Know Your System! – Turning Data Mining from Bias to Benefit through System Parameter Permutation

February 28, 2014
Submitted for Review to the
National Association of Active Investment Managers (NAAIM)
Wagner Award 2014

By
Dave Walton
StatisTrade
dwalton@StatisTrade.com

Abstract

This paper is targeted towards active traders who follow a systematic approach to alpha generation and wish to thoroughly understand the potential risks and rewards expected from a trading system prior to allocating capital. The conclusions are applicable to all timeframes and parameter-based systems though the example specifically discussed herein is an end-of-day trading system. The goal of this paper is to assist the trader in answering two questions: 1) "What is a reasonable performance estimate of the long-run edge of the trading system?" and, 2) "What worst-case contingencies must be tolerated in short-run performance in order to achieve the long-run expectation?" With this

Electronic copy available at: http://ssrn.com/abstract=2423187

information, the trader can make probabilistic, data-driven decisions on whether to allocate capital to the system and once actively trading, whether the system is "broken" and should cease trading.

After explaining assumptions, definitions, and methods, the paper discusses how and why traditional trading system development processes lead to positively biased performance estimates due to the data mining bias (DMB). Without understanding the substantial effects DMB has on historical simulation results, inappropriate or inaccurate conclusions may be reached. Several existing methods of DMB mitigation are briefly examined.

The main focus of the paper is to outline System Parameter Permutation (SPP) and its application as a simple yet effective performance estimation method as an alternative to traditional cross-validation or more complex DMB compensation. In the simplest of terms, SPP is a method to generate sampling distributions of system performance metrics. The method provides a practical means to estimate the performance of the trading system edge as well as to perform statistical significance testing. SPP leverages the statistical law of regression toward the mean and maximizes the use of historical market data.

In system optimization, regression toward the mean indicates that the specific combination of optimized parameter values which led to extreme performance in historical simulation will probabilistically not retain a level of extreme performance in the future. Rather than examining regression toward the mean over time, SPP leverages the optimization inherent in typical parameter-based system development to generate a

Electronic copy available at: http://ssrn.com/abstract=2423187

sampling distribution of performance expectations where the effects of regression toward the mean may be examined across system variants (combinations of parameter values).

After defining SPP, the paper provides instructions for how to apply the method to any type of parameter-based trading system. Specific instructions are included on how to apply the method in order to estimate probable ranges of long-run and short-run performance as well as to test statistical significance. The paper also explains why SPP is effective and how probabilistic information can be extracted in order to facilitate data-driven capital allocation decisions.

One of the strengths of SPP is that the resulting sampling distributions of system performance metrics enable realistic contingency planning based on probabilities. Unlike random resampling methods such as bootstrap or Monte Carlo permutation (MCP), the random variation in SPP comes from the application of a set of slightly varied entry/exit rules on actual market data where trading signals are evaluated using a realistic simulated portfolio. In effect, SPP explores facets of the trading system that would otherwise remain hidden yet are possible in real trading.

For demonstration purposes, SPP is applied to an example rotational system based on relative momentum. After performing traditional optimization and cross-validation on this system, SPP is used to create long-run and short-run performance estimates which are compared to the traditional approach.

The example shows that compared to standard out-of-sample (OOS) cross-validation, SPP provides the trader with much more information. SPP creates long-run

and short-run sampling distributions of system metrics using all available historical market data whereas traditional OOS cross-validation provides only a point estimate on a subset of historical market data. SPP enables probabilistic decision making whereas traditional OOS necessitates a binary pass/fail decision. Thus SPP enables a much deeper understanding of how the trading system may perform going forward.

SPP applied to the relative momentum trading system also shows that the system outperforms its buy-and-hold benchmark over the long run. However, the SPP short-run, worst-case contingency analysis indicates that in order to achieve long-term outperformance, the trader must be willing to accept the possibility of significant underperformance and negative absolute returns in the short-run. With this information, the capital allocation decision may be made probabilistically.

The paper concludes with key takeaways. Ultimately the strength of SPP is the balance of simplicity, ease of implementation, and realism. The author hopes SPP will help the reader increase his odds of success in trading the markets through a deeper understanding of how his trading system functions and provide a more realistic view of expected future performance.

1. Introduction

Prior to putting capital at risk, every trader desires an accurate estimate of the potential risks and rewards expected from a trading system and often employs historical simulation to gain such an understanding. Unfortunately, many traders are subsequently frustrated by poor realized trading system performance that does not live up to overly optimistic expectations. One large and prevalent source of overly optimistic expectations that remains largely misunderstood and underestimated is the data mining bias (DMB).

Even though DMB tends to have a large impact on historical simulation results, mitigation tools available to the average trader are relatively crude. More advanced tools are available to academics and quantitative professionals but are largely too complex for the average trading system developer. This paper attempts to change that by introducing System Parameter Permutation (SPP). With SPP, the average trader is armed with a simple yet powerful tool to effectively mitigate data mining bias and more accurately estimate future trading system performance.

The power of SPP extends beyond mitigating data mining bias however. SPP explores facets of the trading system due to the interaction of system rules, portfolio effects and market data that other methods do not. Thus SPP enables a much deeper understanding of potential risks and rewards prior to allocating capital to a system.

2. Basic Requirements, Definitions and Methods

The only requirement in order to apply SPP as defined in this paper is that the trading system must be completely rules based and use parameters which are optimized

during the development process. This requirement is necessary because SPP makes use of the parameter optimization process and corresponding data. Thus SPP is most applicable to trading systems based on technical analysis.

The following definitions and methods are used heavily throughout this paper:

Quantitative Trading System: A trading system defined by clear, unambiguous, and comprehensive entry and exit rules which can be machine coded. The results of a quantitative trading system can be independently reproduced and verified. Any mention of the term trading system in this paper implies it is quantitative.

The trading system used in this study is long only, meaning no shorting or inverse ETFs were used. In the context of a historical simulation, long-only systems greatly reduce the number of assumptions made regarding ability to borrow shares, share call-backs, dividends paid, and interest charged. The trading system was developed based on published research and simulated using realistic portfolio simulation which accounts for margin available, prioritization of trading signals, position sizing, and portfolio heat limits.

Trading Universe: A total of ten ETFs were used: SPDR S&P 500 ETF (SPY), iShares Russell 2000 ETF (IWM), iShares MSCI Emerging Markets ETF (EEM), iShares Core S&P Mid-Cap ETF (IJH), PowerShares QQQ ETF (QQQ), SPDR Gold Shares ETF (GLD), iShares MSCI EAFE ETF (EFA), iShares 20+ Year Treasury Bond ETF (TLT), iShares US Real Estate ETF (IYR), iShares 1-3 Year Treasury Bond ETF(SHY). These ETFs were chosen because they are the most liquid ETFs covering major asset classes.

