Is Data Snooping responsible for Technical Analysis Rules Success?

Worik TUREI STANTON

August 14, 2013

Acknowledgements

I owe thanks:

Firstly to Timothy for the time, patience, good humour and wise guidance.

The blessing of PT, Axle and AS needing me. For love despite it all - bringing meaning into the longest and darkest days.

Most importantly to my wife – MT. Loving and supporting me. Not just food on the table and beer in the jug, but for the faith in me. I am yours. Truly.

I am not a religious man. But if there are gods, fate or destiny. I am blessed. This little boy who loves to pull apart and struggles to put together, some how the time passed and good things happened. I do not deserve any of it, and I am eternally grateful.

Lastly, not leastly, and most irreverently I must thank my colleagues in fun (for fun is good DrSuess (1960)) in Kill, Martha! You, (and MJ), truly pull the Oxy Train. Puff Puff, Toot!.

Abstract

Data Snooping is often suspected when effective technical analysis rules are found or presented. It is difficult to tell if a result is due to data snooping, so evaluating technical analysis rules often boils down to detecting data snooping and if it has invalidated the results. Herein we look at several algorithms designed to increase (risk–adjusted) returns for investors, and several techniques for detecting or compensating for data snooping.

We find no easy answer to detecting data snooping. Many of the methods we look at are useful, but there is no known way to get around sparse data and the unrepeatable nature of investment decisions. We conclude that data snooping bias is a persistent risk and it is unlikely that there is any effective single solution to the problem. The best that we can do is be aware of the risk of data snooping and to report how we have dealt with the risk as part of our analysis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 What is Data Snooping?

snoop verb

1. investigate, explore, have a good look at, prowl around, nose around, peer into. He's been snooping around her hotel.
2. spy, poke your nose in, nose, interfere, pry (informal). Governments have been known to snoop into innocent citizens' lives.

Data Snooping is when the data that are used to discover models are used to evaluate those same models, or when models are generated continuously until one that fits the data is found. Alternatively it is when researchers use the results of a previous experiment to design the next using the same data set, and report the results of the final experiment. They have chosen to "go with the data".

Data Snooping is particularly hard to avoid in finance where there is a strictly limited set of data for any given period and it is not possible, generally, to go and discover more from exactly the same generating process.

When data are snooped a model is found that, for some measure, is different from what would be expected under the Efficient Market Hypothesis.

Data Snooping appears in different guises, including the following:

Over Fitting Given enough parameters in a model, it may be possible that any pattern of returns can be fitted (i.e., over fitted) with arbitrary accuracy.

Repeated Experimenting When a researcher keeps trying different models until one is found that fits the data to hand. For example in "A Classic Case of Data Snooping for Classroom Discussion" Crack (1999) (discussed in detail in section 3). A parameter is adjusted until the results are "statistically significant".

Model Extension Using results from previous experiments to guide the choice of models and parametrisations in new experiments. For example there are many studies and theories that relate stock return to firm size (for example Ferguson and Shockley (2003)). It is unclear if there is a genuine economic relationship or if it is a simple observation. In these cases, once one paper has made the observation other researches have studied that same effect (Lo and MacKinlay, 1990).

1.2 Why Data Snooping Matters

In economics generally, and finance specifically, there is limited scope for controlled experiments. So there is a great deal of reliance on statistical analysis (Lo and MacKinlay, 1990). There are limited financial data available however, and because what there is is studied very closely and often, as Lo and MacKinlay, pp 432 put it "the axioms of classical statistical analysis are violated routinely". The researcher has very little choice but to proceed and make do with what there is.

For example Faber (2009) published a market timing algorithm claiming a marked decrease in volatility and a (seemingly) significant improvement in mean return. Faber utilised a simple two–part rule: a 200–day moving average versus the S&P500 index evaluated on the last day of the month. If the index is above the moving average (at the end of the month) buy the index, else hold cash. The article is (as of June 2011) ranked as the 3$^{\text{rd}}$ most downloaded article on the SSRN network where the on-line version resides (more than 73,000 downloads). So it has made a mark, but is it true?

It does seem too good to be true that by utilising such a simple algorithm an improvement in performance is readily available. On the face of it it contravenes the Efficient

Market Hypothesis and makes most modern introductory finance textbooks obsolete.

If Faber's results are due to data snooping then we need to know before we act on the advice. If Faber has found a true property of financial markets then perhaps the authors of introductory textbooks on finance need to reinterpret their opinions about the feasibility of market timing. Either way it surely maters. We will look closely at Faber's results in section 4.

1.3 Our Contribution

We take a fresh look at the problem of data snooping. We use neighbourhood analysis described in section 2.5) to look for signs of data snooping. We also take a close look at the application of statistical analysis and its role in detecting or facilitating data snooping.

We utilise three published studies to test our approaches. The first (Crack (1999) in section 3) is an example of data snooping that we use to test our tools. The other two studies (Faber (2009) and Sullivan, Timmermann, and White (1999)) are both in their own ways classic papers and we subject them to closer study than has been done previously.

In the case of Sullivan et al. (1999) there is an algorithm presented to avoid data snooping. The algorithm has been used in many following papers. By applying different techniques to detect data snooping we cast doubt on the efficacy of the algorithm in Sullivan et al. (1999). We note that the algorithm used by Sullivan et al. (1999) is not applicable to the non–stationary data they apply it to. We also find that the experiments presented in that paper are not fully reproducible.

In our application of statistical analysis we find and present a hitherto unnoticed weakness in the applicability of common statistical techniques to time series of financial returns. We doubt that financial returns are actually sampled from a random distribution. Whilst appearing random they are, perhaps, devoid of useful information, but they are not, generally, randomly sampled. (See section 2.3.7).

Chapter 2

Data Snooping Tools

In this section we describe some tools that may be useful in detecting or countering data snooping, and we look at the decision criteria that must be met to use these tools. Decision criteria are the criteria that we use to decide if a rule is data snooping or not. For example if we judge a technical analysis rule by statistical significance using a t–test we may decide to reject as valid rules where the p–value is less than 0.05. In that case $p \leq 0.05$ is a decision criterion.

Before using any tests we must first reproduce the results of the rule on the data where the author claims the results. If we can succeed in that step we proceed to applying the rules we outline in this section.

Before we apply the tools we should ask "what results do we expect?". We can then apply the tools and compare the results to our prior expectations

We then collate the results from applying the tools and draw our conclusions about data snooping. None of the methods we discuss here are definitive and all require judgement. It is not human nature to accept that a result that has been achieved through much labour is in error, especially if success is pleasing. Effective technical analysis rules (usually) take a lot of effort to find and if they are truly effective are very pleasing. So

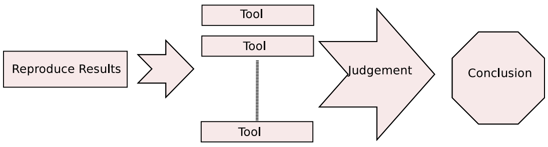


Figure 2.1: Detecting Data Snooping

there are strong incentives to accept the results of a technical analysis rule and to deny that it is data snooping. The point of setting decision criteria is to move as much human judgement as possible to before rule discovery, where it can be more dispassionate. A disadvantage of creating decision criteria is that it is not possible to foresee every eventuality and the criteria will tend to be inflexible.

Satisfying a decision criterion is not proof that the result is not due to data snooping, and failing is not proof that it is. Data Snooping is an unavoidable consequence of the limits to the data available to researchers of financial time series. A rule with strong predictive power will be discovered/created along with spurious rules that are due to data snooping. Tests and criteria for them as outlined in this section are not magic or perfect. They serve to "load the dice" and will filter out some of the more egregious examples of data snooping.

Objective decision criteria cannot be tightly defined until the nature of the actual experiment to be conducted is known. But a set of criteria is outlined below, where possible, for each rule in section 2.1, following.

2.1 Summary of Tools

What tools have been used to try to identify data snooping? Each of the following is introduced and then discussed in detail in the following sections.

Reproducibility If a claim that a strategy produces an excess return cannot be independently verified by another researcher repeating the experiment then there is nothing more that can be done.

Statistical Significance It is not enough to show that a strategy produces an excess return or a better Sharpe ratio than some benchmark. If the performance is statistically significant then it is more credible, and more likely that the performance can be repeated.

Holdouts Partition the data set into clearly defined periods, use one set for developing a model and another set for testing it. A test utilising holdouts is often called an "out of sample" test.

Neighbourhood Analysis Check that a model developed on a set of data is still nearly as good when its parameters are changed slightly. That is, is it robust? Small changes in a model's parameters define a "neighbourhood" around that model. If a good model is truly a model of the underlying data generating process then it is likely that other models in a close neighbourhood are also good.

Other Markets Use data from one market to develop a model and data from another market to test it. If a model is due to data snooping it will work for the market it was developed for and is unlikely to work on others.

Story Telling Is there an economic reason why a model should work? This might be a risk-based argument, or one based on the psychology of market participants.

Feasibility If a researcher ignores transaction costs, liquidity, risk aversion, taxes or other impediments often very interesting looking technical analysis algorithms can be found[1]. A researcher may be concentrating on returns while ignoring increased volatility. While not strictly a data snooping issue in itself many algorithms found by data snooping can be discarded once simple feasibility considerations are taken into account e.g., Trethewey and Crack (2010).

2.2 Reproducibility

The first step in examining a technical analysis rule for data snooping is to reproduce its claimed results. If the description of a rule is to be of any use or interest to us it must contain enough information, or reference enough information, to reproduce in a simulation the properties claimed. This is an elementary point but it is relevant. For example, we found a strategy in a paper published in a reputable journal (Sullivan et al., 1999) that we could not reproduce. (See section 5.6.6). Similarly, Tajaddini and Crack (2012) were unable to reproduce results from a paper published in a respectable journal.

[1] The computer science/artificial intelligence literature is a rich source of such strategies. For example Atsalakis and Valavanis (2009) conclude that the Efficient Market Hypothesis is under threat using a simulation of trading (using neural networks), but no allowance is made for transaction costs.

2.3 Statistical Significance

The level of statistical significance is the probability of rejecting a true null hypothesis. That is, the probability of making a making a Type I Error (Porkess, 2004). Before commencement of an experiment a significance level is specified along with a null hypothesis and a statistical test. At its conclusion the results of the experiment are evaluated using the test. If the test says that the probability of making a mistake by rejecting the null hypothesis being true is less than the level of significance, then the null hypothesis is rejected.

Economic significance without statistical significance cannot be easily distinguished from coincidence. A technical analysis rule that results in higher mean return may be due to the inclusion of an abnormal result, an error or an extremely rare event that should not be modeled[2]. It may be a coincidence.

For example take the following two synthetic return series:

N	X_1	X_2
1	0.629	0.533
2	0.803	0.921
3	0.533	3.000
4	0.981	0.993
5	0.222	0.190
6	0.565	0.742
7	0.997	0.308
8	0.654	0.379
9	0.054	0.087
10	0.671	0.988

The mean of X_1 is 0.6109 and that of X_2 is 0.8141. On the face of it X_2 has a higher mean than X_1. Whilst the raw means are different, is the difference statistically significant?

A t–test (using the built in t–test in the **R** statistical package version 2.11.1) returns

[2] A "...rare event that should not be modeled" is a "Black Swan", see Taleb (2007).

a p–value of 0.484. At conventional levels of significance the null hypothesis that the means are the same cannot be rejected.

The data for both X_1 and X_2 were sampled from a uniform random distribution between 0 and 1. Then the third number in X_2 was replaced with 3. (A true outlier. It is not possible for the number 3 to be in a series using this underlying data generation process, we discuss the definition of outliers further in section 2.9.1).

The relevance to data snooping is that technical analysis rules that pick out an outlier by chance will have a higher mean return. This can be uncovered with a simple statistical test[3].

2.3.1 Weaknesses of Statistical Significance

Although useful, statistical significance, and statistical tests more generally, have weaknesses. Whilst it is true that economic significance without statistical significance has doubtful utility and is likely to be the result of data snooping, it is also true that statistical significance does not necessarily imply economic significance. (McCloskey, 1985).

The mean returns of firm A may be different from those of firm B, with p–value approaching zero, but still they may be very close in an economic sense. The fact the mean returns are different does not imply that each observation is higher, or there is an exploitable pattern in them.

The usual tests for statistical significance are based on the probability that an observation arose by chance. If an observation has been searched for, by repeatedly trialing different models and parametrisations, the assumption that the observation arose by chance is not true, so the statistical significance of the (searched for) result is exaggerated. This search process is one cause of data snooping.

Also if an algorithm can be shown to be affective at extracting excess profits yet these profits cannot be shown to be statistically significant, one must be sure to check that the significance tests are valid and relavent. It is possible to be pleased with profit whilst frustrated with the data. Statistical tests often depend on assumptions made about the

[3] In fact the t–test is not strictly applicable. The underlying data are not normal and N is only 20. The data are independent though. We look more closely at the assumptions of statistical tests in section 2.3.3.

data, such as the distribution or that observations are independent. These assumptions may well turn out to be false (we look more closely at the assumptions in section 2.3.3).

2.3.2 Null Hypotheses

Since statistical significance has its meaning in relation to the null hypothesis, it is important to frame a null hypothesis that is meaningful. The truth value of the null has one binary bit of information and all its meaning is carried in the expression of the null. It can be difficult to frame a useful null.

The procedure of formulating and testing a null hypothesis is not universally accepted as best practise. For instance Armstrong (2007) argues that "An alternative [to null hypotheses] is to focus on the testing of two or more reasonable competing hypotheses. Such testing can be done without the need for significance tests". Alternatively, Armstrong suggests in the same paper, using estimates and confidence intervals[4]. Although this is a subtly different approach it uses the same parametric assumptions as significance tests.

2.3.3 Tests

There are many different tests for significance. They vary widely in the sets of assumptions about the data being tested and in the amount of computation required. The two tests that we will use the most are the standard t–test and the Wilcoxon–Mann test.

We use one–sided and two–sided versions of the tests. When comparing a random value to a fixed value we use a one–sided test and when comparing two random variables we use two–sided tests. For example in section 5.8 we test the returns from technical analysis rules to see if they are greater than zero (a one–sided test) or if they are different from a benchmark (a two–sided test).

With two–sided tests we can either test the two random variable's means (unpaired tests) or their levels (paired tests). For instance in section 4.3.2 we compare the volatility of the returns of a technical analysis rule to that of an index. We compare both the mean volatility over the experimental period with an unpaired test and the volatilities for each period against each other with a paired test.

[4] But it seems to us that using confidence intervals imply using significance tests.

We will use Generalised Methods of Moments (GMM) in section 3 as that is what was used in Crack (1999).

2.3.4 Assumptions of the Tests

We explore the assumptions of the statistical tests in more detail in appendix A

All tests have in common the assumption that the sample is randomly selected.

The t–tests assume the underlying distribution is normal. The one–sided Wilcoxon tests assume a symmetric distribution and the two–sided Wilcoxon tests assume the distributions of the two random variables are identical in shape.

The t–tests assume that the true variance of the sample is unknown. If the true variance is known then the test is still valid, but the Z–test is better (Sheskin, 2004).

2.3.5 Validity of Assumptions for Financial Data

None of the assumptions of the statistical test outlined above hold consistently for all financial data.

The most egregious violation is when we use the tests on price levels. Price levels are not independent; today's price is usually close to yesterday's price, and price levels often have a trend (that is they increase, or occasionally decrease, over time).

Financial market returns are notorious for being non–normally distributed and they often exhibit high but (seemingly) randomly varying degrees of auto–dependence. (Lo and MacKinlay, 2001).

Also it is known that the variance of financial returns series is not constant (Bollerslev et al., 1992); It varies over time. So the returns cannot be described as coming from the same distribution.

For example in section 3 we study a report of the phase of the moon affecting returns. If we look at the 20 years of returns data used in the study we find that the standard statistical tests soundly reject the hypothesis that the returns are normally distributed because of significant kurtosis and skewness[5].

[5] The kurtosis for the EWRETD over the whole period 1977–1996 was 34. On an annual basis over that period the mean annual kurtosis was 5.6, the maximum was 26 (1987) and the minimum 0.16 (1985).

2.3.6 Consequences of Violating the Assumptions of Statistical Tests

Given that the assumptions of the t–test and the Wilcoxon tests are routinely violated in financial data, what then?

Boneau (1960) reports that the t–test is extremely robust to the assumptions of normality and assumption of equal variance being violated, especially when there are a large number of data points[6]. Boneau tested two–sided unpaired t tests and concluded that the tests were robust for sample sizes greater than minuscule. However this was so long as the sample sizes were similar and the distribution parameters of the two samples are the same. This is a point belaboured by Bradley (1978).

In finance when we want to compare the performance of some sort of instrument or strategy (a portfolio or a technical analysis rule) with a benchmark (say an index or zero) it is unlikely that the population parameters are the same.

We have failed to find similar studies into the robustness of the non–parametric Wilcoxon tests.

The continued reliance on p–values and statistical tests when the underlying data do not conform to the underlying assumptions may be more of a cultural phenomenon than a scientific decision.

2.3.7 The Random Nature of Financial Markets

The semi–strong form of the efficient market hypothesis states that all publicly available information is accounted for in the prices of (and therefore returns from) financial securities (Fama and Blume, 1966). Thus there should be no way to utilise publicly available information to earn consistent risk–adjusted excess profits.

The truthfulness of the efficient market hypothesis hypothesis is debated. But it is true that in financial markets there is a large set of agents, of varying intelligence, looking carefully at all information that may possibly affect the prices of securities in the market. The agents act to adjust prices to account for this available information. So we should expect that abnormal profits are few and far between.

[6] Boneau assumed that the assumption of independant errors could easilly be satisfied.

