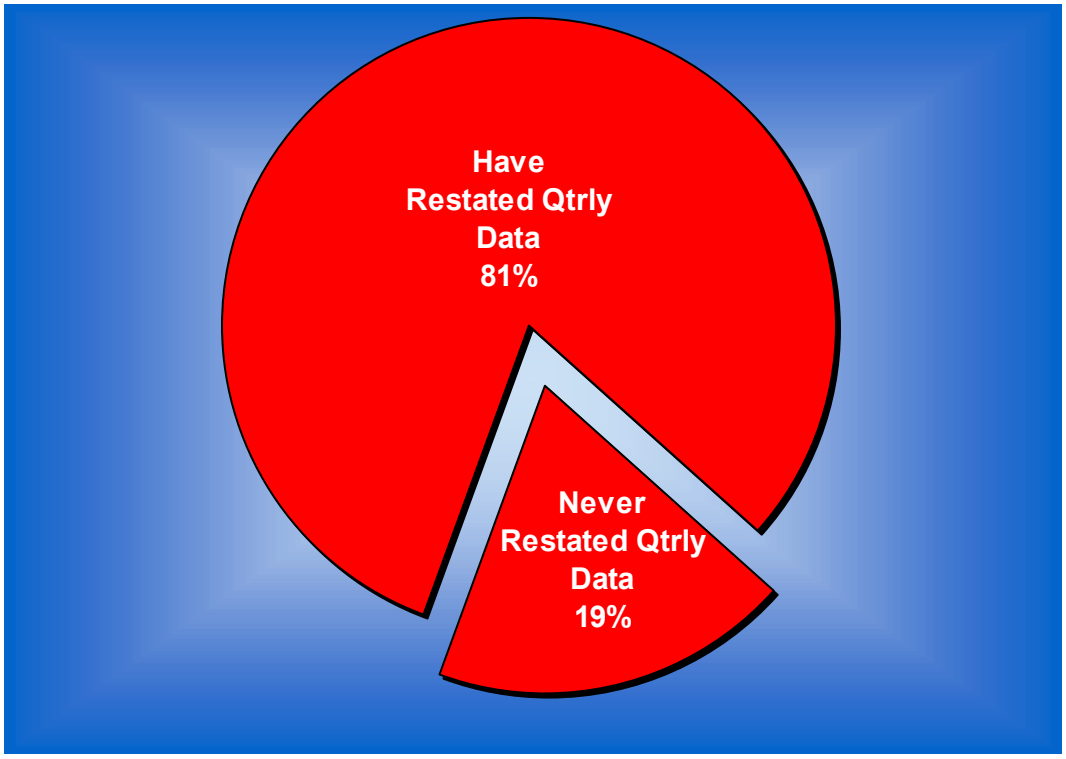


Note that in both Figure 9 and 10, data restatements trail off dramatically to none (0.0%) for the current calendar quarter. This is (obviously, we would hope) because "**RE**statements" are not possible until the next calendar quarter's data is available. Since SEC rules require restatement of both the last and the year ago quarter's data, restatements are not prevalent until the fourth quarterly statement after the original data has been released. (Note also the suspiciously similar shapes between the graph of percent restatements in Figure 10 and share price for the S&P 500 since 1993.)

Another way to think about the prevalence of data restatements is to take a (mental) walk back over time from the present moment to 1987 to see how many companies ever (or never) restated data. Figure 11 presents that look. During the 14-year period January 1987 to December 2000, there were a total of 841 companies in the S&P 500. Of the total, 681 companies restated quarterly data at least once during the 14-year period; 160 companies never restated any quarterly data at any time during the 14-year period….and the 80/20 rule wins again.

Figure 11

Percent of S&P 500 With Any Restated Data
1987 - 2000



Data Differences: *Current (restated) Compustat* vs. *Compustat AFR*

So data restatements exist. How different are restated data values from those *as-first-reported*? And to what degree can differences in restated data values affect factor values and historical research outcomes?

For a start at some answers we aggregated data values by month for the S&P 500 using each of the two datasets. Accordingly, we got two numbers per month: one data value each month using the *Current Compustat* dataset and one data value per month using the *Compustat AFR* dataset. We then compared.

