Data Snooping Bias in Tests of the Relative Performance of Multiple Forecasting Models

Dan Gabriel Anghel[1,2]

[1] *Institute for Economic Forecasting, Romanian Academy*

[2] *Faculty of Finance and Banking and CEFIMO, The Bucharest University of Economic Studies*

This version: 08/05/2019

Abstract. Tests of the relative performance of multiple forecasting models are sensitive to how the set of alternatives is defined. Evaluating one model against a particular set may show that it has superior predictive ability, while changing the number or type of alternatives in the set may show otherwise. This paper focuses on forecasting models based on technical analysis and shows that data snooping bias occurs in tests that dismiss alternatives used by investors and researchers. If all relevant alternatives are not easily observable, then testing for relative model performance becomes problematic and the results should be treated with care.

JEL c1assification: C12, C18, G11, G14, G17

Keywords: Data Snooping; Test Misspecification; Superior Predictive Ability; Relative Performance; Reality Check; Technical Analysis; Efficient Market Hypothesis.

1. Introduction

Searching for better forecasting models is the fundamental objective in many research projects with key theoretical, practical applications. Technological advancements and the rapid growth in computing power have led to a significant increase in the number of investigated alternatives. Analyzing the absolute performance of a model may be desirable in some circumstances, but evaluating relative performance is preferred for obvious reasons. These developments imply that the risk of data snooping[1] is greater than ever before. Data snooping is fast becoming one of the great threats to the advancement of scientific knowledge in the 21st century. How should the relative predictive ability of a new model be tested given existing alternatives? White (2000) argues that tests must account for the data snooping efforts of others. However, considering alternative models that others use is not an established practice in the financial economics literature and an investigation into if and how does this influence reported results has yet to be performed.

In this paper, we investigate how choosing an “unrepresentative” set of alternatives, one which does not account for the data snooping efforts of others, influences the outcomes of tests that evaluate the relative performance of multiple forecasting models. The paper focuses on the literature examining models based on technical analysis, technical trading rules (TTRs) because the number of alternatives is especially large and difficult to exactly pinpoint; but it is relevant for any investigation of the relative performance of multiple forecasting models, regardless of the underlying research topic. First, the paper discusses if the sets of models that are typically used in the literature are representative or not. Second, a simulation exercise is used to show how employing small, unrepresentative sets in seemingly data snooping-free statistical tests generates false discoveries. Third, the potential impact of this particular type of data snooping

---

[1] Data snooping refers to the practice of searching for better performing forecasting models on the same data samples. This increases the chances of founding and using models that just fit the noise in the original data, are lucky, but have little economic significance and perform poorly in out-of-sample applications.

bias on test results and conclusions reported in the TTR literature is evaluated in an extended empirical investigation. Overall, our results show that trading rule universes that are typically used are most likely unrepresentative for what investors and researchers use. Not accounting for the data snooping efforts of others biases test outcomes in favor of showing that some TTRs have statistically significant predictive ability in financial markets. One direct implication of the findings is that positive discoveries of TTR excess performance reported in the literature, which are based on tests that use unrepresentative rule universes, should be treated with more care. More generally, controlling for the data snooping efforts of others is important for obtaining robust results that can withstand the test of time. Also, we argue that evaluating relative (excess) performance becomes problematic when representativeness is ambiguous, relevant alternatives are not fully observable, such as in the case of tests that evaluate forecasting models derived from technical analysis. In this and other similar circumstances, evaluating absolute performance may provide more objective results and may have some merit.

This paper is inspired by the recent discussion in the financial economics literature centered on the impact of test misspecification and cherry-picking results on the robustness of reported results and associated inferences. For example, Kim and Ji (2015) find that the results reported in many surveyed papers become questionable after revised standards for evidence are used instead. Also, they observe strong evidence of publication bias in favor of statistically significant results. Harvey (2017) and Harvey and Liu (2014) discuss the importance of increasing the statistical significance threshold in tests that use widely examined data and controlling for data snooping. Harvey (2017) stresses that “*with the combination of unreported tests, lack of adjustment for multiple tests, and direct as well as indirect p-hacking, many of the results being published will fail to hold up in the future.*” In this paper, we further investigate the extent to which using small, unrepresentative sets in tests of the relative performance of multiple forecasting models can be associated with data snooping and biased results.

In the literature examining models based on technical analysis (TTRs), a typical test defines the set of models, named a “trading rule universe”, simulates trading to measure the relative performance of each model and evaluates the statistical and economic significance of the results. Such a test is biased when not properly handling the associated multiple hypotheses. In his seminal paper, White (2000) introduces the Reality Check (RC) test and solves this issue by accounting for the covariance matrix of the excess returns series generated by the different models. The RC controls for the family-wise error rate and delivers asymptotically valid p-values for evaluating the null hypothesis of no excess performance using an empirical distribution estimated via bootstrap simulation. Modified versions of the RC test have been developed by Hansen (2005), who proposed recentering the test statistic and eliminating poor performing rules in order to improve power, or by Romano and Wolf (2005) and Hsu et al. (2010), who developed step-versions in order to identify all overperforming rules. Alternative tests of relative performance, which are based on Bonferroni bounds and control for the False Discovery Rate (FDR), have been proposed by, among others, Benjamini and Hochberg (1995), Storey (2002) or Barras et al. (2010). For brevity and also because RC-type tests are more frequently used in the literature, our discussion centers on the RC test defined by White (2000), while the Superior Predictive Ability (SPA) test proposed by Hansen (2005) is also considered in a robustness analysis. We thus leave the discussion of FDR-type tests to subsequent research. However, we note that the results of the simulation exercise reported in Section 4 suggest that using unrepresentative universes in statistical tests that employ the FDR strategy may also influence their outcomes. Tests in the RC class seemingly eliminate data snooping by handling the associated multiple hypotheses for the considered fixed set of prediction models. However, White (2000) or Sullivan et al. (1999) argue that the characteristics of the bootstrap-generated distribution used to evaluate the null depend on the size and diversity of the pre-specified universe. Thus, the outcome of RC-type tests should be influenced by the subjective choice that researchers make regarding the composition of the universe. Surprisingly, to the extent of our knowledge, a detailed analysis of if and how this happens has not been performed so far. This paper sets out to fill in this gap.

Because of its specific focus, our paper also directly contributes to the literature examining the excess performance (returns) obtained by TTRs in financial markets, which is one of the most exposed to the risk of data snooping. Put differently, our results have important implications for the literature concerned with the crucial theoretical concept of efficient financial markets (Fama, 1970). There is a widely accepted view that financial prices/returns are not completely random[2]. However, given existing market frictions and other limitations, this does not automatically imply that stock markets are not weak-form efficient and that investors are able to earn economic profits (Jensen, 1978; Timmermann and Granger, 2004). Park and Irwin (2007) provide a comprehensive review of the early literature examining TTR excess performance and find that some favorable evidence exists. For example, 58 out of 92 "modern" studies conclude in favor of technical trading rules being able to earn statistically and economically significant excess returns, this implying that markets are not weak-form efficient in an absolute sense. Most results show that TTRs can predict price movements to a certain extent and can earn excess returns over the buy-and-hold benchmark model. However, in some cases, excess returns disappear after adjusting for trading costs and risk or turn out to be statistically insignificant when accounting for data snooping. More recent papers, including some that employ RC-type tests, report that the excess performance of TTRs in developed stock markets has greatly diminished or even disappeared. Examples include Neuhierl and Schlusche (2010), Bajgrowicz and Scaillet (2012), Shynkevich (2012) and Taylor (2014) for the US market, Ratner and Leal (1999), Fifield et al. (2005) and Marshall and Cahan (2005) for others. Such results imply that markets have become more efficient over time. However, conflicting

---

[2] See Grossman and Stiglitz (1980) for theoretical arguments and Lim and Brooks (2011) for empirical evidence.

findings continue to appear, such as in Urquhart et al. (2015), who find that moving average trading strategies based on signal anticipation yield superior profits to investors in the US, UK, and Japan markets. Also, some authors argue that TTRs remain profitable in emerging stock markets, examples including Metghalchi et al. (2009) for Asian markets, Sobreiro et al. (2016) for BRICS and other 6 markets in Central and Latin America, Metghalchi et al. (2012) for emerging European markets, or Al-Nassar (2014) for stock markets in the Middle-East. Further, TTRs have recently been found to earn some kind of economic profits in tests that reexamine the foreign exchange market (Coakley et al., 2016; Hsu et al., 2016; Zarrabi et al., 2017), the US bond market (Shynkevich, 2016), or the commodity futures market (Han et al., 2016). Are TTRs truly capable of earning significant excess returns after being investigated by investors and researchers for so many years? Although it is possible that some markets are not efficient, or even adaptive (Lo, 2004), recent evidence has revealed that test misspecification, methodological limitations and publication bias may play a role in shaping the conclusions in the literature. In this paper, we further investigate if using unrepresentative rule universes in tests contributes to creating a skewed, more favorable picture of TTR excess performance.

The remainder of the paper is structured as follows. Section 2 analyzes if trading rule universes typically used in the literature are representative or not and presents an alternative that should better account for what investors and researchers use. Section 3 discusses the RC test. Section 4 presents the results of a simulation exercise that investigates if and how using unrepresentative universes biases test results. Section 5 presents an extensive empirical investigation that evaluates the potential impact of this particular type of data snooping on the results of tests focusing on TTR relative performance. Section 6 concludes.

2. Trading Rule Universes

Investment professionals use technical analysis to make investment decisions in financial markets (e.g., Taylor and Allen, 1992; Menkhoff, 2010; Scott et al., 2016). Many researchers also evaluate the performance of TTRs (e.g., Park and Irwin, 2007). Even though there is no indication on the exact number and type of rules that are used, we know that stakeholders routinely mine financial prices data in search of better forecasting models, this resulting in a very large set of considered alternatives over time. Thus, we can safely assume that the total number of rules in the "true" universe is quite large.

Table 1. Independent subsets of trading rules in 686k

No.	Name (Symbol)–Technical Analysis Indicator	Indicator Type	Number of trading rules
1	Accumulation Swing Index (ASI)	momentum	210
2	Arms Ease of Movement (EMV)	momentum	840
3	Aroon Oscillator (AO)	standardized momentum	10,507
4	Balance of Market Power (BMP)	standardized momentum	39,207
5	Bollinger Oscillator (%b)	momentum	12,402
6	Center of Gravity Oscillator (COG)	momentum	252
7	Chaikin Money Flow (CMF)	standardized money flow	25,258
8	Chaikin Oscillator (CO)	money flow	6,174
9	Chande Momentum Oscillator (CMO)	standardized momentum	27,969
10	Commodity Channel Index (CCI)	momentum	616
11	Demand Index (DI)	standardized money flow	25,258
12	Detrended Price Oscillator (DPO)	momentum	672
13	Dynamic Momentum Index (DYMOI)	standardized momentum	37,584
14	Filter (F)	momentum	51
15	Inertia Indicator (INI)	standardized momentum	22,464
16	Kase Convergence Divergence (KCD)	momentum	43,141
17	Kase Peak Oscillator (KPO)	momentum	8,624
18	Klinger Volume Oscillator (KVO)	money flow	6,174
19	Know Sure Thing (KST)	momentum	5,488
20	Linear Regression Slope (LRS)	momentum	371
21	Market Volume Impact (MVI)	money flow	252
22	Money Flow Index (MFI)	money flow	24,978
23	Moving Average Convergence Divergence (MACD)	momentum	4,704
24	New Relative Volatility Index (NRVI)	standardized momentum	30,331
25	On Balance Volume (OBV)	money flow	210
26	Plus DM vs. Minus DM crossover (DMI)	standardized momentum	441
27	PI Opinion Oscillator (PI)	standardized momentum	7,107
28	Polarized Fractal Efficiency (PFE)	standardized momentum	60,426
29	Random Walk Index for High prices (RWI)	momentum	450
30	Rate of Change (ROC)	momentum	672
31	Relative Momentum Index (RMI)	standardized momentum	48,600
32	Relative Strength Index (RSI)	standardized momentum	10,864
33	Relative Vigor Index (RVig)	standardized momentum	60,426
34	Relative Volatility Index (RVI)	standardized momentum	16,859
35	Runs Indicator (R)	momentum	11
36	Stochastic Momentum Index (SMI)	standardized momentum	33,250
37	Stochastic Oscillator (%k)	standardized momentum	1,769
38	Stochastic RSI Oscillator (SRSI)	standardized momentum	16,859
39	The Quantitative Candlestick (Qstick)	momentum	840
40	Triple Exponential Smoothing (TRIX)	momentum	3,402
41	True Strength Index (TSI)	standardized momentum	60,426
42	Ultimate Oscillator (UO)	standardized momentum	22,842
43	Vortex Oscillator (VX)	standardized momentum	7,114
44	Williams Variable Accumulation Distribution (WVAD)	money flow	210

Are trading rule universes typically used in the literature representative? We answer this question by first constructing a rule universe that should better represent the "true" one used by others. Trading rules that researchers typically test–such as filters, moving averages, or the MACD and RSI indicators–are incorporated first. This initial selection is supplemented by other trading rules obtained from looking at the practitioner-oriented literature[3], keeping in mind their popularity, applicability in our context, and distinctiveness. The resulting set of trading rules–denoted thereafter as *686k*–contains a total of 686,304 TTRs and can be divided into independent subsets based on the technical analysis indicators used for their construction, as presented in Table 1. Appendix A in the supplementary materials provides extensive details regarding our choices, the terminology employed, the ways in which specific trading rules are constructed, and the qualitative improvements in terms of the diversity that this new universe has over the ones previously used in the literature. To the extent of our knowledge, this is by far the largest and most representative trading rule universe ever considered. For example, Sullivan et al. (1999) use 7,846 TTRs, Zarrabi et al. (2017) use 7,650 TTRs, Neuhierl and Schlusche (2010) use 10,256 TTRs, Shynkevich (2012) uses 12,937 TTRs, Hsu et al. (2016) use 21,000 TTRs, Shynkevich (2016) uses 27,000 TTRs, and Coakley et al. (2016) use 113,148 TTRs. Of course, the "true" rule universe is very difficult, if not impossible, to observe; so even this construction might still not be representative[4]. However, as the number and type of TTRs are based on what others use and are greatly diversified, the divergence should be significantly reduced.