Historical Simulation Timeframe: The trading system historical simulation period for this analysis begins 11/19/2005 and ends 5/31/2013. The start date was selected to allow roughly one year of market data history for all traded ETFs (GLD started trading 11/19/2004) prior to any entry signal.

Input Data: Daily OHLC market data were used from Norgate Premium Data and adjusted for splits. Market data were not adjusted to include dividends in order to avoid non-linear, a posteriori distortion of technical indicator-based trading signals that use percentages (Kaufman (2013)). To be as realistic as possible, historical dividend data were used from Yahoo! Finance[1] and dividend payments were injected into the portfolio as cash per the applicable ex-dividend date.

Transaction Costs and Fees: A $0.01 per share per side allowance was made for commissions as well as a 0.05% estimate per side for slippage. Where applicable, margin interest was charged daily at a rate of 1.5% + the Fed Funds daily rate[2] (varied between 0.04% and 5.41% in the simulation period). Data for the Fed Funds daily rate was taken from the public website of the Federal Reserve Bank of New York[3]. All order types used were Market-On-Open, Market-On-Close, or Market-On-Stop. Thus fees and slippage were modeled towards what a retail trader might expect to see.

Output Data: Four different system metrics were evaluated: 1) compounded annual return including dividends, 2) max drawdown, 3) annualized information ratio[4] (vs.

[1] http://finance.yahoo.com/
[2] From the current (January 2014) schedule of fees for margin interest from Interactive Brokers <$500,000 borrowed.
[3] http://www.newyorkfed.org/
[4] Defined as expected active return (system return – benchmark return) divided by tracking error: $IR = \frac{E[R_S - R_b]}{\sqrt{var[R_S - R_b]}}$.

dividend re-invested SPY ETF), and 4) annualized standard deviation of daily returns. For cross-validation of historical simulations, traditional Out-Of-Sample (OOS) testing was used for comparison to the SPP performance estimation method discussed in this paper. In OOS testing, market data was split into 80% training and 20% validation sets, with the validation set comprising the most recent data.

3. Data Mining Bias

Many traders are familiar with the idea that future trading system performance is likely to be worse than was seen in historical simulation. However, the origins of this performance degradation are often not well understood. One significantly large cause is the DMB, also commonly known by other names such as curve-fitting, over-fitting, data snooping, or over-optimization. DMB is built-into the typical system development process and yet largely remains unknown, misunderstood, and/or ignored.

This may be understandable for retail traders with limited knowledge of statistics. However, Bailey et al. (2013) note that professional publications also tend to disregard or gloss over the effects of DMB. Unfortunately ignoring the problem doesn't eliminate the consequence which is that the trading system fails to live-up to performance expectations in cross-validation or worse in live trading.

3.1 Understanding Data Mining Bias

To understand DMB, one must first recognize its two preconditions which are inherent to the system development process: 1) randomness and, 2) a multiple comparison procedure in the search for the best system rules. The interaction of

randomness and the search process is unique to the system rules evaluated and the historical market data and results in inflated performance metrics.

The first precondition of DMB, randomness, means the random walk component of market data. In any sequence of trades, the result of system rules acting on the random walk component is equally likely to be favorable (good luck) or unfavorable (bad luck). Thus realized trading system performance consists of two components of unknown relative magnitude: the inherent edge and luck. Periods of good and bad luck cause variability around the long-run expected performance due to the system edge.

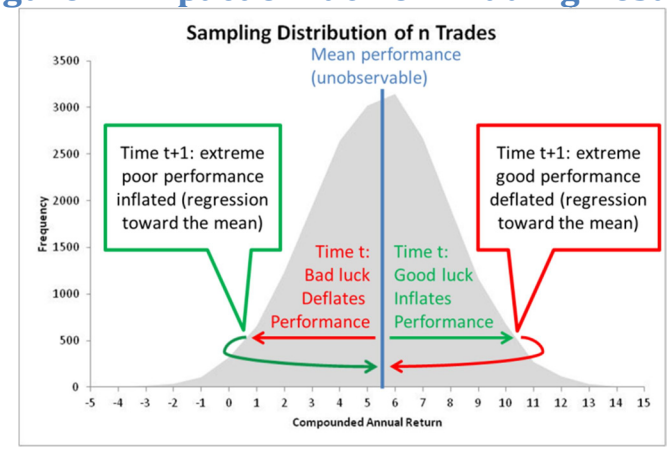
The second precondition of DMB is the multiple comparison and selection process inherent to the typical system development process. At each stage of development, system rules and parameters exhibiting the best performance are selected from historical simulation results. This selection process is known as data mining. Because of the random component in measured performance, the selected rules are guaranteed to have taken advantage of good luck. The probability that a favorable result is due to chance alone increases with the number of combinations tested.

Almost all trading system development platforms support multiple types of search optimization algorithms and thus lead the developer, perhaps unknowingly, into a data mining venture. The process of data mining to find the best performing system rules is not the problem however. Data mining in attempt to find the best (in meeting the objectives of the trader) entry/exit rules and best combination of parameters is a natural,

intuitive process. In fact, Aronson (2007) mentions that data mining is the "preferred method of knowledge acquisition" when employing technical analysis.

The real problem is not considering that the performance of the chosen system rules is inflated by good luck and that the same amount of good luck is not likely to repeat in the future. In fact, the statistical law of regression toward the mean[5] indicates that extreme performance in historical simulation will be probabilistically followed by performance closer to the unknown, long-run level of performance of the inherent edge. This is illustrated in figure 1.

Figure 1: Impact of Luck on Trading Results



3.2 The Consequences of Data Mining Bias

DMB has two consequences: inflated performance metrics and inability to perform statistical inference using standard methods. Both consequences can lead to improper decision making.

A logical question is how large DMB might be. Although the magnitude of the DMB is specific to the analyzed trading rules and market data, it can be quite large. For

[5] "Regression toward the mean" is a statistical law, not to be confused with the financial term "mean reversion" which assumes that observed high and low prices are temporary and that price will tend to move to the average over time.

6402 simple trading rules data mined on the S&P500 index over 25 years of historical data, Aronson (2007) found that the level of annual return needed to overcome DMB was approximately 15% at the significance level of $\alpha = 0.05$ and none of the examined rules had any statistically significant edge.

Further, attempting to test the statistical significance of performance metrics using standard statistical inference procedures is not valid when the data contains systematic error (DMB is systematic error). Sound statistical inference in the context of data mining requires the use of a sampling distribution which includes the effect of good luck.

3.3 Mitigating Data Mining Bias

Data mining bias is systematic; it is inherent to the typical system development process. DMB cannot be lessened or eliminated by evaluating via the "best" system performance metrics or by performing a "perfect" historical simulation (properly modeled transaction costs, clean and accurately adjusted market data, no look-ahead bias, no hindsight bias, no survivorship bias, properly modeled portfolio effects, omissions and contingencies considered, etc.). The only viable methods to estimate performance or test significance in the presence of DMB are those that consider systematic error.