From this perspective the sequence of prices, hence returns, in a financial market are encoded efficiently by the market to be the best representation available of the relevant information. It is random in the sense that if it had any predictable patterns in it that would constitute publicly available information that market participants could use, and they would “consume” that information removing it from the market. In the real world of taxes, transaction costs and agent’s risk aversion this is not as simple as “market clearing” or “price discovery”. But because the agents (market participants) are presumably intelligent these “distortions” are incorporated into their models and accounted for.

This is analogous to the Algorithmic Information Theory (AIT) concept of randomness. In AIT a number Γ is random, relative to a computer C, if there is no set of inputs to that computer, smaller than Γ, that produces Γ as the output of C[7](Calude, 1994).

It is not a perfect analogy. For one thing the “computer” in this case, the market and its participants, is not constant. In finance the “computer” reprogrammes itself continuously. Also the inputs are possibly not regular. It may not be possible to describe the set of all information.

But this is different from the concept of randomness that involves sampling from a, possibly unknown, distribution.

The question is: Are the two forms of randomness isomorphic? Can I hope to find a parametric distribution, of arbitrary complexity, for the return series that is smaller than the series itself?

Empirically the answer at first glance would appear to be “yes, of course”. There are many empirical properties of financial markets that have been observed. For instance if there is a large price movement on one day, we are more likely to see a large movement the next day. Variance is not globally constant but it is locally constant.

2.3.8 Decision Criteria

To have any meaning the statistical test to be used, and the level of statistical significance required, must be decided beforehand.

A statistical test must be appropriate to the data and the assumptions of the statistical

[7] This is analogous to compression. Γ is incompressible.

test. Financial data often exhibit high kurtosis, distribution of returns have high peaks and long tails, and the time series of returns often have non–constant variance as well as elements of auto–correlation. It is unlikely that all the assumptions of any test can be fully satisfied.

It could be argued that because financial time series are the output of a "machine", the parts of being the market and all agents in the market, that the data are not drawn from any distribution. (See section 2.3.7 above). The objective of the machine is to find the "true price" of securities. The true price is a Platonic concept, and is unknown. The difference between the true price and the market price is the error in the machine's operation and we can expect that to be normally distributed. But since we do not know the true price we cannot discover the error. Neither the series of market prices nor market returns need be normally distributed. This implies that statistical tests on market prices (and their derivatives such as market returns) cannot be valid. Any assumptions about the distribution must by definition be violated.

We discussed the statistical tests we will be using and the applicable assumptions in section 2.3.3

2.4 Holdouts

If an effect observed in a sample of data is truly a characteristic of the system then it will persist, unless these characteristics change. A holdout sample is a set of data that is available to the modeler during the model construction phase but it is not used in tuning the parameters of the model. These data can then be used to test the model.

2.4.1 Why Holdouts Might Not Work

Holdouts can detect that an algorithm does not work (by demonstrating its ineffectiveness). But in the case where there are a lot of algorithms to test it cannot rule out data snooping. When there is a large enough number of algorithms tested over a training set, we can expect by random chance some will be effective. If the original set of algorithms tested is large enough, the set of effective algorithms will itself be large enough that we can expect a proportion of them to be effective over the holdout set also (Kuang et al., 2010).

For example consider a study that finds 1,000 economically useful trading rules statistically significant, by chance, at the 5% level, and where holdout data are available. When the study is repeated on the holdout data $0.05 \times 1,000 = 50$ rules are expected to be judged significant at the 5% level on the holdouts. At that point it will still be unknown which of the 50 successful rules are deemed successful due to data snooping.

2.4.2 Decision Criteria

To implement a "holdout test" at least three questions must be considered.

Which data? The holdout set must be specified. In section 3 we look at an algorithm for dividing index returns into higher and lower return periods based on the phase of the moon. The experiment was completed a decade ago so we can choose holdout data that came after the paper was published. For practitioners looking to devise trading algorithms it is not feasible to use data after the period of interest, as the period of interest starts tomorrow.

Another possible solution is to dig deeper and use data from before the sample used to develop the model (which we do in section 3 also). This is feasible unless there is a reason to believe that the underlying characteristics have changed over the period (this is related to section 2.7.1 on Story Telling). In the general case it should be possible to divide the data into two continuous and consistent sections.

Another possible solution is to "stripe" the data. Using interleaved periods of data to train and test the model. For instance to investigate a model for "day of the week" effects then use even weeks to tune the model and odd weeks to test it.

How much data? The whole problem of data snooping in finance arises because of the paucity of data. Every series of economic events occurs precisely once and will not be seen again. Thus it is important to conserve data, as they will not be repeated and what we have is all we will have. That is new data are not necessarily from the same data generating process. The bigger the holdout segment the less data there are to develop the model. There is no general answer to how much is enough, except that both the training and holdout set must be large enough to be able to draw sensible conclusions.

How much of it must agree? Ideally the holdout and training data would be in complete agreement. If a trading strategy generates $X\%$ excess return in training and $Y\%$ excess return for the holdout then X and Y should be close for us to believe the strategy works.

2.5 Neighbourhood Analysis

Financial markets are noisy. This implies that they are approximate in their behaviour. A strategy for trading a market (in a technical analysis sense) is a parameterised model. If such a model has superior performance on financial markets then, if the model is "robust", a small change to the model specification (the parameters) should result in a small change to its performance. This is a definition of "robust".

If the model is "brittle" in its parameters (a small change in a parameter causes a major change in the model's performance) then the model is a description of the data, rather than a description of underlying properties of the data. That suggests the model's performance is due to data snooping.

Given that financial markets are noisy, any measurement is composed of the value of the thing being measured and some random component. Any model must estimate the underlying value of the thing in consideration within the tolerance of the randomness. This describes a "region" for any parameter measured in finance, a "noise region". It may not be possible to clearly define the boundaries of that region but generally there will be points that are known to be within and without the region.

For example there may be an economic reason to use a one–year moving average in a market. To implement this suppose a technical analyst uses a 251–day moving average on a set of data spanning 25 years and generates a promising rule. The parameter of this rule is 251. The 250–day and 252–day moving averages are all in the same region as the

251–day moving average[8]. The 20–day moving average is not. So we would expect that the performance of the rule with a parameter of 250 or 252 would be about the same as the 251–day rule, but we would not expect the performance of the 20–day parametrisation of the rule to be similar.

If a developed model fits the data well, but upon changing a parameter a small amount (so the parameter is within the region) the model fits the data badly then that implies that the model describes that particular data set and not the underlying properties. That is, it is data snooping.

2.5.1 Decision Criteria

In neighbourhood analysis we are looking for brittle boarders of the effective region. To do this a smooth region in the parameter space around the optimum settings for the proposed model must be established and in this space we must define how stable we expect the results to be.[9]. Hence anti–hoc we must decide:

- How wide does the neighbourhood have to be to accept the result is not due to data snooping?
- If a small change to the model specification should result in a small change to its performance, (as described above), what sort of change in performance for change in parameters is acceptable? How steep can the slope be?

Before commencing an experiment the experimenter must identify the parameters that define neighbourhoods in the search space and the objective function to be used. Beyond that the decision about the neighbourhoods discovered is a subjective one, or requires further work to uncover more of the nature of neighbourhoods. But often an experimenter will have little choice about defining a neighbourhood.

[8] For the NZSE50 data there is an average of $251\frac{1}{3}$ trading days between 2001–2009 inclusive.

[9] This is not always possible. Take for example the rule "15% of Mondays suffer losses greater than 12%". It is not clear how a smoothly changing neighbourhood region could be defined because "Monday" is the parameter and cannot be varied without losing meaning

2.6 Other Markets

Given that people are mostly the same over most of the world, and similar economies exist that respond in similar manners to similar events, if an effect is observed in one market it should, with few exceptions, also be observed in others. Thus data are available to test a model, customised on one market, on another. It is possible that data snooping may create a self fulfilling prophecy. For instance a trader in New York, say, may notice that returns are particularly good in the first week of January. It may be data snooping but traders in other markets may well buy more stocks on the first trading day of the year, thereby spreading the effect.

Clearly not all people are the same everywhere and there are clear differences between markets. For instance, and obviously, bank holidays do not fall on the same days everywhere and "Islamic" societies tend to take Fridays off and "Christian" societies take Sundays off. But there are deeper cultural differences. For instance involving lucky numbers (Brown and Mitchell, 2008) or religious festivals (Bialkowski et al., 2009).

Also similar markets can be quite well correlated. A spurious effect on one market may be observed simultaneously in other markets. It might be spread to other markets by traders' subjective beliefs formed by watching the first market.

2.6.1 Decision Criteria

The other markets chosen to test a rule must be both similar and dissimilar. Similar in that it is reasonable to assume that the same underlying forces operate in both. Dissimilar in that the two markets are not so correlated that anomalies in one appear in the other. For example if two markets have largely overlapping participants then some pattern of trading that is due to a meme among the traders, a transient phenomenon that is difficult if not impossible to model, may well appear in both markets.

We have the same question as in section 2.4.2 on Holdouts: how many to use? How many other markets are sufficient? How many markets have to show the same effects, or not, for a data snooping diagnosis?

There can be no objective answer to those questions. Zero other markets is too few, and the more that exhibit the same results the more confident we are that the results are not due to data snooping. The fewer that do the less confident we will be.

What can be done objectively is to decide which other markets will be examined before embarking on the test taking into account what we know about the links between the markets and the correlation of the markets.

2.7 Story Telling

Prices in financial markets move for reasons. We do not believe in magic. Fundamentally it is because traders are buying and selling securities. But traders make those decisions for, not always rational, reasons[10].

So it is reassuring, when a technical trading rule is discovered, to have a reason for its effectiveness. A story. Without such a thing, it is hard to believe that a rule is anything other than a data snooped description of the input data. The lack of a story does not, however, mean that it is data snooping. There may be an unknown economic reason for the rule being effective.

An example of such a story is in Chan et al. (2000) where momentum effects in stock market indexes from a number of countries are found. The story is said to be that momentum is partly driven by post-earnings announcement drift which is in turn driven by under reaction to news.

Computer science literature is replete with examples of trading rules discovered by neural networks, evolutionary algorithms and other “black box” methods. Quite apart from a tendency to ignore trading costs or feasibility (see section 2.8) they often do not have any economic justification.

On the other hand chartists generally have economic reasons for their charts (often to do with market over/under reactions and mean reversion) which on close examination turn out to be spurious[11]. The existence of an economically plausible story is no guarantee that a rule is not a spurious result of data snooping. Indeed the researcher may “over fit the story” by searching for any possible explanation (e.g., Crack (1999)).

[10] See Poteshman and Serbin (2003) for an example of provably irrational behavior in option markets. A particular set of circumstances allowed the authors to prove that some trades were irrational. Or see Rashes (2001) for evidence of traders being temporarily confused by similar ticker symbols. One can only wonder how many other irrational trades there are where it cannot be proved?

[11] There does exist some evidence of some effective charting rules, see for instance Lo et al. (2000).

2.7.1 Decision Criteria

A story cannot be effectively quantified, but that which cannot be measured need not be discarded. Subjective tests can have value.

A rule that has no economic intuition, no story, is less attractive than one that does. A false rule found using data snooping will not have a story, except due to coincidence.

It is human nature to believe results that have been worked hard for. A researcher that discovers a (seemingly) useful technical analysis algorithm will be receptive to reasons why it should work. So to avoid this confirmation bias the story must be known before the rule is found. That is not to say that a story found subsequently must be false, it is just to say that such a “post–hoc” story is not the evidence against data snooping that it would be if it were known ante–hoc.

For a rule that exploits unknown properties or economic laws there can be no story, clearly. But a rule that is found using search guided by a story suits us better than a rule found by searching randomly.

2.8 Feasibility

The Efficient Market Hypothesis (EMH) is that “...any information that could be used to predict stock performance should already be reflected in stock prices” Bodie, Kane, and Marcus (2005, p. 370). The implication is that there do not exist any technical analysis rules, utilising only price/return information, that can be used to gain consistent excess risk adjusted profits from the market after accounting for risk aversion, transaction costs and taxes.

We commonly find that reports of the failure of the EMH fail to consider the feasibility of their trading method. It is not enough to be able to predict prices better than the market, one needs to be able to act on the information. There must be a feasible trading strategy.

The mechanism for the market price adjusting to reflect the information available is investors exploiting (i.e., trading on) the information available to them. Since price information is available at negligible cost it follows from the EMH that the information available in prices cannot be used to gain excess profits (Bodie et al., 2005).

This implies that there can be no profitable technical analysis rule. But there have been many papers that uncover profitable technical analysis rules. For instance in Chavarnakul and Enke (2008) which used neural networks to optimise a volume based trading rule. The optimised rule was reported to perform better than buy and hold, violating the EMH. Yet there was no consideration of transaction costs (as acknowledged by the authors) nor was there consideration of feasibility (e.g., if in the time between generating the signals and implementing the trades the prices moved to remove profits). Without considering these problems the algorithm cannot be said to contradict the EMH.

2.8.1 Decision Criteria

Judging the feasibility of a technical analysis rule after the fact (often many years after the fact in long studies) is not trivial. For two of the studies we look at here we use index data that goes back to 1928, the original studies used data from the 19th century, and times were different then.

To be a feasible strategy there must have been securities to trade. The three studies we consider all use index data, but indexes are not available for trade, and index proxies, in the form of index funds, were not accessible to traders until the 1970s[12]. There was nothing to stop index tracking portfolios being formed before then, except for practical considerations.

Additionally the trading must be profitable after accounting for transaction costs. Ignoring transaction costs is a form of data snooping. When researchers ignore transaction costs they are removing important data from their data sets so that their models are better. In our experience many researchers and many studies do ignore transaction costs, or address them tangentially.

More subtly a rule may generate a return stream that has a positive mean excess return, but the return stream of the rule is more volatile than the return stream of the benchmark it is compared to. When considering feasibility it is important to consider what it is that market participants want and need. Ignoring volatility is also a form of

[12] The Vanguard fund started trading in 1975 and was the first index fund (Bogle, 2006).

data snooping in the same way that ignoring transaction costs is.

To be useful, and to violate the EMH, a trading strategy must (a) be profitable when back tested (including transaction costs) and (b) have a risk/reward ratio that is acceptable. Searching through the space of all possible rules will uncover a lot of rules that, on the face of it, are attractive. But most (if not all) will be infeasible. Such infeasible rules do not need to be examined, further, for data snooping.

It is not possible to precisely quantify what transaction costs will be, as they depend on specific market conditions on the trading day[13]. One approach to transaction costs (and the approach we take in section 5) is to look at the break even transaction cost. We compare a break even transaction cost with what would be reasonable transaction costs from the period. This is an unsatisfactory method. As Bajgrowicz and Scaillet (2009) point out break even transaction costs are considered ex–post so are hard to use when designing rules. Additionally there is no clear "reasonable" transaction cost for any period, let alone one that is many decades in the past. But even so, we can think of no better approach.

It is not possible to define an "acceptable" return/volatility relationship. Every investor has a unique tolerance for variability in returns. The way we deal with the relationship between volatility and returns is to report a Sharpe ratio when we need to. In section 3 the assertion we test is that of greater returns and lower volatility, and we test both independently. The Sharpe ratio is useful, but it is arbitrary. Yet we are again forced to use it as there is no better approach we can think of.

To assess profitability and volatility levels a benchmark must be chosen. Without a benchmark there is no way to illustrate the effectiveness, or not, of a rule. At the very least a benchmark of zero is implied, investing your wealth in a vault. But in the real world that is generally a negative real return.

The EMH postulates that no rule can beat the market (Bodie et al., 2005), so some concept of a market portfolio is needed. For example a portfolio of T–Bills to benchmark a zero-beta long–short fund, or a portfolio of stocks to benchmark a long-only stock fund.

[13] Models of transaction costs do exist, see for example Trethewey and Crack (2010).

2.9 Outlier Analysis

Often the out performance of a model is found to be due to a small handful of significant days. For instance a model of the New York Stock Exchange index (NYSE) in 1987 that is "Stay out of the market in months beginning with 'O' " appears on the face of it ridiculous. But it would be significantly better than buy and hold over that period because of the market crash in October of that year[14].

Those occasions can be viewed as outliers. If there is an algorithm that suggests being long some trading days and short others, and those long days coincide with market rises and the short with falls, such that the rises and falls are in the tails of the distribution of returns there can be two interpretations:[15]

- The algorithm stumbled onto those days by accident.
- The algorithm correctly identified those days. It describes the underlying economic process.

Deciding which interpretation is true is difficult. If a hypothesised algorithm explains the extreme result (the month begins with "O" is not a good explanation) then the experimenter would tend to accept that the algorithm describes the underlying economic process. (See section 2.7 on "Story Telling".)

If the hypothesised algorithm works only for a sub–period of the available data then that result may be explainable by unusual data points in the "good" subset (see the section on "Holdouts". Section 2.4) and the experimenter would tend to accept the algorithm as a happy accident.

2.9.1 What is an Outlier?

"An outlier is an observation that is far removed from the others in the set." Porkess (2004). That is a subjective definition that does not do much to help deal with a data set.

[14] It is an interesting observation that the "stay out of the market in October" works well in many years.

[15] There is no need to assume a parametric distribution. The top and bottom 1% (say) of historical moves can define the tails.

Porkess helps by suggesting two possible definitions of outliers:

1. Points two standard deviations away from the mean.

2. Points more than 1.5× the inter quartile range away from the top or bottom quartile

However there is nothing unusual about data points meeting one or both of these criteria, which implies there is no such thing as outliers in financial markets. Rather they are a natural and important part of the system. In this context we do not view "outliers" as points to be removed. But rather we view them as phenomenon to be acknowledged. They should be removed only if they are a some sort of measurement error, such as a data entry mistake (such as a misplaced decimal point) or a failure of the data capture mechanism (such as a programming error).