Figure 12 shows the difference in aggregate Sales for the S&P 500 between *Current Compustat* data and *Compustat AFR* data (using *AFR* data as the base) by month from 1987 through 2000. As you can see, *Current Compustat* and the *as-first-reported* (*Compustat AFR*) datasets are different. Aside from an interesting, eye-balled general

correlation with the state of the economy over time, there is little pattern to when negative versus positive variances occur. Restated (*Current Compustat*) data values were higher than the *as-first-reported* (*AFR*) values in 74 of the 168 months. These positive variances ranged from $22.7 million in December 2000 to an extreme of $46.3 billion in October 1998. In the 94 months (of the 168 total) in which restated Sales data values were lower than *as-first-reported*, the variances ranged from a low of -$95.1 million in February 1991 to an extreme of -$38.7 billion in December 1988. The absolute value of the variance in *Current* versus *as-first-reported* Sales of $85 billion represents 9% of the S&P 500 average aggregate Sales over the time period. Remember, though, that in any given month on average fewer than 1 in 3 (29%) of the S&P 500 companies' values are responsible for 100% of this $85 billion variance.

Figure 12

S&P 500 in Aggregate
Sales - Difference in *Restated* vs. *As First Reported* Data
By Month January 1987 to December 2000

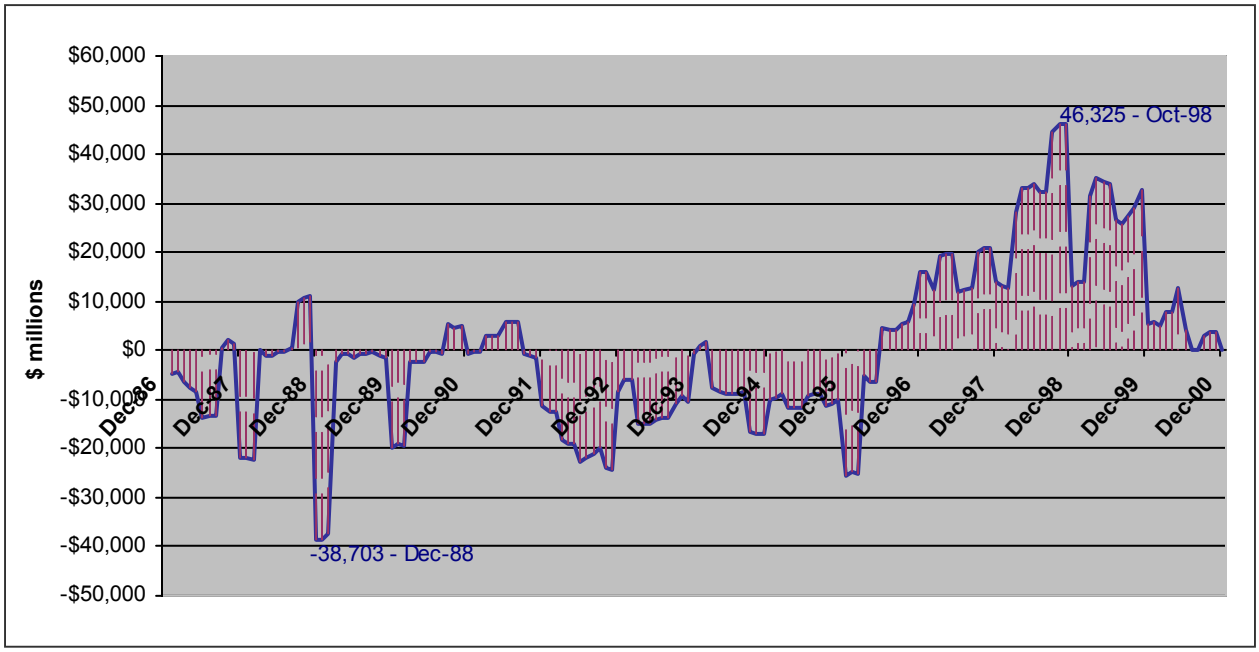
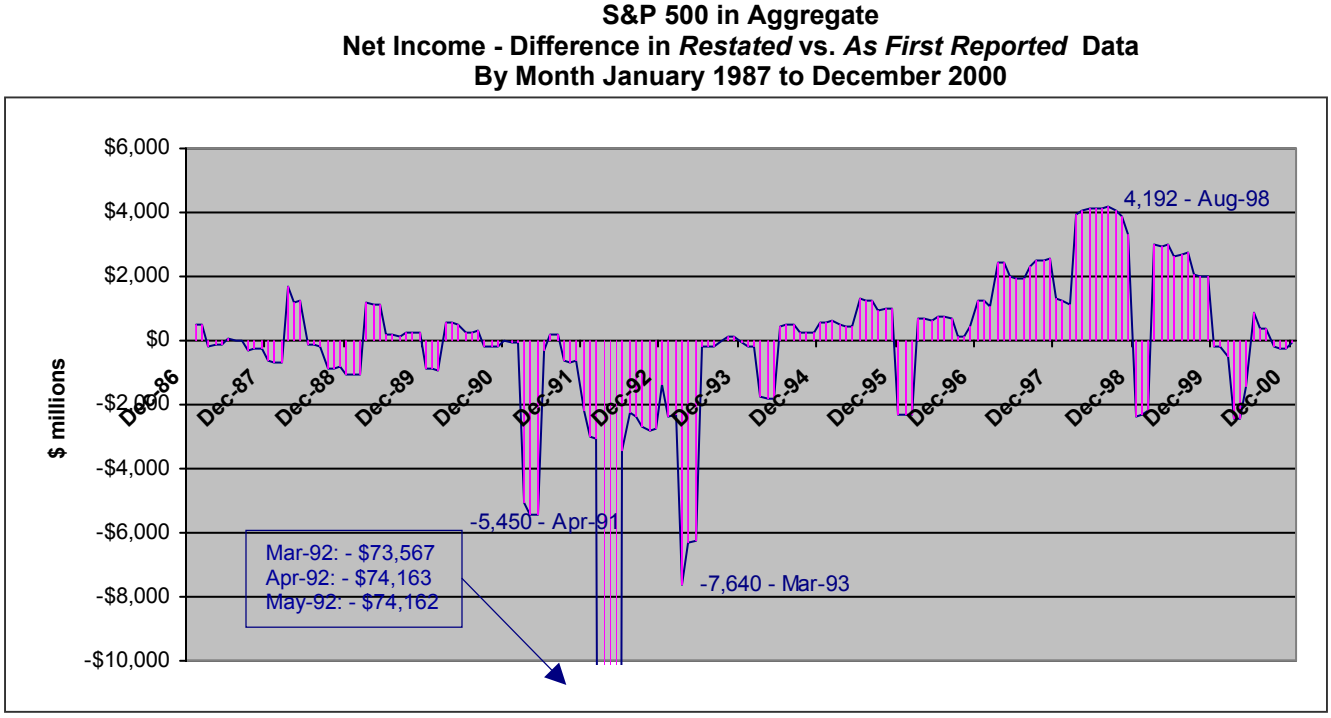


