TICK DATA

Survivorship Bias in the Development of Equity Trading Systems

A review of the Survivorship Bias inherent in current practices of defining research universes for testing equity trading systems

By Thomas Neal Falkenberry, CFA
SVP Tick Data Division, Nexa Technologies
10134-G Colvin Run Rd.
Great Falls, VA 22066
Tel: 703.757.1370
tnf@tickdata.com
Type of software used: return calculations provide courtesy of Linda Lang at Autumn Wind Asset Management using Bloomberg's TRA function.

Equity traders take great efforts to ensure that the results from their system backtesting generalize well in the future on unseen data. These efforts usually focus on controlling the curve-fitting nature of trading system development. They include walk-forward testing, the use of in-sample and out-of-sample security sets, and the use of in-sample and out-of-sample trading environments to name a few. In the end, the trader hopes to avoid the bane of trading system development: the failure of real-time results to live up to backtested results.

While techniques to control bias such as those mentioned above are widely practiced, few traders pay attention to another source of bias that slips subtlety into the process but has a dramatic impact, survivorship bias. Survivorship bias enters the system development process early in response to the seemingly harmless question, “across which securities should I test my trading concepts?” This question is typically given very little thought. Often, it is resolved by acknowledging that an appropriate universe for testing is one that is defined by liquidity requirements. “I am a big hedge fund, each of my positions averages x so I can only trade

securities with liquidity or market cap of y. Therefore, my research universe will consist of all stocks currently meeting these liquidity constraints." The trader obtains historical price data, and perhaps supporting fundamental data, for these stocks and proceeds to test his or her trading strategy. Once refined, the models are then executed in real-time across the same static symbol set. After some period of time the trader decides to reoptimize or modify the rules of the trading strategy. More often than not, the trader goes back to the same universe, perhaps updated for current prices, and repeats the process.

After many years of working with individual and institutional traders, we find the practice described above to be dominant. Traders define research universes based on the liquidity measures of actively traded securities. The measures most often used are market capitalization (or membership in a market index where membership itself is based on market capitalization) or some measure of money flow (i.e. price x average volume). The most common practice is to define a research universe as constituents of a market index, the S&P 500 Index® or the Russell 1000 Index ®, for example.

This paper will outline the survivorship bias inherent in conducting trading system research on a universe of securities comprised of members of an index, or ranking system, where membership is based on capitalization of other proxy measures of liquidity. We will define a sample research universe and attempt to measure a portion of the survivor bias inherent in that universe. We believe five years is a representative time window across which traders are testing their strategies. As such, we will attempt to measure the bias across the five year period 12/31/98 - 12/31/03.

The Problem

Universes where membership is based upon capitalization, such at the Russell and S&P indices, reward (include) companies whose stock prices have been outperforming, i.e. rising in relative ranking based on capitalization, and punish (remove) companies whose stock prices have fallen such that their market capitalization not longer qualifies for inclusion in the index. For example, the 1050th company in market cap experiences relative outperformance versus existing members of the index and its market cap increases in rank to 990th. That stock then becomes a member of the index on the next index rebalancing date. In the meantime, an underperforming company that was a member of the index is crowded out as its market cap now falls below the 1000th largest. The outperforming stock is in and the underperforming stock is out. A trader that defines his/her universe on the day following such a theoretical event will test his/her trading strategy only on the outperforming company and will never see the impact of the underperforming stock on the strategy's results. This is survivorship bias.

However, it gets worse. Real-time practice begins to disconnect from simulation almost immediately. The next stock that rises up the ranks of capitalization to merit inclusion in the index will not be added to the universe. Again, I am assuming the universe, once defined, remains static. The underperforming company that was just crowded out of the index is not removed from the universe. As a result, in real time the trader <u>is not</u> trading the outperforming company, <u>is</u> trading the underperforming company, and both are in direct opposition to what was done in simulation.

There are two means by which survivor bias enters our research universe:

- The bias of excluding companies from a research universe that met membership criteria historically but did not on the day the universe was defined. (Type 1 Bias). For example,

Enron, Worldcom, Global Crossing, and endless dot com blowups maintained substantial weights in the Russell 1000 during the first half of our five year test period. However, by virtue of defining the universe as of 12/31/03, these companies, and their negative downside performance, has been excluded from our universe since these companies dropped out of the index. Here, by excluding "drop outs" the trader is unknowingly, and unrealistically, assuming he can filter out future "drop outs" from the universe in real-time, and hence, can shield his trading systems from such events.