2.9.2 Outliers of known Parametric Systems

If the underlying distribution, and its parameters, are known then data that originates from a different distribution can possibly be identified. This is illustrated in figure 2.2 page 29. The green circles plot standard log-normal observations with the sign randomly flipped. The black crosses plot standard normal observations.

However return series for financial data are not from known distributions. The distributions have fat tails with infrequent and irregular extreme values during booms and busts, e.g. in October 2008 there were several extreme moves.

Financial data have been described better as being from a mixture of distributions. See for example Johanson and Sornette (1998) where the authors identify two distributions of returns, the common occurrence and a different distribution in crashes. Thus the crashes can be viewed as distributional outliers. Or see Press (1967) which suggests that the distribution of returns from some stocks in the Dow Jones Industrial Averages index is a "Poisson mixture of normal distributions". We do not see how being from a mixture of distributions gets us anywhere. True a mix of distributions fits the data better but it still leaves us in the same pickle. Is an "outlier" a true outlier or a datum from an ordinary distribution? One "outlier" by itself is probably a true outlier (for example it could be an error) but it is also not very interesting. It is unclear how this can be applied to financial data.

2.9.3 Outliers of Non–Parametric Systems or when Parameters Unknown

Not knowing the parameters of the underlying distribution, nor having an analytical reason to reject some points, we might use a non–parametric approach.

The most extreme values can be removed from the data set, on the assumption that those values are likely to be outliers from a process other than that that is being modeled.

It is often stated that most of the benefit of investing in markets is had on very few best days (Malkiel, 1999, pp. 163). Testing that influence of outliers on VWRETD in figure 2.3, it can be seen that at the 1% level, removing outliers increases mean returns. That is negative outliers dominate. The effect is even more marked at the 10% level.

It remains unclear that there is any characteristic of market prices that we could use to define “outliers”. However we choose to look at extreme values we cannot rule out the possibility that we are simply looking at a natural part of the market.

2.9.4 Errors

Gathering data is an imperfect process, and there will be errors. For example Tajaddini and Crack (2012) find multiple data errors in two foreign exchange data sets. Misreported data can be true outliers. Whatever the mechanisms driving prices, a mistake in reporting those prices is not a valid mechanism. Examining extreme points often will uncover errors[16].

2.9.5 Cupolas

In recent years there has been some excitement in the finance community about copulas. Copulas are a method of converting a group of univariate distributions (several random

[16] Sometimes obvious errors are not extreme points. For example if the bid is greater than the ask price.

variables) into a multivariate distribution (a single function of several random variables). The univariate distributions are the 1–dimensional marginal distributions of the copula.

The motivation is to find a better method of modeling dependence among different securities than correlation (see for instance Rodriguez (2007, pp 401) that uses copulas to model "daily returns from five East Asian stock indices during the Asian crisis, and from four Latin American stock indices during the Mexican crisis,"). Copulas can capture non–linear dependencies and as Rodriguez puts it "...[copulas] are able to capture tail behaviour without the need of using discretion to define extreme outcomes." Rodriguez (2007, pp 403)

Better statistical methods of modelling dependency will provide better description of observed phenomenon in financial markets. Observations of data outside the range expected will lead to even better models[17].

Imagine the next generation model, that is unknown today. A model that in back testing can predict all of the observed phenomena, including phenomena not used to set the model's parameters. If in subsequent time periods data are uncovered (by the passage of time) that defy the predictions of the model, in what sense are those data outliers? Clearly using the definition supplied by the model they are, but that is scant consolation to investors losing wealth. The model, no matter how good, is allways an imperfect description. The true process remains unknown.

2.9.6 Decision Criteria

It is entirely unclear if there is any sort of outlier analysis that can be useful for detecting data snooping. It can be used to explain the performance of a rule, but that simply defers the problem. A rule's efficacy may be due to identifying true outliers (from a different distribution). But that is not to distinguish a rule due to data snooping, that is effective only on the training set from a true rule that describes some property of the underlying distribution.

[17] Rodriguez (cited above Rodriguez (2007, pp 401)) says: "I find evidence of changing dependence structures during periods of financial turmoil. I also test whether these changes are best described using copulas with tail dependence and asymmetry compared to symmetric and also to tail-independent copulas." This is typical, after the event it is possible to describe it. But a description is not a model.

2.9.7 Outliers, Summary

- However defined, outliers are interesting points.
- Except for errors there is no justification for removing outliers.
- Given the importance of extreme values in financial series, predicting and exploiting outliers is an acceptable goal for a technical analysis algorithm.

2.10 Literature Review

In Lo and MacKinlay (1990) the authors credit Aldous (1989) for coining the term "data snooping statistics", "where you have a family of test statistics $T(a)$ whose null distribution is known for fixed a, but where you use the test statistic $T = T(a)$ for some a chosen using the data" Lo and MacKinlay (1990, p433). Lo and MacKinlay examine data snooping biases in finance and find "...that even mild forms of data snooping can change inferences substantially." (Lo and MacKinlay, 1990, p435)

Concern about data snooping is not limited to the finance literature. Ioannidis (2005) claims to show that because of data snooping it is the case that most research findings, in medicine, are false.

Data Snooping takes many forms. Crack (1999) purposely sets out to use data snooping to show a "relationship" between phases of the moon and S&P 500 returns[18]. This is a clear case of searching for a relationship until one is found. In Leippold and Lohre (2012) a recent example of simultaneously testing several hypotheses is presented.

Arguably the most important paper published on the subject of data snooping is Sullivan et al. (1999). They used an algorithm (introduced in White (2000)) based on the "Stationary Bootstrap" introduced in Politis and Romano (1994). This was further refined in Hansen (2005), Romano and Wolf (2005) and Hsu and Kuan (2005).

The original papers by Sullivan et al. and the techniques developed therein have been extensively cited in the Finance literature since, becoming something of a "gold standard" for correcting for data snooping bias. See for example Qi and Wu (2005) where a similar study to Sullivan et al. (1999) was performed on the foriegn exchange market, or Hsu

[18] We look closely at Crack (1999) in chapter 3

and Kuan (2005) where Sullivan et al.s study was expanded to include more sophisticated trading methodologies.

There has been criticism too. Bajgrowicz and Scaillet (2009) criticise Sullivan et al. for not accounting for transaction costs and for a lack of out of sample success for the techniques presented that were claimed to be free of data snooping bias. In chapter 5 we make similar criticisms of Sullivan et al. (1999) and also question the assumptions behind their use of the stationary bootstrap.

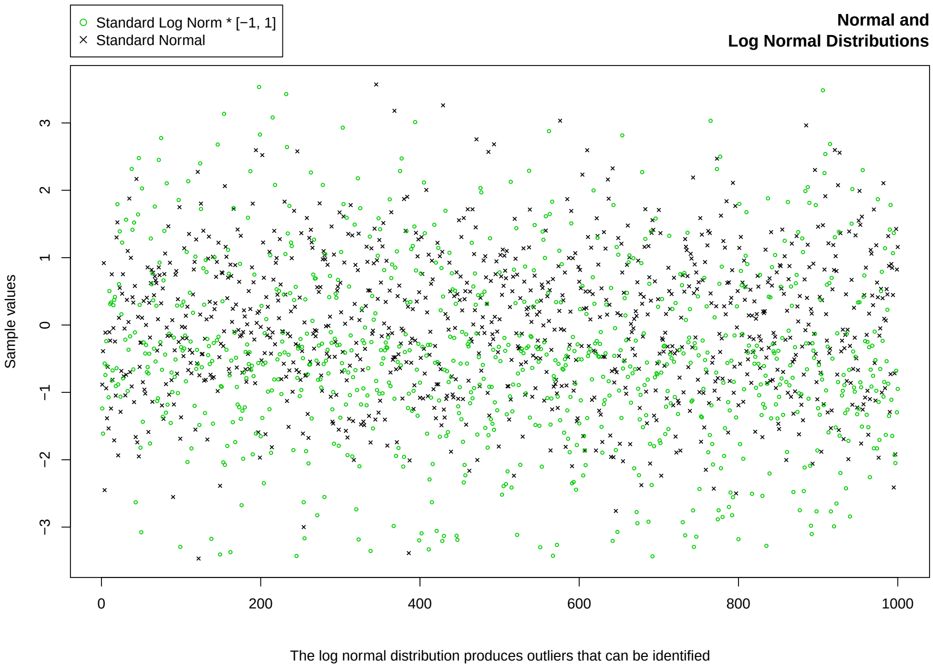


Figure 2.2: When two known distributions are mixed the outliers generated by one may be apparent.

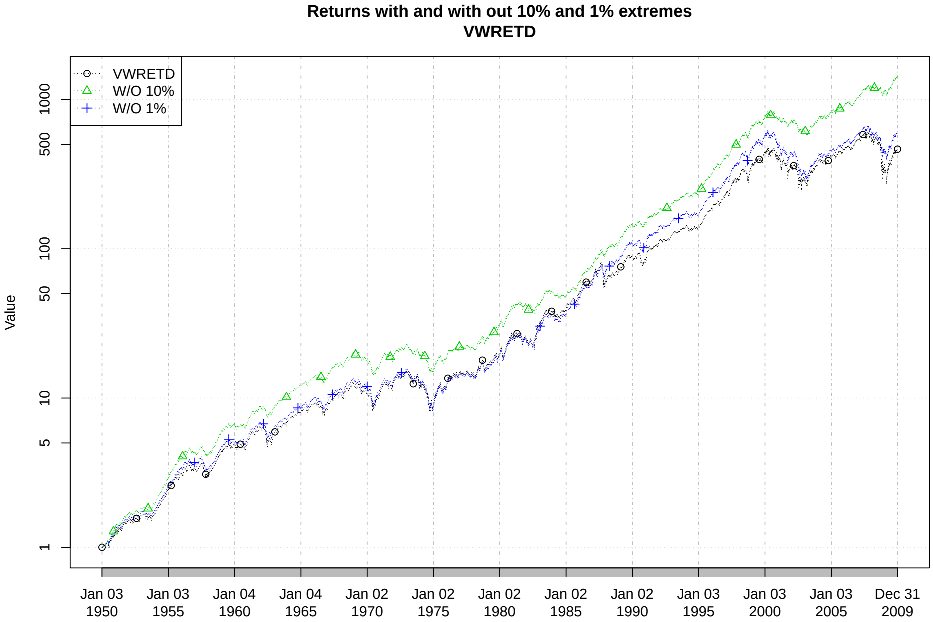


Figure 2.3: The effect of removing outliers from the VWRETD series. Three value series: With no change (black circles), with the returns in the most extreme 1% truncated to the 1% quantiles, positive and negative (blue plus signs) and similarly for the 10% quantiles (green triangles).

Chapter 3

A Classic Case: Lunar Investing

The lunatic is on the grass...
The lunatic is in the hall...
The lunatic is in my head...
Got to keep the loonies on the path

Waters (1973)

In this section we reexamine a study where the author claimed to be data snooping on purpose.

In Crack (1999) a technique is given for using the phase of the moon to divide daily returns and volatility (variance) of the combined NYSE/AMEX and NASDAQ equally weighted and value weighted indexes into periods of higher returns/low volatility and lower returns/higher volatility. Statistical significance was used as the objective function. It was found that the periods around the new and full moons had significantly smaller daily returns and significantly higher volatility than days in the rest of the month[1].

3.1 Profitable Moon

The natural null hypothesis is that the moon has no influence over investor behaviour[2]. The alternative given by Crack (1999) is:

[1] Kathy Yaun, Lu Zheng and Qiaoqiao Zhu published a paper (Yuan et al., 2006) in the Journal of Empirical Finance in 2006 that found returns were about 3% to 5% higher close to the full moon than to the new moon for global indexes. Ilia D. Dichev and Troy D. Janes published a paper (Dichev and Janes, 2003) that found that returns were higher nearer the new moon than the full moon for all USA indexes and for 23 out of 24 other countries.

[2] There is a lot of folklore concerning the moon and human moods and behaviours. However Crack (1999) looked into the available psychology literature at the time of his paper and found no scientific support for lunar effects on human behaviour.

> ...the period of time near the new moon or the full moon is “bad” relative to the interphase period (the period away from new and full moon). ...my lunar behavioral hypothesis is simply that near new moon or full moon, volatility will be higher and returns will be lower than during the interphase period.

The question is “Was Crack data snooping?” (as he claimed).

3.1.1 Economic Intuition

Suspending disbelief and rejecting the null hypothesis that the moon has no detectable effect on investment returns, what are the parameters of the moon’s influence? When testing methods for detecting the differing periods of lunar influence and differentiating real rules from rules purely a result of data snooping it will be important to have some idea of what makes sense economically.

For instance: If a particular phase of the moon is associated with higher returns when compared to the rest of the lunar cycle, say the new moon with none of the moon’s face illuminated, then it is reasonable to believe that the day before and the day after that phase will not be among the worst returns. The moon’s level of illumination is at a particular value for a vanishingly small period of time. To the human observer the phase will seem to last two or three days and the descriptive terms (full, first/second quarter, new etcetera) are approximations.

Any real effect should be apparent in more than one market. We are supposing some unknown deep mechanism here in human psychology and/or physiology. It is reasonable then to expect that the effect would not be specific to a particular geographic market.

So when considering, ante–hoc, what data snooping detection tools to use from section 2, we can expect neighbourhood analysis (section 2.5) and other markets (section 2.6) to be effective. Also we should not be able to conclusively reject the “story”. That is, we must not be able to prove that the moon has no effect on human mood.

3.1.2 Algorithm in “A Classic Case...”

The “full” and “new” moons are described with the parameter P, the proportion of the moon’s face illuminated by the sun.

The periods near and far from the new and fill moon are defined using a parameter α, $0 < \alpha < 0.5$. The "new/full" (NF) period is when $P < \alpha$ or $P \geq 1 - \alpha$, the "interphase" (IP) period (the rest of the lunar cycle) $\alpha \leq P < 1 - \alpha$[3].

The data used by Crack are for 1 January 1977 – 31 December 1996. The values for α reported in Crack (1999) are $0.20, 0.25, 0.30, 0.35, 0.40$. It was found that without exception mean return was higher and volatility (variance) lower in the "new/full" than in the "interphase" period. Setting α at 0.3 gave the most significant results.

3.2 Reproduction

The first step was to reproduce Crack's results. It is important to assure ourselves that the results that were presented are accurate.

The return data were taken from the indexes VWRETD and EWRETD ([VE]WRETD) published by the Center for Research in Securities Prices (CRSP). These are respectively value and equally weighted composite indexes for the NYSE/AMEX and NASDAQ. The TRETURNS field was used (being total returns, including dividends).

The algorithm was reimplemented in **R**. See section B page 133 for the code that produced these results. The original Matlab code is in section B.1 page 141[4].

The data that are now on the United States Naval Meteorology and Oceanography Command's web site differs from that obtained from Crack on two days. On 8–January–1984 and 19–September-1996. 8–January–1984 was not a trading day and 19–September–1996 was, but is not included in the original experiment so these inconsistancies in the data are of no consequence in this section.

For this reproduction the moon data did not differ in any meaningful way from Crack's.

The results for the repeat of the experiment are in tables 3.1 and 3.2.

The results are very close to the original. The best value for α is 0.3 and the p–values at that α are significant at 5%. This indicates that the period when the moon is less than 30% full or more than 70% full has higher returns and lower volatility than the other periods. On the face of it, there is a "lunar effect".

[3] Crack (1999) has a typo where $P = 1 - \alpha$ is in both the "interphase" and "new/full" periods.

[4] Professor Crack made the original lunar phase data he used available, and also the Matlab code that he used.

α	N_{NF}	N_I	$\mu_{NF} \times 253$	$\mu_I \times 253$	Z–stat	(p val.)	$\sigma_{NF} \times \sqrt{253}$	$\sigma_I \times \sqrt{253}$	Z–stat	(p val.)
0.20	2948	2108	0.1016	0.1819	−1.1464	0.1258	0.1390	0.1215	1.2909	0.0984
			(0.0454)	(0.0468)			(0.0188)	(0.0073)		
0.25	3334	1722	0.1029	0.1974	−1.3418	0.0898	0.1372	0.1212	1.3123	0.0947
			(0.0425)	(0.0500)			(0.0174)	(0.0077)		
0.30	3703	1353	0.1017	0.2263	−1.7939	0.0364	0.1349	0.1237	1.1264	0.1300
			(0.0386)	(0.0539)			(0.0161)	(0.0090)		
0.35	4056	1000	0.1192	0.1994	−1.0311	0.1512	0.1336	0.1252	0.8841	0.1883
			(0.0394)	(0.0568)			(0.0154)	(0.0092)		
0.40	4390	666	0.1249	0.2018	−0.9081	0.1819	0.1325	0.1288	0.3721	0.3549
			(0.0354)	(0.0704)			(0.0149)	(0.0098)		

Table 3.1: Results of Crack's algorithm applied to the VWRETD (value weighted) index from 1977-01-01 to 1996-12-31. The first column (α) defines how the lunar month is divided into near new or full (NF) or "interphase" (I). The next two columns (N_{NF} and N_I) are the number of days in both sets (there are 5,056 daily returns over all). The next two columns ($\mu_{NF} \times 253$ and $\mu_I \times 253$) record the mean returns for the periods (defined by α) annualised. The Z–Stat and p–value columns are the results from the GMM test with the null hypothesis that the returns in the NF period and the I period are the same. The four final columns apply to the standard deviation with the means, Z–stat and p–value.