One such method is to estimate performance and perform significance testing on an independent data sample which effectively is looking at system performance after regression toward the mean has occurred; this is known as cross-validation. Another method is to perform significance testing by creating a sampling distribution of maximum means that reflects the role that good luck plays in data mining; this is known as bias

compensation. Yet another method is to calculate a deflation factor for data mining bias which is applied to measured performance metrics.

Aronson (2007) explains each of these methods in detail. The key strengths and weakness of each is summarized in table 1.

Table 1: Comparison of Methods to Mitigate DMB

	Strengths	Weaknesses
Cross-Validation	Ease of use, allows statistical inference, provides performance estimate	Inefficient use of market data, smaller sample size reduces accuracy
Bias Compensation	Allows statistical inference, efficient use of market data	Complex, special software + large database required, no performance estimate
Bias Deflation	Provides performance estimate, efficient use of market data	Possibly inaccurate, large database required, statistical inference not possible

4. System Parameter Permutation

Each method of DMB mitigation described in the previous section has certain limitations or complexities. This section offers an alternative method named System Parameter Permutation (SPP). SPP provides a practical means of estimating the performance of a trading system as well as statistical significance testing. SPP is not subject to data mining bias[6] and uses standard trading system optimization approaches that are already built-into generally available system development software packages.

SPP provides much more than a method to mitigate DMB however. SPP enables the trader to objectively determine: 1) the performance of the inherent edge expected in the long-run, and 2) the worst-case performance expected in the short-run. With this information, the trader can make data-driven decisions on whether to allocate capital to

[6] To ensure the absence of DMB, SPP must be conducted as a standalone process (not to compare systems). Any ex post selection based on performance has the potential to introduce DMB as discussed in section 3.

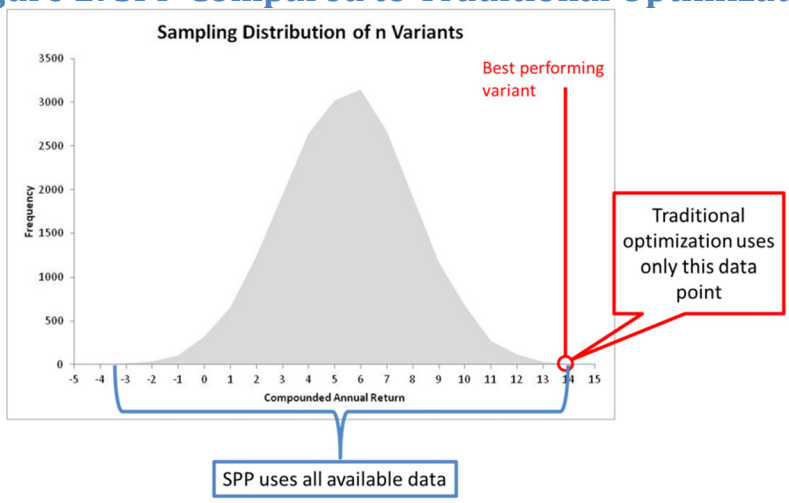
the system and once actively trading, whether the system is "broken" and should cease trading.

4.1 System Parameter Permutation Defined

In the simplest of terms, SPP generates sampling distributions of system performance metrics by leveraging the system optimization process. Each point in a given distribution is the result of a historical simulation run that accurately modeled portfolio effects. Via sampling distributions, the trader may evaluate a system based on any desired performance metrics. SPP then uses the descriptive statistics of the sampling distributions to arrive at performance estimates and measures of statistical significance.

Unlike standard optimization, SPP does not simply choose the best set of parameters but rather uses all of the performance data available for all sets of parameters evaluated during optimization. Whereas traditional optimization picks the best set of parameters and discards the rest, SPP makes use of all available information. Figure 2 illustrates the difference.

Figure 2: SPP Compared to Traditional Optimization



For each system metric of interest, the output of SPP is a sampling distribution that includes trade results from all system variants (combinations of parameter values) where the median serves as the best estimate of true system performance. This is very different than cross-validation or data mining bias compensation which use the result of a single sequence of trades in order to estimate system performance.

The median performance is used as the best estimate of future performance for several reasons: 1) the median is not subject to data mining bias because no selection is involved; 2) no assumptions of the shape of the distribution are required; and 3) the median is robust in the presence of outlier values.

4.2 Steps of System Parameter Permutation

In order to generate sampling distributions of system variant performance metrics, the set of parameter ranges under which the trading system is expected to function is determined ex ante in preparation for optimization. Methods to choose the parameter ranges and observation points are beyond the scope of this paper; however Kaufman (2013) and Pardo (2008) are suggested for further research into these topics. SPP follows these general steps:

1) Parameter scan ranges for the system concept are determined by the system developer.
2) Each parameter scan range is divided into an appropriate number of observation points (specific parameter values).
3) Exhaustive optimization (all possible parameter value combinations) is performed using a realistic portfolio-based historical simulation over the selected time period.

4) The simulated results for each system variant are combined to create a sampling distribution for each performance metric of interest (e.g. CAR, max drawdown, Sharpe ratio, etc.). Each point on a distribution is the result of a historical simulation run from a single system variant.

Figure 3: General Steps of System Parameter Permutation

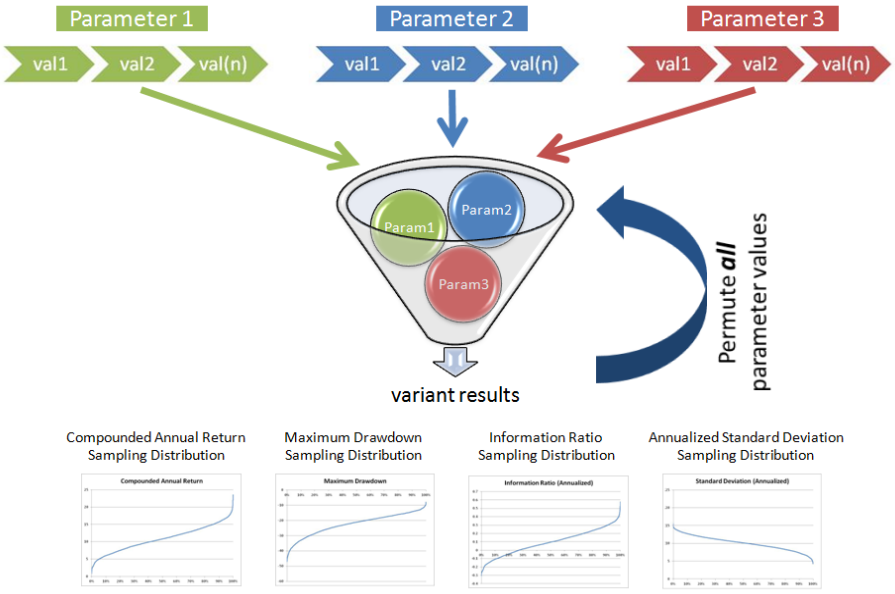


Figure 3 illustrates the process. In this case, sampling distributions for four performance metrics are shown for illustration. Any number of specific performance metrics may be selected by the trader for his specific objectives. The cumulative distribution function (CDF) for each metric may be examined directly and may be used for performance estimation and statistical inference.