We test the representativeness assumption by estimating the effective span (defined in Sullivan et al., 1999) of the small, independent rule universes contained by *686k* and comparing them to the span of a rule universe representative for what is used in the literature. We choose the universe used by Sullivan et al. (1999)–denoted thereafter as STW–as the benchmark

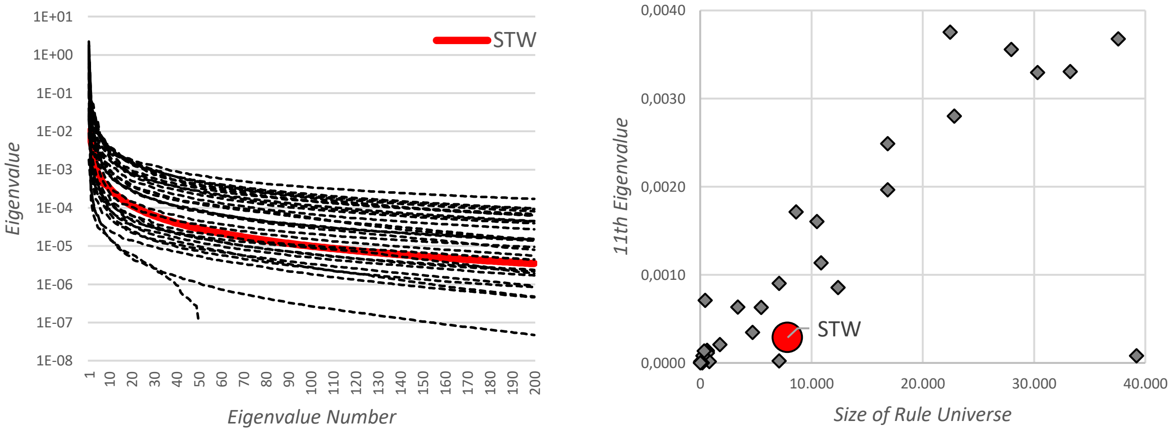
---

[3] We search publications such as the Journal of Technical Analysis, the International Federation of Technical Analysts Journal, or the Technical Analysis of Stocks & Commodities–The Traders' Magazine. We also examine technical analysis books, such as Wilder (1978) or Colby (2002).

[4] Some papers incorporate TTRs derived from various artificial intelligence and computer optimization algorithms (e.g., Brabazon el al., 2012). Also, hedge funds and other skilled investors may additionally incorporate more sophisticated rules, or may use combinations of rules from different areas, such as fundamental analysis, behavioral finance and so on. We implement a conservative approach and disregard such alternatives in order to avoid hindsight bias (Timmermann and Granger, 2004) and also because they require more expertise and a higher implementation cost, which makes them accessible to only a small fraction of investors.

because of its medium-to-large size compared to what is typically used, because of its visibility, and also because data on its span are reported by the authors (Figure 1 in Sullivan et al., 1999, p. 1660). Similar to STW, the spans of our universes are estimated on daily closing price data for the Dow Jones Industrial Average (DJIA) index from 1897 to 1986.

Figure 1. Effective span of trading rule universes



Panel A. Eigenvalues 1 to 200

Panel B. Size vs. 11th eigenvalue

NOTE. This figure reports the first 200 eigenvalues of the covariance matrix of excess returns for TTRs in 31 restricted rule universes reported in Table 1, alongside the rule universe used by Sullivan et al. (1999), designated as STW. Eigenvalues are sorted in descending order. The total number of nonzero eigenvalues represent the effective span of the universe.

The results are reported in Figure 1 and show that the STW universe is dominated in terms of span by 19 of the 31 considered rule universes[5], including some that contain fewer than 7,846 rules. For example, the rule universe generated using the MACD indicator only has 4,704 rules, yet its span is on average 2.08 times higher compared to STW[6]. The universe with the largest span is generated by the DYMOI indicator; it contains 37,584 rules and has a span that is on average 34.06 times higher compared to STW. In general, the results show that the span of the considered universes is positively correlated with their size. Adding prediction models to the analysis extracts more information from the data and implies that the additional practitioner-oriented TTRs incorporated into *686k* to account for the data snooping efforts of others retain sufficient orthogonality compared to the ones that have been used in previous

---

[5] The span is not estimated for 13 of the 44 universes because some require traded volume data that is not available for the DJIA index throughout the considered sample, while for the others computational demand is too high.

[6] This is computed as the average ratio between the first 200 eigenvalues and proxies the ratio between the total number of nonzero eigenvalues assuming similar decays of the eigenvalue series (this is observed in our results).

studies. The span of the *686k* universe is not estimated but should be much larger compared to all of the other smaller universes, which are contained within. Overall, the results show that omitted trading rules that are considered by investors and researchers do generate payoffs that increase the span of rule universes, this implying that the representativeness assumption fails for rule universes typically used in the literature. Section 4 explores if and how this exposes the analysis to data snooping risk and biases test results. In preparation, Section 3 introduces the RC test and discusses some of its characteristics in the context of evaluating prediction models based on technical analysis.

3. The Reality Check (RC) test

A technical analysis indicator (denoted x) is a function $f_x\colon \mathbb{R}^{n_x} \to \mathbb{R}$ that measures a certain characteristic of price movements[7]. On the other hand, a technical trading rule (TTR) is a mathematical statement, based on the values of one or more indicators, used to make investment decisions. Evaluating a TTR is equivalent to making a prediction about how prices will move in the future over a specified interval, which is typically set to one observation. Investors that use TTRs aim to earn statistically significant excess returns by mechanically trading the market. A TTR (denoted k) is represented using a "signal function" $\delta_{k,t}\colon \mathbb{R}^{p_k} \to \{0, 1\}$ that indicates the expected direction of price movements and recommends the appropriate market position[8]. A trading rule universe is a collection of $K \in \mathbb{N}^*$ technical trading rules.

---

[7] For example, the Moving Average Convergence/Divergence (MACD) indicator measures price momentum and is defined as $f_{MACD}\colon \mathbb{R}^{n_{MACD}} \to \mathbb{R}, f_{MACD,t}(m,n,P) = EMA_t(m,P) - EMA_t(n,P)$, where $EMA_t(m,P)$ denotes an exponential moving average of the price series, with smoothing factor $\frac{2}{m+1}$, computed at time $t = \overline{1,T}$. The MACD takes $n_{MACD} = 3$ parameters, the price vector P and the integers m, n representing the length of the "lookback window" for the two averages. In practical applications, the price series is omitted from the definition and the MACD is considered as having $n_{MACD} = 2$ parameters.

[8] For example, the trading rule "$MACD\_1$" can be defined using the signal function $\delta_{MACD\_1,t} = \mathbb{1}_{\{MACD_t(12,26,P)>0\}}$, where $\mathbb{1}_{\{\cdot\}}$ represents the indicator function. A value of 1 predicts that prices will rise, while 0 that they won't. Thus, this rule instructs the investor to go long when the MACD(12,26) takes positive values and to stay out of the market otherwise. Trading rules can be extended to incorporate short positions (in this case, the signal function will be able to take an additional value: $\delta_{k,t}\colon \mathbb{R}^{p_k} \to \{-1, 0, 1\}$), a flexible money management strategy that can partially open/close positions ($\delta_{k,t}\colon \mathbb{R}^{p_k} \to [-1, 1]$), or margin trading ($\delta_{k,t}\colon \mathbb{R}^{p_k} \to [-L, L]$, where L is the leverage defined as the inverse of the margin requirement).

The Reality Check test evaluates the null hypothesis that the best performing TTR in a universe has no superiority over the benchmark, i.e. that its average excess return is not positive and statistically significant. It does so by considering the associated multiple hypotheses and controlling for the Family-wise Error Rate. The test first computes the loss function associated with each TTR by multiplying its signal function to the market log-return (ζ_t):

$$L\left(\zeta_t,\ \delta_{k,t-1}\right) = -\delta_{k,t-1}\zeta_t, \qquad t = \overline{1,T}, \qquad k = \overline{1,K} \tag{1}$$

Using *t–1* for the signal function eliminates contemporaneous trading and controls for the look-ahead bias. Considering the buy-and-hold rule as the benchmark ($\delta_{0,t} = 1, t = \overline{1,T}$) and a sample length of T observations, the excess return series ($d_{k,t}$) and the average excess return (d_k) for each TTR are then calculated:

$$d_{k,t} = L(\zeta_t,\ \delta_{0,t-1}) - L(\zeta_t,\ \delta_{k,t-1}), \qquad t = \overline{1,T}, \qquad k = \overline{1,K} \tag{2}$$

$$\bar{d}_k = \frac{1}{T}\sum_{t=1}^{T} d_{k,t}\,, \qquad k = \overline{1,K} \tag{3}$$

The test statistic is defined as the maximum average excess return (T_n^{RC}) and is evaluated using an empirical distribution ($T_{b,n}^{RC*}$) that is estimated via bootstrap simulation with B iterations. The asymptotically valid p-value ($\hat{p}_{RC}$) is directly computed to evaluate the null:

$$T_n^{RC} = \max\left(n^{1/2}\bar{d}_1, \dots, n^{1/2}\bar{d}_K\right), \qquad n = T \tag{4}$$

$$T_{b,n}^{RC*} = \max\left(n^{1/2}\bar{d}_{b,1}^{*}, \dots, n^{1/2}\bar{d}_{b,K}^{*}\right), \qquad n = T, \qquad b = 1..B \tag{5}$$

$$\hat{p}_{RC} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{\left\{T_{b,n}^{RC*} > T_n^{RC}\right\}}\,, \qquad n = T \tag{6}$$

In its original specification, the RC test does not account for transaction costs. One way to consider them would be to compute the ex-post break-even cost for the best TTR in the universe and to compare it with actual market costs, such as in Metghalchi et al. (2012). However, this approach can bias the test in favor of TTRs that have higher cost-free performance, but possibly lower cost-adjusted performance. To correct for this potential problem, we use an adjusted specification that directly incorporates trading costs into the loss function:

$$L(\zeta_t,\ \delta_{k,t-1}) = c_{k,t} - \delta_{k,t-1}\zeta_t, \qquad t = \overline{1,T}, \qquad k = \overline{1,K} \tag{7}$$

Eq. (7) is used instead of eq. (1) and also enables the incorporation of liquidity cost into the analysis, which is often overlooked, even though it can potentially bias results in favor of showing TTR excess performance more often. The trading cost incurred by rule *k* at time *t* is:

$$c_{k,t} = \mathbb{1}_{\{\delta_{k,t-1} \neq \delta_{k,t-2}\}}(0.5\% + l_t), \qquad t = \overline{1,T}, \qquad k = \overline{1,K} \tag{8}$$

The cost is positive when a trade is executed (if the value of the signal function changes) and zero otherwise. When a trade occurs, a fixed broker fee of 0.5% is added to the liquidity cost (l_t), which is defined based on the daily price range:

$$l_t = \begin{cases} \ln\left(\dfrac{H_t}{C_t}\right), \delta_{k,t-1} > 0 \\ \ln\left(\dfrac{C_t}{L_t}\right), \delta_{k,t-1} = 0 \end{cases}, \qquad t = \overline{1,T}, \qquad k = \overline{1,K} \tag{9}$$

Adjusting for the liquidity cost in this way is equivalent to simulating trading at the least favorable prices of the day: buy trades are executed at the maximum price and sell trades are executed at the minimum price. This definition may overestimate the actual liquidity costs incurred by traders, but it is should additionally incorporate the price impact cost, which may be important especially in low liquid markets.