α	N_{NF}	N_I	$\mu_{NF} \times 253$	$\mu_I \times 253$	Z–stat	(p val.)	$\sigma_{NF} \times \sqrt{253}$	$\sigma_I \times \sqrt{253}$	Z–stat	(p val.)
0.20	2948	2108	0.2138	0.2812	−1.3483	0.0888	0.1060	0.0918	1.4516	0.0733
			(0.0534)	(0.0458)			(0.0141)	(0.0058)		
0.25	3334	1722	0.2129	0.2981	−1.7112	0.0435	0.1043	0.0921	1.4135	0.0788
			(0.0514)	(0.0468)			(0.0130)	(0.0060)		
0.30	3703	1353	0.2087	0.3328	−2.2135	0.0134	0.1036	0.0908	1.8331	0.0334
			(0.0497)	(0.0525)			(0.0120)	(0.0067)		
0.35	4056	1000	0.2225	0.3206	−1.6877	0.0457	0.1022	0.0924	1.6304	0.0515
			(0.0483)	(0.0548)			(0.0114)	(0.0074)		
0.40	4390	666	0.2290	0.3267	−1.6423	0.0503	0.1016	0.0915	1.0098	0.1563
			(0.0458)	(0.0623)			(0.0116)	(0.0057)		

Table 3.2: Results for the EWRETD (equally weighted) index from 1977-01-01 to 1996-12-31. The columns are the same as in table 3.1

3.2.1 Detecting Data Snooping. Decision Criteria

Now we are satisfied that we can reproduce the results that Crack got, we ask "is it data snooping?". In this section the tools from section 2 for detecting data snooping are considered, ante–hoc, to prepare for considering whether Crack's results are due to data snooping.

3.2.1.1 Statistical Significance

The objective function of Crack's study was statistical significance. Thus by the measure he used his results were significant.

The test that Crack used for statistical significance was implemented by "Generalised Method of Moments" (GMM, Arnold and Crack (1999))[5]. The GMM results should be compared to the results from other tests.

According to Crack (1999) the GMM tests "with Newey–West standard errors" "...are therefore robust to both auto–correlation and heteroskedasticity in the moments".

In addition to repeating the GMM tests, a standard t–test and a Wilcoxon–Mann non–parametric test will be used.

As discussed in section 2.3.6 the standard t–test is robust to non–normality, with the amount of data we have, but is not robust to auto correlation. The Wilcoxon–Mann test being a non–parametric test is robust to non–normality, auto–correlation and heteroskedasticity.

3.2.1.2 Holdouts

To implement holdouts we need to: select which holdout data to use, how much holdout data to use and what level of agreement we need between the holdout period and test period.

Crack's study used data from 1977–1996 inclusive. There are, from CRSP, data available from 1926 to 2009. Crack's study covered 20 years, and since there are 20 years of data available from before Crack's study we will use the period 1957–1976 as the first holdout period. That answers both which and how much. As for the agreement: If the

[5] GMM is not a widely used test outside of economics.

moon truly influences returns in the manner uncovered by Crack then the same influence should be present before and after the period covered by Crack. We will use a second holdout period of 1997–2009. It is conceivable that after uncovering an inefficiency in the market (the lower returns near the new and full moon) traders exploited and eliminated the effect. In which case we expect to see the lunar effect in the 1957–1976 period but not in the 1997–2009 period.

In Summary:

- We use the two periods 1957–1976 and 1997–2009 as holdouts.
- We expect to see the same statistically significant lunar influenced returns in both holdout periods. Allow for the possibility that the effect is only apparent in the first holdout period hypothesising that the discovery of the lunar influence alerted traders who moved to eliminate it.

3.2.2 Neighbourhood Analysis

Crack's algorithm divided the lunar month using α (see section 3.1.2). For each possible value of α, say α_i we can define another, $\alpha_{i+1} > \alpha_i$ such that α_{i+1} is the smallest value of α greater than α_i that divides the months differently from α_i. If IP_i and NF_i are the IP and NF sets defined by α_i, then α_{i+1} is the smallest α such that $\alpha_{i+1} > \alpha$ and $(IP_i \neq IP_{i+1}$ or $NF_i \neq NF_{i+1})$.

Because there are a finite number of trading days, the set of α_is is finite and defines neighbourhoods in Crack's analysis. The sets NF_i and IP_i are neighbours of sets NF_{i+1} and IP_{i+1}, and α_i is "next to" α_{i+1}. All the αs are in table 3.7 on page 43.

Ante–hoc there is no practical criteria to set over what the transition between "neighbours" should look like, or how to distinguish brittle transitions. All that we can do is learn, post–hoc, and train our judgement.

3.2.3 Outlier Analysis

In section 2.9.6 we gave up on using outlier analysis for detecting data snooping. But it still may be revealing to analyse the outliers. We may learn the reason the algorithm performs as it does: we may find the trading days responsible.

3.2.4 Other Markets

As discussed in section 2.6.1, we want to use markets that are correlated (so the same underlying processes operate) but not too correlated (so one market is essentially the same as the other).

The VWRETD and EWRETD indexes are NYSE and NASDAQ composite indexes. So we should look for markets that were well developed in the 1977–1996 period.

Also for comparison Crack's investigation of the VWRETD and EWRETD indexes the other markets need to have been in operation over the same period.

The choice of which markets to use was also guided by availability. In the end all developed market indexes with data covering the 1977–1996 period that are available currently via the Otago University Bloomberg terminal were used. These six indexes selected for the Other Markets test are in two groups of three. The first group are indexes from the USA and overlap explicitly with Crack's data. There are members of the overlapping indexes that are also included in the CRSP data set. The second group are from foreign countries and, whilst they may be correlated to a greater or lessor extent (see table 3.10) they do not specifically overlap.

Overlapping Market Indexes

CCMP The NASDAQ composite index (thus a component of both VWRETD and EWRETD that Crack used). Capitalisation weighted index of three NASDAQ markets: Global Select, Global Market and Capital Select. `http://www.nasdaq.com/symbol/ccmp`

INDU The Dow Jones Industrial Average. A price–weighted average of the 30 blue–chip stocks that are generally the leaders in their industry.

`http://www.djaverages.com/index.cfm?go=industrial-overview`

SPX Standard and Poor's 500. A capitalisation weighted index of 500 stocks representing all major industries in the USA. `http://www2.standardandpoors.com/spf/pdf/index/SP_500_Factsheet.pdf`

Non–Overlapping Market Indexes

HSI The Hang Seng Index. Free–float capitalisation weighted index of companies from the Stock Exchange of Hong Kong[6]. `http://www.hsi.com.hk/HSI-Net/`

NKY The Nikkei–225 Stock Average. A price weighted average of the 225 top rated Japanese companies listed in the first section of the Tokyo Stock Exchange. `http://www.bloomberg.com/quote/NKY:IND`

SPTSX A capitalisation weighted index designed to measure market activity of stocks listed on the Toronto Stock Exchange. `http://www.standardandpoors.com/indices/sp-tsx-60/en/us/?indexId=spcadntx--caduf--p-ca-l--`

The results of using these markets are described in section 3.6.

[6] "Free–float" refers to the fact that when calculating the weights for the capitalisation weighting only shares that are not closely held (e.g., by company insiders) are counted.

3.2.5 Story Telling

According to folklore the moon's phase has an effect on human moods and behaviour. Thus it is feasible, if one chooses to believe the tales, that the phase of the moon could influence markets via the moods of market participants. Crack found that returns were higher away from full and new moons. The story could be that around new and full moons investors are more nervous and bearish.

However Crack reported that there is no scientific evidence of any effects of the moon and human mood. If such an effect were to be found in financial markets it would be a new effect.

Still there is no positive reason to reject the hypothesis of the moon effecting human moods. But we could find no stories that would imply that the "interphase" period would give better returns and volatility.

3.2.6 Feasibility

There is no direct way of trading on the value of the indexes VWRETD or EWRETD. However if the volatility differences exist and are not priced then there are potential option trading strategies that exploit that miss–pricing[7]. If lunar months can be divided into high and low return portions than moving into high β stocks in high return periods and into low β stocks in low return periods could potentially be profitable.

3.3 Implementing Statistical Significance

As laid out in section 3.2.1.1 there are three tests to be compared:

1. Generalised Method of Moments tests.

2. t–test for two independent samples.

3. Wilcoxon–Mann test.

Statistical significance was the objective function for Crack's study, and we have already reproduced his results. So we know that using the test he used, and at his best α

[7] Studying implied volatility using options, and looking for correlations with lunar cycles is another approach.

the results were significant. Our objective here is then to look at statistical significance with the other two tests as well. A statistical significance test is a mathematical analogy of an opinion. We are getting second and third opinions.

Using the following values for α the statistical significance of testing the mean returns in the IP period and NF period (the null hypothesis being that the returns for the two periods are the same) for all three tests was calculated: 0.15 0.17 0.19 0.21 0.24 0.26 0.28 0.30 0.32 0.34 0.36 0.39 0.41 0.43 0.45 and plotted in figure 3.1.

From this we can see again that the GMM test clearly indicated significantdifference in mean returns at Crack's optimal α. The t–test agrees with Crack for EWRETD but only just for VWRETD. The Wilcoxon–Mann test has a different opinion and does not reject the null hypothesis that there is no difference in mean returns for the IP and NF periods. However the Wilcoxon–Mann test, being non–parametric, has less power and is therefore more prone to accepting a false null.

We cannot do mathematics by holding elections. All we can say that is that the results are not as satisfactory as Crack found when viewed through the lens of other statistical tests.

But we can notice that Crack used one–sided tests. We have followed Crack's techniques quite closely, and have done one–sided tests too (which explains the high p–values in tables 3.3 – 3.5). But Crack is looking for a difference in returns between the interphase and new/full periods. There is no anti–hoc information as to which returns should be higher. So in this case the use of one–sided tests is itself data snooping. If Crack had used two–sided tests then fewer significant results would have been found.

3.4 Implementing Holdouts

In table 3.3 and 3.4 the results for the later holdout period (1997 – 2009) are shown[8]. The mean return results for 1997–2009 are both the opposite of what Crack (1999) found (i.e., $\mu_{NF} > \mu_N$) is not significant. There is no clear pattern for the standard deviation of returns. Moreover, and damningly, the results are not statistically significant even using

[8] Lunar data from Meteorology and Command (2010) had to be used as the previous data that Crack used did not extend after 1996, when he made the file.

the GMM test. And given that we are using statistical significance from the GMM test as our objective criteria we can say that the holdouts, after Crack's experiment, are a failure[9].

In tables 3.5 and 3.6 the results for the earlier period (1957 – 1977) are presented. The results are similar to the results for the later period with NF returns consistently higher that IP returns, no clear pattern for standard deviation and no detected significant p–values using GMM.

So it is clear that the results are not significant for the period following Crack's study and importantly the results are not significant for the period immediately preceding Crack's study.

3.5 Implementing Neighbourhood Analysis

In table 3.7 is a representative sample of all possible αs. For example taking the first two αs in table 3.7: 0.2599514 and 0.2699923. The division of the days into IP and NF by 0.2599514 is different from the division of the days into IP and NF by 0.2699923. But there is no $\alpha^{'}$ such that $0.2599514 < \alpha^{'} < 0.2699923$ and the division of days into IP and NF using $\alpha^{'}$ differs from that produced by α equals 0.2599514 or α equals 0.2699923.

So the neighbourhood around Crack's reported "best α", 0.3, consists of 10 unique points (αs) inside the region defined by the nearest two other αs Crack looked at, 0.25 and 0.35.

We ran Crack's study using these αs and the results are in table 3.8 and figure 3.2. The regions around Crack's best α (0.3) are significant for mean return for both VWRETD and EWRETD. The volatility improvements are significant around Crack's best α for the Equally weighted index but none is significant for the Value weighted index.

The near neighbourhoods are similar in performance. So there is no evidence of data snooping from Neighbourhood Analysis.

[9] Consistently positive results in the absence of statistical significance should make us question relying on statistical significance measures as an objective function. We discussed this in section 2.3.1.

α	N_{NF}	N_I	$\mu_{NF} \times 253$	$\mu_I \times 253$	Z–stat	(p val.)	$\sigma_{NF} \times \sqrt{253}$	$\sigma_I \times \sqrt{253}$	Z–stat	(p val.)
0.20	1906	1366	0.1191	−0.0333	1.4832	0.9310	0.2143	0.2111	0.5104	0.3049
			(0.0642)	(0.0937)			(0.0289)	(0.0270)		
0.25	2166	1106	0.1416	−0.1132	1.7272	0.9579	0.2148	0.2091	1.1124	0.1330
			(0.0588)	(0.1344)			(0.0279)	(0.0282)		
0.30	2399	873	0.0912	−0.0427	0.8653	0.8066	0.2141	0.2099	0.7932	0.2138
			(0.0588)	(0.1455)			(0.0277)	(0.0290)		
0.35	2631	641	0.0871	−0.0742	0.8949	0.8146	0.2133	0.2117	0.1887	0.4251
			(0.0500)	(0.1837)			(0.0271)	(0.0318)		
0.40	2843	429	0.0864	−0.1493	1.1730	0.8796	0.2118	0.2206	−0.7195	0.7641
			(0.0526)	(0.2086)			(0.0268)	(0.0358)		

Table 3.3: Results for the VWRETD (value weighted) index from 1997-01-01 to 2009-12-31. This is after the period covered by Crack's study (Crack (1999)) and shows that the effects noticed by Crack disappears in the period after his study. (The columns are the same as for table 3.1 page 34)

α	N_{NF}	N_I	$\mu_{NF} \times 253$	$\mu_I \times 253$	Z–stat	(p val.)	$\sigma_{NF} \times \sqrt{253}$	$\sigma_I \times \sqrt{253}$	Z–stat	(p val.)
0.20	1906	1366	0.2267	0.1142	1.2416	0.8928	0.1794	0.1763	0.8123	0.2083
			(0.0712)	(0.0949)			(0.0286)	(0.0281)		
0.25	2166	1106	0.2354	0.0708	1.2444	0.8933	0.1787	0.1769	0.4467	0.3276
			(0.0640)	(0.1326)			(0.0277)	(0.0296)		
0.30	2399	873	0.1989	0.1273	0.5540	0.7102	0.1788	0.1762	0.5715	0.2838
			(0.0641)	(0.1366)			(0.0277)	(0.0301)		
0.35	2631	641	0.2019	0.0891	0.7313	0.7677	0.1785	0.1764	0.2400	0.4052
			(0.0601)	(0.1686)			(0.0273)	(0.0329)		
0.40	2843	429	0.2003	0.0437	0.8975	0.8153	0.1768	0.1861	−0.7205	0.7644
			(0.0630)	(0.1904)			(0.0270)	(0.0374)		

Table 3.4: Results for the EWRETD (equally weighted) index from 1997-01-01 to 2009-12-31. This is after the period covered by Crack's study (Crack, 1999) and shows that the effects noticed by Crack disappears in the period after his study. (The columns are the same as for table 3.1 page 34)

α	N_{NF}	N_I	$\mu_{NF} \times 253$	$\mu_I \times 253$	Z–stat	(p val.)	$\sigma_{NF} \times \sqrt{253}$	$\sigma_I \times \sqrt{253}$	Z–stat	(p val.)
0.20	2931	2084	0.0887	0.0573	0.4909	0.6882	0.1163	0.1170	−0.1787	0.5709
			(0.0431)	(0.0550)			(0.0086)	(0.0094)		
0.25	3312	1703	0.0856	0.0565	0.4414	0.6705	0.1162	0.1173	−0.3308	0.6296
			(0.0426)	(0.0575)			(0.0085)	(0.0094)		
0.30	3673	1342	0.0783	0.0685	0.1476	0.5587	0.1178	0.1131	0.7859	0.2160
			(0.0425)	(0.0575)			(0.0090)	(0.0095)		
0.35	4032	983	0.0799	0.0582	0.3041	0.6195	0.1170	0.1149	0.3406	0.3667
			(0.0394)	(0.0677)			(0.0087)	(0.0106)		
0.40	4364	651	0.0681	0.1265	−0.6185	0.2681	0.1166	0.1168	−0.0377	0.5151
			(0.0392)	(0.0884)			(0.0086)	(0.0115)		

Table 3.5: Results for the VWRETD (value weighted) index from 1957-01-01 to 1976-12-31. This is before the period covered by Crack's study (Crack, 1999) and shows that the effects noticed by Crack was not apparent in the period before his study. (The columns are the same as for table 3.1 page 34)

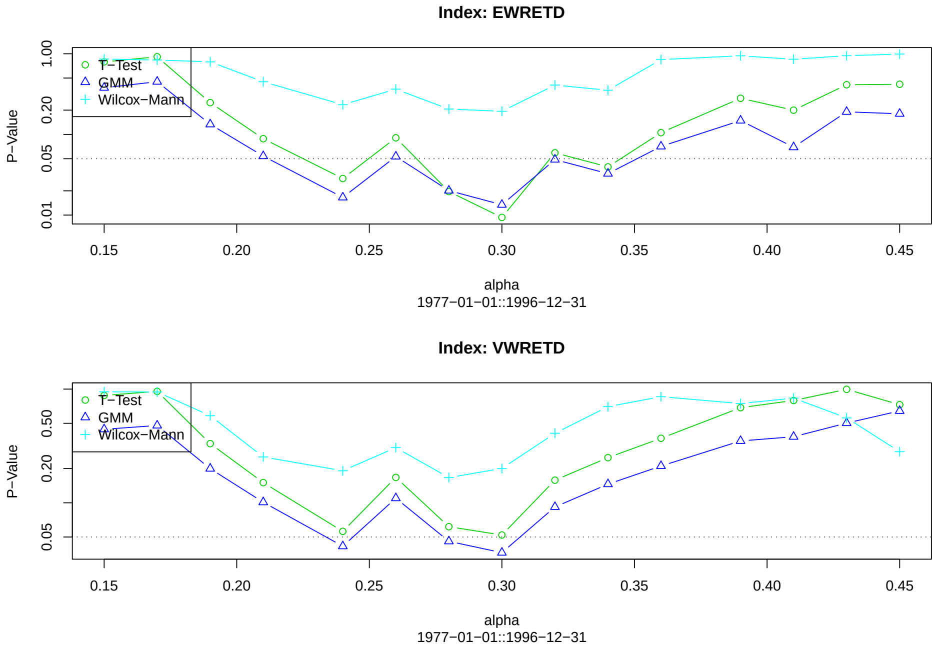


Figure 3.1: Significance for various αs and three tests: Wilcoxon–Mann, GMM and the t–test. The null hypothesis is that mean returns for IP days equals the mean return for NF days.