Conversely, companies such as GTE, Warner-Lambert, Texaco, BestFoods, Bankers Trust, Biogen, and American Stores were also index members during the majority of our five year test period. However, these companies are excluded from our research universe because they were acquired and did not trade as of the date we defined our universe. These companies represent a source of upside return to our universe as they were all acquired at handsome premiums. Bias works in both directions.

- The bias of including companies in a research universe that did not meet membership criteria until recently (Type 2 Bias). Companies such as Michaels Stores, Carmax, ImClone Systems, Pulte Homes, Sandisk, BEA Systems, and eBay are stocks that demonstrated terrific performance in the periods prior to their inclusion in the index. A trader that defined a research universe several years ago would not have included these companies, as they had not yet met the criteria for membership. A trader that defines the universe today will include these companies, but will invariably include them further back in time than the period in which they actually met membership criteria. The error of **"inclusion prior to qualification"** adds the performance of companies who later met membership requirements to simulation where that performance, by definition, is one of relative strength since it is what led to index membership.

For example, Carmax, Inc. (KMX) was added to the Russell 1000 Index® in December, 2002. The stock was added because its relative performance was strong and it rose up the ranks of US companies by capitalization from 3,734th in 2000 to 590th in 2002. The trader that includes KMX in today's universe and commits the error of "**inclusion prior to qualification"** gets the benefit in simulation of prior year's returns of 477% in 2001 and 66% in 2000. Yet, this is not what would have been done in real-time trading during 2000 and 2001. No such error is possible in real-time. A trader cannot add in real-time the stock today that does not meet universe criteria but will next year due to its forthcoming strong relative performance. This small, seemingly insignificant, procedural oversight introduces a bias into the returns available in simulation versus those than can be expected in real-time. The magnitude of the bias is much larger than is commonly perceived. Consider a case whereby a trader's criteria for inclusion in a research universe produces the results labeled "Table 1." on page 12.

When companies were screened in 2001 stocks A, B, and C were the only companies that met the criteria and therefore represented the trader's universe for both historical research and future real-time trade execution. If the screen is rerun in 2003 using the same criteria stocks A, C, and D are the only stocks that qualify.

Assume stock B goes bankrupt in 2002, i.e. Enron. In 2003, stock B is excluded from the universe that the trader will use in simulation. Yet, the same criteria applied in 2001 would have included this stock in both simulation and in real-time trading. This is Type 1 bias. The 2003 trader's simulation has been sheltered from an event that the 2001 trader would have had to address (with real money on the line). And what if Stock C is 2004's Enron. Then the 2003 trader has a real-time problem that will not be faced by the 2005 trader who re-screens companies only to find that Stock C failed to satisfy the criteria and is excluded from simulation.

Now, assume stock D is Carmax. In 2004, the trader includes Carmax in his backtesting even though this stock would not have been eligible for real-time trading at the time the trader is allowing it in simulation. This is Type 2 bias. The simulation gets the opportunity to capture large returns that were not available in real-time.

This paper will attempt to estimate the magnitude of survivor bias from this later source, Type 2 bias resulting from including companies in a research universe that did not meet the criteria for inclusion across the full period being tested in simulation.

The Test

For purposes of this paper, we will define our universe as all current members of the Russell 1000 Index. We identified all companies in the index as of December 31, 1998 and December 31, 2003. We then compared membership lists to determine the companies that were:

- common to the index on both periods (Group A),
- added to the index after December, 1998 but who were already publicly traded as of December, 1998 (Group B),
- added to the index but whose IPO occurred after December, 1998 (Group C).

It is necessary to differentiate between companies added to the index that were already public (Group B) versus those whose IPO occurred after 12/31/98 (Group C) as our measurement of survivor bias will require that we measure total returns across the common time period 12/31/98 – 12/3103. Group C companies experienced IPOs on different dates, and hence, it is not possible to directly compare their returns to Group A. The breakdown of the Russell 1000 into these three groups is depicted on page 12 as “Illustration 1.”