In order to ensure the SPP result is not biased, care must be taken to thoughtfully select parameter scan ranges ex ante. If SPP is repeated multiple times by changing the parameter scan ranges in attempt to get a better result, data mining is at work and the SPP estimate may become positively biased. Since the intent of SPP is the avoidance of bias,

such a practice would be counterproductive. Thus it is important that the system developer start the system development process with this consideration in mind.

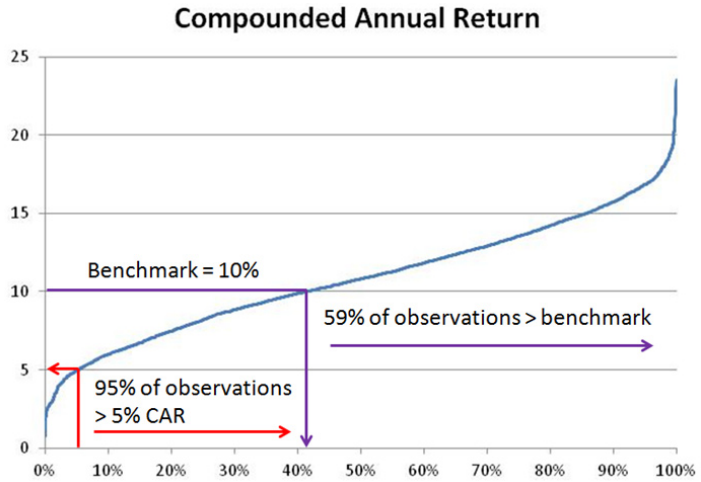
4.2.1 SPP Estimate of the Long–Run Performance of the Trading System

The trader would like to answer the question: "What is a reasonable performance estimate of the long-run edge of the system?" SPP can effectively answer this question.

To generate long-run performance estimates, sampling distributions are produced as described above using all available market data. The use of all available market data enables the best approximation of the long-run so the more market data available, the more accurate the estimate. For each performance metric of interest, the median value is used as the best, unbiased performance estimate.

The trader may also be interested in testing the statistical significance of the SPP long-run performance estimates either in terms of absolute returns or relative to a benchmark. Because SPP generates complete sampling distributions, estimated p-values and confidence levels may be observed directly from the CDF as illustrated in figure 4.

Figure 4: Using the Cumulative Distribution Function for Statistical Inference



The example in figure 4 shows that in 95% of cases, the true value lies in the confidence interval above the level of 5% compounded annual return (CAR); this is equivalent to a p-value of 0.05. Depending on the objectives of the trader, this may or may not be satisfactory. If the trader is interested in outperforming a benchmark with a CAR of 10%, the picture is a bit different. In only 59% of cases does the true value lie in the confidence interval above the benchmark return; this is not statistically significant.

4.2.2 Short-Run Performance Estimate and Worst-Case Contingency Analysis

Whereas the long-run performance estimate indicates what may be expected from the system edge long term, short-run variability may be significant. Thus, the trader would also like to answer the question: "What worst-case contingencies must be tolerated in short-run performance in order to achieve the long-run expectation?" Once the short-run time period is specified, SPP can effectively answer this question.

The duration of the short-run time period is dependent on the preferences and psychology of the trader and/or clients. Chekhlov et al. (2003) mention that the typical drawdown duration tolerated by clients of managed account practitioners ranges from 1-2 years at the most. In any case, the trader needs to determine the duration of the short-run time period that best fits the trading objectives. In general, shorter duration periods have wider ranges of expected performance.

The following steps explain how to perform SPP for the short-run time period:

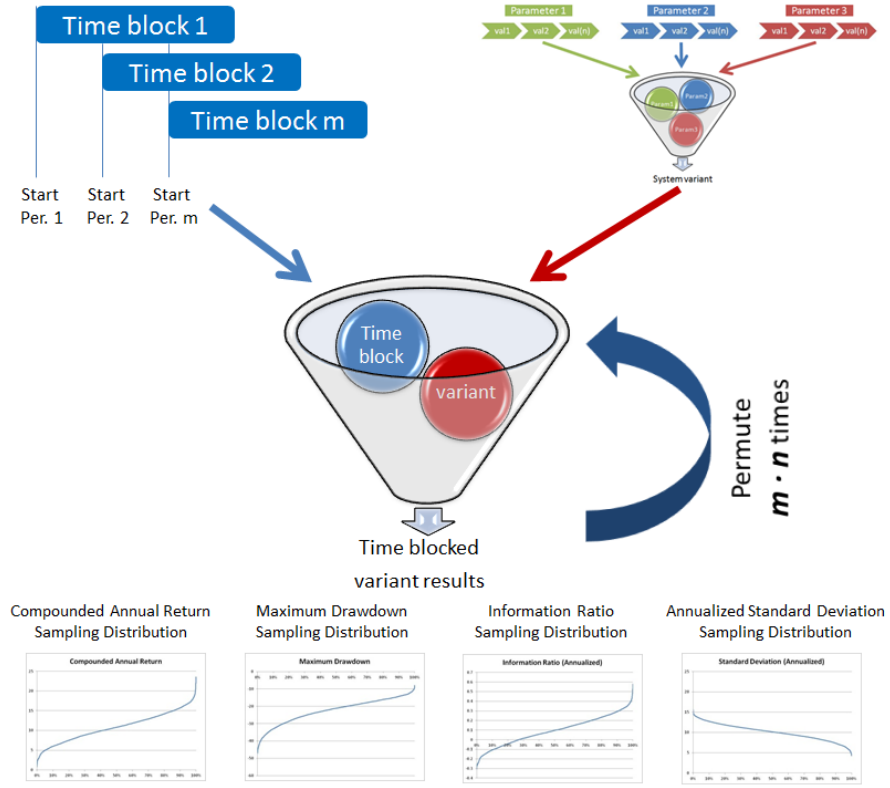
1) All available market data is split into blocks equal in length to the short-run time period ($\boldsymbol{t}$). Each time block may overlap with the previous block depending on the

timeframe of trading signals (such as any month within a year or any hour within a day). This results in some number of time blocks ($\boldsymbol{m}$).

2) Steps 1-4 of the general SPP process are performed on all $\boldsymbol{m}$ time blocks separately. Thus if a system has $\boldsymbol{n}$ combinations of parameter values, a total of $\boldsymbol{m} \cdot \boldsymbol{n}$ optimization permutations are performed on a historical time period of length $\boldsymbol{t}$ in order to generate the sampling distribution for each performance metric of interest over the selected short-run timeframe.