α	N_{NF}	N_I	$\mu_{NF} \times 253$	$\mu_I \times 253$	Z–stat	(p val.)	$\sigma_{NF} \times \sqrt{253}$	$\sigma_I \times \sqrt{253}$	Z–stat	(p val.)
0.20	2931	2084	0.1418 (0.0574)	0.1157 (0.0627)	0.4703	0.6809	0.1126 (0.0067)	0.1142 (0.0100)	−0.2661	0.6049
0.25	3312	1703	0.1419 (0.0585)	0.1097 (0.0625)	0.5453	0.7072	0.1128 (0.0069)	0.1142 (0.0099)	−0.2756	0.6086
0.30	3673	1342	0.1357 (0.0574)	0.1180 (0.0597)	0.3308	0.6296	0.1142 (0.0072)	0.1107 (0.0110)	0.4485	0.3269
0.35	4032	983	0.1372 (0.0547)	0.1052 (0.0730)	0.5024	0.6923	0.1134 (0.0071)	0.1126 (0.0121)	0.0949	0.4622
0.40	4364	651	0.1243 (0.0557)	0.1754 (0.0772)	−0.6648	0.2531	0.1131 (0.0071)	0.1145 (0.0139)	−0.1524	0.5606

Table 3.6: Results for the EWRETD (equally weighted) index from 1957-01-01 to 1976-12-31. This is before the period covered by Crack's study Crack (1999) and shows that the effects noticed by Crack was not apparent in the period before his study. (The columns are the same as for table 3.1 page 34)

$$\alpha \in \left\{ \begin{array}{cc} 0.2599514 & 0.2699923 \\ 0.2799991 & 0.2899573 \\ 0.2999561 & 0.3099852 \\ 0.3199322 & 0.3299898 \\ 0.3399466 & 0.3499958 \end{array} \right\}$$

.

Table 3.7: The αs in Crack's algorithm (Crack, 1999) divide the period 1977–1996 into periods of "near new and full moon" (NF) and "interphase moons" (IP). In this table are all the αs such that $0.25 < \alpha < 0.35$ and such that there is a change in the way the α divides the days into NF and IP sets.

EWRETD (Equally Weighted)

α	N_{NF}	N_I	$\mu_{NF} \times 253$	$\mu_I \times 253$	Z–stat	(p val.)	$\sigma_{NF} \times \sqrt{253}$	$\sigma_I \times \sqrt{253}$	Z–stat	(p val.)
0.35	4023	1033	0.2218	0.3202	−1.6182	0.0528	0.1022	0.0928	1.4879	0.0684
			(0.049)	(0.0548)			(0.0115)	(0.0072)		
0.3399	3959	1097	0.2224	0.3124	−1.4866	0.0686	0.1022	0.0932	1.2997	0.0968
			(0.0487)	(0.0567)			(0.0117)	(0.0069)		
0.33	3872	1184	0.2192	0.316	−1.6292	0.0516	0.1028	0.0918	1.523	0.0639
			(0.0492)	(0.055)			(0.0119)	(0.0066)		
0.3199	3815	1241	0.2153	0.3236	−1.8986	**0.0288**	0.1028	0.0923	1.6044	0.0543
			(0.0485)	(0.0555)			(0.0118)	(0.0071)		
0.31	3743	1313	0.2101	0.3325	−2.1656	**0.0152**	0.1033	0.0914	1.7387	**0.041**
			(0.0495)	(0.0531)			(0.0119)	(0.0069)		
0.30	3666	1390	0.2113	0.3225	−1.9762	**0.0241**	0.1035	0.0916	1.6949	**0.0451**
			(0.0492)	(0.0541)			(0.0121)	(0.0065)		
0.29	3605	1451	0.2088	0.3242	−2.172	**0.0149**	0.1038	0.0911	1.7428	**0.0407**
			(0.0495)	(0.0514)			(0.0123)	(0.0064)		
0.28	3533	1523	0.2149	0.3044	−1.7832	**0.0373**	0.1035	0.0926	1.4172	0.0782
			(0.0498)	(0.0492)			(0.0124)	(0.0065)		
0.27	3452	1604	0.2178	0.2937	−1.5168	0.0647	0.1037	0.0928	1.3526	0.0881
			(0.0504)	(0.0481)			(0.0127)	(0.0063)		
0.26	3375	1681	0.2144	0.2971	−1.73	0.0418	0.1041	0.0923	1.4148	0.0786
			(0.0507)	(0.0465)			(0.0129)	(0.0062)		

VWRETD (Value Weighted)

α	N_{NF}	N_I	$\mu_{NF} \times 253$	$\mu_I \times 253$	Z–stat	(p val.)	$\sigma_{NF} \times \sqrt{253}$	$\sigma_I \times \sqrt{253}$	Z–stat	(p val.)
0.35	4023	1033	0.1211	0.1894	−0.8616	0.1945	0.1338	0.1248	0.9127	0.1807
			(0.04)	(0.0573)			(0.0155)	(0.0091)		
0.3399	3959	1097	0.1204	0.1879	−0.8451	0.199	0.1337	0.1257	0.7611	0.2233
			(0.0396)	(0.0598)			(0.0158)	(0.0088)		
0.33	3872	1184	0.1108	0.2145	−1.3811	0.0836	0.1344	0.1237	0.9701	0.166
			(0.0394)	(0.0564)			(0.016)	(0.0084)		
0.3199	3815	1241	0.1077	0.2192	−1.5691	0.0583	0.1342	0.1251	0.9769	0.1643
			(0.0379)	(0.0561)			(0.0157)	(0.0094)		
0.31	3743	1313	0.0988	0.2386	−2.0551	**0.0199**	0.1348	0.1237	1.1415	0.1268
			(0.0379)	(0.0537)			(0.0159)	(0.0091)		
0.30	3666	1390	0.1028	0.2202	−1.6678	**0.0477**	0.135	0.1237	1.1241	0.1305
			(0.0383)	(0.0555)			(0.0162)	(0.0087)		
0.29	3605	1451	0.1015	0.2184	−1.734	**0.0415**	0.1356	0.1227	1.2201	0.1112
			(0.0392)	(0.0514)			(0.0164)	(0.0084)		
0.28	3533	1523	0.1057	0.2031	−1.4499	0.0735	0.1357	0.1229	1.1851	0.118
			(0.0402)	(0.0497)			(0.0167)	(0.0083)		
0.27	3452	1604	0.1098	0.1895	−1.1526	0.1245	0.1362	0.1225	1.2003	0.115
			(0.0411)	(0.0503)			(0.017)	(0.008)		
0.26	3375	1681	0.1035	0.1984	−1.3796	0.0839	0.1369	0.1216	1.2928	0.098
			(0.0416)	(0.0496)			(0.0173)	(0.0078)		

Table 3.8: Results for the CRSP VWRETD (value weighted) and EWRETD (equally weighted) indexes, from 1977-01-01 to 1996-12-31 for αs in the neighbourhood of Crack's best α. The significant p–values (in bold) are in the region of Crack's best α. The closest neighbours are significant.

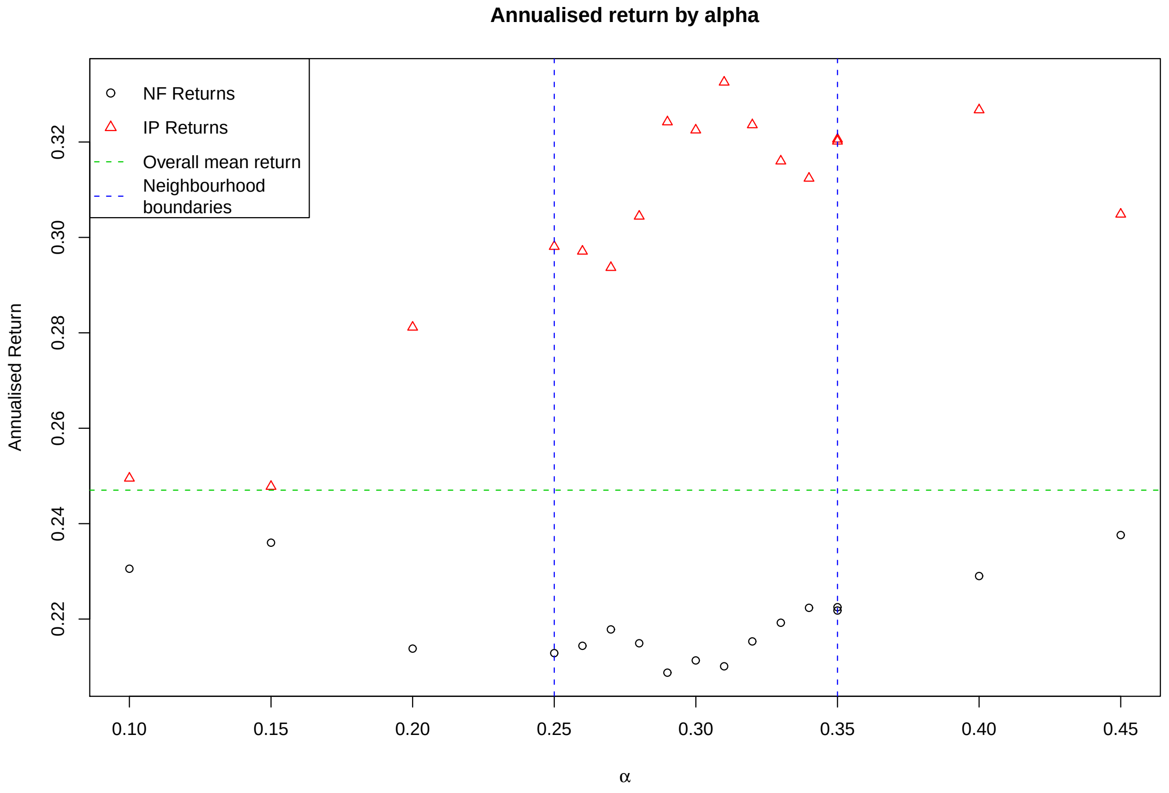


Figure 3.2: The mean EWRETD returns separated into NF and IP over the full range of αs. The “interesting” neighbourhood, described in the text and in table 3.7, is highlighted and all αs in that neighbourhood are plotted.

3.6 Implementing Other Markets

For the reasons outlined in section 3.2.4 we are going to look at Crack's results in other markets.

There are six markets we are going to look at and we will run the exact same study as Crack. We summarise the results in table 3.9.

Interestingly two of the six markets (CCMP and SPTSX) we looked had a significant difference between the NF and IP sets using the same criteria Crack used in his paper. A third market (SPX) was almost but not quite significant at 5%.

Given that the CCMP, which is a NASDAQ composite index, forms part of the CRSP [VE]RWETD indexes it is not too surprising that it should show some of the same features.

SPTSX is an index of stocks traded on the Toronto stock exchange, not part of the CRSP data, and the SPX index does have some commonality with CRSP, yet the Canadian SPTSX is a better fit than the U.S. SPX to Crack's hypothesis about the moon and returns.

It is interesting to look at the correlation of the return series for all those indexes, in table 3.10. SPX has the highest correlation with the CRSP value weighted index followed by (in descending order) INDU, CCMP then SPTSX all with correlations over 70%. The HSI and lastly the NKY both have correlations below 20%.

For the equally weighted CRSP index the correlations are also in two groups. The most correlated is CCMP (91%) followed by SPX, SPTSX and INDU (in that order) all over 70%. HSI, correlated 22%, and NKY at 19% form the other group.

In terms of agreeing with Crack's prescription for higher returns and lower volatility the p–values are in table 3.9.

The best three are all significant, or nearly so, at 5% for the mean return but only CCMP is significant for volatility. In terms of correlated markets there are four that are reasonably highly correlated ($> 70\%$) with the CRSP indexes.

So, even for well correlated markets the effect that Crack observed is not universal. It fails to appear at all in either of the Asian markets.

So looking at these other markets offers no evidence to refute the hypothesis that Crack was data snooping.

Index	p–value$_{\mu}$	p–value$_{SD}$
CCMP	0.0174	0.0492
SPTSX	0.0411	0.1001
SPX	0.0521	0.1366
INDU	0.1932	0.1407
HSI	0.6575	0.0802
NKY	0.7178	0.3437

Table 3.9: The p–values for the separation of the mean returns and standard deviations of the NF and IP sets at $\alpha = 0.3$ for the six "other markets"

	CCMP	HSI	INDU	NKY	SPTSX	SPX	VWRETD	EWRETD
CCMP	1.000							
HSI	0.180	1.000						
INDU	0.730	0.139	1.000					
NKY	0.164	0.198	0.103	1.000				
SPTSX	0.677	0.178	0.660	0.157	1.000			
SPX	0.779	0.136	0.961	0.103	0.681	1.000		
VWRETD	0.856	0.156	0.943	0.125	0.730	0.982	1.000	
EWRETD	0.912	0.224	0.717	0.188	0.722	0.755	0.836	1.000

Table 3.10: Correlation of the returns in the other markets used to compare with Crack's results. The date range is the same 1977–1996 period that Crack used.

CCMP	The NASDAQ composite index (thus a component of both VWRETD and EWRETD that Crack used). Capitalisation weighted index of three NASDAQ markets: Global Select, Global Market and Capital Select.
HSI	The Hang Seng Index. Free–float capitalisation weighted index of companies from the Stock Exchange of Hong Kong.
INDU	The Dow Jones Industrial Average. A price–weighted average of the 30 blue–chip stocks that are generally the leaders in their industry.
NKY	The Nikkei–225 Stock Average. A price weighted average of the 225 top rated Japanese companies listed in the first section of the Tokyo Stock Exchange.
SPTSX	A capitalisation weighted index designed to measure market activity of stocks listed on the Toronto Stock Exchange.
SPX	Standard and Poor's 500. A capitalisation weighted index of 500 stocks representing all major industries in the USA.
VWRETD	
EWRETD	

3.7 Conclusion

Given the tests that we used the results for the lunar hypothesis can be summarised thus:

Data Snooping Test	No Evidence of Data Snooping	Inconclusive	Evidence of Data Snooping
Statistical Significance		X	
Holdouts			X
Neighbourhood Analysis	X		
Other Markets			X

Only the neighbourhood analysis showed no evidence of data snooping. It appears that as the α enters the "good zone" the performance gradually improves and as α leaves the region the performance gradually degrades.

When testing for statistical significance it appears that Crack choose a test well suited to his purposes. Considering figure 3.1 (page 43) the t–test would have been almost as good, but the Wilcoxon–Mann shows no statistical significance. Only a very small proportion of the samples had a significant result from the point of view of T–test and Wilcoxon–Mann test but about a third were significant from the point of view of GMM. Clearly GMM was a better choice for Crack to use, and there is evidence that the p–values where due to data snooping. So the phenomenon that Crack reported (periods of better returns and lower volatility in the lunar month) appears to be real in the period Crack looked at.

The "Holdout" test was very clear. The 1977–1996, the period studied by Crack, show no sign of the properties Crack observes during 1977–1996.

The "Other Markets" test was also fairly clear. In the two Asian markets the indexes showed none of the lunar properties while two of the four North American indexes showed similar patterns in returns, and a third was almost significant at the 5% level (using GMM). Only one of the North American markets showed a significant difference in volatility between the IP and NF sets of returns. The four daily returns to the North American indexes are all fairly significantly correlated with the CRSP data that Crack used, but even then the results are equivocal. If all the correlated markets had shown the same properties Crack found in the CRSP data we would have been suspicious, but when correlated markets do not show the same properties we should reject the properties as data snooping.

In the final analysis it does look to us that Crack did find a way of partitioning daily returns, in one market, according to the phase of the moon that separated them into high return and low volatility days and lower return higher volatility days. But we do think that the statistical p–values are due to data snooping. Also the absence of the effect in other markets and in holdout data illustrates that the effect in itself is due to data snooping. Crack has described the data, not the data generating process. He claimed to be data snooping and we agree with him.

3.8 Explain why it (appeared to) Work

Given that it is clear that Crack was data snooping, the question remains what was the out–performance of the IP returns caused by? Is it just random chance that days with slightly better returns fell into the IP set? Or are there unusual days (outliers) in the IP or NF sets that skew the means?

One way to get a feel for this is to look at the days with the most extreme returns and see if they tend to occur in the IP or NF sets.

To this end we have done further analysis using the 1976–1997 data with α set to 0.3. But this time we partition the data according to daily returns.

Our objective is to find whether the success of Crack's approach was due to the distribution of some outliers, or whether it was due to a wider distribution of returns.

We order the data by absolute daily return. We then divide it into two sets using a cutoff value. One set (labeled "Below" in table 3.11) includes all the days where the absolute daily return is less than the cutoff, and the other (labeled "Above") where absolute daily returns are above the cutoff. We then compare these two sets in the same manner that Crack did. We run Crack's algorithm, with $\alpha = 0.3$, twice. Once on the data below and once on the data above the cutoff.

We repeat with nine cutoffs evenly spaced, such that each new cutoff moves about 370 data points to the below set from the above set.