To estimate the magnitude of Type 2 survivor bias, we will compare the cumulative five year returns of those stocks that were common to the index on both December, 1998 and December, 2003 (Group A) with those companies that were added to the index after December, 1998 but who were already publicly traded (Group B). Our hypothesis is that Group B stocks will exhibit stronger performance than Group A stocks. If this is true, and the trader executes trades in real-time against the same universe used in simulation, then the return differential between Groups A and B demonstrates a bias in simulated results over what can be expected in real time because of the error of "**inclusion prior to qualification".** In other words:

- Group B stocks demonstrate strong relative performance prior to their inclusion in the index. ***(this is what moves them into the index to begin with)***
- The trader includes them in simulation but includes them, and their upward biased returns, for the full period under study rather than just the period from which they were actually added to the index. (***inclusion prior to qualification*****)**
- The trader takes no such efforts to revise the universe against which trades are being executed in real-time for new Group B stocks ***(real-time practice diverges from simulation)***
- Real-time results that do not measure up to simulated results.

To test this hypothesis, we calculated total returns for all 656 companies in Group A and all 242 companies in Group B from the close on December 31, 1998 until the close on December 31, 2003. We assumed dividends were reinvested into additional shares at the opening of the day following the ex-dividend date. Transaction costs were assumed to equal zero. Consider the statistics on page 12 labeled "Chart 1."

The immediate impression is striking. The median 5 year cumulative return for a stock added to the index was 136.35% versus 36.98% for a company already in the index. Likewise, the median return for Group B is 4 times as large as Group A, 220.40% versus 53.35%. Standard deviation is

notably higher, but the volatility is "good" volatility, i.e. a large number of stocks producing staggering returns of 700%-1100% with very few stocks producing large negative returns. A histogram of the returns of the two groups can be found on page 13 labeled "Group A Cumulative Returns" and "Group B Cumulative Returns."

An additional observation is the large difference between the mean five year cumulative returns of Group A and Group B, 53.35% and 220.49%, respectively, and the actual Russell 1000 Index return for the same period of –1.88%. Where did the negative return come from? Collectively, Groups A and B represent 898 of the 991 companies in the index as of December 31, 2003. Together, their five year average return is 98.39%, almost exactly 100 percentage points higher than the actual index. The answer is two fold:

- The index is capitalization weighted and large caps significantly underperformed small caps during the period.
- More importantly, the large negative returns of index "drop outs" (Enron, WorldCom, Global Crossing, dot coms) are included in the actual index returns but are not included in Group A or Group B returns as these companies were not in the index on the date we created the universe. This is Type 1 bias as defined on page 2 of this paper.[1]

Research results, particularly those from long-only systems, that compare returns generated from our sample universe against broad indices is very likely to lead to a false sense of accomplishment by the trading system developer. A more reasonable benchmark against which to compare universe results would be the average or median return of the universe itself. A strategy return of 40% looks good against the actual index's return of –1.88% but looks dismal compared to the biased universe's average return of 98%.

We then conducted a two-tailed test to determine if the mean returns between the two groups were statistically significant. We used a significance level of 0.01. Our null hypothesis is as follows:

- Group B's mean return – Group A's mean return = 0. There is no difference between the mean returns of companies added to the index versus those that were already in the index.

The calculated z statistic is 7.12 versus a critical value of 2.58. Therefore, there is sufficient evidence (at 0.01) to support the claim that Group B's mean return is significantly higher than Group A's. **Stocks that were added to the index exhibited greater returns over the full test period than stocks that were already in the index.** Since the trader is using a static universe future Group B stocks will not be added to the trader's opportunity set. Yet, these stocks, and their upwardly biased returns, were used in simulation. As a result, simulated results are upwardly biased compared to that which is available in real-time.

An incorrect conclusion from this test would be to assume that the reason Group B stocks exhibited greater performance is because they were added to the index. Group B stocks outperformed Group A because the test period encompassed a period of strong small cap performance relative to large cap. Capitalization indices always "pull from the bottom". If GE and Microsoft merge, the 1001st company by market cap gets pulled into the index. Group B stocks are the stocks that were pulled into the index. By definition, they are smaller cap companies.

The cause for Group B outperformance, however, is irrelevant for purposes of this paper. Whatever the cause, small cap strength versus large cap, growth versus value, companies that benefit from a falling dollar versus those that do not, cyclical versus non-cyclical, the fact remains that the decision to utilize a static universe and include companies in simulation during

periods that they were not eligible for trading per the criteria used to establish membership in the trader's universe will favorably bias backtesting results. This bias can be expected to be a greater problem for long-only traders than short-only or long-short traders.