Figure 13 shows a similar graph (*Current Compustat* data versus *Compustat AFR* data) for the same time period using the S&P 500 as a single portfolio. This time Net Income data values were used. Restated Net Income was higher than *as-first-reported* in 91 of the 168 months; lower in 77 months. The positive variances ranged from $ 20.2 million in July 1987 to an extreme of $ $4.2 billion in August 1998. The negative variances ranged from -$19.6 million in September 1993 to an extreme of -$ 74.2 billion in April 1992. The large negative variances in Net Income in 1992-93 correspond with the time when FASB forced recognition of accrued pension liabilities. The absolute value of the

variance in Net Income over the entire time period, $70 billion, represents 120% of the $58.5 billion S&P 500 average aggregate Net Income. Again, remember that in any given month, on average, only 29% of the S&P 500 companies are responsible for 100% of this Net Income variance.

Figure 13

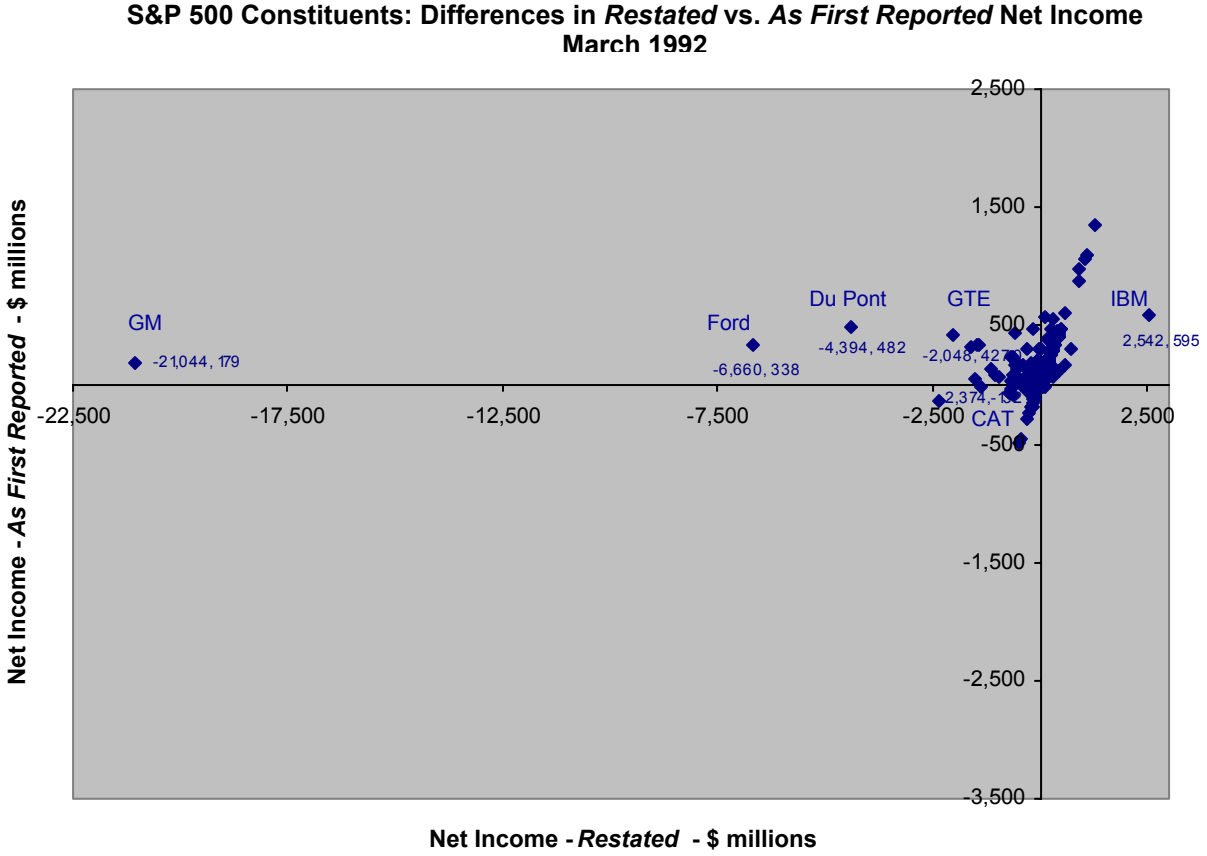


In summary, when you hit a month in which either a Sales or a Net Income data value has different restated data than was first-reported, there is no way to predict either the direction or magnitude of the variance, much less which companies' data values are responsible for the difference.

Figure 14 presents, in scatter-plot form, the results of a bore-down analysis of Net Income taken from one of these months, March 1992, for the constituents of the S&P 500. The X-axis shows Restated Net Income data values taken from *Current Compustat*. The Y-axis is Net Income *as-first-reported* taken from *Compustat AFR* data. Note the cluster of data points around the X-Y axis and the number of points on, or close to, a 45-degree angle line. In fact, for 308 of the 500 companies (62%) the Restated Net Income numbers were the same as the *as-first-reported* values. One-in-three companies (163 of them) showed lower Restated than *as–first-reported* Net Income numbers. The remaining 5% of the companies (28) had higher Restated than *as-first-reported* Net Income values. As can be seen in Figure 14, the extremes are extreme; we've attempted to label a couple of the data points. For instance, General Motors originally reported $ 179.3 million in Net

Income for the first quarter of 1992. The (restated) Net Income number for GM for Q1Y92 taken from the latest cut of the Current Compustat data shows a $21 billion dollar loss. At the other extreme is IBM which originally reported $595 million in first quarter 1992 Net Income. The latest available figure using *Current Compustat* data for IBM's Q1Y92 Net Income is $2.5 billion, four-and-a-quarter times greater than was actually reported. As stated previously, market price is never restated. A comparison of GM's and IBM's P/E ratios for March 1992 yield wildly different conclusions using *as-first-reported* data than using *Current Compustat* (restated) data.