Our null hypothesis is that the effect noticed by Crack is distributed throughout the data and it is not caused by outliers. We will reject the null hypothesis if we can find a cutoff where the effect is present in data below the cutoff and not in the data above.

Looking at the results in table 3.11 we can see that our null hypothesis cannot be rejected. It looks like extreme value “outliers” are not responsible for the effect. In fact, it seems that the data with smaller absolute returns are more likely to be responsible for the effect. The p–values for differences in mean and variance are almost all small for the below sets and almost all not small for the above sets. It could be that the extreem values produce higher standard errors in the estimates of mean and standard deviation and this gives higher (i.e., less significant) p–values there.

EWRETD Data

Cutoff	Below	#NF	#IP	μNF	μIP	p–value$_\mu$	σNF	σIP	p–value$_\sigma$
0.008618	Below	372	134	-0.692	0.504	0.007	0.271	0.228	0.001
	Above	3331	1219	0.309	0.314	0.442	0.058	0.059	0.825
0.006286	Below	724	288	0.027	0.710	0.003	0.213	0.173	0.002
	Above	2979	1065	0.253	0.231	0.793	0.048	0.047	0.230
0.004934	Below	1106	412	0.197	0.674	0.004	0.179	0.151	0.012
	Above	2597	941	0.214	0.183	0.899	0.040	0.040	0.695
0.003988	Below	1479	545	0.295	0.633	0.004	0.158	0.135	0.011
	Above	2224	808	0.151	0.130	0.835	0.034	0.034	0.422
0.003227	Below	1853	677	0.307	0.555	0.010	0.143	0.124	0.018
	Above	1850	676	0.110	0.111	0.485	0.028	0.028	0.196
0.002501	Below	2234	802	0.299	0.510	0.010	0.132	0.115	0.026
	Above	1469	551	0.071	0.075	0.394	0.022	0.022	0.324
0.001830	Below	2607	935	0.278	0.463	0.011	0.123	0.108	0.034
	Above	1096	418	0.043	0.042	0.508	0.016	0.017	0.674
0.001220	Below	2969	1079	0.256	0.414	0.013	0.115	0.101	0.031
	Above	734	274	0.016	0.015	0.575	0.011	0.011	0.776
0.000614	Below	3332	1222	0.232	0.369	0.014	0.109	0.095	0.030
	Above	371	131	0.003	-0.009	0.988	0.006	0.005	0.050

VWRETD Data

Cutoff	Below	#NF	#IP	μNF	μIP	p–value$_\mu$	σNF	σIP	p–value$_\sigma$
0.012239	Below	363	143	-0.153	0.865	0.007	0.348	0.293	0.026
	Above	3340	1210	0.129	0.151	0.330	0.084	0.082	0.186
0.008781	Below	740	272	0.024	0.648	0.024	0.270	0.240	0.069
	Above	2963	1081	0.121	0.120	0.510	0.067	0.066	0.197
0.006841	Below	1120	398	0.104	0.562	0.010	0.231	0.210	0.114
	Above	2583	955	0.101	0.086	0.667	0.054	0.055	0.696
0.005302	Below	1473	551	0.138	0.456	0.020	0.207	0.186	0.087
	Above	2230	802	0.078	0.069	0.626	0.044	0.043	0.030
0.004111	Below	1860	670	0.148	0.420	0.016	0.187	0.171	0.130
	Above	1843	683	0.055	0.036	0.778	0.035	0.035	0.315
0.003075	Below	2226	810	0.149	0.361	0.023	0.172	0.158	0.116
	Above	1477	543	0.031	0.025	0.601	0.028	0.027	0.028
0.002224	Below	2611	931	0.137	0.321	0.024	0.160	0.148	0.155
	Above	1092	422	0.018	0.017	0.515	0.020	0.020	0.242
0.001482	Below	2979	1069	0.125	0.287	0.027	0.150	0.139	0.150
	Above	724	284	0.004	-0.002	0.635	0.013	0.013	0.471
0.000720	Below	3341	1213	0.113	0.253	0.032	0.142	0.130	0.135
	Above	362	140	0.002	-0.006	0.820	0.006	0.007	0.841

Table 3.11: What causes the effect noticed by Crack? Here we sort the returns by absolute value to see if it is the extreme values causing the effect. The top table is EWRETD data and the lower table is VWRETD. We then divide it into equally sized subsets defined by nine cutoffs. Then we apply Crack's algorithm to the data below and the data above the cutoff, separately. It looks to us as if the effect noticed by Crack is apparent in the data with the extreme values removed, and therefore not in the extreme values themselves.

Chapter 4

An Interesting Case: Market Timing

"A Quantitative Approach to Tactical Asset Allocation" (Faber, 2009) published in 2007 and updated in 2009 proposed a simple method of market timing. At the end of every month a 10–month moving average of the level of the S&P 500 is compared to the level of the S&P500 index and if the index (including dividends) is above the moving average buy the index, if it is below then hold cash[1].

4.1 Faber's Claim

Faber argues that the main benefit of his timing algorithm is in improving Sharpe ratios by reducing volatility. Although, he says, if followed 1900 – 2008 his algorithm produces an increase in annualised returns (table 4.1).

	S&P 500	Timing
Annualised Return	9.21%	10.45%
Volatility	17.87%	12.01%
Sharpe (4%)	0.29	0.54
Maximum Drawdown	−83.66%	−50.31%
Best Year	52.88%	52.40%
Worst Year	−43.86%	−26.87%

Table 4.1: S&P 500 Total Returns vs. Timing Total Returns (1900–2008). Taken from Faber (2009). The "4%" refers to an assumed risk free rate for calculating the Sharpe Ratios.

[1] Faber uses the 90–day T–Bill, as do we. Given the one month reevaluations, using 30–day instruments would be more realistic, but will make little difference.

4.2 Reproducing Faber's Results

Faber used the S&P 500 index including gross dividends and the 90–day US T–bill rate for the cash rate.

4.2.1 Data

The data used to reproduce Faber's results were obtained from the Otago University Bloomberg terminal. Using the monthly S&P500 index and the field "TOT_RETURN_INDEX_GROSS_DVDS".

The data from the Otago University Bloomberg terminal start in 1926. Faber's data started in 1900, so we could not reproduce, or test, all of Faber's results. But we have enough of the data, covering multiple business cycles, to test Faber's claims.

For calculating volatility the simple returns were calculated using the "returns" function from the R library fBasics (Wuertz and Rmetrics, 2010). We calculated a variance deviation (σ_M^2) using monthly data then an annualised volatility $\sigma_A = \sqrt{12 \times \sigma_M^2}$.

The 90–Day bill rate that Faber used as a risk free rate was downloaded from the Federal Reserve website (Reserve, 2010).

4.2.2 Reproduced Results

In his Exhibit 2 Faber reports the total return, volatility, max drawdown, worst year and Sharpe ratio[2] for the S&P 500 from 1973–2008. Reproducing the results using data from the Otago University Bloomberg terminal gave similar but not quite identical results. See table 4.2.

Faber does not carefully define drawdown. We have struggled to find a precise definition of what a drawdown is in the context of finance and portfolio values. In Magdon-Ismail et al. (2004) it is described as "...the largest drop from a peak to a trough". It seems the intention is to describe "the maximum possible loss a buy and hold investor can experience by choosing the worst times to enter and exit the market" (our description). We would define "loss" as a negative return. This is the definition that we use.

[2] Faber assumed a risk free rate of 6%, so it was not really a Sharpe ratio

	Faber	Attempted Reproduction
Return	9.26%	9.29%
Volatility	15.54%	15.55%
Max Drawdown	−44.73%	−40.67%
Best Year	37.58%	34.06% (1995)
Worst Year	−36.77%	−32.97% (2008)
Best 12 Months	Not reported	61.18% (Start 1982–06–30)
Worst 12 Months		−38.94% (Start 1973–09–28)
Sharpe(6%)	0.21	0.21

Table 4.2: Faber's Exhibit 2 and our reproduced figures.

The differences in best and worst year are odd. It would have been helpful if Faber had included which years he found. In reproducing Faber's figures it was found that the best and worst year were 1995 and 2008 respectively. Whilst the best and worst 12 month periods started in 30 June 1982 and 30 September 1973 respectively.

When it comes to reproducing the "quantitative" results, Faber's timing algorithm, we can only attempt part. Faber starts his main simulation in 1900 whilst the data available from the Otago University Bloomberg terminal start in 1926.

Figure 4.1 is taken from Faber's Exhibit 7 (Faber (2009, pp 12)).

In table 4.3 page 56 is Faber's data for the worst ten years (in terms of absolute return) for the period 1900–2008. We only have data for the risk free rate from 1934 so half of the years identified by Faber there is no reproduction.

4.2.3 Summary

Our reproduction of Faber's results is not perfect. We do not get exactly the numbers that he did. In some cases Faber was a little unclear, for example not precisely defining drawdown,[4] and being unclear about exactly which time periods some of his results covered. We did, however, manage to reproduce the effects that Faber described. The timing strategy does produce slightly better returns and reduced volatility. So we have enough material to work with in the following sections.

[4] But then who does precisely define "drawdown"? There does not seem to be any precise and widely accepted definition of this often used phrase

	Faber's Results		Reproduction	
	S&P 500	Timing	S&P 500	Timing
1931	−43.86%	1.41%	No Data	
2008	−36.77%	1.33%	−32.97%	1.37%
1937	−35.26%	−7.65%	−37.11%	−12.45%
1907	−29.61%	0.09%	No data	
1974	−26.47%	8.16%	−25.94%	7.48%
1917	−25.26%	2.51%	No Data	
1930	−25.26%	2.51%	No Data	
2002	−22.10%	−4.62%	−20.94%	−4.78%
1920	−16.69%	−4.80%	No Data	
1973	−14.69%	−15.36%	−13.4%	−14.2%

Table 4.3: Faber's Exhibit 8 "S&P 500 Ten Worst Years vs. Timing, 1900-2008" and reproduced from available data . Interestingly it seems that the S&P 500 index was first published in 1957 S&P 500. From 1928–1957 a daily index of 90 stocks was published[3] Wilson and Jones (2002). In 1957 it was expanded to 500 stocks and became the S&P 500 Wilson and Jones (2002).

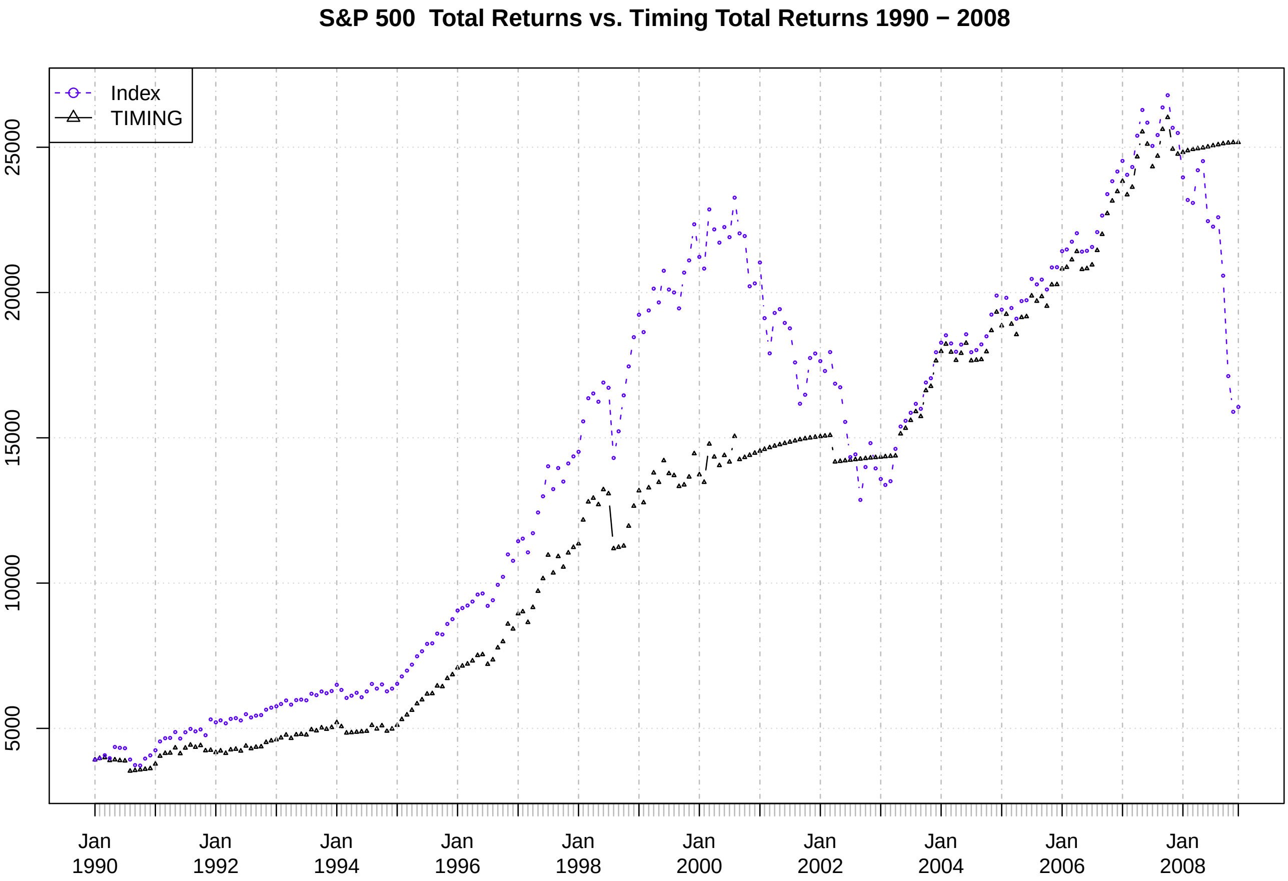


Figure 4.1: S&P 500 Total Returns vs. Timing Total Returns (1990–2008)

4.3 Statistical Significance

4.3.1 Decision Criteria

We will use a 5% level of significance for all our tests in this section. We will use the Wilcoxon–Mann non–parametric test and a standard two–sample t–test.

We test a timing portfolio with level P_T that is generated by investing one dollar at the start of the period (1931) and reinvesting the returns from Faber's timing algorithm each month. The index portfolio with level P_I is similar but uses the monthly returns of the S&P 500 for the period (buy and hold).

We will use paired and unpaired tests over volatility, return and Sharpe ratios. Paired tests compare individual observations and examine if one value (Faber's timing algorithm or buy and hold of the index) is mostly higher or lower than the other. Unpaired tests compare the overall means[5].

We used a t–test for the paired and unpaired tests and a Wilcoxon signed rank test for non–parametric testing of the paired data and Wilcoxon rank sum test (equivalent to the Mann-Whitney test) for the unpaired data. The tests we used were as implemented in **R** (Ihaka and Gentleman, 2010).

We calculate volatility on an annual basis by calculating the standard deviation of the monthly returns for both Faber's timing portfolio and the buy and hold portfolio.

When we also test the Sharpe ratios we calculate them for each year using the 12 monthly returns to calculate a volatility for that year and using 90–day T–Bills for the risk free rate[6].

We test the following statistics, as reported in table 4.4:

Level We test if the level of the timing portfolio fund is less than or equal to the level of the index ($P_T \leq P_I$). It only makes sense to test the level in the paired case as it is non–stationary and the long term mean of the price level of the S&P 500 does not

[5] It is possible to have a time series, A, which is mostly larger than another time series, B (using the same observation times) yet the mean value of B is greater than the mean value of A. If at most times t $A_t > B_t$ but when $B_t > A_t$ B is much greater than A

[6] Faber reports Sharpe ratios calculated using a constant risk free rate. (Faber (2009, pp 21 Exhibit 15, pp 22 Exhibit 16, pp 28 Exhibit 21)) We think that using the same risk free rate the algorithm uses, when it is out of the market, for calculating Sharpe ratios is more appropriate than using a constant rate.

have a clear interpretation. The null hypothesis is that the level of Faber's timing portfolio not greater than the level of the buy and hold portfolio on each month.

Volatility We do a paired comparison of volatility of the monthly returns for each year. The null hypothesis is that the volatility in the returns of Faber's timing portfolio is not less than for the buy and hold portfolio. Do an unpaired test of the null hypothesis that the the mean of the monthly volatilities, for the complete series, is not less for Faber's timing portfolio than for the buy and hold portfolio (that is we test the null: $H_0 : \ SD_{S\&P} - SD_{Faber} \geq 0$).

Returns A paired test of the null hypothesis that the returns of the buy and hold portfolio are greater than those of Faber's timing portfolio, and an unpaired test of the null hypothesis that he mean return of the buy and hold portfolio is not less than that of Faber's timing portfolio.

Sharpe ratios A paired test of the null hypothesis that the Sharpe ratios (calculated annually) of the buy and hold portfolio are greater than those of Faber's timing portfolio, and an unpaired test of the null hypothesis that he mean of the Sharpe ratios for the buy and hold portfolio is greater than that of Faber's timing portfolio.

The tests will be performed on data for the whole period (1931–2009) and a moving 10–year window for the unpaired tests, although we report discrete decades for brevity.

4.3.2 Results

Tables 4.4 and 4.5 present the *p*–values from the t–test and the Wilcoxon–Mann test.

For the complete period neither test rejected any of the null hypotheses except for the volatility ($V_I \leq V_T$) which was clearly rejected.

For each decade we mostly do not reject the null that the level of Faber's timing index is less or equal to the index.

We do mostly reject the null that the timing algorithm has worse volatility. The t–test failed to reject the null for one observation. But the tests rejected the null for the mean volatility (column "Vol UP") for five of nine decades in the Wilcoxon–Mann test and four of nine for the t–test.