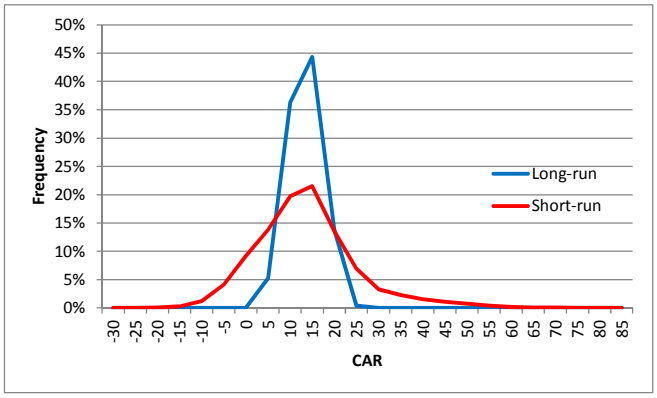
Figure 5 below illustrates the process. Again, sampling distributions for four performance metrics are shown for illustration. Any number of specific performance metrics may be selected by the trader for his specific objectives.

Figure 5: SPP Application for Short-Run Performance Estimation



The sampling distributions resulting from this process each contain many more individual samples with a higher degree of variation than were generated via the SPP long-run performance estimate process. However each sample has a shorter simulation timeframe and thus a fewer number of closed trades contained in each sample. With fewer closed trades per sample, the standard error associated with each sample increases. As the standard error per sample increases, so does the variation of the sampling distribution. The increased variation can be seen in the respective probability density functions as shown in figure 6 below.

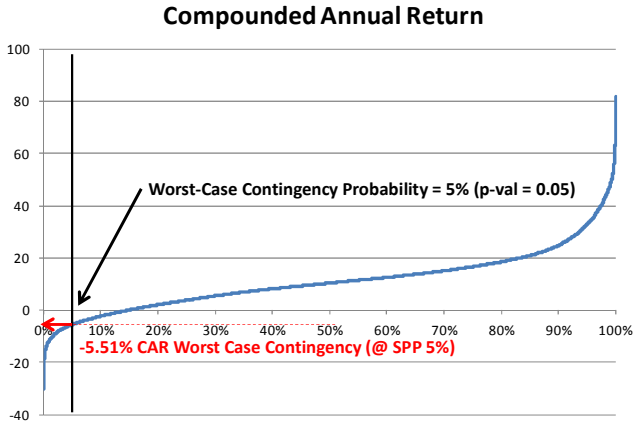
Figure 6: Increased Variation in Short-Run Performance



With sampling distributions, the trader may make a probabilistic, data-driven decision of whether to risk capital on the system. To do so, the trader determines a probability level he determines to be highly improbable but tolerable as his worst-case (common levels are 5% or 1%). Alternatively, the trader may specify the worst-case in terms of the least favorable but tolerable level of performance. Whatever worst-case probability or level of performance is chosen, the CDF of the short-run system metrics of choice are examined as in figure 7. If the worst-case contingency at the respective

probability cannot be tolerated by the trader or clients, capital should not be allocated to the system.

Figure 7: Evaluating the "Worst-Case Contingency"



The example in figure 7 indicates that if the trader cannot tolerate a 5% probability of a realized compounded annual return of -5.51% over the short-run time period chosen, the system should not be traded. If the worst-case contingency is tolerable and capital is allocated to the trading system, the same (or different) worst-case probability(s) or level(s) of performance may be used to determine whether the system is "broken" and if trading should cease.

The stop trading decision should be made when the system has been traded for the duration of the short-term period selected. Thus if one year was selected as the short-term time period, the stop trading decision should only be made at the one year mark. Again using the example in figure 7, if realized performance is worse than -5.51% over a year of trading, the trader may decide to stop trading the system because the ex ante worst case contingency was violated. Any timeframe or probability may be used in this decision. Thus SPP enables the trader to add an objective method of risk control to his trading plan.

4.3 Why System Parameter Permutation is Effective

Using traditional optimization, all performance metrics for the system are derived from the single (best) sequence of trades selected during the optimization process. In order to generate a distribution of contingencies, randomization techniques employing resampling such as bootstrap or Monte Carlo Permutation (MCP) are commonly used.

There are several problematic assumptions made by resampling methods but two are of particular interest here: 1) the result of a single historical simulation is representative of the future distribution of trade results; 2) real world portfolio effects combined with position sizing are accurately modeled. The discussion of data mining bias already showed that assumption number one is problematic. Assumption number two is also problematic; portfolio effects such as buying power, dynamic inter-symbol correlation, and autocorrelation would likely not allow some of the resampled results to occur in real trading. Likewise, this type of randomization does not explore trades unseen in the original, single sample sequence of trades that may have occurred under slightly different conditions. This is a natural consequence of random resampling.

Unlike random resampling, the random variation in SPP originates from the application of a set of slightly varied entry/exit rules on actual market data where trading signals are evaluated using a realistic simulated portfolio. In effect, SPP explores facets of the trading system that would otherwise remain hidden yet are possible in real trading.

SPP produces reliable estimates of trading system performance by: 1) leveraging the statistical law of regression toward the mean, and 2) extracting maximum information

from available market data. For #1, the use of a large number of combinations of parameter values thoroughly examines various ways randomness may affect the system and thus estimates the effects of regression toward the mean. For #2, the use of all available market data ensures that performance results contain the smallest standard error possible and that the system has been exposed to the most varied market conditions possible. Both are explored in more detail.

4.3.1 How SPP Leverages Regression toward the Mean

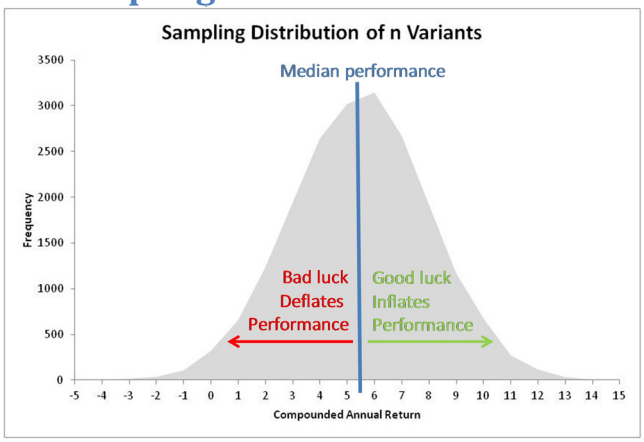
In system optimization, regression toward the mean indicates that the specific combination of optimized parameter values which led to extreme performance in historical simulation will probabilistically not retain a level of extreme performance in the future. The section on data mining bias showed that extreme performance tends to regress toward the mean level over time as the impact of luck tends to change. It is instructive to examine the mechanics of how luck affects system variant performance.