Figure 14



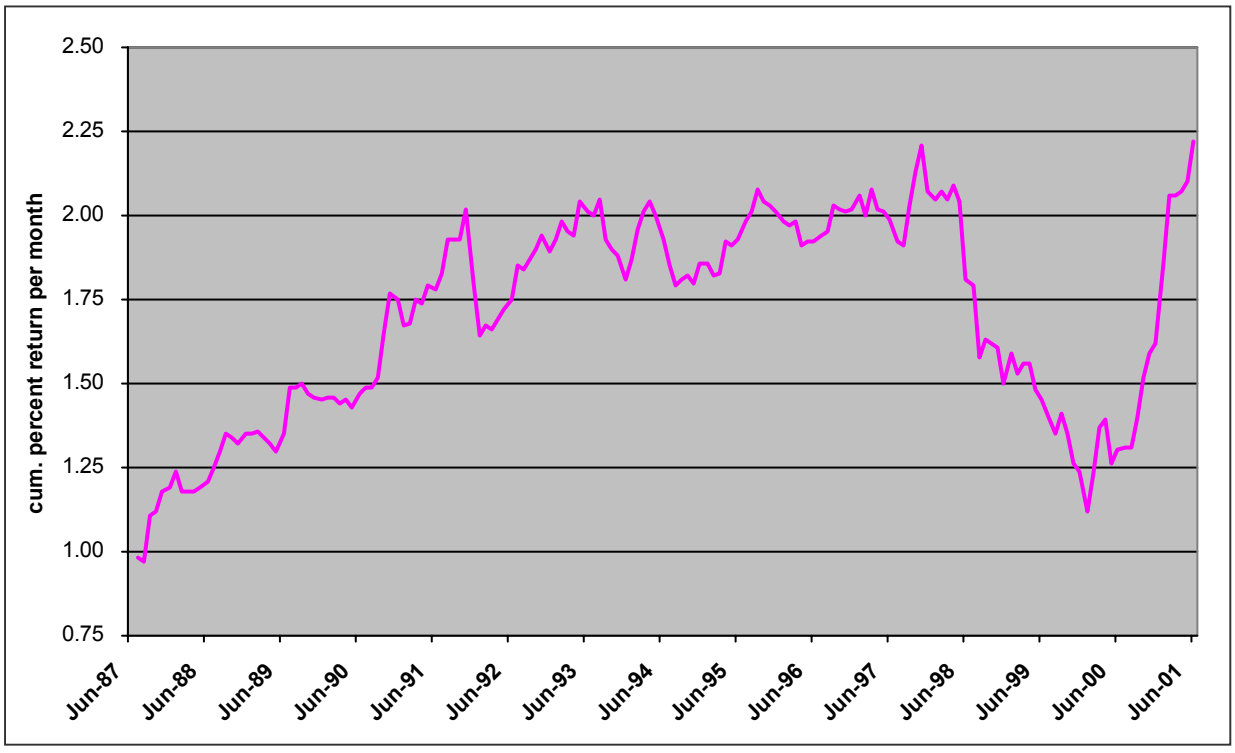
So far we've looked at *as-first-reported* versus restated differences in raw data items. Let's move on to examine factor construction differences using the *Current Compustat* versus *Compustat AFR* datasets.

We calculated an Earnings-to-Price factor ("E-to-P") for each company in the S&P 500 by month from June 1987 forward using four-quarter trailing fully diluted earnings per

share excluding extraordinary items with an assumed reporting lag of two months. We then formed two sets: the top 50 and the bottom 50. We subtracted the returns for the two portfolios. Then we calculated the product of 1+ the difference in returns to get a cumulative wealth index over time. Figure 15 shows the results of these calculations using *Current Compustat* (restated) data. So what we've shown is that Value (High E/P) did dramatically better than Growth (Low E/P) between 1987 and late 1991. For the next six years Value did somewhat better than Growth. From November 1991 until January 2000, Growth dramatically out performed Value. Since January 2000 Value has quite precipitously outperformed Growth. (We doubt at this point there will be gasps of surprise from anyone.)