What To Do

In an earlier paper on filtering high frequency data I argued that it is necessary to filter real time data the same as historical data[2]. The same applies to building a universe across which to test trading strategies. Maintain common practices in real-time as are used in simulation. This starts by removing the static properties of the universe.

- Identify the criteria for inclusion in your research universe. To be clear, this paper has not argued against using index membership as valid criteria for inclusion. Rather, it has argued that using a static capitalization-based universe introduces upside bias to simulated returns by including newly added companies in the research universe while no such effort is made in real-time.

- Identify the frequency at which universe membership can change using your criteria. If you choose Russell 1000 Index® membership as your criteria then your universe is subject to annual rebalancing each July as that is the date the index is rebalanced. If your criteria are the stocks with the greatest money flow as measured over the most recent 10 days, then the frequency at which your universe's membership can change is daily.

- Recreate "as of" universe membership lists on each date membership changes. In our test five year period, we would compile five "as-of" membership lists. Consolidate all membership lists into a Comprehensive Research Universe.

- Build a database that identifies the date range for which each security met membership criteria and is eligible to be traded in simulation. As noted, Carmax became an index member in 2002.

Its extraordinary returns of 2001 and 2000, that drove it to index membership, are not eligible for backtesting.

- Obtain pricing and related data on all companies in the Comprehensive Research Universe. Obtaining information of dead companies is challenging.

The trader has now created a ***dynamic, comprehensive research universe*** that includes all companies that met his or her criteria within the time frame that they met that criteria. The trader has also linked real-time practices to simulated practices. Unfortunately, many hours of work has also been added to the process. However, based on the evidence of this paper, any research using static capitalization-based universes may, in the end, be far more painful.

Graphics

Table 1.

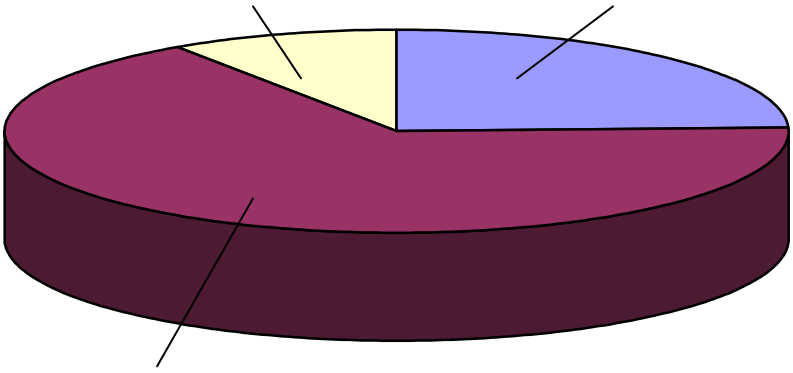
Universe'01	Universe '03	
Stock A	Stock A	
Stock B		← Type 1 Bias
Stock C	Stock C	
	Stock D	← Type 2 Bias

Illustration 1.

Russell 1000 Membership as of December 31, 2003

Group C – 93 companies added to the index whose IPOs were post 12/31/98

Group B – 242 companies added to the index that were public as of 12/31/98.