The critical step when performing the RC test consists in estimating the empirical distribution of excess returns using eq. (5). As a bootstrap simulation is used, the empirical distribution estimated to evaluate the null hypothesis asymptotically converges to the distribution of the test statistic. However, its characteristics depend on the choices made regarding the data sample and trading rule universe, which can be considered two sources of risk for the analysis. The risk associated with the choice of data is related to the number of observations and to the ergodicity of the underlying data generating process. Both are exogenous to the researcher but may be controlled by either extending the sample to increase power or, if the process is not ergodic, by employing structural break tests and then performing the test on subsamples for which the process is stable. On the other hand, the size and diversity of the trading rule universe are fully controlled by the researcher and improperly specifying them may introduce data snooping bias in the analysis. Section 4 discusses if and how this happens.

4. Unrepresentative universes and data snooping: a simulation exercise

In this section, we test the hypothesis that restricting the size and diversity of trading rule universes increases the number of false discoveries, i.e. that data snooping bias occurs when the relative performance of forecasting models is tested using unrepresentative universes. The analysis centers on estimating the number of false discoveries of RC tests performed for trading rule universes of varying sizes, on simulated random data on which TTRs should have no superior predictive ability. The *686k* universe defined in Section 3 is considered as the benchmark. Data snooping bias can thus be estimated by the change in false discovery rates between tests that employ small, restricted universes and tests that employ the benchmark.

Six random generated data sets are constructed and used, each based on a discretized zero-drift Geometric Brownian Motion process, assuming different volatility parameters: $\sigma \in \{0.15; 0.20; 0.25; 0.30; 0.35; 0.40\}$. Each data set has 4,000 years of price and volume data, each year consisting of roughly $n = 260$ observations (days). Given an initial fixed price of $C_0 = 1{,}000$, the next day's closing price is $C_t = C_{t-1}e^{\sigma\epsilon_t\sqrt{\tau}}$, the daily price range is $R_t = \sigma C_t \epsilon_t' \sqrt{\tau}$, the high (maximum) price is $H_t = C_t + u_t R_t$, the low (minimum) price is $L_t = C_t - (1 - u_t)R_t$ and the opening price is $O_t = L_t + (H_t - L_t)u_t'$, where ϵ_t and ϵ_t' are independently drawn from a standard normal distribution, u_t and u_t' are independently drawn from a standard uniform distribution and $\tau = n^{-1}$. The daily volume is $V_t = ce^{x_t}$, where $c = 1{,}000$ is fixed and x_t is independently drawn from a standard normal distribution.

The tests are performed on subsamples of predetermined lengths. We consider *1 month, 1 quarter, 1 year* and *4 years* to evaluate potential differences in data snooping bias by sample length. For each data set, a total of 4000 distinct tests are performed, 1000 for each type of sample length[9]. Each test has two stages and proceeds as follows. In the first stage, a single-rule universe is constructed using the "luckiest" rule in *686k*, i.e. the rule that generates the highest excess return relative to the buy-and-hold benchmark rule. Then, its performance is evaluated using the RC test at standard significance levels of 1%, 5%, and 10%. The distribution of the RC test statistic is estimated via the stationary bootstrap procedure of Politis and Romano (1994), by resampling random blocks of data of average length $q = 1/\sqrt[4]{n}$ (this is based on the recommendation of Hall et al., 1995) directly from the excess return series. Resampling blocks of data accounts for the autocorrelation in market returns and it is not particularly useful in this exercise. However, it is useful for the empirical investigation in Section 5 and it's also employed here for consistency. The number of bootstrap iterations is set to $B = 1000$.

In the second stage, consecutively larger rule universes are constructed and tested on the same sample by adding TTRs to the initial single rule universe, until the entire *686k* benchmark universe is tested. New TTRs are added in an order consistent with Table A1 from Appendix A in the supplementary materials. With each additional rule, the distribution of the RC test statistic is re-estimated by resampling an additional 100 times from the excess return series of the new rule[10]. This procedure assures that adding new alternatives does not change the actual trading rule that is evaluated by the RC test, which is always the "luckiest" one in each sample. Instead, the characteristics of the RC distribution used to evaluate the excess performance changes, with the result of potentially influencing test outcomes. Because the

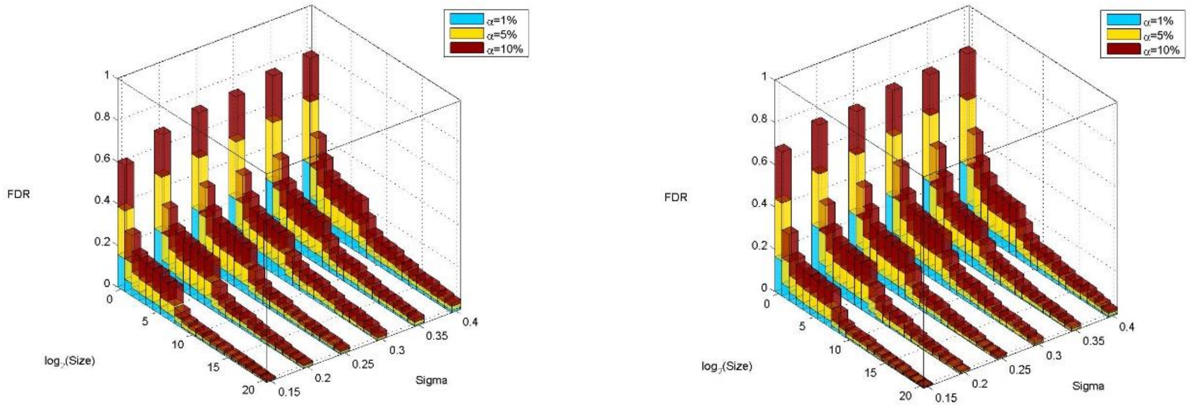
---

[9] When using a sample length of 4 years, all 4,000 years of simulated data are used. When using smaller sample lengths, only the first 1000 periods of that type are used. For example, when using 1 quarter as the sample length, the first 1,000 quarters from the randomly generated data sets are used, which amount to 250 years in this case.

[10] The number of bootstrap iterations is restricted to 100 in the second stage to reduce computational demand, which would otherwise be significant. In a preliminary analysis using a limited sample, 1000 simulations are used to verify the robustness of this choice. The results show that the test outcomes do not materially change.

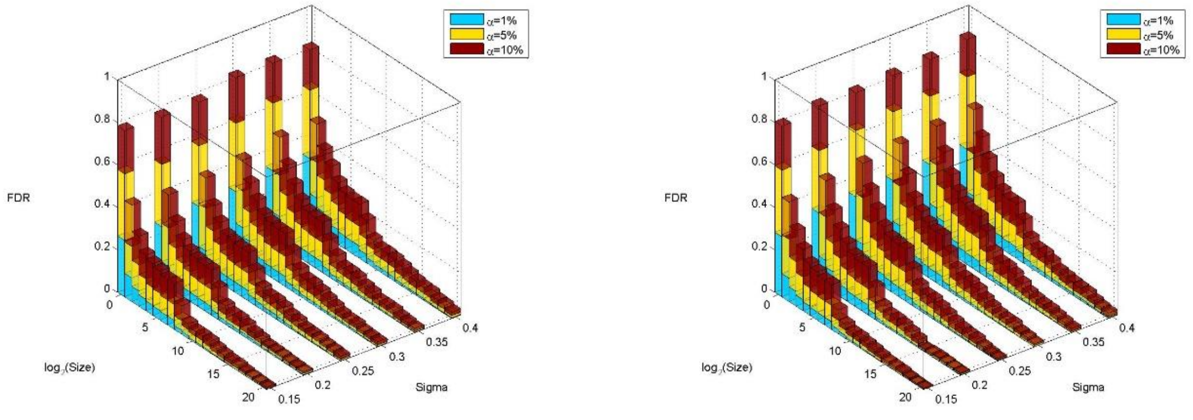
random nature of the data series used, no TTRs should have statistically significant predictive ability, all RC null hypotheses are true and any null rejection constitutes a false discovery[11]. The aim of the exercise is to estimate the data snooping bias arising in RC tests when reducing the size and representativeness of rule universes, controlling for various market conditions and testing assumptions.

Figure 2. False discovery rates (FDR) and the size of the prediction model universes



Sample length: 1 month

Sample length: 1 quarter



Sample length: 1 year

Sample length: 4 years

An overview of false discovery rates estimated in the simulation exercise is shown in Figure 2 for all rule universes with a size of $2^m, m = \overline{1,19}$, alongside the benchmark *686k* universe, which is shown by approximating its size to 2^{20} for illustrative purposes. Numerical

---

[11] As Benjamini and Hochberg (1995) point out, these testing conditions assure that controlling the FWER is equivalent to controlling the FDR. Thus, the simulation provides an estimate of data snooping bias in both types of tests.

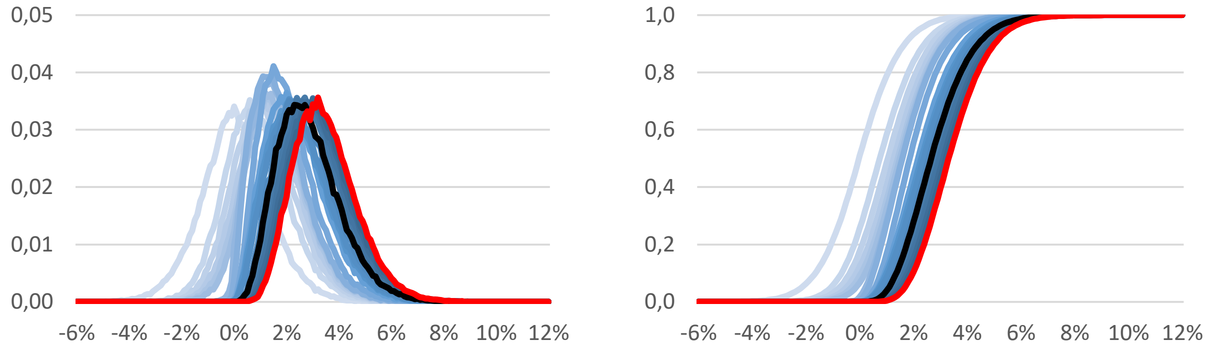
results are shown in Appendix B of the supplementary materials for the largest 11 universes of these, alongside absolute and relative estimates of the amount of data snooping bias. Several interesting findings are worth noting. First, false discoveries significantly increase when the size of the trading rule universes decreases. This shows that data snooping bias does occur in RC-type tests when rule universes are small, unrepresentative. The result is robust to the choice of sample length, the volatility of the data generating process or test significance level. The number of false discoveries becomes especially large for universes that contain less than 2^{11} trading rules but are significantly higher compared to the benchmark *686k* universe even for the other restricted versions. For example, the average difference in false discovery rates between the top-10 largest restricted universes and the benchmark is between 0.39 and 2.62 percentage points depending on test conditions, which amounts to an increase of 180% to 975% in relative terms. This implies that the size and diversity of prediction model universes, particularly their representativeness for the data snooping efforts of others, have a very important influence on the outcomes of tests that examine the relative performance of multiple forecasting models. In particular, using unrepresentative universes biases tests that evaluate TTR excess performance in favor of showing a more favorable picture. This highlights the need to exercise more caution when analyzing existing positive evidence regarding the economic relevance of TTRs. Our results do not invalidate previous findings, but instead, show that additional tests are required to investigate their robustness to changes in the size and diversity of rule universes by accounting for TTRs that investors and researchers use.

Second, the effects of using unrepresentative universes in terms of data snooping bias are larger when the volatility of the data generating process increases, irrespective of the size of the selected universe, subsample length or test significance level. This result shows that evaluating the relative performance of forecasting models in more volatile markets reduces the relevance of test results. Put differently, it shows that prediction models become “luckier” in markets where volatility is high and implies that tests of their superior predictive ability in these conditions are exposed to additional data snooping bias. This triggers additional concerns for existing old and new evidence in favor of TTRs being economically relevant in markets associated with high volatility and low liquidity, such as small-cap sector stocks (e.g., Shynkevich, 2012), emerging stock markets (e.g., Metghalchi et al., 2012), emerging market currencies (e.g., Hsu et al., 2016), or markets in which prices experience persistent declines (bear markets). Moreover, the results trigger questions regarding findings related to "anomalous" asset pricing factors based on technical analysis on portfolios sorted by volatility, such as the one proposed by Han et al. (2013). These and other similar results should be treated with more care until additional robustness tests that control for the correlation between market volatility and data snooping bias are considered.