Figure 15

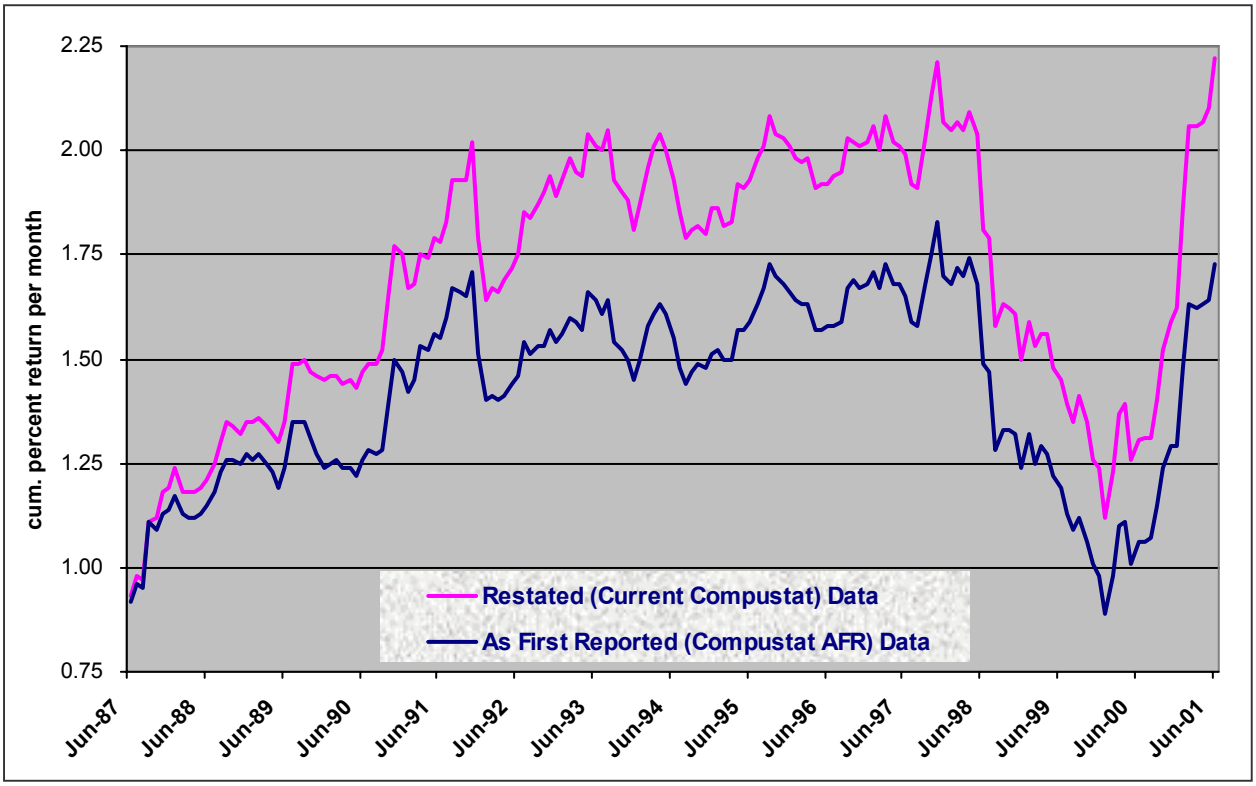
***Current Compustat* (restated) Data**
Cumulative Return of the
Top 50 "Value" (Low E/P) *less* Top 50 "Growth" (High E/P) Stocks
in the S&P 500 by Month – June 1987 to June 2001



We then reran the identical analysis using *as-first-reported* data values from the *Compustat AFR* dataset. Figure 16 shows the results from both restated data and *as-first-reported* data analyses.

Figure 16

***Current Compustat* (restated) Data vs. *Compustat AFR* Data**
Cumulative Return of the
Top 50 "Value" (Low E/P) *less* Top 50 "Growth" (High E/P) Stocks
in the S&P 500 by Month – June 1987 to June 2001



The graphs clearly measure the same phenomenon. But given the difference in the curves (particularly in the 1987-90 period and over the last 14 months), each dataset is telling a slightly different story. Furthermore, using the restated data dramatically overstates the magnitude of the of the E/P cumulative differences effect.

Over the last 7 years, the maximum amount of overstatement between the two datasets analyses was 28% found in June 2001. The minimum amount of overstatement was 20% found in March 1997.

A look at the underlying data of each analysis reveals both differences in the rank order and in the composition of the Growth versus Value portfolios using the restated versus the *as-first-reported* datasets.

On a month-by-month basis comparing the composition of the Value and the Growth portfolios shows substantial change from one dataset to the other in the rank order of the companies in the portfolios as shown in Figure 17. Since we were using the Low and High E/P 50's as portfolios, changes in rank order obviously don't affect our analysis of

differences in cumulative return between the two datasets. But they certainly could make a difference in your analysis depending on what you were doing and where you drew your x-tile boundaries.

Figure 17

Analysis of Rank Order Changes
Low and High E/P Portfolios
Restated (*Current Compustat*) Data vs. As-First-Reported (*Compustat AFR*) Data
Change: # of Companies / Month

	Maximum # - month	Minimum # - month	Median Jun 87 – June-00
No Change in Rank Order	56 - Apr-01	3 - May-89	20
Change in Rank Order >5	32 - Jan-91	6 - Sep-98	11

As (if not more) important than the changes in the rank order of the High and Low portfolios' members are the changes in the composition of the portfolios using restated (*Current Compustat*) as opposed to *as-first-reported* (*Compustat AFR*) data. It is these changes in portfolio composition that account for the differences in returns in Figure 16. An analysis of the members in each of the two sets of portfolios constructed using the two datasets reveals that it is differences in a relatively small number of companies that account for the large difference between the two datasets' returns.

