Avoiding data mining in forex systems

How do you know if your trading strategy is capitalizing on a genuine market dynamic or just lucking out?

BY CURRENCY TRADER STAFF

One of the biggest concerns when building trading systems is the reliability of trading signals. Given enough time, it's possible to identify many spurious relationships in a financial time series (a problem known as data mining) that can lead to systems that fail under new market conditions.

This happens because sometimes there is no causal relationship between signals and the subsequent price action, and as a result the strategy's profitability in testing — even over long periods — is merely the result of random chance. In other words, sometimes a system can just get lucky in testing, even for a relatively long time. As a result, it's imperative to be able to determine whether your trading logic has a genuine causal relationship with the currency pair's you're trading, is merely manifesting a spurious correlation

The good news is that there are analysis techniques that can help us judge whether a given trading system variable is causally correlated with the profit it generates.

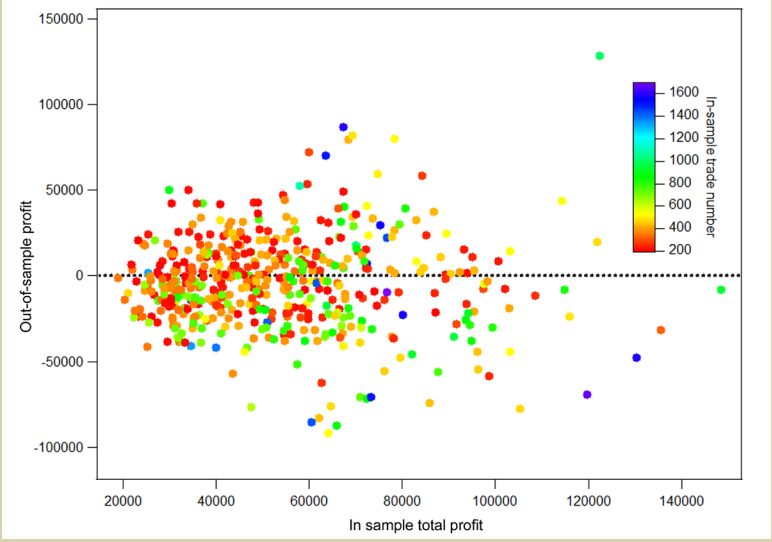
Pure data mining

To judge if a certain variable — such as the closing price — can lead to the creation of systems that can succeed under unknown market conditions (i.e., future data), it's important to first understand what happens when we generate trading systems using a variable that is based on a spurious correlation to the financial time series.

We'll do this by generating trading strategies using a random variable (0 to 1) in the Euro/U.S. dollar (EUR/USD) pair from Jan. 1, 1991 to Jan. 1, 2001 (the in-sample data period), and then analyze the characteristics of these results when applied to data from Jan. 1, 2001 to Aug. 8, 2012 (the out-of-sample period). Keep in mind this variable is completely random, so any "relationship" we think we find is actually meaningless. (All tests used Kantu, a parameter-less price-pattern creation engine discussed in previous articles.) For all tests, 500 systems were created, each of which generated at least 20 trades per year and a correlation coefficient (R^2) greater than 0.9.

Figure 1 shows the relationship between in-sample and out-of-sample profits for the strategies generated using a random variable. It's obvious that despite the complete lack of causality in the systems, there are nonetheless a few that — by random chance — achieve significant levels of profitability in the unknown market conditions represented by the out-of-sample data

FIGURE 1: RANDOM PROFITABILITY



Even some systems with signals generated by a random variable were able to produce profitable results in both in-sample and out-of-sample periods.

period. This demonstrates why "validation" through a successful out-of-sample test does not negate the possibility of spurious results based on data mining. Simply, the possibility always exists that good results will be produced in the out-of-sample because of data mining.

Figure 2 shows the equity curve of one of these random-variable systems. It was tested on 10 years of in-sample data and then validated on 11 years of out-of-sample data — but we know with absolute certainty the correlation is absolutely spurious and will lead to losses going forward. However, neither trading frequency nor any other system performance statistic can be used to identify cases of spurious correlation, because outliers that match these criteria will always exist.

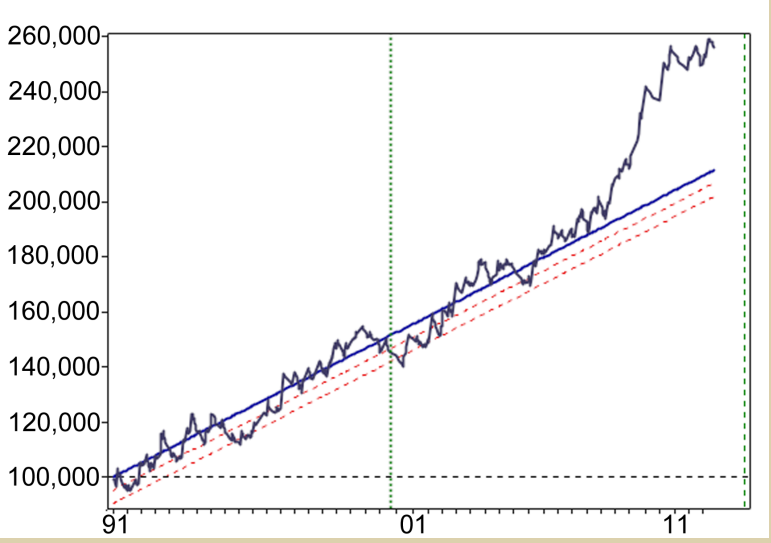
The good news is that these results reveal some key characteristics of systems generated from variables with no causal connection to the underlying data. First, the probability of success in out-of-sample conditions is less than 50% (in the preceding example it was 47%). This is a function of the bid-ask spread, which causes a slight negative bias. Second, the average of out-of-sample result is negative. Third, there is no preceding example the correlation coefficient between the in-sample profit and out-of-sample profit was -0.025, which indicates no correlation. Finally, total trades from the in-sample period are negatively correlated with out-of-sample profitability because the negative expectation generated by the spread causes overall results to become increasingly negative as the number of trades increases.