In general, good luck involves some combination of catching favorable market moves and avoiding adverse market moves. One way luck affects system performance is through the interaction of parameters on market data. Parameter values control the exact timing of entry and exit signals; one combination of parameters may generate a very favorable set of entry and exit signals where other similar combinations may generate much less favorable signals on the same market data.

SPP generates a distribution of performance results from a large number of individual historical simulation runs that use the same market data applied to different

combinations of parameter values. The distribution includes the results from many slightly different entry/exit signal combinations across simulation runs. With a large number of samples, the impact of regression toward the mean is seen to varying degrees over the distribution of system variant performance results as shown in figure 8 below.

Figure 8: Sampling Distribution Generated by SPP



Another way luck affects system performance is through the interaction of the timing of market entry/exit signals and portfolio effects such as buying power, dynamic inter-symbol correlation, and autocorrelation. As demonstrated by Krawinkel (2011) randomly skipped trades can have a large impact on realized system performance. Yet, this phenomenon remains largely unrecognized and underexplored. SPP thoroughly and realistically explores this effect through the distribution of performance results.

In SPP, one combination of parameter values may capture a certain set of trades whereas a slight variation in parameter values may capture trades not previously seen and/or skip others that were previously captured. Through this interaction, SPP includes the effects of randomly skipped and included trades. Again the impact of regression toward the mean is seen to varying degrees over the distribution of performance results.

4.3.2 How SPP Extracts Maximum Information from Available Market Data

SPP minimizes standard error of the mean (SEM) by using all available market data in the historical simulation. As sample size increases, SEM decreases proportional to the square root of the sample size due to the mathematical identity: SEM = $s/\sqrt{n}$.

Although the use of all available market data is not a unique feature of SPP, it is one of its strengths. In contrast, traditional cross-validation methods split market data in some way. The effect on SEM of such a split can be large. Table 2 shows the approximate percentage increase of SEM for various data splitting schemes over SPP.

Table 2: Increase of SEM for Market Data Splits

	50/50 Split	80/20 Split	90/10 Split
In-sample	41%	12%	5%
Out-of-sample	41%	124%	216%

Further, Inoue and Kilian (2002) found that OOS and IS tests are equally reliable in the presence of data mining once proper critical values are used and that IS (using all market data) tests have power advantages when there is "unmodelled structural change in the parameter of interest" (a change in market conditions). The use of all available market data ensures that the system has been exposed to the most varied market conditions possible in historical simulation. Doing so cannot guarantee that future market conditions will be similar to those seen historically but any sort of data split ensures loss of information and thus less representative performance results. The most information rich historical simulation uses all available market data.

5. Practical Example of SPP Applied to a Model Trading System

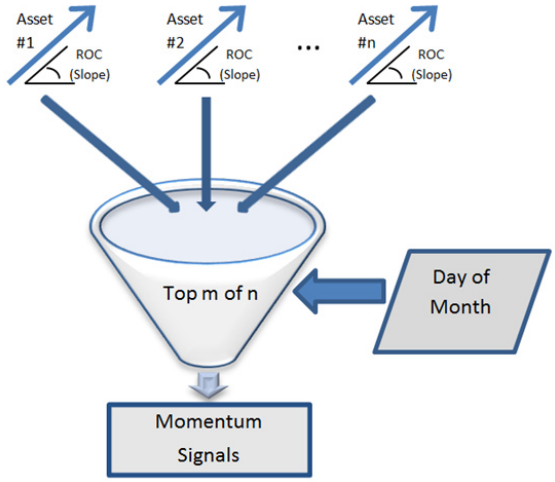
The relative momentum concept in the style of Blitz and Van Vliet (2008) was chosen to create an example system because significant research has validated these types of strategies within and across many different asset classes (Asness et al (2009)). Further, a large amount of post-publication, out-of-sample validation exists for relative momentum (Asness et al (2009)) thus confirming its viability.

5.1 Generalized System Model

The relative momentum trading system concept is based on the observation that the best performing assets or asset classes in the current period tend to continue their outperformance in the next period. Research indicates that momentum measured over 3-12 months tends to show the largest edge.

The generalized system model defines how momentum is measured, the number of assets to comprise the portfolio, and the timing of asset selection. In the interest of risk management, a catastrophic stop-loss is added to the general model as well. Thus the generalized system model shown in figure 9 contains four parameters.

Figure 9: Relative Momentum Generalized System Model



The ROC indicator was chosen to measure momentum as the percentage change over the look-back period. The timing of asset rotation was chosen to be once per month on a specific day in relation to the last trading day of month. Finally a catastrophic stop loss as a percentage of the entry price was introduced for risk management.

The parameter scan ranges were defined in light of the generalized system concept. The portfolio composition was limited to the top 2-5 assets out of 10 in a balance between momentum and diversification. The ROC look-back length was varied in increments of 10% starting from 60 trading days (~3 months) up to 251 trading days (~1 year). The date of entry/exit rotation chosen was the last trading day of month +/- 5 trading days. Finally the stop loss was varied from 10% to 20% in increments of 2%. The system details are shown in table 3.

Table 3: Relative Momentum Trading System Details

System Component	Indicator	Minimum	Maximum	Step	# Values
# Assets Held	N/A	2	5	1	4
Momentum Rank	ROC(a)	60	251	10%	16
Rotation Time Period	Last DOM + b	-5	5	1	11
Stop Loss Point	% of entry price	10%	20%	2%	6

Exhaustive optimization of the above scan ranges resulted in 4224 combinations of parameter values. The method of position sizing used was equal margin per position. A standard 100% maintenance margin requirement was used along with a 5% cash safety buffer. This allowed up to 95% of trading capital to be used to take entry signals.

5.2 Optimization Results and Out-of-sample Calibration

This section discusses traditional OOS testing applied to the trading system. The OOS analysis is used for comparison purposes to SPP. The trading system was

optimized using the annualized information ratio (vs. the dividend reinvested SPY ETF) as the fitness function in order to maximize benchmark outperformance. The OOS test used 80% of available market data in-sample and the remaining 20% was reserved for out-of-sample calibration. Results are shown in table 4.

Table 4: Relative Momentum System OOS Results

	IS	OOS	OOS % of IS
Compounded Annual Return	22.41%	14.47%	65%
Maximum Drawdown	-18.06%	-9.6%	53%
Annualized Standard Deviation	19.08%	11.96%	63%
Annualized Information Ratio	0.70	-1.20	-171%

Using this method, the OOS performance metrics serve as the only unbiased estimates in setting expectations for future performance and thus also serve as the determinant of whether to risk capital on the system. Standard practice in OOS testing dictates that a system passes cross-validation if OOS performance is >= 50% of IS performance. In this case, the majority of the system metrics are above the desired threshold yet the information ratio for the OOS segment is much below. Therefore this system fails traditional cross-validation due to the unacceptable OOS information ratio.