Third, false discoveries vary with the length of the data sample. Particularly, they increase when the data sample is longer in tests involving universes larger than about 2^{11} trading rules, while the opposite occurs in tests involving universes that contain fewer rules. However, both absolute and relative differences in false discoveries compared to the benchmark universe generally increase with the length of the data sample, irrespective of the size of the rule universe, the volatility of the data generating process or test significance level. This finding shows that the effects of using unrepresentative rule universes in terms of data snooping bias get stronger when longer data samples are used. It implies that TTRs are able to fit more of the noise in the data, are "luckier", in extended samples and that the examination of TTR excess performance should be complemented by robustness tests on shorter time intervals to control for data snooping.

To show how data snooping bias arises in tests of the relative performance of multiple forecasting models, Figure 3 presents how the empirical distribution of the test statistic (representing maximum average excess returns), estimated via bootstrap simulation and used to evaluate the RC null hypothesis, changes with the size of the rule universe for the specific case of $\sigma = 0.30$ and a sample length of *1 year*. Table 2 presents some statistics for the empirical distribution, the average p-value for the associated RC test and the proportion of false discoveries (FD) obtained at standard significance levels.

Figure 3. Size of rule universes and the empirical distribution of the RC test statistic



Panel A. Probability density

Panel B. Cumulative density

NOTE. This figure shows the empirical distribution of the RC test statistic (maximum distributions of excess returns) estimated in RC tests that use rule universes of size $2^m, m = \overline{1,19}$. *The large 686,304 rule universe is distinctively depicted using a red line. An intermediate* 2^{13} *rule universe, comparable in size to what researchers typically use, is depicted using a black line.*

Table 2. The empirical distribution of the test statistic and outcomes of RC tests

Panel A: Characteristics of maximum distribution of excess returns–1000 tests on simulated random data, σ=0.30, sample size=1 year

Size of rule universe	2^{0}	2^{2}	2^{4}	2^{6}	2^{8}	2^{10}	2^{12}	2^{14}	2^{16}	2^{18}	686,304
Average	0.0000	0.0111	0.0150	0.0171	0.0201	0.0243	0.0251	0.0280	0.0297	0.0322	0.0343
Std. Dev.	0.0133	0.0122	0.0118	0.0107	0.0105	0.0119	0.0119	0.0120	0.0118	0.0119	0.0119
Skewness	0.0723	0.3361	0.3736	0.7294	0.8194	0.6509	0.6459	0.6724	0.6365	0.5931	0.5707
Excess Kurtosis	0.4059	0.3360	0.4097	0.6126	0.8198	0.4206	0.4297	0.5201	0.5010	0.4181	0.4052
Jarque-Bera stat.	773.4	2353.4	3025.9	10430.4	13991.5	7797.7	7722.0	8661.3	7798.4	6590.7	6112.7

Panel B: Results for the associated RC tests–1000 tests on simulated random data, σ=0.30, sample size=1 year

Size of rule universe	2^{0}	2^{2}	2^{4}	2^{6}	2^{8}	2^{10}	2^{12}	2^{14}	2^{16}	2^{18}	686,304
Average p-value	0.0527	0.1816	0.2607	0.2928	0.3602	0.4688	0.4882	0.5580	0.6040	0.6660	0.7143
	(0.0019)	(0.0051)	(0.0068)	(0.0075)	(0.0085)	(0.0090)	(0.0090)	(0.0093)	(0.0093)	(0.0091)	(0.0088)
	[27.99]	[35.32]	[38.38]	[38.88]	[42.17]	[52.33]	[54.03]	[60.21]	[64.76]	[73.52]	[81.22]
FD* (1%)	32.6%	10.7%	6.1%	5.7%	4.2%	2.1%	2.0%	1.4%	1.0%	0.5%	0.2%
FD* (5%)	65.3%	26.2%	18.6%	16.7%	11.4%	6.1%	5.7%	4.2%	3.4%	2.4%	1.6%
FD* (10%)	83.6%	42.8%	29.6%	26.3%	20.8%	11.3%	10.3%	7.1%	6.0%	4.4%	3.2%

*Note. Standard errors in round parenthesis; t-statistics in square parenthesis. *FD denotes the proportion of Type I errors (false discoveries).*

The results show that the maximum excess return distribution used to evaluate the RC null hypothesis moves to the right as more TTRs are added to the universe, which is consistent with the hypothesis that the additional rules have residual orthogonality (Arellano-Valle and Genton, 2008; Hartigan, 2014). This implies that trading rules that are used by practitioners should not be considered ex-ante as irrelevant. All distributions are rightly skewed and, thus, not normal. The distribution estimated using *686k* dominates the other alternatives at every quantile, this showing that it has a significantly higher effective span. Put differently, the analysis shows that the distributions shift to the left when TTRs are removed from the benchmark universe. This implies that restricting the size of rule universes by not fully considering what investors and researchers use decreases the effective span (informativeness) of universes, lowers test critical values, generates false discoveries, and biases test results.

Finally, the simulation exercise shows that false discoveries are not eliminated when the extended *686k* rule universe is used in tests, even though they are significantly reduced compared to when using all other restricted alternatives. For example, when using universes that can be considered large for what has been recently used in the literature (such as the ones containing between 2^{13} and 2^{17} rules), false discoveries in RC tests occur 63%-600% more often, depending on the significance level. Nevertheless, this finding generates some important questions for tests of the relative performance of multiple forecasting models. How does the "true", representative universe look like? What would happen if more prediction models would be added to the analysis? When should one stop adding alternatives to the universe? Situations in which the full set of alternatives used by practitioners and researchers is very difficult, if not impossible, to observe, such as in the case of prediction models based on technical analysis, do not allow providing satisfactory answers to these questions. This exposes the associated statistical tests to ambiguity risk and decreases their scientific relevance. It also makes testing for relative model performance subjective, problematic. Because of this, evaluating the absolute performance of forecasting models may provide more objective results and may have some merit in circumstances where the set of alternatives is unclear.

5. Data snooping bias and TTR excess performance: an empirical investigation

In this section, we discuss the results of an empirical exercise designed to evaluate data snooping bias in tests that use real stock market data. We consider daily price and volume data for individual stocks listed in all markets tracked by Thomson Reuters Eikon on November 14, 2013, which have at least 5 listings. In total, there are 81 markets that serve 88 countries. For each, we select up to 40 companies that are part of the main market index. For indices that contain more companies, only 40 of them are randomly select[12]. For indices that contain less, all companies in the index are selected and the list is supplemented (if possible) using other listings in the descending order of their market capitalization. This results in a sample of 2579 stocks, for which all available historical trading price and volume data are retrieved up to November 14, 2013. A summary of the data sample is presented in Appendix C in the supplementary materials.

The empirical exercise first evaluates the excess performance of TTRs that are part of the benchmark *686k* rule universe using the RC test. The tests are performed on non-overlapping one-year intervals for all stocks in the sample to enable the investigation of temporal variations in data snooping bias. The results in Section 4 show that intermediate amounts of data snooping bias are expected when this sample length is used. Subsamples that have less than 65 observations are excluded because of insufficient liquidity. To estimate the data snooping bias associated with restricting the size and diversity of the rule universes, the 44 small, unrepresentative universes formed using individual technical analysis indicators are also tested and the results are compared to the benchmark. As shown in Section 2, these universes have similar characteristics to the ones that are typically employed in the literature, in terms of both size and effective span. If data snooping biases the results of tests that use real stock market data, then the null rejection rates should be higher in the tests that use the restricted universes compared to the ones that use the benchmark. Further, the differences would be an indication of the number of false discoveries that is due to data snooping.

---

[12] The companies are ordered by name and then every *[N/40]* in the list is drawn, where N represents the total number of stocks in the index and *[x]* represents the integer part of *x*.

5.1.Data snooping bias and TTR performance in stock markets around the world

In total, 34,887 tests are performed for the *686k* universe and 1,535,116 tests are performed for the rule universes constructed using individual indicators. Table 3 provides a summary of the results. In the case of the *686k* universe, which is presented in Panel A, prediction models derived from technical analysis indicators generate positive cost adjusted excess returns in 34,678 tests, which amount to 99.4% of the total. However, when considering statistical significance, the RC null hypothesis is rejected only 227 times (0.65% of the total) at the 10% level, 96 times (0.27% of the total) at the 5% level, and 14 times (0.04% of the total) at the 1% level. The results obtained using the restricted rule universes, which are reported in Panel B, show that the RC null hypothesis is rejected 13,525 times (0.88% of the total) at the 10% level, 5,725 times (0.37% of the total) at the 5% level, and 1,132 times (0.07% of the total) at the 1% level. In all cases, the rate of null rejections is about two times higher compared to the benchmark results, with the effect being stronger as we decrease the confidence level. Null rejections are inflated 2 times at the 1% level, 1.81 times at the 5% level, and 1.8 times at the 10% level. These results show that data snooping arising from using small, unrepresentative universes in RC tests inflates the number of null rejections and, thus, biases results at any selected confidence level. Ultimately, it causes a significant amount of false discoveries and incorrectly skews conclusions in favor of showing that TTRs display superior predictive ability.

Results grouped by both stock and year are reported in Panel C of Table 3. This enables the analysis of instances when at least one of the 44 tests that use unrepresentative rule universes rejects the null hypothesis and evaluates the extreme scenario in which independent researchers test different rule universes on the same data sample without controlling for data snooping and then make inferences based on a meta-analysis of their results. The RC null hypothesis is rejected at least once for 773 stock-years (2.21% of the total) at the 10% level, 337 stock-years (0.96% of the total) at the 5% level, and 70 stock-years (0.20% of the total) at the 1% level.

This shows that considering the positive results of others without considering their data snooping efforts can significantly increase the number of false discoveries. Particularly, the evidence in favor of TTRs being economically profitable is inflated 7.93 times at the 1% level, 6.61 times at the 5% level, and 6.16 times at the 10% level.

Table 3. Summary statistics of RC test results on real stock market data

Panel A. Results when using 686k in tests

Statistic	Value	Percent of total
i. Number of tests	34,887	100.00%
ii. Number of tests in which TTRs obtained positive excess returns	34.678	99.40%
iii. Number of tests with statistically significant positive excess returns (RC test)		
10% confidence level	*227*	*0.65%*
5% confidence level	*96*	*0.27%*
1% confidence level	*14*	*0.04%*
iv. Likelihood of TTRs to repeat positive excess returns⊥		
*Conditional on TTR indicator class***	*2284*	*6.54%*
*Conditional on TTR indicator class and strategy****	*1572*	*4.50%*
*Conditional on TTR indicator class, strategy and parameters*****	*21*	*0.06%*
v. Likelihood of TTRs to repeat significant performance (10% significance)†		
*Unconditional**	*5*	*0.01%*
*Conditional on TTR indicator class***	*1*	*0.00%*
*Conditional on TTR indicator class and strategy****	*0*	*0.00%*
*Conditional on TTR indicator class, strategy and parameters*****	*0*	*0.00%*

Panel B. Results when using the 44 small, unrepresentative rule universes in tests

Statistic	Value	Percent of total
i. Number of valid tests	1,534,970	100.00%
ii. Number of tests in which TTRs obtained positive excess returns	1,071,904	69.83%
iii. Number of tests with statistically significant positive excess returns (RC test)		
10% confidence level	*18,132*	*1.18%*
5% confidence level	*7,867*	*0.51%*
1% confidence level	*1,242*	*0.08%*
iv. Likelihood of TTRs to repeat best performance⊥		
*Conditional on TTR indicator class, strategy and parameters*****	*74,502*	*4.85%*
v. Likelihood of TTRs to repeat significant performance (10% significance)†		
*Unconditional**	*496*	*0.03%*
*Conditional on TTR indicator class, strategy and parameters*****	*28*	*0.00%*

Panel C. Results when testing the 44 restricted rule universes–aggregated at the stock-year level

Statistic	Value	Percent of total
i. Number of subsamples	34,887	100.00%
ii. Number of tests in which TTRs obtained positive excess returns	25,878	74.17%
iii. Number of tests with statistically significant positive excess returns (RC test)		
10% confidence level	*1,399*	*4.01%*
5% confidence level	*635*	*1.82%*
1% confidence level	*111*	*0.31%*
iv. Likelihood of TTRs to repeat best performance⊥		
*Conditional on TTR indicator class***	*6,965*	*19.96%*
*Conditional on TTR indicator class and strategy****	*6,866*	*19.68%*
*Conditional on TTR indicator class, strategy and parameters*****	*1,555*	*4.45%*
v. Likelihood of TTRs to repeat significant performance (10% significance)†		
*Unconditional**	*70*	*0.20%*
*Conditional on TTR indicator class***	*7*	*0.02%*
*Conditional on TTR indicator class and strategy*****	*7*	*0.02%*
*Conditional on TTR indicator class, strategy and parameters*****	*5*	*0.01%*