Studying normal trading variables

Because this information lets us know exactly what systems generated by spurious correlations look like, we can now evaluate whether systems generated with real, non-random trading variables are likely to be either spuriously or causally correlated with our underlying time series.

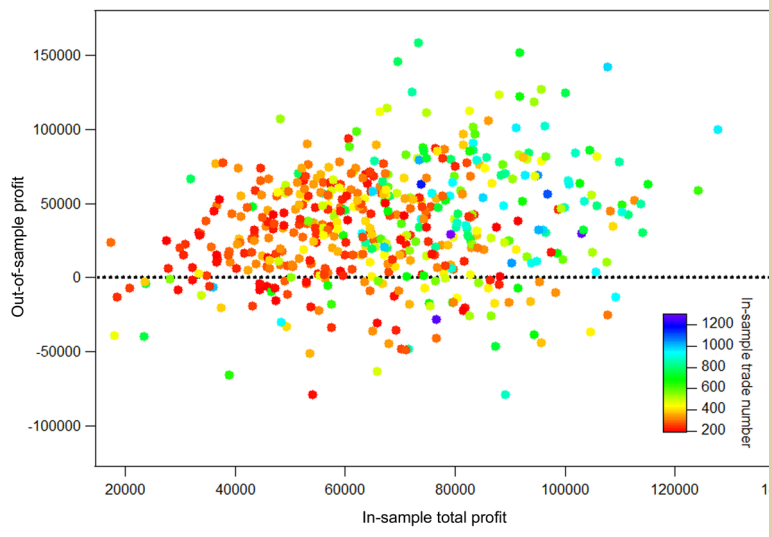
Figure 3 shows the in-sample and out-of-sample profit relationship of 500 systems generated using opening price relationships in the

FIGURE 2: SPURIOUS EQUITY CURVE



Although this system's equity curve looks great (and spans in-sample and out-of-sample data), the correlation is absolutely spurious and will lead to losses in future trading.

FIGURE 3: OPENING PRICE SIGNALS



The systems in this test, which used the EUR/USD opening price instead of a random variable, produced much stronger relationships between the in-sample and out-of-sample periods.

EUR/USD. In this case, the probability of being profitable in the out-of-sample period increased to 83% (from 47% in the initial experiment), while the correlation between the in-sample and out-of sample profits increased from -0.025 to 0.21.

Also note the correlation between the total number of trades and the out-of-sample profit is now positive (0.18), implying we now have an expectation to be profitable as we trade more, which previously was not the case. The average of the out-of-sample results is now also positive, while it was negative in the initial experiment.

Certainly these numbers could also be achieved by a random variable — imagine we were just extremely lucky with our data-mining — but the overall probability of this happening is less than one in a billion. As a result, we can state that price patterns generated using the opening price have a high probability of having an underlying causal relationship with the EUR/USD price series, and therefore are likely to offer some predictive edge in the future.

Table 1 compares the statistics of using a random variable to generate trading systems vs. using the opening price, the closing price, or both. There is a very important difference between the results of the systems created using the random variable and those created using the close or open. This doesn't mean data mining doesn't happen with systems generated using our time series — otherwise the in-sample/out-of-sample correlations would be much higher — but it does mean the overall probability of this happening is lower (i.e., there is a predictive edge within our overall results), while there is no edge in systems created using a random variable.

Identifying causality in your trading strategies

This type of experiment allows you to evaluate whether a certain type of trading system – using a given level of complexity and a given set of variables — is giving you profitable results because of a real relationship with your price series or a spurious relationship that has a high chance of resulting in losses in the future.

It's clear that conducting simple out-of-sample tests is not enough: You need to establish whether several different relationships of the variables (generated systems) lead to better results in the out-of-sample period than a variable that leads purely to spurious correlations. You can change the type of system (complexity) and the types of variables used to create a strategy (indicators, fundamental data, price patterns, etc.) to determine the type of combination that generates the highest causality within a large pool of trading systems.

The important thing is to base your conclusions on a series of systems generated with the same overall idea rather than the out-of-sample results of a single strategy — which, as previously shown, can easily lead to data mining.

Daniel Fernandez is an active trader focusing on forex strategy analysis, particularly algorithmic trading and the mathematical evaluation of long-term system profitability. *For more information on the author, see p. 4.*

TABLE 1: RANDOM VS. REAL

Signals from	OS profit	Avg. OS profit	IS/OS profit (R)	IS total trades / OS profit (R)
Random variable	47%	$-3,925	-0.025	-0.14
Open	83%	$34,745	0.21	0.18
Close	84%	$37,003	0.13	0.13
Open and close	86%	$32,623	0.16	0.15

Trading systems generated using the opening and closing prices displayed much different characteristics than those generated using a random variable.