The tests consistently fail to reject the null for the tests on returns and for the Sharpe ratios (except for one Mann–Whitney result, but with so many results a solitary type I error is not unexpected).

Date Range	Level	Vol Paired	Vol UP	Ret Paired	Ret UP	Sharpe Paired	Sharpe UP
1932–2009	1.000	**0.000**	**0.000**	0.838	0.727	0.162	0.251
1932–1941	1.000	**0.008**	**0.020**	0.867	0.797	0.508	0.503
1942–1951	1.000	**0.006**	0.078	0.738	0.604	0.601	0.518
1952–1961	1.000	**0.033**	0.245	0.908	0.647	0.928	0.633
1962–1971	**0.000**	**0.009**	**0.009**	0.490	0.494	0.571	0.526
1972–1981	**0.000**	0.054	**0.046**	0.364	0.428	0.273	0.281
1982–1991	1.000	**0.006**	0.129	0.940	0.741	0.982	0.812
1992–2001	1.000	0.110	0.172	0.473	0.488	0.214	0.283
1999–2008	**0.000**	**0.025**	**0.010**	**0.028**	0.085	0.074	0.056

Table 4.4: p–values for t–tests on S&P 500, Faber's timing portfolio and buy and hold of the index (bold indicates significance less than five percent). The first column, "Level", is the p–value of the test of the level of Faber's timing portfolio (P_T) Vs. the index (P_I). $H_0 : P_T \leq P_I$. The next two columns are the p–values of the tests of the timing portfolio volatility (V_T) Vs. the index volatility (V_I). The "Vol Paired" column is a paired test (t–test for two dependant samples) with $H_0 : V_T \geq V_I$ and the "Vol UP" $H_0 : \mu_{V_T} \geq \mu_{V_I}$ is an unpaired test (t–test for independent samples). The fifth and sixth columns are similar tests of returns. They compare the returns of the timing portfolio R_T with the returns of the index. Column "Ret Paired" tests $H_0 : R_T \leq R_I$ and column "Ret UP" tests $H_0 : \mu_{R_T} \leq \mu_{R_I}$. The last two columns test the Sharpe ratios, S_T and S_I for the timing algorithm Sharpe ratio and the index Sharpe ratio respectively. Column "Sharpe Paired" tests $H_0 : S_T \leq S_I$ and the column "Sharpe UP" tests $H_0 : \mu(S_T) \leq \mu(S_I)$.

Date Range	Level	Vol Paired	Vol UP	Ret Paired	Ret UP	Sharpe Paired	Sharpe UP
1932–2009	1.000	**0.000**	**0.000**	0.369	0.883	0.921	0.388
1932–1941	1.000	**0.002**	**0.012**	0.495	0.744	0.590	0.698
1942–1951	1.000	**0.004**	0.083	0.179	0.697	0.545	0.466
1952–1961	1.000	**0.022**	0.248	0.689	0.757	0.918	0.635
1962–1971	**0.000**	**0.010**	**0.006**	0.457	0.606	0.850	0.568
1972–1981	**0.000**	**0.016**	0.072	0.144	0.261	0.936	0.568
1982–1991	1.000	**0.010**	0.062	0.813	0.770	0.960	0.717
1992–2001	1.000	**0.009**	0.182	0.674	0.682	0.639	0.398
1999–2008	**0.000**	**0.014**	**0.022**	0.097	0.340	**0.020**	0.068

Table 4.5: p–values for Wilcoxon–Mann tests on S&P 500, Faber's timing and the buy and hold portfolios (bold indicates significance less than five percent). The first column, "Level", is the p–value of the test of the level of Faber's timing portfolio (P_T) Vs. the index (P_I). $H_0 : P_T \leq P_I$. The second two columns are the p–values of the tests of the timing portfolio volatility (V_T) Vs. the index volatility (V_I). The "Vol Paired" column is a paired test (Wilcoxon matched pairs signed ranks test) with $H_0 : V_T \geq V_I$ and the "Vol UP" $H_0 : \mu_{V_T} \geq \mu_{V_I}$ is unpaired test (Mann–Whitney U). The fifth and sixth columns are similar tests of returns. They compare the returns of the timing portfolio R_T with the returns of the index. Column "Ret Paired" tests $H_0 : R_T \leq R_I$ and column "Ret UP" tests $H_0 : \mu_{R_T} \leq \mu_{R_I}$. The last two columns test the Sharpe ratios, S_T and S_I for the timing algorithm Sharpe ratio and the index Sharpe ratio respectively. Column "Sharpe Paired" tests $H_0 : S_T \leq S_I$ and the column "Sharpe UP" tests $H_0 : \mu(S_T) \leq \mu(S_I)$.

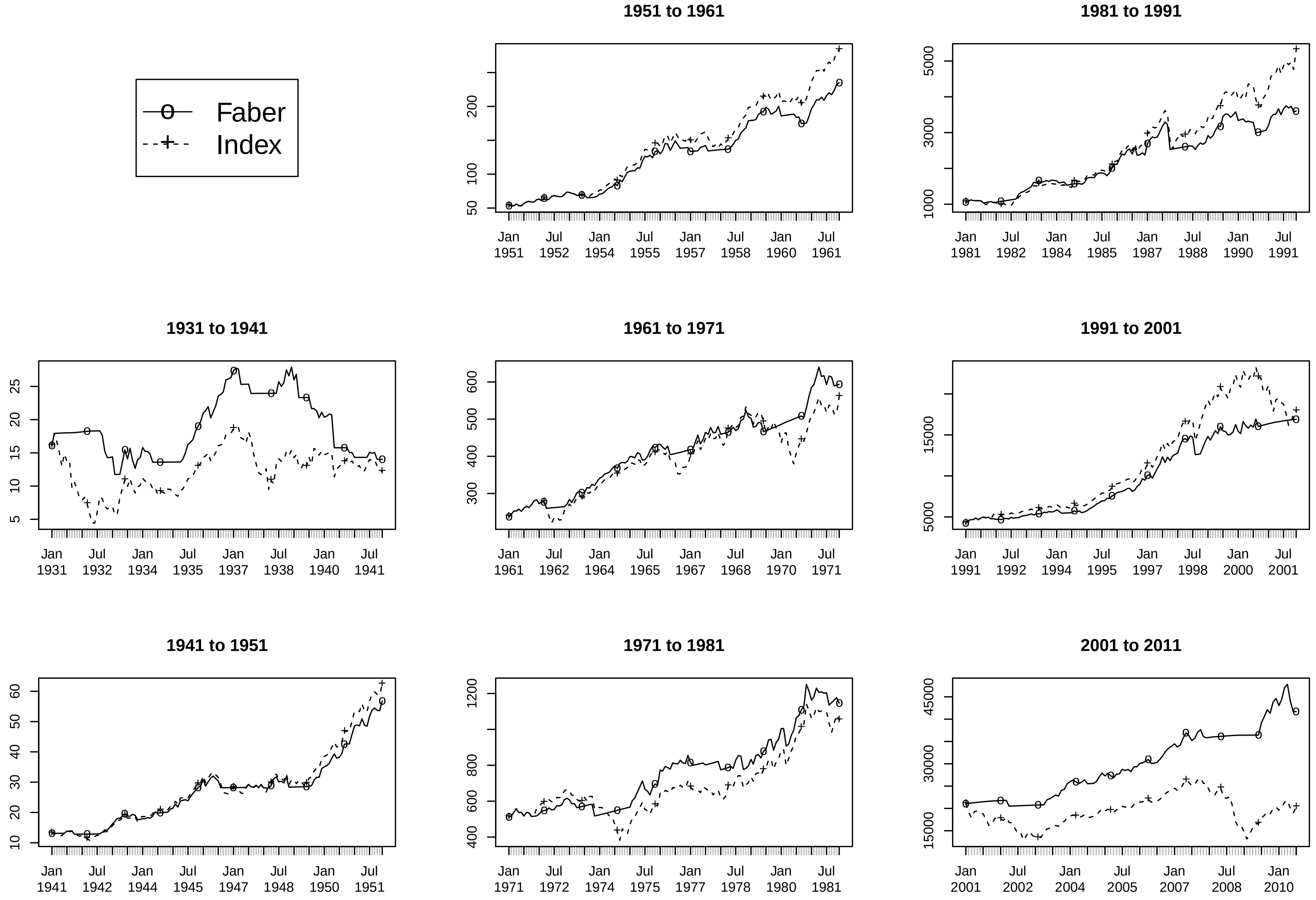


Figure 4.2: Faber vs. S&P500 in every decade since 1931. Each 10 year period begins anew with Faber's portfolio equal to the index, then we simulate trading the portfolio using Faber's rule over the decade.

4.3.3 Interpretation

The tests for the level are a bit pointless. It is clear which periods the timing portfolio performs better than the index by inspection. (See figure 4.2).

For volatility it is better news for Faber. The null hypothesis that for each year the volatility of the timing algorithm is worse (greater than) the volatility of the index $V_T \geq V_I$, ("Vol Paired" column of tables 4.4 and 4.5), year by year, is rejected in all Mann–Whitney cases and in all but two cases by the t–test. It is rejected for the complete data set (1932–2009) also.

The null hypothesis that the mean volatility of the index over a decade is better (lower) than the timing algorithm's mean volatility over each decade, $\mu V_T \geq \mu V_I$, ("Vol UP" column of tables 4.4 and 4.5) is rejected except for the decades 1952–1961 (both tests), 1972–1981 (Mann–Whitney only), 1982–1991 (t–test only) and 1992–2001 (both tests). It is rejected for the complete data set (1932–2009) also.

This is a good story for Faber.

For returns it is a constant story. For both tests for all periods there is no evidence to reject the null hypothesis that the timing algorithm returns are lower than the index returns.

For the Sharpe ratios there is almost no evidence that the Sharpe ratios of Faber's timing algorithm are better than those of the index. This is especially bad news for Faber, as his claim is that there is a large reduction in volatility yet similar returns.

4.4 Neighbourhood Analysis

The data we have are monthly. So the moving average was a 10–month moving average. We can look at the 9–month and 11–month moving averages to see if the the results are robust in near neighbourhoods.

4.4.1 Results

The results of the neighbourhood analysis are in table 4.6. The results for the 9–month and 11–month moving averages are very similar to what Faber got for his 10–month moving average.

The maximum drawdown for all three moving average horizons is the same for the same period 1987-08-31 to 1987-10-31.

The 10–month period Faber choose has the lowest return and does not have the lowest volatility or best Sharpe ratio which indicates that Faber was not knowingly data–snooping around the period (but our data is not quite the same as Faber's).

				Drawdowns			
Months		Return	Volatility	Start	End	Amount	Sharpe
9	Faber	0.1097	0.1183	1989-12-31	1990-07-31	0.0819	1.193
10	Faber	0.1067	0.1200	1987-08-31	1988-07-31	0.2051	1.179
11	Faber	0.1152	0.1214	1987-08-31	1988-07-31	0.2051	1.233
N/A	S&P 500	0.0929	0.1566	2000-08-31	2002-08-31	0.3798	0.682

Table 4.6: Comparing Faber's timing algorithm for the S&P 500 1973–2008 using 9–month, 10–month (Faber's original choice) and 11–month moving averages.

4.5 Other Markets

We will look at some other markets and see if Faber's results for 1973–2008 hold there. We have Bloomberg data, as described in section 3.2.4. Additionally we have to locate a risk free rate for the markets we investigate.

We will look at markets from the USA (CCMP; Nasdaq Composite), Canada (SPTSX; Toronto Stock Exchange) and Japan (NKY; Nikkei 225) as these are the markets for which we have both price data and risk free rates.

We will reproduce the results, look at statistical tests,[7] and do Neighbourhood Analysis.

[7] For more details on these statistical tests using sub–periods see appendix C.

Test	Market	Date Range	Level	Vol Paired	Vol UP	Ret Paired	Ret UP	Sharpe Paired	Sharpe UP
t–test	CCMP	1972–2009	0.000	0.000	0.001	0.326	0.399	0.218	0.264
	SPTSX	1973–2008	0.000	0.000	0.000	0.074	0.205	0.215	0.273
	NKY	1973–2008	0.000	0.000	0.000	0.162	0.253	0.137	0.145
Wilcoxon–Mann	CCMP	1972–2009	0.000	0.000	0.001	0.309	0.602	0.913	0.541
	SPTSX	1973–2008	0.000	0.000	0.000	0.649	0.549	0.798	0.505
	NKY	1973–2008	0.000	0.000	0.000	0.264	0.387	0.829	0.304

Table 4.7: t–test and Wilcoxon–Mann test on Faber's algorithm in other markets. The first column, "Level", is the p–value of the test of the level of Faber's timing portfolio (P_T) vs. the index (P_I). $H_0 : P_T \leq P_I$. The second two columns are the p–values of the tests of the timing portfolio volatility (V_T) vs. the index volatility (V_I). The "Vol Paired" column is a paired test with $H_0 : V_T \geq V_I$ and the "Vol UP" $H_0 : \mu_{V_T} \geq \mu_{V_I}$. The fifth and sixth columns are similar tests of returns. They compare the returns of the timing portfolio R_T with the returns of the index. Column "Ret Paired" tests $H_0 : R_T \leq R_I$ and column "Ret UP" tests $H_0 : \mu_{R_T} \leq \mu_{R_I}$. The last two columns test the Sharpe ratios, S_T and S_I for the timing algorithm Sharpe ratio and the index Sharpe ratio respectively. Column "Sharpe Paired" tests $H_0 : S_T \leq S_I$ and the column "Sharpe UP" tests $H_0 : \mu(S_T) \leq \mu(S_I)$.

4.5.1 CCMP: A convenient choice

The first market index we will examine is CCMP. It is the NASDAQ Composite Index. The description of CCMP from Bloomberg says it is "...a broad–based capitalization–weighted index of stocks in all three NASDAQ tiers". The level of the index is displayed in figure 4.3. It is a convenient choice as being from the USA we can use the same risk free rate as is used for analysing the S&P 500.

In table 4.7 the t–test and the Wilcoxon–Mann test results are presented. For the statistical tests we have fewer years than we had for the S&P 500 data. The results for the level reject the null that the level of Faber's algorithm is less than or equal to the level of the index. That null that the volatility of Faber's algorithm is greater than or equal to the volatility of the index is rejected also. The null that the returns and Sharpe ratio of Faber's algorithm are less than or equal to those of the index is also rejected.

In table 4.8 the general results for CCMP using the 10–month moving averages and its neighbours are displayed along with the data from reproducing Faber's results on this market. The results are better than for the S&P 500 index that Faber used. There is a marked improvement in Sharpe ratios, volatilities and even returns. Also the 11–month moving average is better indicating that Faber has not likely used data snooping to decide on the 10–month horizon of the moving average.

These results (for the CCMP index) are similar to, and an improvement on, the results Faber obtained on the S&P500 index.

4.5.2 SPTSX

The SPTSX is a capitalisation weighted index of the Toronto Stock Exchange. We obtained the data from the Otago University Bloomberg terminal. The level of the index and of the Faber timing algorithm are displayed in figure 4.4 page 68. In table 4.8 page 66 there are the results for applying Faber's algorithm using the 10–month moving average and the 9–month and 11–month month moving averages. As above for CCMP all results are better for the timing algorithm than for the index alone. The results of the statistical tests are very similar to those for CCMP in table 4.7.

4.5.3 NKY

The Nikkei 225 (NKY) index gave similar results for the previous two indexes in reproduction, for Neighbourhood Analysis and statistical tests.

Market	Method	MA	Annualised		Drawdown			Sharpe
		Months	Return	Volatility	Start	End	Amount	
CCMP	Faber	9	0.08498	0.1676	2000-02-29	2002-11-30	0.4706	0.6014
	Faber	10	0.1002	0.1633	2000-02-29	2002-01-31	0.3663	0.7403
	Faber	11	0.1068	0.1647	2000-02-29	2002-01-31	0.2959	1.0850
	Index		0.0722	0.2271	2000-02-29	2002-08-31	0.7200	0.5166
NKY	Faber	9	0.0569	0.1133	1989-12-31	1998-07-31	0.2341	0.8617
	Faber	10	0.0523	0.1169	1989-12-31	1998-07-31	0.3439	0.9583
	Faber	11	0.0501	0.1174	1989-12-31	1998-07-31	0.3545	0.9365
	Index		0.0151	0.1923	1989-12-31	2008-10-31	0.7796	0.3919
SPTSX	Faber	9	0.0940	0.1214	1994-01-31	1994-10-31	0.1269	0.3776
	Faber	10	0.0936	0.1207	1994-01-31	1994-09-30	0.1061	0.3771
	Faber	11	0.0947	0.1246	2000-08-31	2002-05-31	0.1829	0.4193
	Index		0.0564	0.1690	2000-08-31	2002-08-31	0.4122	0.1619
SPX	Faber	9	0.1097	0.1183	1989-12-31	1990-07-31	0.0819	1.193
	Faber	10	0.1067	0.1200	1987-08-31	1988-07-31	0.2051	1.179
	Faber	11	0.1152	0.1214	1987-08-31	1988-07-31	0.2051	1.233
	Index		0.0929	0.1566	2000-08-31	2002-08-31	0.3798	0.682

Table 4.8: Neighbourhood analysis on Other Markets

4.5.4 Summary of Other Markets

The results seem consistently better in other markets than they do in the case of the S&P 500 that Faber originally used.

4.6 Was Faber Data Snooping?

In Faber's original example (using the S&P 500 index) the results are equivocal. It is not clear that Faber's timing algorithm can reliably produce a better Sharpe ratio than that available by buying and holding the index. It is clear that a penalty is paid in returns for the better volatility.

In the other three markets that we examined it is clear that there were statistically significant lower volatilities, but the improvement in Sharpe ratios was not significant (see table 4.7).