*NOTE: ⊥The number of tests in which TTRs outperform the benchmark in two consecutive years, expressed as a percent of the total number of tests. †The number of tests in which the null hypothesis of no economic profitability is rejected at the 10% level in two consecutive years, expressed as a percent of the total number of tests. *Unconditional–estimated for any TTR in the rule universe. **Conditional on TTR indicator class–estimated for TTRs that are based on the same technical analysis indicator (entry/exit strategy, parameter values might vary). ***Conditional on TTR indicator class and strategy–estimated for TTRs that are based on the same technical analysis indicator and the same entry/exit strategy (parameter values might vary). ****Conditional on TTR indicator class, strategy, and parameters– estimated for TTRs that are identical in terms of all aspects, including parameter values.*

Results grouped by rule universe are reported in Table 4. For the restricted universes at the 10% significance level, rejection rates vary from a minimum of 0.22% to a maximum of

2.99%, with a median (mean) of 1.19% (1.18%). This is significantly higher compared to the null rejection rate obtained when using the benchmark *686k* rule universe, which is 0.65%. This pattern replicates when analyzing test results at the 5% and 1% levels: both median and average rejection rates are 1.8-2 times higher when the restricted universes are used, even though the rules in the benchmark generate positive excess returns more often. The maximum difference is recorded for the smallest rule universe, which is the one generated by the Runs Indicator. In this case, rejection rates are 4.6-7.3 higher, depending on the significance level. There are 6 cases in which rejection rates are lower for restricted rule universes compared to the benchmark. However, these occur only for the least profitable indicators, which generate positive excess returns less than half of the time. This implies that rejection rates are low in tests that use them not because of lower data snooping, but because they are not able to consistently predict price movements. All other results show that null rejections in tests that use restricted rule universes are significantly higher compared to those that use the benchmark. As rule universes that are typically used in the literature resemble the restricted ones that are used here, while the benchmark *686k* universe should be more representative, our results show that data snooping has a significant influence on the outcomes of tests that examine the relative performance of prediction models in empirical setups. Specifically, data snooping biases test results and skews conclusions in favor of the showing that prediction models perform better than they truly do.

Table 4. Null Rejection Rates aggregated by trading rule universe

Trading rule universe	**rPR***	**NRR, α=0.10**	**NRR, α=0.05**	**NRR, α=0.01**
Accumulation Swing Index	26.14%	1.72%	0.66%	0.07%
Arms Ease of Movement	24.92%	0.78%	0.35%	0.07%
Aroon Oscillator	88.28%	1.41%	0.65%	0.08%
Balance of Market Power	83.45%	1.20%	0.49%	0.06%
Bollinger Oscillator	76.55%	1.36%	0.57%	0.08%
Center of Gravity Oscillator	44.17%	0.25%	0.08%	0.00%
Chaikin Money Flow	89.51%	1.21%	0.54%	0.08%
Chaikin Oscillator	68.94%	1.03%	0.45%	0.07%
Chande Momentum Oscillator	93.03%	1.29%	0.55%	0.09%
Commodity Channel Index	46.03%	1.00%	0.41%	0.08%
Runs Indicator	33.23%	2.99%	1.42%	0.29%
Demand Index	87.24%	1.61%	0.74%	0.10%
Detrended Price Oscillator	81.79%	0.79%	0.35%	0.07%
Dynamic Momentum Index	96.58%	0.80%	0.32%	0.04%
Filter	50.70%	2.04%	0.87%	0.16%
Inertia Indicator	92.64%	1.30%	0.57%	0.08%
Kase Convergence Divergence	97.91%	1.43%	0.62%	0.08%
Kase Peak Oscillator	94.66%	0.98%	0.41%	0.06%
Klinger Volume Oscillator	46.99%	0.22%	0.09%	0.01%

Know Sure Thing	80.78%	1.15%	0.47%	0.06%
Linear Regression Slope	66.59%	1.04%	0.47%	0.06%
Market Volume Impact	36.71%	0.40%	0.17%	0.03%
Money Flow Index	87.85%	1.33%	0.57%	0.08%
Moving Average Convergence Divergence	88.01%	1.02%	0.43%	0.06%
New Relative Volatility Index	90.47%	1.10%	0.49%	0.05%
On Balance Volume	8.85%	0.47%	0.22%	0.04%
Plus DM vs. Minus DM crossover	57.29%	1.63%	0.69%	0.12%
PI Opinion Oscillator	84.45%	1.29%	0.55%	0.08%
Polarized Fractal Efficiency	95.40%	1.30%	0.52%	0.08%
Random Walk Index for High prices	50.43%	0.97%	0.42%	0.08%
Rate of Change	70.23%	0.83%	0.37%	0.06%
Relative Momentum Index	92.47%	1.63%	0.72%	0.11%
Relative Strength Index	84.86%	1.33%	0.55%	0.09%
Relative Vigor Index	94.78%	1.73%	0.75%	0.11%
Relative Volatility Index	87.84%	1.13%	0.47%	0.07%
Stochastic Momentum Index	92.97%	1.77%	0.76%	0.11%
Stochastic Oscillator	57.48%	0.93%	0.43%	0.07%
Stochastic RSI Oscillator	57.10%	0.93%	0.38%	0.05%
The Quantitative Candlestick	20.76%	0.55%	0.23%	0.04%
Triple Exponential Smoothing	75.74%	1.50%	0.70%	0.08%
True Strength Index	89.57%	1.64%	0.72%	0.11%
Ultimate Oscillator	79.49%	1.18%	0.51%	0.07%
Vortex Oscillator	89.43%	1.19%	0.48%	0.06%
Williams Variable Accumulation Distribution	10.20%	0.28%	0.11%	0.01%
SUMMARY RESULTS FOR THE 44 RESTRICTED RULE UNIVERSES				
Minimum	*8.85%*	*0.22%*	*0.08%*	*0.00%*
Maximum	*97.91%*	*2.99%*	*1.42%*	*0.29%*
Median	*81.29%*	*1.19%*	*0.49%*	*0.07%*
Average	*69.83%*	*1.18%*	*0.51%*	*0.08%*
Std. Deviation	*25.88%*	*0.51%*	*0.23%*	*0.04%*
BENCHMARK RESULTS–686k	*99.40%*	*0.65%*	*0.27%*	*0.04%*

NOTE. *rPR is the rate of positive returns, which is defined as the number of tests for which the excess average return of the best rule was higher compared to the benchmark, divided by the total number of tests performed. The Null Rejection Rate (NRR) is the number of tests that reject the null hypothesis of no economic profitability at the α confidence level, expressed as a percentage of the total number of tests performed.

Our results also allow reconsidering the economic profitability of TTRs while controlling for the data snooping bias arising from using small, unrepresentative universes. The tests that use the extended *686k* rule universe show that there are very few instances when TTRs have economic relevance and that the null rejection rates (Table 3, Panel A, section iii) are similar and even lower compared to the ones reported in Section 4 for tests that employ the same universe and same sample length, but random generated data (Table B3 in the supplementary materials). This can be observed irrespective of the choice for confidence level and volatility parameter and shows that the positive results are within the bounds of randomness. Additionally, the likelihood of TTRs replicating significant performance in consecutive periods is close to zero (Table 3, Panel A, section v), which shows that even if they can generate economic profits from time to time, this is not a robust property. From a theoretical perspective, the results show that deviations from market efficiency are rare, insignificant and most likely random. This implies that stock markets are efficient at pricing information obtained using technical analysis indicators and supports the weak-form Efficient Market Hypothesis of Fama (1970). From a practical perspective, the results show that TTRs do not have any economic relevance and are not able to help investors earn significant, systematic excess returns from trading on stock markets around the world.

5.2.Data snooping bias and TTR performance in restricted samples

To get a better understanding of the impact of data snooping bias from using small, unrepresentative universes in limited data samples, the results are grouped and analyzed by year and by stock market. As before, null rejection rates are compared for the *686k* universe, the 44 restricted universes and the restricted universes grouped by both stock and year.

Table 5. Null Rejection Rates aggregated by year

	Number of tests	***686k***			**Restricted rule universes**			**Restricted rule universes (aggregated by stock-year)**		
Year	**using 686k**	**NRR, α=0.10**	**NRR, α=0.05**	**NRR, α=0.01**	**NRR, α=0.10**	**NRR, α=0.05**	**NRR, α=0.01**	**NRR, α=0.10**	**NRR, α=0.05**	**NRR, α=0.01**
1979	4	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1980	18	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1981	86	2.33%	1.16%	1.16%	2.69%	2.16%	1.10%	8.13%	3.48%	1.16%
1982	94	0.00%	0.00%	0.00%	1.28%	0.14%	0.00%	5.31%	2.12%	0.00%
1983	97	0.00%	0.00%	0.00%	0.39%	0.00%	0.00%	1.03%	0.00%	0.00%
1984	133	2.26%	0.75%	0.00%	2.90%	1.65%	0.11%	4.51%	3.00%	1.50%
1985	202	0.50%	0.00%	0.00%	0.50%	0.20%	0.00%	1.48%	0.49%	0.00%
1986	214	0.00%	0.00%	0.00%	0.10%	0.00%	0.00%	1.40%	0.00%	0.00%
1987	247	0.81%	0.00%	0.00%	1.02%	0.53%	0.01%	2.83%	0.80%	0.80%
1988	278	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1989	293	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
1990	343	2.04%	1.17%	0.00%	3.20%	1.68%	0.29%	9.32%	4.37%	1.16%
1991	406	0.00%	0.00%	0.00%	0.22%	0.00%	0.00%	1.97%	0.00%	0.00%
1992	461	0.65%	0.22%	0.00%	1.19%	0.45%	0.00%	4.77%	2.16%	0.00%
1993	560	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.17%	0.00%	0.00%
1994	660	0.45%	0.15%	0.00%	0.79%	0.22%	0.00%	2.72%	0.75%	0.00%
1995	759	1.05%	0.13%	0.00%	1.29%	0.30%	0.00%	3.42%	1.71%	0.00%
1996	907	0.44%	0.44%	0.00%	0.87%	0.48%	0.07%	3.19%	1.32%	0.33%
1997	997	0.60%	0.10%	0.00%	1.35%	0.46%	0.02%	4.91%	2.00%	0.20%
1998	1083	1.29%	0.37%	0.00%	1.95%	0.79%	0.04%	6.18%	3.13%	0.46%
1999	1159	0.26%	0.26%	0.00%	0.42%	0.25%	0.00%	1.63%	0.60%	0.08%
2000	1265	1.26%	0.79%	0.24%	2.33%	1.14%	0.28%	8.14%	3.55%	0.86%
2001	1353	0.81%	0.22%	0.00%	1.19%	0.39%	0.04%	3.47%	2.06%	0.22%
2002	1432	0.28%	0.00%	0.00%	0.94%	0.36%	0.00%	4.60%	1.53%	0.13%
2003	1535	0.00%	0.00%	0.00%	0.04%	0.00%	0.00%	0.58%	0.13%	0.00%
2004	1620	0.00%	0.00%	0.00%	0.02%	0.00%	0.00%	0.30%	0.12%	0.06%
2005	1686	0.00%	0.00%	0.00%	0.05%	0.00%	0.00%	0.83%	0.29%	0.05%
2006	1784	0.22%	0.00%	0.00%	0.33%	0.08%	0.00%	1.56%	0.56%	0.05%
2007	1941	0.10%	0.10%	0.05%	0.22%	0.10%	0.06%	1.08%	0.46%	0.10%
2008	2043	4.89%	2.40%	0.34%	8.53%	4.05%	0.77%	23.25%	12.28%	2.59%
2009	2102	0.10%	0.00%	0.00%	0.06%	0.00%	0.00%	0.57%	0.14%	0.00%
2010	2223	0.09%	0.00%	0.00%	0.31%	0.11%	0.00%	1.66%	0.67%	0.08%
2011	2283	0.74%	0.31%	0.04%	1.97%	0.73%	0.07%	8.49%	3.54%	0.43%
2012	2308	0.35%	0.04%	0.00%	0.57%	0.23%	0.00%	2.38%	0.95%	0.12%
2013	2311	0.22%	0.13%	0.04%	0.34%	0.15%	0.05%	1.29%	0.51%	0.08%
Minimum		0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Maximum		4.89%	2.40%	1.16%	8.53%	4.05%	1.10%	23.25%	12.28%	2.59%
Median		*0.26%*	*0.00%*	*0.00%*	*0.50%*	*0.20%*	*0.00%*	*1.97%*	*0.67%*	*0.06%*
Average		*0.62%*	*0.25%*	*0.05%*	*1.06%*	*0.48%*	*0.08%*	*3.46%*	*1.51%*	*0.30%*
Std. Deviation		*0.99%*	*0.49%*	*0.20%*	*1.58%*	*0.82%*	*0.23%*	*4.35%*	*2.27%*	*0.56%*

NOTE. The Null Rejection Rate (NRR) *is the number of tests that reject the null hypothesis of no economic profitability at the* α *confidence level, expressed as a percentage of the total number of tests performed.*

The results aggregated by year are reported in Table 5 and show that RC tests that use the benchmark *686k* rule universe reject the null hypothesis less often compared to tests that use restricted rule universes in any time interval. Depending on the year and on the way the analysis is conducted, the excess performance of TTRs in tests that use unrepresentative rule universes are inflated by up to 24 times. For example, null rejections increase 2.26 times in 2008 at the 1% level, 5.75 times in 2012 at the 5% level and 3.44 times in 2010 at the 10% level. If more researchers conduct independent tests using unrepresentative rule universes and then draw conclusions based on a meta-analysis of their results, false discoveries increase by 975% in 2011 at the 1% level, 2275% in 2012 at the 5% level and 1744% in 2010 at the 10% level. Overall, the results show that data snooping biases test results in favor of TTR excess performance, irrespective of the period in which the analysis is performed.