5.3 SPP Long-Run Estimate of System Performance

Next, SPP was performed on the system as specified in section 4.2.1. In contrast to the previously described method, SPP uses all available market data and when applied to the same system, provides much more information. Table 5 shows the traditional OOS results/estimates compared to the respective SPP estimates and to the buy-and-hold benchmark (SPY ETF with dividends reinvested). The goal in employing this system is to

outperform the benchmark and thus the statistical significance of outperformance for each system metric (via the equivalent p-value) is also shown.

The data in table 5 may be used by the trader to decide whether to allocate capital to the system. For example, the trader may ask "Is an unbiased estimate of realizing a 8.94% CAR sufficient reward to compensate for the risk of a -24.22% drawdown and an annualized 15.61% standard deviation? Is a p-value of 0.10 for CAR significant enough to be confident in outperforming the benchmark?" These questions may be answered via the SPP generated sampling distributions.

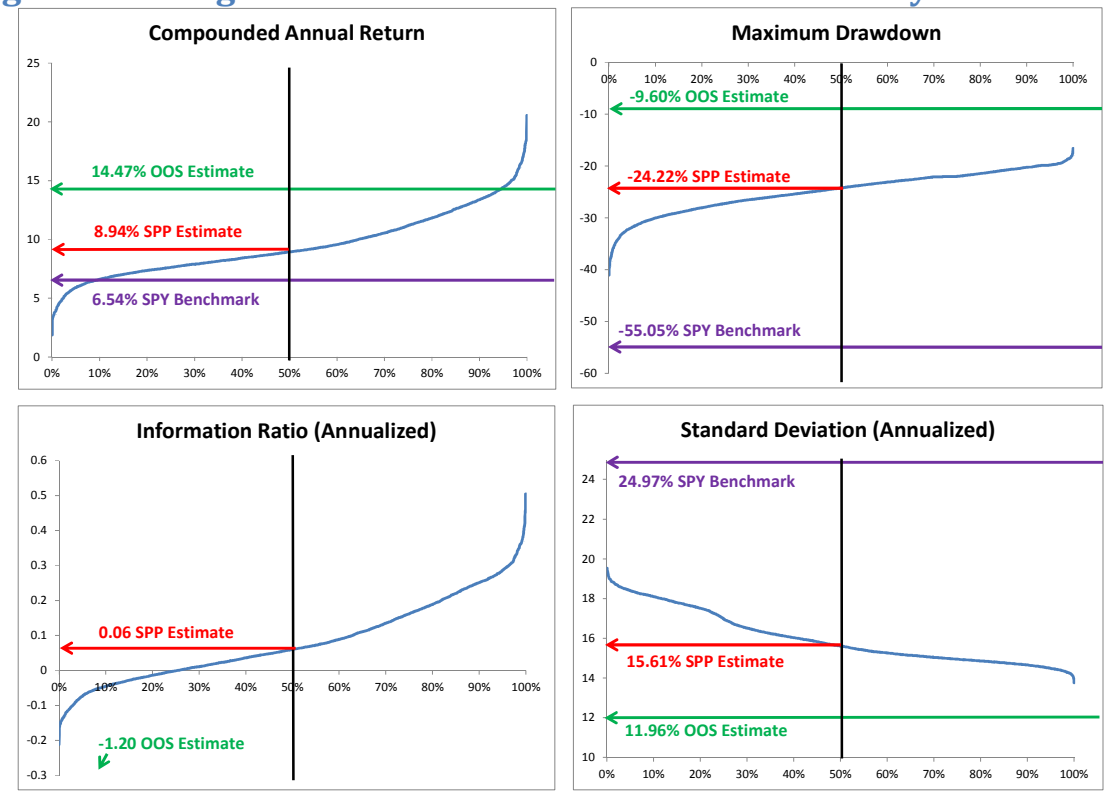
Table 5: Long-Run SPP Estimate vs. OOS and Benchmark

	Compounded Annual Return	Maximum Drawdown	Annualized Information Ratio	Annualized Standard Deviation
Cross-Validation OOS Estimate	14.47%	-9.60%	-1.20	11.96%
SPP Estimate of Long-Run Perf.	8.94%	-24.22%	0.06	15.61%
SPY Benchmark	6.54%	-55.05%	N/A	24.97%
Equiv. P-Val for Outperformance	0.10	0.00	0.25	0.00

The results in table 5 are taken from specific points along the SPP sampling distributions. For the four system metrics examined in this example, the CDFs (blue) are shown in figure 10 from which the trader can make further probabilistic estimates. The system metrics are shown on the y-axis of the charts and the cumulative probabilities on the x-axis.

Additionally, the SPP estimate (red), OOS estimate (green) and benchmark (purple) are overlaid onto each CDF chart. The vertical black line highlights the median of the sampling distribution. The intersection point of the CDF and the benchmark as measured along the x-axis is the value of the equivalent p-value from table 5.

Figure 10: Long-Run SPP Generated CDFs for Selected System Metrics



5.4 SPP Worst-Case Contingency Analysis for Calendar Year Performance

The next analysis uses the calendar year as the short-term time period of interest. The historical market data were divided into seven blocks, for each of the full calendar years present in the data. The process from section 4.2.2 was completed on this data in order to evaluate the expected worst-case contingency for any calendar year period.

Table 6: Calendar Year SPP Worst-Case Contingency vs. OOS and Benchmark

	Compounded Annual Return	Maximum Drawdown	Annualized Information Ratio	Annualized Standard Deviation
Cross-Validation OOS Estimate	14.47%	-9.60%	-1.20	11.96%
Worst-Case Contingency (@SPP 5%)	-12.98%	-23.95%	-1.45	21.67%
SPY Benchmark Minimum	-36.27%	-47.04%	N/A	10.03%
SPY Benchmark Maximum	22.8%	-7.63%	N/A	41.92%