There was only one month since June 1987 in which the High E/P 50 and Low E/P 50 companies were identical when constructed using *Current Compustat* as opposed to *Compustat AFR* data. That month was April 2001. The median number of companies missing from each dataset's 100 combined companies' portfolios was 8 per month in the 14-year period June 1987 to June 2001. The maximum number of company differences in one dataset as compared to the other for any single month was 17 in August 1989. And here, in Figure 18, are those differences. Note that two companies, Eastman Kodak and Interco, switch from being in the High E/P 50 portfolio (using *AFR* data) to being in the Low E/P portfolio (using *Current Compustat* data). (That's fodder for an interesting "oops" if one's intent was to go long on one portfolio and short the other.)

Figure 18

Low and High E/P Portfolios – Missing Companies
April 1989

	In *Restated*/ Not in *AFR*	In *AFR*/ Not in *Restated*
High E/P Portfolio	1. CNA Financial Corp. 2. Goodrich (B.F.) Co. 3. NACCO Industries 4. SAFECO Corp. 5. Security Pacific 6. Temple-Inland Inc.	**1. Eastman Kodak** 2. General Dynamics 3. Holding Corp. **4. Interco Inc.** 5. Mead Corp. 6. NL Industries
Low E/P Portfolio	7. Burlington Northern Inc. 8. CSX Corp. **9. Eastman Kodak** **10. Interco Inc.** 11. Trinova Inc. 12. M/A-Com Inc. 13. Travelers Corp. 14. P-HM Corp. 15. USF&G 16. Unilivar NV 17. Warner Communications	7. Advanced Micro Devices 8. Detroit Edison Co. 9. Enserch Corp. 10. Genentech Inc. 11. General Signal Corp. 12. Lin Broadcasting 13. Monarch Machine Tool Co. 14. Service Corp. Int'l. 15. Williams Co's Inc. 16. Zayre Corp 17. Computervision Corp.

So what does this really say? Unless you are careful to use the right dataset, you risk misaligning an historical portfolio construction task with a now-forward forecasting dataset and end up buying the wrong stocks.

The Two Looks of Compustat: Structure and Use Strategy

The *Compustat AFR* dataset was built to be used as a companion tool to *Current Compustat* data. For three good reasons we can think of (and any number additional we're sure you could think of) it makes sense to use the two datasets in tandem.

Reason #1: The *Compustat AFR* data is constructed and presented as a superset of *Current Compustat*. *Compustat AFR* data is reported for a company for a time period if, and only if, the *as-first-reported* value for the company for the time period is different from the *Current Compustat* (restated) value. As a result, the *Compustat AFR* dataset does not, by itself, contain Compustat quarterly data for all Compustat companies. If there have been no data restatements, then *as-first-reported* and *restated* data are identical and the only data values for the company for that time period are found in the *Current Compustat* data. In addition, as noted previously, the data for the most recent reporting period for a company cannot (by definition) be re-stated; it's just now been "stated" for the first time. So *Compustat AFR* data, by definition, can not contain the most current period's data for a company. You will need to find those most recent quarterly data values in the *Current Compustat* dataset.

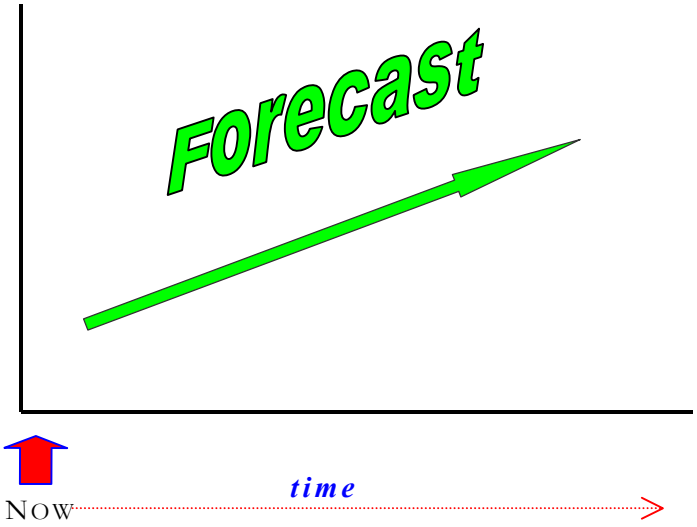
Reason #2: Sooner or later, most of you will find yourselves (like most Quants we've run across) both *forecasting* (needing *Current Compustat* data) and *backcasting* (needing *Compustat AFR* data). To use one but not the other dataset, risks finding yourself in the situation of having Mutt but missing Jeff, having *day* needing *night*, finding Ying but wanting Yang, …… you get the idea.

Reason #3: That's the way the data was developed (with *Compustat AFR* as an add-on, superset requiring *Current Compustat*) and that's the way S&P sells it. *Current Compustat* is the base, required dataset; *Compustat AFR* is an add-on.