Implications for the discussion regarding market efficiency can be obtained from analyzing the results obtained using the *686k* universe, which show temporal variations in the excess performance of TTRs in our sample. Periods in which prediction models derived from technical analysis indicators have low success rates (stock markets are more efficient) relate to calm and favorable (positive) market conditions, while periods in which strategies are able to earn economically significant excess returns (stock markets are less efficient) relate to periods of financial, macroeconomic, social instability. The most successful year for TTRs is by far 2008, the climax of the most recent financial crisis. About half of all RC null rejections originate in this year alone. Other periods of financial market instability rank high, such as the European sovereign debt crisis around 2011, the dot-com bubble burst at the beginning of the current millennia, or the Asian financial crisis around 1998. Also, the average excess returns earned by the best performing trading strategies and the null rejection rates of RC tests are lower in the first half of the sample, which generally corresponded to a period of more stable and rising markets. These results are consistent with existing evidence that find rising return predictability when prices decline (e.g., Lim and Brooks, 2011) and seem to support the Adaptive Market Hypothesis of Lo (2004). They also hint that TTRs may have some merit as a risk management aid in timing exit points around the onset of a bear market. However, the results in Section 4 show that data snooping bias rises when markets are more volatile. Thus, our tests cannot distinguish between true significant results and false discoveries, as the null rejection rates are comparable to the ones obtained using similar conditions in the simulation exercise (Table B3 in the supplementary materials). Further, null rejection rates are very low from an economic perspective even in 2008, not exceeding 4.89% at the 10% level, 2.4% at the 5% level and 1.16% at the 1% level. As a result, we argue that these results rather support the Efficient Market Hypothesis and reinforce the conclusion that TTRs lack economic relevance when used for trading on stock markets around the world.

Table 6. Null Rejection Rates aggregated by stock market

	Number of tests using	**Extended rule universe (686k)**			**Restricted rule universes**			**Restricted rule universes (aggregated by stock-year)**		
Market	**686k**	**α=0.10**	**α=0.05**	**α=0.01**	**α=0.10**	**α=0.05**	**α=0.01**	**α=0.10**	**α=0.05**	**α=0.01**
AE	295	0.34%	0.00%	0.00%	1.64%	0.56%	0.01%	7.79%	2.37%	0.33%
AR	600	1.50%	0.17%	0.00%	2.10%	0.65%	0.00%	5.50%	3.00%	0.16%
AT	301	0.00%	0.00%	0.00%	0.27%	0.06%	0.00%	3.32%	0.66%	0.33%
AU	865	0.00%	0.00%	0.00%	0.08%	0.00%	0.00%	1.27%	0.23%	0.00%
BA	75	9.33%	6.67%	5.33%	10.46%	7.74%	5.43%	16.00%	12.00%	8.00%
BE	352	0.28%	0.00%	0.00%	0.85%	0.47%	0.01%	2.27%	1.13%	0.56%
BG	143	2.10%	0.70%	0.00%	4.96%	2.93%	0.15%	7.69%	6.29%	1.39%
BH	260	3.85%	2.31%	0.38%	4.89%	2.62%	0.81%	12.30%	6.92%	1.92%
BR	574	0.17%	0.17%	0.00%	0.34%	0.19%	0.00%	1.56%	0.52%	0.00%
BRVM	120	0.83%	0.00%	0.00%	0.45%	0.01%	0.00%	3.33%	0.83%	0.00%
CA	1060	0.00%	0.00%	0.00%	0.02%	0.00%	0.00%	0.75%	0.00%	0.00%
CH	342	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.29%	0.00%	0.00%
CL	532	1.13%	0.75%	0.00%	2.16%	0.85%	0.29%	7.14%	3.00%	0.93%
CN	349	0.00%	0.00%	0.00%	0.52%	0.08%	0.00%	6.01%	2.29%	0.00%
CO	168	1.19%	0.00%	0.00%	0.96%	0.64%	0.00%	2.38%	1.19%	0.00%
CY	222	0.00%	0.00%	0.00%	3.53%	0.92%	0.04%	11.26%	4.95%	0.00%
CZ	136	0.00%	0.00%	0.00%	0.18%	0.06%	0.01%	0.73%	0.00%	0.00%
DE	662	0.00%	0.00%	0.00%	0.04%	0.01%	0.00%	0.90%	0.30%	0.00%
DK	419	0.00%	0.00%	0.00%	0.19%	0.02%	0.00%	2.62%	0.71%	0.00%
EE	154	2.60%	1.95%	0.00%	4.06%	2.72%	0.42%	6.49%	4.54%	1.94%
EG	325	2.77%	1.54%	0.00%	3.99%	1.79%	0.17%	12.92%	5.53%	0.61%
ES	578	0.00%	0.00%	0.00%	0.15%	0.03%	0.00%	2.59%	0.86%	0.00%
FI	436	0.23%	0.00%	0.00%	0.39%	0.05%	0.00%	2.98%	1.60%	0.00%
FR	926	0.11%	0.00%	0.00%	0.20%	0.09%	0.00%	1.29%	0.43%	0.10%
GR	624	1.12%	0.00%	0.00%	2.06%	0.78%	0.03%	8.17%	3.04%	0.16%
HK	758	0.00%	0.00%	0.00%	0.22%	0.02%	0.00%	0.65%	0.52%	0.00%
HR	214	0.47%	0.00%	0.00%	2.12%	0.75%	0.07%	7.94%	3.73%	0.00%
HU	183	0.00%	0.00%	0.00%	0.44%	0.11%	0.00%	2.73%	0.54%	0.00%
ID	567	0.71%	0.35%	0.00%	1.08%	0.56%	0.02%	3.35%	1.05%	0.35%
IE	412	0.73%	0.49%	0.00%	1.95%	0.94%	0.04%	5.58%	2.91%	0.48%
IL	664	0.30%	0.00%	0.00%	1.14%	0.27%	0.00%	3.61%	1.50%	0.00%
IN	470	0.00%	0.00%	0.00%	0.08%	0.00%	0.00%	1.70%	0.42%	0.00%
IQ	153	2.61%	0.00%	0.00%	2.80%	1.27%	0.00%	7.84%	3.92%	0.00%
IS	59	1.69%	1.69%	0.00%	1.57%	1.27%	0.00%	3.38%	3.38%	0.00%
IT	566	0.00%	0.00%	0.00%	0.21%	0.03%	0.00%	3.71%	1.23%	0.00%
JO	487	0.21%	0.00%	0.00%	1.32%	0.22%	0.00%	5.95%	1.64%	0.20%
JP	981	0.00%	0.00%	0.00%	0.07%	0.00%	0.00%	1.01%	0.20%	0.00%
KE	560	0.00%	0.00%	0.00%	0.75%	0.10%	0.00%	6.78%	2.14%	0.00%

KR	705	0.00%	0.00%	0.00%	0.02%	0.00%	0.00%	0.85%	0.14%	0.00%
KW	572	0.17%	0.00%	0.00%	1.52%	0.31%	0.00%	5.41%	2.62%	0.17%
KZ	45	4.44%	2.22%	2.22%	4.74%	2.52%	1.91%	11.11%	8.88%	2.22%
LB	115	0.00%	0.00%	0.00%	1.87%	0.98%	0.11%	6.95%	4.34%	2.60%
LK	609	1.64%	0.82%	0.00%	2.36%	1.29%	0.17%	6.23%	3.77%	0.98%
LT	257	3.11%	2.33%	0.00%	4.16%	2.50%	0.60%	8.17%	4.66%	1.55%
LV	182	3.85%	2.75%	1.10%	4.08%	3.12%	1.36%	6.04%	4.39%	2.74%
MA	400	0.50%	0.25%	0.25%	1.34%	0.61%	0.18%	4.00%	2.25%	0.50%
MU	147	0.68%	0.00%	0.00%	2.21%	0.54%	0.00%	10.20%	2.72%	0.00%
MX	485	0.62%	0.21%	0.00%	0.93%	0.56%	0.00%	3.09%	1.44%	0.20%
MY	963	0.52%	0.00%	0.00%	0.83%	0.32%	0.00%	3.32%	1.34%	0.20%
NA	193	1.04%	0.00%	0.00%	0.83%	0.29%	0.00%	4.14%	2.59%	1.55%
NG	217	2.30%	0.00%	0.00%	2.11%	0.47%	0.00%	5.99%	2.76%	0.00%
NL	833	0.84%	0.12%	0.00%	1.00%	0.52%	0.09%	2.52%	1.44%	0.36%
NO	622	0.32%	0.16%	0.00%	0.65%	0.12%	0.00%	3.37%	1.28%	0.00%
NZ	547	1.10%	0.73%	0.00%	1.32%	0.89%	0.12%	3.10%	1.64%	0.91%
OM	381	3.15%	1.84%	0.00%	3.85%	1.95%	0.11%	10.49%	5.24%	0.78%
PE	492	3.05%	1.42%	0.41%	4.44%	1.83%	0.43%	10.56%	5.28%	0.81%
PH	505	0.79%	0.40%	0.00%	1.75%	0.62%	0.07%	5.34%	2.57%	0.19%
PK	538	0.56%	0.19%	0.00%	1.97%	0.60%	0.02%	5.94%	2.60%	0.18%
PL	364	0.55%	0.27%	0.00%	0.55%	0.37%	0.01%	1.64%	0.54%	0.27%
PT	545	0.37%	0.00%	0.00%	2.09%	0.66%	0.00%	8.44%	4.22%	0.00%
QA	407	0.25%	0.00%	0.00%	0.64%	0.25%	0.01%	2.94%	0.98%	0.24%
RO	458	1.53%	0.87%	0.00%	2.93%	1.50%	0.15%	6.11%	3.93%	1.09%
RS	102	1.96%	0.00%	0.00%	5.35%	1.50%	0.04%	13.72%	7.84%	0.00%
RU	149	3.33%	2.67%	0.00%	4.24%	2.34%	0.12%	8.72%	5.36%	1.34%
SA	415	0.00%	0.00%	0.00%	0.44%	0.12%	0.00%	3.37%	2.16%	0.00%
SE	666	0.00%	0.00%	0.00%	0.26%	0.07%	0.00%	1.95%	0.30%	0.00%
SG	686	0.15%	0.00%	0.00%	0.53%	0.19%	0.01%	2.18%	0.87%	0.14%
SI	82	2.44%	1.22%	0.00%	5.70%	3.70%	1.24%	13.41%	9.75%	3.65%
SK	72	0.00%	0.00%	0.00%	0.85%	0.00%	0.00%	4.16%	0.00%	0.00%
TH	701	0.71%	0.29%	0.14%	1.36%	0.58%	0.14%	5.13%	2.28%	0.28%
TN	313	0.00%	0.00%	0.00%	0.30%	0.00%	0.00%	2.55%	0.31%	0.00%
TR	636	0.31%	0.00%	0.00%	0.42%	0.14%	0.00%	1.41%	0.62%	0.00%
TW	729	0.00%	0.00%	0.00%	0.18%	0.01%	0.00%	2.19%	0.54%	0.00%
TZ	69	1.45%	0.00%	0.00%	3.22%	0.79%	0.09%	14.49%	5.79%	4.34%
UA	153	7.19%	1.96%	0.00%	8.49%	3.59%	0.11%	26.79%	14.37%	1.30%
UK	821	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.48%	0.00%	0.00%
US	1072	0.00%	0.00%	0.00%	0.04%	0.00%	0.00%	0.65%	0.18%	0.00%
VE	138	0.00%	0.00%	0.00%	1.79%	1.18%	0.08%	2.89%	2.89%	1.44%
VN	227	4.85%	3.52%	0.88%	6.89%	4.09%	1.31%	13.65%	8.81%	3.96%
ZA	682	0.15%	0.00%	0.00%	0.06%	0.00%	0.00%	0.73%	0.14%	0.00%
Minimum		*0.00%*	*0.00%*	*0.00%*	*0.01%*	*0.00%*	*0.00%*	*0.29%*	*0.00%*	*0.00%*
Maximum		*9.33%*	*6.67%*	*5.33%*	*10.46%*	*7.74%*	*5.43%*	*26.79%*	*14.37%*	*8.00%*
Median		*0.42%*	*0.00%*	*0.00%*	*1.11%*	*0.50%*	*0.00%*	*3.86%*	*2.16%*	*0.15%*
Average		*1.10%*	*0.51%*	*0.13%*	*1.83%*	*0.87%*	*0.20%*	*5.45%*	*2.77%*	*0.64%*
(Std. Deviation)		*1.67%*	*1.07%*	*0.66%*	*2.05%*	*1.26%*	*0.68%*	*4.56%*	*2.83%*	*1.26%*