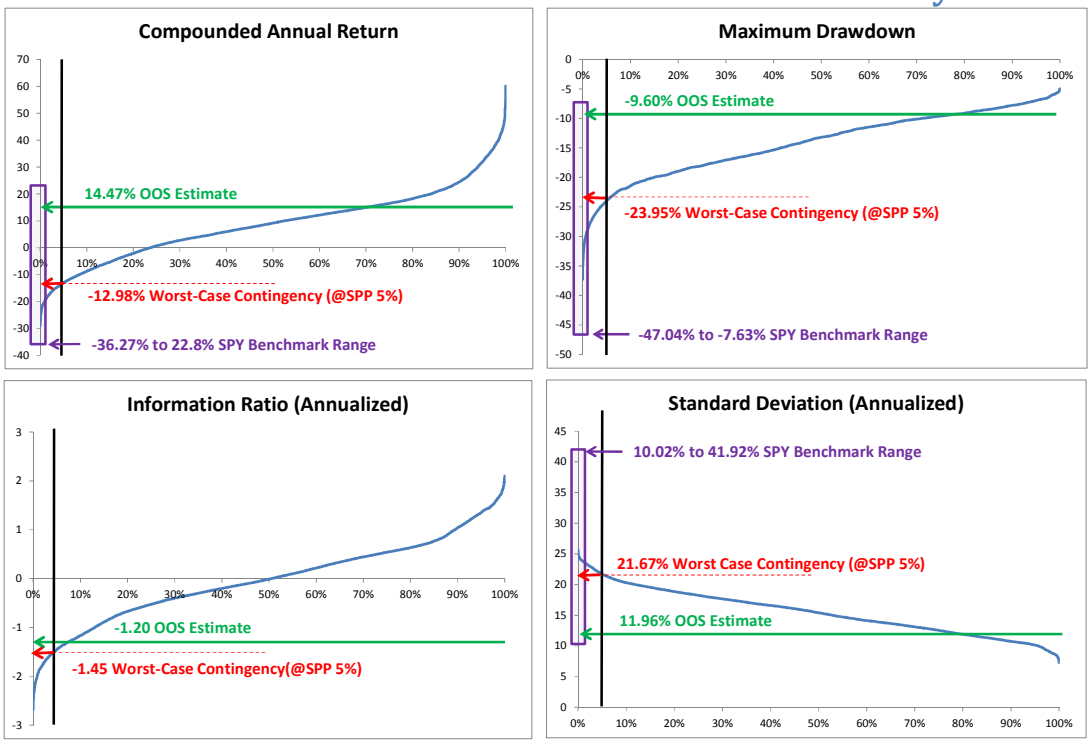
In this case, the SPP 5th percentile (equivalent to p-value = 0.05) was chosen as the worst-case contingency probability. Table 6 shows the same OOS results/estimate from

above compared to the SPP worst case contingency and to the range of the buy-and-hold benchmark over each of the seven full calendar years in the historical simulation period.

The trader must decide whether the worst-case contingency for calendar year performance shown in table 6 is tolerable in order to achieve the long-run SPP performance expectations of the trading system shown in table 5 (previous section). For example, the trader must be prepared to accept a 5% probability of realizing a -1.45 annualized information ratio (significantly underperforming the benchmark) in any given calendar year while at the same time, achieving negative absolute returns (-13% CAR).

Figure 11 shows the CDFs (blue) for the four chosen system metrics as well as the SPP estimate (red) and OOS estimate (green) overlaid. The vertical black line highlights the 5th percentile of the sampling distribution (worst-case contingency probability chosen) and the calendar year range of the benchmark is shown by a purple bar on the y-axis.

Figure 11: Calendar Year SPP Generated CDFs for Selected System Metrics



5.5 Discussion of Results

The above example showed that, compared to standard OOS cross-validation, SPP provides the trader with much more information. SPP creates long-run and short-run sampling distributions of system metrics using all available historical market data whereas traditional OOS cross-validation provides only a point estimate on a subset of historical market data. SPP enables probabilistic decision making whereas traditional OOS necessitates a binary pass/fail decision. Thus SPP enables a much deeper understanding of how the trading system may perform going forward.

SPP applied to the relative momentum trading system demonstrates outperformance over the buy-and-hold benchmark in the long-run with varying degrees of statistical significance for different system metrics. Specifically, an equivalent p-value of 0.10 for CAR outperformance is marginally significant. In contrast, an equivalent p-value of 0.00 for max drawdown outperformance is highly significant and these results together indicate that a strength of relative momentum is avoidance of large drawdowns.

However, the SPP worst-case contingency analysis for calendar year performance demonstrated that in order to achieve long-term outperformance, the trader must be willing to accept the possibility of significant underperformance in any calendar year. With this information, the capital allocation decision may be made probabilistically.

6. Conclusions

It is essential for any trader to thoroughly understand what to expect from a trading system before allocating capital. Without knowledge of the probable ranges of

performance expected in the future, the trader or client is prone to abandon a good system in the stress of an unexpected drawdown or period of underperformance. Even worse, capital may be allocated on the basis of inflated expectations gained from traditional evaluation methods when the system should be discarded in the light of the probabilistic information that SPP is able to provide.

The majority of traditional system development approaches provide a single, point estimate of performance and/or measure of statistical significance based on a single sequence of trades. With the limited information from such a point estimate, the capital allocation decision is difficult at best. In contrast, SPP produces sampling distributions of system metrics that allow more realistic contingency planning based on probabilities.

Ultimately SPP offers a simple, easy to use, yet realistic method to estimate future system performance. It is the balance of these three factors that is the true strength of the method. Thus SPP is broadly applicable by traders and system developers of varying backgrounds and adds value in real-life practice.

The trading system example showed that SPP provides a clear, balanced picture of expected system performance where standard cross-validation did not. The example also demonstrated that the relative momentum trading system is likely to outperform the buy-and-hold benchmark over the long run but that in order to achieve long-term outperformance, the trader must be willing to accept the possibility of significant underperformance in any given year. With this information, the capital allocation decision may be made probabilistically.

7. References

Aronson, David R., 2007, *Evidence-based Technical Analysis: Applying the Scientific Method and Statistical Inference to Trading Signals*, John Wiley & Sons, Inc., Hoboken, NJ, 544p.

Asness, Clifford S., Tobias J. Moskowitz, and Lasse J. Pedersen, 2009, "Value and Momentum Everywhere," AFA 2010 Atlanta Meetings Paper, http://ssrn.com/abstract=1363476 (December 1, 2013).

Baily, David H., Jonathan M. Borwein, Marcos López de Prado, and Qiji Jim Zhu, 2013, "Pseudo-mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-sample Performance," http://ssrn.com/abstract=2308659 (December 20, 2013).

Blitz, David C. and Pim Van Vliet, 2008, "Global Tactical Cross-Asset Allocation: Applying Value and Momentum Across Asset Classes," *Journal of Portfolio Management* 35 (1), p. 23-38.

Chekhlov, Alexei, Stanislov Uryasev, and Michael Zabarankin, 2003, "Portfolio Optimization with Drawdown Constraints," http://www.ise.ufl.edu/uryasev/files/2011/11/drawdown.pdf (December 20, 2013).

Inoue, Atsushi, and Lutz Kilian, 2002, "In-Sample or Out-of-Sample tests of Predictability: Which One Should We Use?," Working Paper No. 195, European Central Bank Working paper Series (November 2002).

Kaufman, Perry J., 2013, *Trading Systems and Methods, + Website, 5th Edition*, John Wiley & Sons, Inc., Hoboken, NJ, 1232p.

Krawinkel, Thomas, 2011, "Buying Power – The Overlooked Success Factor," NAAIM Wagner Award 2011.

Pardo, Robert, 2008, *The Evaluation and Optimization of Trading Strategies 2nd Edition*, John Wiley & Sons, Inc., Hoboken, NJ, 334p.