Conclusion

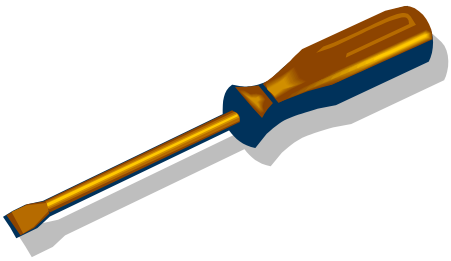
#1: This a hammer.

This is a *now-forward* forecast ……



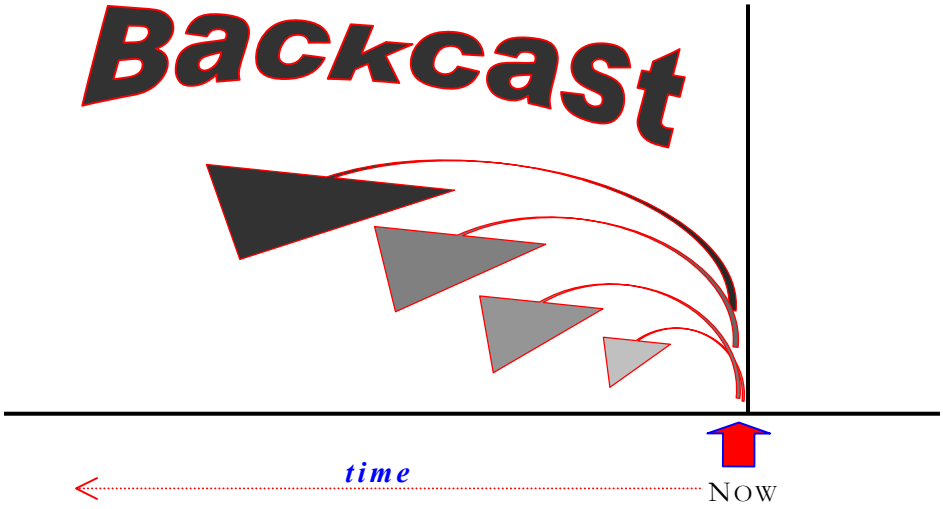
…… which should be based on ……

Current (restated) Compustat Data



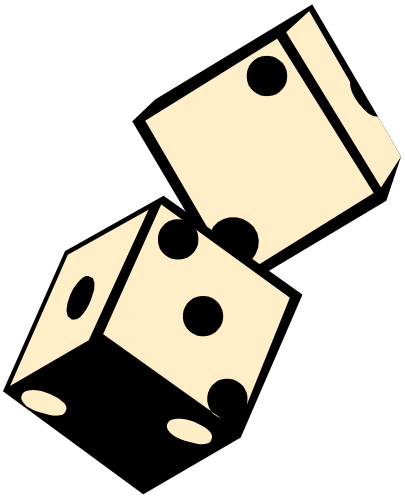
#2: This a screwdriver. It's not a hammer. It's a different tool.

This is a *from-now-back* backcast. It's different too...



.....and it should be based on .....

Compustat AFR (as-first-reported) Data



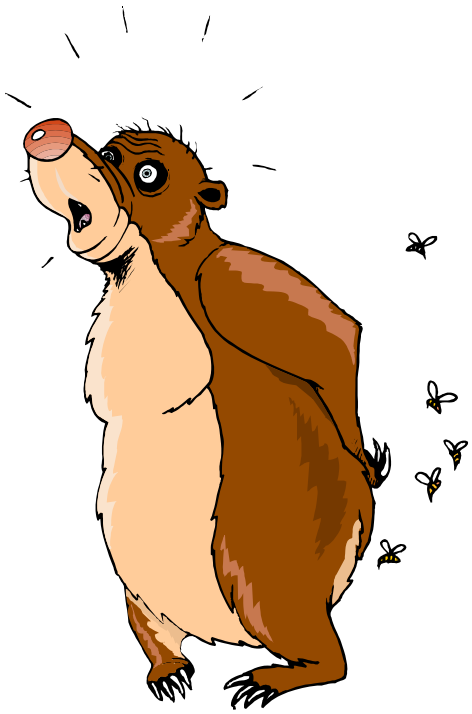
#3: These represent the potential risks and rewards from quantitative equities research.

The purpose of this paper has been to provide guidance in aligning the appropriate tools with intended tasks (i.e., #1 and #2)

…so that when tumbling the numbers (in #3), ……

…. one maximizes the gain….

…and minimizes the pain….



…. from performing historical quantitative research.