NOTE. The Null Rejection Rate (NRR) *is the number of tests that reject the null hypothesis of no economic profitability at the* α *confidence level, expressed as a percentage of the total number of tests performed.*

The results aggregated by stock market are reported in Table 6. When analyzing the excess performance of TTRs in the *686k* universe, they show that some asymmetries also exist between the different markets in our sample. At the 10% confidence level, no RC null rejections occur for 26 stock markets, which are mainly developed ones. On other hand, TTRs earn excess returns 9.33% of the time in Bosnia and Herzegovina, 7.19% of the time in Ukraine, 4.85% of the time in Vietnam, 4.44% of the time in Kazakhstan, and at rates between 1% and 4% in other 50 (mainly developing) stock markets. As we lower the confidence level towards 1%, TTRs become unprofitable in all but 8 stock markets. At the intermediate 5% confidence level, TTRs are profitable in 31 of the 80 markets, with success rates ranging between 0.12% and 2.75%; two outliers exist–Vietnam with 3.52% and Bosnia and Herzegovina with 6.67%. Overall, our findings agree with the literature showing that TTRs are not relevant in developed markets and are more informative and more profitable in smaller, less developed stock markets (countries). However, our analysis shows that previous results should be treated with care, as they may be impacted by data snooping bias do not necessarily imply that some markets are not efficient. First, given that volatility tends to be higher in emerging and frontier markets, TTRs are "luckier" and a higher rate of RC null rejections is expected. Second, the null rejection rates obtained in tests using *686k* are mostly similar or lower compared to the ones obtained in the simulation exercise presented in Section 4, with the exception of only 8 markets, which are very small and have important trading barriers for investors; this increases the likelihood that any superior information gained from prediction models based on technical analysis may not be used in actual trading. Third, the rates of null rejections do not surpass 10% even in these markets and are generally low from an economic perspective. Because of this, we argue that these results are explained by both the Efficient Market Hypothesis and the data snooping bias arising in tests of the relative performance of multiple prediction models from using small, unrepresentative rule universes. Our conclusion thus departs from views in the literature that support the excess performance of TTRs even in small, less developed markets.

Comparing the tests conducted using restricted rule universes to the ones that use the benchmark shows that data snooping is the factor that causes this disagreement. Specifically, RC tests that use the *686k* universe reject the null hypothesis less often for all markets in the sample, and the excess performance of TTRs is significantly inflated in some markets. For example, positive discoveries for Bahrain increase 2.13 times at the 1% level in tests that use the restricted rule universes, compared to the ones that use *686k*. Similarly, positive discoveries for The Netherlands increase 4.33 times at the 5% level, while positive discoveries for Kuwait increase 8.94 times at the 10% level. In the extreme, but not entirely unrealistic scenario in which researchers conduct independent tests using unrepresentative universes and then draw conclusions based on a meta-analysis of the reported results, positive discoveries for Bahrain increase 5.05 times at the 1% level, positive discoveries for Argentina increase 17.65 times at the 5% level, while positive discoveries for Kuwait increase 31.82 times at the 10% level. Overall, the results show that data snooping is an important factor that influences the results of RC tests and skews conclusions in favor of incorrectly supporting TTR excess performance in restricted data samples.

5.3.Robustness analysis

In this section, we evaluate the robustness of the results reported in Sections 5.1 and 5.2 to changes in the testing methodology. Particularly, data snooping bias is reevaluated using the SPA test of Hansen (2005). The rate of null rejections is computed and compared for tests that use the benchmark *686k* rule universe and tests that use 43 smaller, unrepresentative universes (the same as before are used, except that the universes generated by the Filter and Runs indicators are merged). For this exercise, only data from 18 emerging stock markets in Central and Eastern Europe (listed in Table 7, Panel C) is considered, because the results in Section 5.2 show that this is where RC test null rejections predominantly occur. A total of 4,208 tests are performed using the benchmark rule universe and 180,944 tests using the 43 alternatives.

Table 7. Null rejection rates and data snooping bias in SPA tests

Panel A: NRR aggregated by rule universe				**Panel B: NRR aggregated by year**							
				686k				**Restricted rule universes**			
Rule universe	**α=0.1**	**α=0.05**	**α=0.01**	**Year**	**α=0.1**	**α=0.05**	**α=0.01**	**Year**	**α=0.1**	**α=0.05**	**α=0.01**
%b	3.35%	1.64%	0.38%	1991	0.00%	0.00%	0.00%	1991	3.15%	3.15%	1.37%
%k	2.99%	1.52%	0.38%	1992	0.00%	0.00%	0.00%	1992	1.95%	1.67%	0.37%
AO	3.49%	1.64%	0.36%	1993	0.00%	0.00%	0.00%	1993	13.24%	12.34%	9.48%
ASI	23.08%	21.22%	11.86%	1994	2.70%	0.00%	0.00%	1994	4.90%	4.53%	2.14%
BMP	3.49%	2.21%	1.00%	1995	0.00%	0.00%	0.00%	1995	3.98%	3.67%	2.28%
CCI	4.25%	2.28%	0.57%	1996	0.00%	0.00%	0.00%	1996	8.34%	6.58%	3.51%
CMF	2.85%	1.66%	0.76%	1997	0.00%	0.00%	0.00%	1997	5.59%	5.41%	3.61%
CMO	2.88%	1.45%	0.36%	1998	0.00%	0.00%	0.00%	1998	5.13%	3.62%	1.47%
CO	2.50%	1.50%	0.76%	1999	0.00%	0.00%	0.00%	1999	7.35%	6.98%	4.67%
COG	6.63%	5.30%	1.21%	2000	1.41%	0.70%	0.70%	2000	5.26%	3.32%	1.72%
DI	3.54%	2.00%	0.48%	2001	1.84%	0.61%	0.00%	2001	3.65%	2.74%	1.46%
DMI	2.26%	1.07%	0.24%	2002	2.25%	0.56%	0.00%	2002	5.62%	3.78%	1.20%
DPO	0.95%	0.48%	0.14%	2003	0.00%	0.00%	0.00%	2003	4.72%	4.45%	2.63%
DYMOI	2.14%	0.97%	0.21%	2004	0.00%	0.00%	0.00%	2004	6.14%	5.80%	3.87%
EMV	19.77%	18.23%	9.91%	2005	0.41%	0.00%	0.00%	2005	5.74%	5.29%	3.38%
F	5.44%	3.49%	0.88%	2006	0.00%	0.00%	0.00%	2006	5.15%	4.67%	2.77%
INI	2.66%	1.40%	0.45%	2007	0.66%	0.33%	0.33%	2007	5.72%	5.32%	3.27%

KCD	1.28%	0.55%	0.12%
KPO	1.73%	0.67%	0.14%
KST	1.52%	0.67%	0.12%
KVO	4.56%	2.54%	0.45%
LRS	2.57%	1.05%	0.21%
MACD	1.05%	0.40%	0.05%
MFI	3.07%	1.45%	0.33%
MVI	4.06%	2.76%	0.67%
NRVI	2.92%	1.52%	0.31%
OBV	43.96%	43.13%	31.44%
PFE	3.16%	1.45%	0.33%
PI	2.73%	1.28%	0.29%
Qstick	22.24%	21.06%	11.64%
RMI	4.33%	2.26%	0.45%
ROC	1.90%	0.88%	0.19%
RSI	3.42%	1.64%	0.48%
RVI	2.71%	1.31%	0.33%
RVig	3.26%	1.78%	0.88%
RWI	5.58%	4.61%	2.14%
SMI	3.33%	1.50%	0.29%
SRSI	4.23%	2.19%	0.40%
TRIX	3.11%	1.45%	0.21%
TSI	3.64%	2.04%	0.38%
UO	2.57%	1.43%	0.36%
VX	2.61%	1.33%	0.40%
WVAD	40.54%	39.42%	26.19%
Min	*0.95%*	*0.40%*	*0.05%*
Max	*43.96%*	*43.13%*	*31.44%*
Median	*3.16%*	*1.52%*	*0.38%*
Average	*6.24%*	*4.85%*	*2.53%*
St.Dev.	*9.44%*	*9.49%*	*6.52%*
Benchmark–*686k*	0.97%	0.40%	0.07%

2008	7.59%	4.11%	0.32%	2008	20.54%	10.93%	2.66%
2009	0.00%	0.00%	0.00%	2009	4.40%	4.14%	2.52%
2010	0.00%	0.00%	0.00%	2010	4.32%	4.02%	2.48%
2011	1.13%	0.00%	0.00%	2011	4.61%	2.71%	1.14%
2012	0.00%	0.00%	0.00%	2012	5.17%	4.50%	2.61%
2013	0.00%	0.00%	0.00%	2013	3.96%	3.60%	1.95%
Min	*0.00%*	*0.00%*	*0.00%*	*Min*	*1.95%*	*1.67%*	*0.37%*
Max	*7.59%*	*4.11%*	*0.70%*	*Max*	*20.54%*	*12.34%*	*9.48%*
Median	*0.00%*	*0.00%*	*0.00%*	*Median*	*5.15%*	*4.45%*	*2.52%*
Average	*0.78%*	*0.27%*	*0.06%*	*Average*	*6.03%*	*4.92%*	*2.72%*
St.Dev.	*1.70%*	*0.87%*	*0.17%*	*St.Dev.*	*3.83%*	*2.47%*	*1.79%*

Panel C: NRR aggregated by stock market

Extended rule universe (686k)				**Restricted rule universes**			
Country	**α=0.1**	**α=0.05**	**α=0.01**	**Country**	**α=0.1**	**α=0.05**	**α=0.01**
BA	9.33%	5.33%	0.00%	BA	13.95%	10.45%	6.48%
BG	1.40%	0.70%	0.00%	BG	7.63%	4.91%	2.18%
CY	1.80%	0.45%	0.00%	CY	4.92%	2.67%	1.04%
CZ	0.00%	0.00%	0.00%	CZ	5.35%	4.79%	2.77%
EE	1.30%	1.30%	0.00%	EE	5.68%	4.20%	1.43%
GR	0.80%	0.32%	0.00%	GR	6.28%	4.99%	2.66%
HR	0.47%	0.00%	0.00%	HR	5.17%	3.61%	1.66%
HU	1.09%	0.00%	0.00%	HU	6.40%	5.71%	3.43%
LT	1.56%	0.39%	0.00%	LT	5.86%	4.04%	1.49%
LV	2.75%	1.10%	1.10%	LV	5.67%	4.36%	1.97%
PL	0.27%	0.00%	0.00%	PL	6.54%	6.06%	3.60%
RO	0.22%	0.22%	0.00%	RO	5.09%	3.53%	1.51%
RS	0.00%	0.00%	0.00%	RS	6.36%	3.28%	1.44%
RU	0.66%	0.66%	0.66%	RU	5.62%	4.14%	2.08%
SI	3.66%	1.22%	0.00%	SI	11.17%	9.36%	4.23%
SK	0.00%	0.00%	0.00%	SK	2.07%	1.65%	0.90%
TR	0.00%	0.00%	0.00%	TR	6.80%	6.34%	4.12%
UA	1.96%	0.65%	0.00%	UA	7.14%	3.62%	0.87%
Min	*0.00%*	*0.00%*	*0.00%*	*Min*	*2.07%*	*1.65%*	*0.87%*
Max	*9.33%*	*5.33%*	*1.10%*	*Max*	*13.95%*	*10.45%*	*6.48%*
Median	*0.95%*	*0.35%*	*0.00%*	*Median*	*6.07%*	*4.28%*	*2.02%*
Average	*1.52%*	*0.69%*	*0.10%*	*Average*	*6.54%*	*4.87%*	*2.44%*
St.Dev.	*2.20%*	*1.24%*	*0.29%*	*St.Dev.*	*2.53%*	*2.17%*	*1.46%*

NOTE. The Null Rejection Rate (NRR) *is the number of tests that reject the null hypothesis of no economic profitability at the* α *confidence level, expressed as a percentage of the total number of tests performed.*

Table 7 reports the results, which show that data snooping bias has an even more significant impact on tests that use the SPA methodology. Particularly, compared to tests that use the benchmark *686k* rule universe, null rejection rates for unrepresentative universes occur on average (median) 6.4 to 35.4 (3.2 to 5.3) times more often, depending on the selected confidence level. When aggregating the data by year and by stock market, the relative differences in null rejection rates are similar. Moreover, many years and markets exist for which tests that use unrepresentative rule universes reject the null hypothesis, while tests that use the benchmark do not. Overall, the results support earlier conclusions regarding the influence of data snooping bias on tests that analyze the relative performance of multiple forecasting models. In particular, they show that the excess performance of TTRs is also overstated when using small, unrepresentative universes in the SPA test.