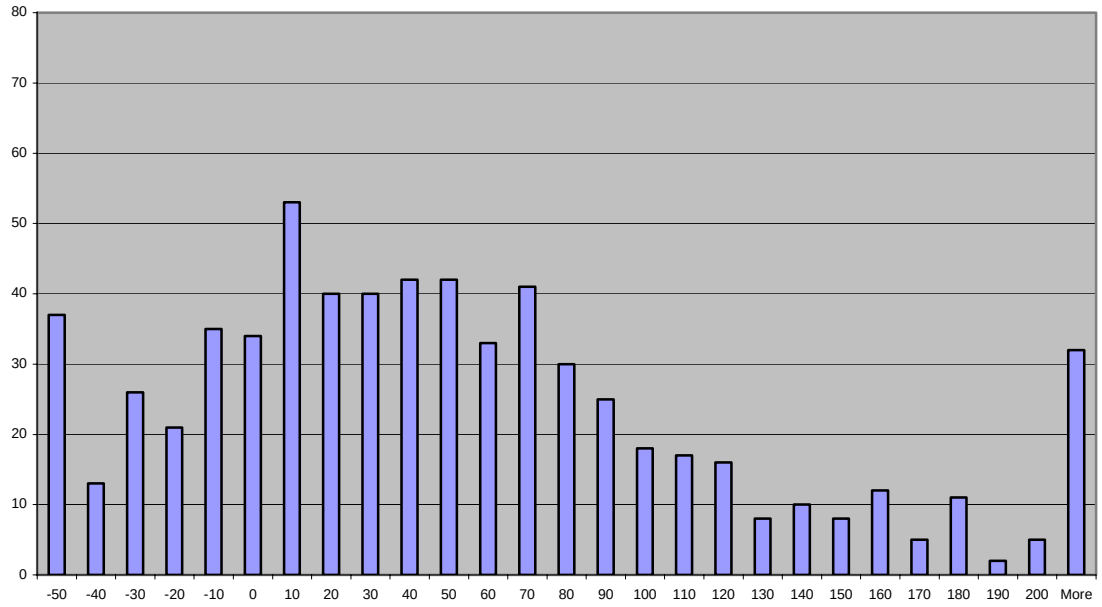
Group A – 656 companies common to the Index on both dates.

Chart 1.

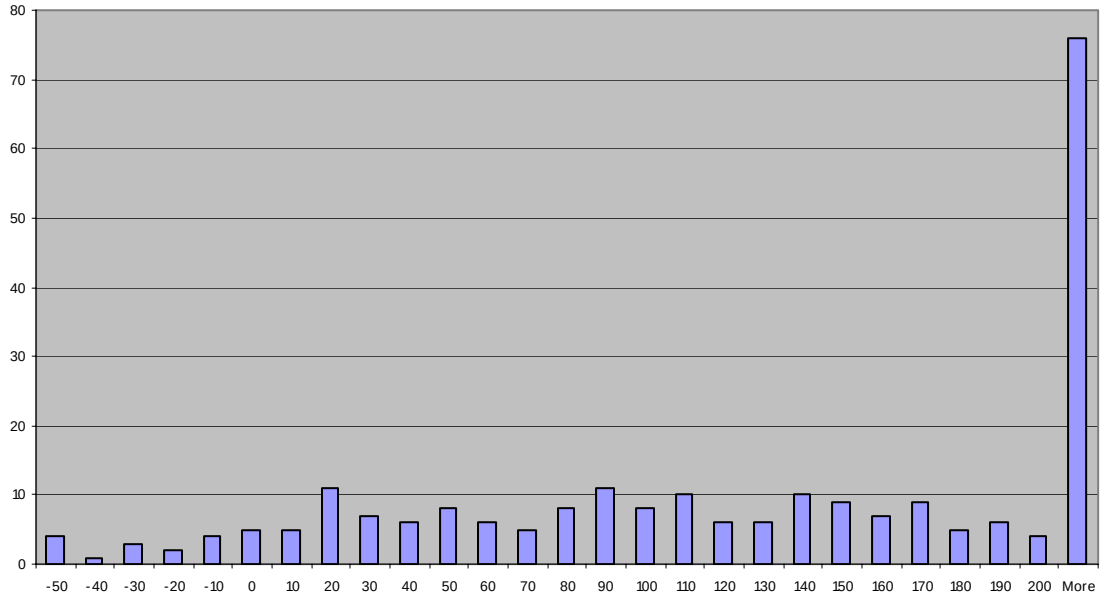
The statistics:

	Group A	**Group B**
# Observations	656	242
Mean Cumulative Return	53.35	220.40
Median Cumulative Return	**36.98** ←	**136.35** ←
Standard Deviation	101.38	359.82

Group A Cumulative Returns.



Group B Cumulative Returns.



End Notes

[1] Interestingly, the survivorship bias introduced by Type I Bias (eliminating "drop-outs") may help explain the relative difficulty many traders experience in developing short-only trading strategies. If drop outs are eliminated there are fewer big winners in simulation that may exist in real-time. In other words, where the omission of drop outs overstates simulated results versus real-time for the long-only trader, the omission may actually understate simulated results for the short-only trader.

[2] Falkenberry [2003] "Filtering High Frequency Data", Tick Data, Inc.