Even though the SPA test is more powerful and rejects the null of no economic profitability more often in this sample compared to the RC test, analyzing the excess performance of TTRs in this context yields that null rejection rates remain scarce in tests employing the *686k* universe. Particularly, only 41 tests (0.97%) reject the null at the 10% confidence level, 17 tests (0.40%) reject the null at the 5% confidence level, and 3 tests (0.07%) reject the null at the 1% confidence level. Because these results are obtained for some of the smallest stock markets in the entire sample, they provide additional support for the Efficient Market Hypothesis and earlier conclusion that TTRs are not relevant from an economic perspective when used by investors for trading in stock markets around the world.

6. Conclusions

This paper performs a novel investigation into how choosing small sets of prediction models, which do not account for what investors and researchers use, introduces data snooping bias in statistical tests that examine their relative performance. The paper focuses on the Reality Check (RC) test of White (2000) and on the literature concerned with the excess performance of models derived from technical analysis, technical trading rules (TTRs). Our analysis shows that the effective span of rule universes is positively correlated with their size. Even though trading rules universes employed in the literature are becoming larger, they do not typically account for the data snooping efforts of others and are yet to get near to the size of the "true" universe that can be safely presumed to be used by investors and researchers. This hypothetical construction is unobservable, but should easily contain millions of prediction models.

In a simulation exercise conducted on randomly generated data, we find that using small, unrepresentative rule universes increases false discoveries and biases the outcomes of RC-type tests in favor of showing that some forecasting models have superior predictive ability. TTRs are "luckier" and data snooping bias is stronger when the size or diversity of the rule universe are restricted, when the volatility of the underlying data generating process increases, or when the length of the data sample increases. In an empirical exercise that uses a more representative rule universe of 686,304 models as a benchmark, we find that 44 smaller, unrepresentative universes–comparable in size and information span to the ones that are typically employed in the literature–overestimate the economic relevance of TTRs by 1.8-2 times on average, depending on the data sample, the characteristics of the trading rule universe, or the test significance level. In the extreme, but not entirely unrealistic case in which independent researchers perform tests using unrepresentative rule universes and then draw conclusions based on a meta-analysis of their results, the excess performance of TTRs can be inflated by 6.16-7.93 times on average, and even by as much as 32 times.

Our findings have several important implications. First, they contribute to the recent debate that highlights the need to thoroughly investigate and mitigate data snooping, as a way to increase the relevance and reliability of published results. In particular, we argue that previous findings showing TTRs to be relevant from an economic perspective in some financial markets should be treated with more care. Trading rules can appear relevant in our tests that use random data and their "luckiness" increases in some setups (e.g., higher market volatility, longer data samples). As a consequence, existing evidence should be reexamined using tests that control for the data snooping efforts of others. New tests should also consider and control for this problem. In our own reevaluation of the economic relevance of TTRs, we find no significant evidence to support that they are able to earn excess returns when used for trading in stock markets around the world, after subtracting transaction costs and adjusting for data snooping bias using a more representative 686,304 prediction model universe. This finding is robust to the choice regarding the test significance level, the way the data is aggregated or analyzed, and the selected testing methodology. Thus, relative to a broad set of technical trading rules, prices in stock markets around the world incorporate information efficiently. From a theoretical perspective, this supports the weak-form Efficient Market Hypothesis of Fama

(1970), as opposed to the Adaptive Market Hypothesis of Lo (2004) that has recently gained importance (Lim and Brooks, 2011). From a practical perspective, the findings show that technical trading rules have limited use for making investment decisions, at least when they are used independently from other forecasting methods. Investors seeking to use technical analysis should be better off with passively managing their portfolios.

Second and more generally, our results imply that data snooping bias occurs and is significant in statistical tests that evaluate the relative performance of multiple forecasting models without accounting for relevant alternatives. Moreover, when all relevant alternatives are not observable, tests are impacted by ambiguity risk, testing for relative performance becomes problematic, and results should be treated with more care. Testing for absolute performance avoids this problem and provides more objective results, but evaluating relative performance remains necessary to answer some important scientific questions. This implies the need to develop new testing methodologies that control for the subjective choice regarding the set of alternatives that are used in tests that evaluate their relative performance.

References

Al-Nassar, N. S., 2014. The profitability of trading rules in stock markets: Evidence from GCC countries. PhD Thesis, RMIT University.

Arellano-Valle, R. B., Genton, M. G., 2008. On the exact distribution of the maximum of absolutely continuous dependent random variables. Statistics and Probability Letters 78(1), 27-35.

Bajgrowicz, P., Scaillet, O., 2012. Technical trading revisited: False discoveries, persistence tests, and transaction costs. Journal of Financial Economics 106(3), 473-491.

Barras, L., Scaillet, O., Wermers, R., 2010. False discoveries in mutual fund performance: Measuring luck in estimated alphas. The Journal of Finance 65(1), 179-216.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological), 289-300.

Brabazon, A., Dang, J., Dempsey, I., O'Neill, M., & Edelman, D. (2012). Natural computing in finance: a review. In Rozenberg, Grzegorz; Bäck, Thomas; Kok, Joost N.(eds.). Handbook of Natural Computing: Theory, Experiments and Applications, 1707-1735.

Coakley, J., Marzano, M., Nankervis, J., 2016. How profitable are FX technical trading rules?. International Review of Financial Analysis 45, 273-282.

Colby, R.W., 2002. The Encyclopedia Of Technical Market Indicators, Second Edition. McGraw-Hill.

Fama, E. F., 1970. Efficient capital markets: A review of theory and empirical work. The Journal of Finance 25(2), 383-417.

Fifield, S.G., Power, D.M., Donald Sinclair, C., 2005. An analysis of trading strategies in eleven European stock markets. The European Journal of Finance 11(6), 531-548.

Grossman, S. J., Stiglitz, J. E., 1980. On the impossibility of informationally efficient markets. The American Economic Review 70(3), 393-408.

Hall, P., Horowitz, J.L., Jing, B.Y., 1995. On blocking rules for the bootstrap with dependent data. Biometrika 82(3), 561-574.

Han, Y., Yang, K. and Zhou, G., 2013. A new anomaly: The cross-sectional profitability of technical analysis. Journal of Financial and Quantitative Analysis, 48(5), pp.1433-1461.

Han, Y., Hu, T., Yang, J., 2016. Are there exploitable trends in commodity futures prices?. Journal of Banking and Finance 70, 214-234.

Hansen, P.R., 2005. A Test for Superior Predictive Ability. Journal of Business and Economic Statistics 23(4), 365-380.

Hartigan, J. A., 2014. Bounding the maximum of dependent random variables. Electronic Journal of Statistics 8(2), 3126-3140.

Harvey, C. R., 2017. Presidential address: The scientific outlook in financial economics. The Journal of Finance 72(4), 1399-1440.

Harvey, C. R., Liu, Y., 2014. Evaluating Trading Strategies. The Journal of Portfolio Management 40(5), 108-118.

Hsu, P. H., Hsu, Y. C., Kuan, C. M., 2010. Testing the predictive ability of technical analysis using a new stepwise test without data snooping bias. Journal of Empirical Finance 17(3), 471-484.

Hsu, P. H., Taylor, M. P., Wang, Z., 2016. Technical trading: Is it still beating the foreign exchange market?. Journal of International Economics 102, 188-208.

Jensen, M.C., 1978. Some Anomalous Evidence Regarding Market Efficiency. Journal of Financial Economics, 6(2/3), 95-101.

Kim, J. H., Ji, P. I., 2015. Significance testing in empirical finance: A critical review and assessment. Journal of Empirical Finance 34, 1-14.

Lim, K.P., Brooks, R., 2011. The evolution of stock market efficiency over time: A survey of the empirical literature. Journal of Economic Surveys 25(1), 69-108.

Lo, A.W., 2004. The Adaptive Markets Hypothesis. The Journal of Portfolio Management 30(5), 15-29.

Marshall, B. R., Cahan, R. M., 2005. Is the 52-week high momentum strategy profitable outside the US?. Applied Financial Economics 15(18), 1259-1267.

Menkhoff, L., 2010. The use of technical analysis by fund managers: International evidence. Journal of Banking and Finance 34(11), 2573-2586.

Metghalchi, M., Du, J., Ning, Y., 2009. Validation of moving average trading rules: Evidence from Hong Kong, Singapore, South Korea, Taiwan. Multinational Business Review 17(3), 101-122.

Metghalchi, M., Marcucci, J., Chang, Y. H., 2012. Are moving average trading rules profitable? Evidence from the European stock markets. Applied Economics 44(12), 1539-1559.

Neuhierl, A., Schlusche, B., 2010. Data snooping and market-timing rule performance. Journal of Financial Econometrics, 9(3) 550-587.

Park, C.H., Irwin, S.H., 2007. What do we know about the profitability of technical analysis?. Journal of Economic Surveys 21(4), 786-826.

Politis, D.N., Romano, J.P., 1994. The stationary bootstrap. Journal of the American Statistical Association 89(428), 1303-1313.

Ratner, M., Leal, R. P., 1999. Tests of technical trading strategies in the emerging equity markets of Latin America and Asia. Journal of Banking and Finance 23(12), 1887-1905.

Romano, J. P., Wolf, M., 2005. Stepwise multiple testing as formalized data snooping. Econometrica 73(4), 1237-1282.

Scott, G., Carr, M., Cremonie, M., 2016. Technical Analysis: Modern Perspectives. CFA Research Foundation Reviews 11(1), 1-36.

Shynkevich, A., 2012. Performance of technical analysis in growth and small cap segments of the US equity market. Journal of Banking and Finance 36(1), 193-208.

Shynkevich, A., 2016. Predictability in bond returns using technical trading rules. Journal of Banking and Finance 70, 55-69.

Sobreiro, V. A., da Costa, T. R. C. C., Nazário, R. T. F., e Silva, J. L., Moreira, E. A., Lima Filho, M. C., Kimura, H., Zambrano, J. C. A., 2016. The profitability of moving average trading rules in BRICS and emerging stock markets. The North American Journal of Economics and Finance 38, 86-101.

Storey, J.D., 2002. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(3), 479-498.

Sullivan, R., Timmermann, A., White, H., 1999. Data-snooping, technical trading rule performance, and the bootstrap. The Journal of Finance 54(5), 1647-1691.

Taylor, M. P., Allen, H., 1992. The use of technical analysis in the foreign exchange market. Journal of International Money and Finance 11(3), 304-314.

Taylor, N., 2014. The rise and fall of technical trading rule success. Journal of Banking and Finance 40, 286-302.

Timmermann, A., Granger, C. W., 2004. Efficient market hypothesis and forecasting. International Journal of Forecasting 20(1), 15-27.

Urquhart, A., Gebka, B., Hudson, R., 2015. How exactly do markets adapt? Evidence from the moving average rule in three developed markets. Journal of International Financial Markets, Institutions and Money 38, 127-147.

White, H., 2000. A reality check for data snooping. Econometrica 68(5), 1097-1126.

Wilder, J. W., 1978. New concepts in technical trading systems. Trend Research.

Zarrabi, N., Snaith, S., Coakley, J., 2017. FX technical trading rules can be profitable sometimes!. International Review of Financial Analysis 49, 113-127.