Judging from the fact that Faber would have done better to use a different index, and a different moving average we can say with some certainty that he was not data snooping.

As an aside, we have not been able to show that Faber's algorithm does not work. Some doubt remains but it does seem effective, at least some of the time in some markets, and when it is not effective it is not terrible. This is a surprising result. Conventional wisdom has it that market timing is damaging to returns (Malkiel, 1999, pp. 163), yet Faber's algorithm is very simple. Some may say that Faber's algorithm is not truly market timing but is trend following (see for instance Holcomb (2011)) as it makes no attempt to predict market movements but reacts to them. In our opinion this is splitting hairs. Faber's algorithm attempts to get out of bear markets before they fully develop (Faber (2009, pp 11)) and to be in the market at other times. To us that is market timing.

Data Snooping Test	No Evidence of Data Snooping	Inconclusive	Evidence of Data Snooping
Statistical Significance		X	
Holdouts		N/A	
Neighbourhood Analysis	X		
Other Markets	X		

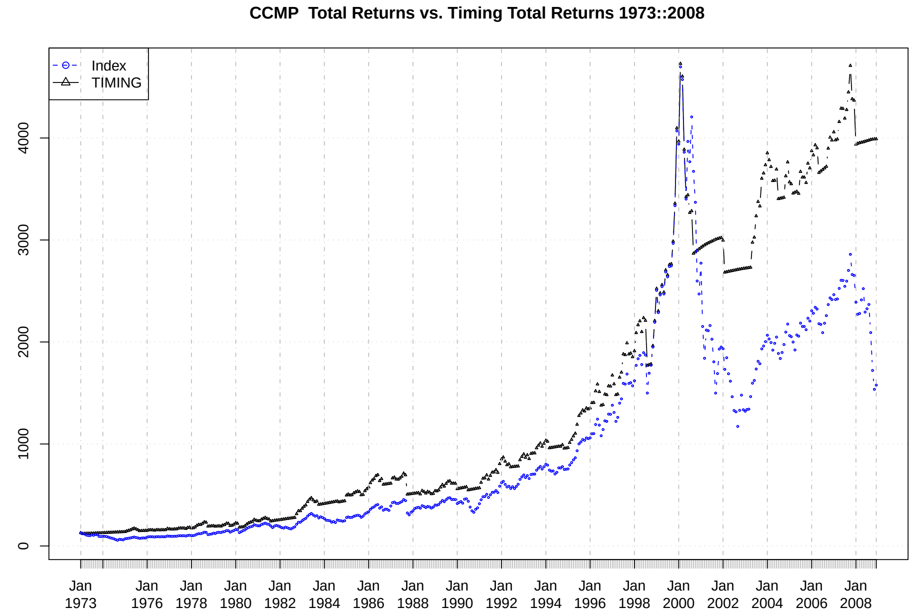


Figure 4.3: The CCMP Index and Faber's timing algorithm for 1973–2008

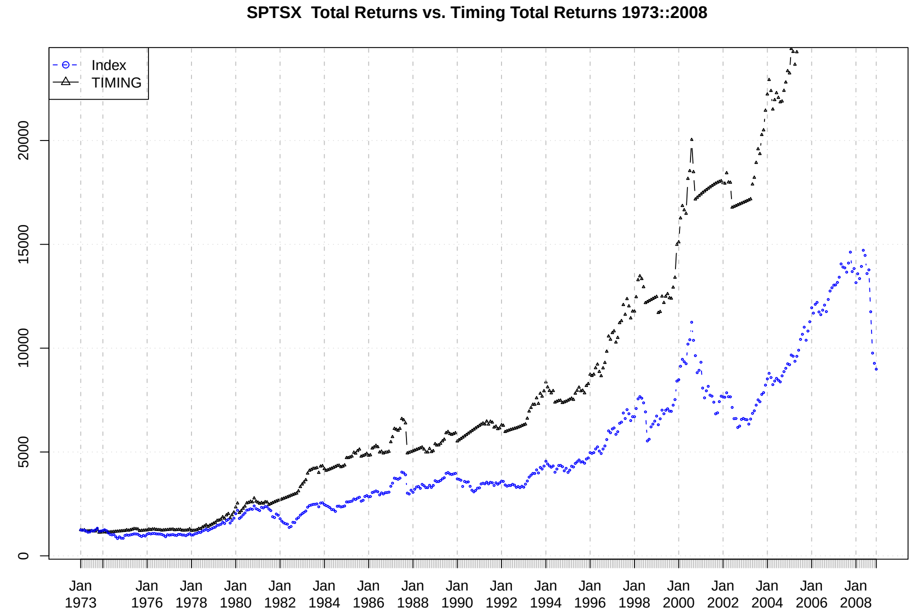


Figure 4.4: The SPTSX Index and Faber's timing algorithm for 1973–2008

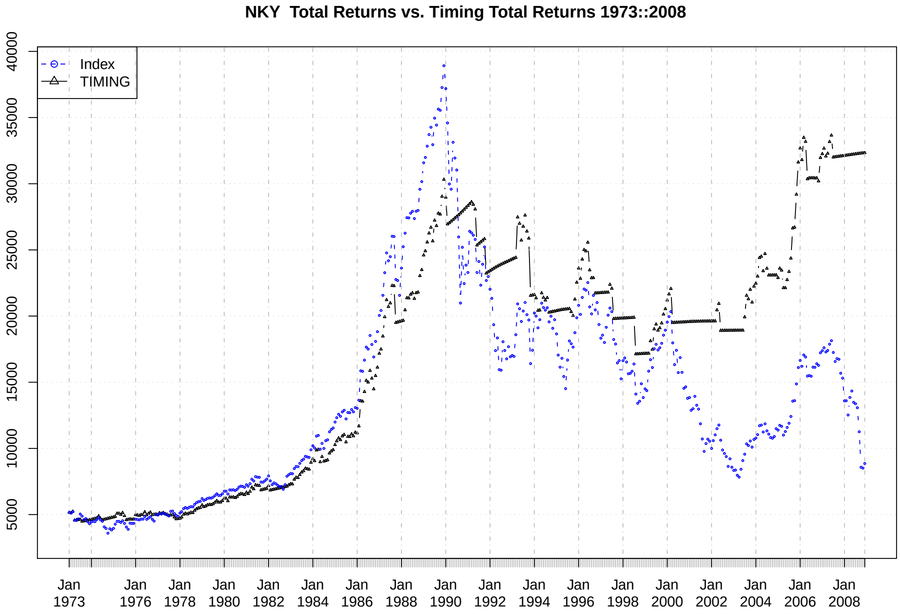


Figure 4.5: The NKY Index and Faber's timing algorithm for 1973–2008

Chapter 5

A Comprehensive Case: Searching for Rules

In this section we look closely at the paper "Data-Snooping, Technical Trading Rule Performance, and the Bootstrap" published in The Journal of Finance in October 1999 and authored by Ryan Sullivan, Allan Timmermann, and Halbert White (Sullivan et al., 1999). The authors examined a large number of technical analysis rules on the Dow Jones Index from 1897 to 1996. The authors base their work on an earlier paper from 1992 (Brock et al., 1992) which did a similar study with a smaller set of rules.

Sullivan et al. (1999) claim to have found that several of the rules were "effective", in that the mean annual return of the rules was close to, or better than, 20%. They also claim to have proved that the apparent success was not due to data snooping using a bootstrapping algorithm. The bootstrap algorithm is described in White (2000), which is a companion paper.

We will attempt to apply our techniques to the same rules as Sullivan et al. (1999) to see if we agree with their results. We have chosen to study this paper as it directly addresses data snooping, and it is extensively cited[1].

[1] Google Scholar reports more than 500 citations, SSRN more than 100.

5.1 The Rules, Data and Technique in Sullivan et al. (1999)

Sullivan et al. (1999) looked at five classes of rules:

- Filter Rules
- Moving Averages
- Channel Breakouts
- On Balance Volume Averages
- Support and Resistance

The rules are trialed on the daily closing prices of the Dow Jones Industrial Average Index from 1897 to 1996. We look more closely at the rules in section 5.4

Sullivan et al. (1999) divided the data into sub-periods, mostly the same as in Brock et al. (1992):

1897 – 1914 The first sub–sample and it ends with the closing of the stock exchange for the First World War. (In Brock et al. (1992) this sub–period ended in July 1914 and the second started in January 1915)

1915–1938 This sub–sample is chosen to include both the "rise of the twenties" and the depression of the 1930s.

1939–June 1962 This sub–sample is chosen to include the Second World War and it ends when the Centre for Research in Securities Prices (CRSP) begins its daily price series

July 1962 – 1986 The most recent period in Brock et al. (1992) covers the data from the start of the CRSP daily data to the end of their data set

1987–1996 Sullivan et al. (1999) added this sub–period to cover the gap between their study and that of Brock et al. (1992)

1897–1986 This period is the complete period covered in Brock et al. (1992)

1897–1996 This is the complete period of the Sullivan et al. (1999) study.

In each period each rule was trialed separately using many different parametrisations, generating 7,500 rules[2].

The rules were evaluated using two criteria: Mean return and Sharpe ratio. For each sub–period there are (up to) two best rules, one with the highest mean return and one with the highest Sharpe ratio.

The process of rule discovery (generating 7,500 rules and testing them all against the same 110 years of data, 1897 – 1996, being about 30,000 observations) is data snooping. In Sullivan et al. (1999) it is asserted that the bootstrapping algorithm can adjust for data snooping and give an accurate picture of the true worth of the rules. The authors use the bootstrapping algorithm (White, 2000) to assign p–values to the rules, testing the null hypothesis of no predictive superiority over holding a risk free asset.

This is similar to what we are doing herein. We are interested in the question: Is a rules performance due to data snooping? In Sullivan et al. (1999) the claim is that that question is answered definitively.

5.1.1 Risk Free Rates

To calculate Sharpe ratios, and returns when out of the market, Sullivan et al. (1999) used three risk free rates for three periods.

1897–1925 Interest rate for 90-day stock exchange time loans[3]

1926–June 1954 One–month T–bill rates reported form CRSP

July 1954 to 1996 Daily Federal Funds rate

Since we only have data from 1928 (as we describe in section 5.6.1) we use the One–month t–bill rates (we get our data from the website of Aswath Damodaran[4]) and the daily Federal Funds rate from the Federal Reserve Bank of St Louis's website[5].

[2] Strictly speaking they did not generate 7,500 rules, but 7,500 parametrisations of the five rules.

[3] As reported in Banking and Monetary Statistics, 1914-1941 (1943).

[4] `http://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/histret.html`

[5] `http://research.stlouisfed.org/fred2/series/FEDFUNDS`

5.2 Our Approach

In section 5.6 we implement the rules that Sullivan et al. (1999) reported as being best using the two criteria: mean return and Sharpe ratio. We model a portfolio starting at time zero with one dollar and follow the signals generated by the rules. Thus we use our total wealth (initially one dollar) to be long the index when the signal is 1, short the index, to the value of our (current) total wealth, when the signal is -1 and in the risk free asset when the signal is 0.

We then apply our methods from section 2 to ascertain if we can detect data snooping.

5.3 Our Expectations and What We Found

We expected to find that the algorithms described as "best" by Sullivan et al. (1999) would nominally "make money" when applied to the Dow Jones index. That is to say that a simulated portfolio, short, neutral or long the index following the instructions of one of the rules selected by Sullivan et al. (1999), would be more profitable than buying and holding the index.

We are interested to see if such rules could violate the Efficient Market Hypothesis by returning feasible consistent risk adjusted returns, in excess of holding the index.

We were disappointed to find that the rules in Sullivan et al. (1999) were not described precisely. We did our best to reconstruct the rules from the description but in two cases (Channel Breakouts section 5.4.4 and Filter section 5.6.7) cases we found that the rules as described performed poorly on the data in the time period where Sullivan et al. (1999) reported they were the best rules.

Examining the "Reality Check" bootstrapping algorithm, as described in White (2000), we found that the assumptions required are violated using long-term financial data. White's "Reality Check" is based on the "Stationary Bootstrap" described in Politis and Romano (1994). It is a requirement for that algorithm that the underlying data be stationary. The daily return of the Dow Jones Industrial Averages, the index used in Sullivan et al. (1999) has not been stationary over the 80 years of our study. We discuss this and other problems with the Reality Check in section 5.12.1.

Lastly the best algorithms that Sullivan et al. (1999) find switch between long, neutral and short with great frequency. So we calculated a break even transaction cost for each one and found that they are not practical for most of the periods in question, and probably not practical for any. If a technical analysis algorithm cannot return excess trading profits it cannot be said to violate the Efficient Market Hypothesis. (We discuss this in section 2.8 above).

5.4 The Rules from Sullivan et al. (1999)

Here we present the rules as described in Sullivan et al. (1999) that were found by them to be "best", those from table 5.1 (page 79). We do not examine the support and resistance rules as they were not found to be the best, for data we have, by Sullivan et al. (1999).

Sullivan et al. (1999) do not explicitly describe the rules with equations, instead they give English language summaries and references split between the body of their article and an appendix. We do our best to reconstruct exactly the rules they used and claimed to be effective.

5.4.1 Moving Average Rules

Sullivan et al. (1999) found moving average rules to be the best for the 1915–1938 sub–period (using both criteria), the 1939–1986 (with Sharpe Ratio as the criterion) and for the 90–year and 100–year periods both starting in 1897 and finishing in 1986 or 1996 (using both criteria). The moving average rules found to be best used one moving average, compared to the level of the index. The rule gave a long signal when the price is above the moving average and a short signal when the price goes below the moving average. A "band" filter was tested where the price has to go above or below the moving average by a percentage.

Sullivan et al. (1999) tested rules with two moving averages but none of those made it into the category of "best".

The moving average rule appears in equation 5.1.

$$S_{MA_t} = \begin{cases} 1 & P_t > MA(n) \times (1+b) \\ -1 & P_t < MA(n) \times (1-b) \\ 0 & \begin{cases} MA(n) \times (1-b) \leq P_t < MA(n) & if\ long \\ MA(n) \times (1+b) \geq P_t > MA(n) & if\ short \end{cases} \end{cases} \tag{5.1}$$

Where S_{MA_t} is the signal (1 for long, -1 for short and 0 for neutral) and $MA(n)$ is an n–day moving average.

Sullivan et al. (1999) state that their best rules use $n \in \{2, 5\}$ and $b \in \{0, 0.001\}$. See table 5.2.

5.4.2 On Balance Volume Averages

Sullivan et al. (1999) found that an On Balance Volume Average (OBV) rule gave the best results when judged by mean return for the period 1939–1986[6]

The OBV works by keeping a running total of the volume traded. If the price goes up the volume for that day is added, if the price goes down it is subtracted. A moving average of the daily volume tally is calculated then compared with the tally to generate the long/short signal. The rule is in equation 5.2.

$$\begin{aligned} S_{OBV_t} &= \begin{cases} 1 & OBV_t > MA_{OBV}(t,n) \\ -1 & OBV_t < MA_{OBV}(t,n) \end{cases} \\ OBV_t &= \begin{cases} OBV_{t-1} + VOL_t & P_t > P_{t-1} \\ OBV_{t-1} - VOL_t & P_t < P_{t-1} \end{cases} \\ MA_{OBV}(t,n) &= \frac{\sum_{i=t-n}^{t-1} OBV_i}{n} \end{aligned} \tag{5.2}$$

where n is the number of days in a moving average. Sullivan et al. (1999) state that $n = 2$ gave the best results. See table 5.2.

[6] They counted this as two periods split at June/July 1962. The same rules were found to be best for both those sub–periods. Hence in our simulations we combine the two periods into one period 1939–1986.
.

5.4.3 Filter Rules

Sullivan et al. (1999) found a filter rule worked best for the period 1987–1996 using mean return criterion.

The filter rule works by producing a long signal when the price moves $x\%$ above the low point while short, and a short signal when the price moves $x\%$ below the high point while long.

Sullivan et al. (1999) add an extra parameter, y, the liquidation parameter. When the price moves $y\%$ above a low (if short) of $y\%$ below a high (if long) go neutral. It follows that $y \leq x$.

The filter rule is in equation 5.3.

$$S_{F_t} = \begin{cases} +1 & P_t \geq L_t \times (1+x) \\ 0 & \begin{cases} P_t \leq H_t \times (1-y) & \text{while long} \\ P_t \geq L_t \times (1+y) & \text{while short} \end{cases} \\ -1 & P_t \leq H_t \times (1-x) \end{cases} \tag{5.3}$$

$$H_t = \text{Highest value while long index}$$

$$L_t = \text{Lowest value while short index}$$

where x is the "initiation" parameter. The change in security price ($x\times$ price) required to initiate a position, and y is the "liquidation" parameter. The change in security price ($y \times$ price) required to liquidate a position.

Sullivan et al. (1999) state that their best rule used $y = 0.10$ and $x = 0.12$. See table 5.2

5.4.4 Channel Breakouts

Sullivan et al. (1999) found that a channel breakout rule was best in the 1987–1996 period using the Sharpe ratio criterion.

A channel is defined (in Sullivan et al. (1999)) when the high over the previous n days is within x percent of the low in the same period. When the price breaks out of the channel a long or short signal is generated if the price goes above or below the channel, respectively.

We look at the Channel Breakout Rule, and some of the difficulties in implementing it, in section 5.6.8 below.

Given what we have read in Sullivan et al. (1999),[7] we define a channel breakout in equation 5.4

$$\begin{aligned}
&P_{m_{t,n}} = \frac{(max([P_{t-n},...,P_{t-1}])+min([P_{t-n},...,P_{t-1}]))}{2}\\
&P_{h_{t,n}} = P_{m_{t,n}} \times (1+\tfrac{x}{2})\\
&P_{l_{t,n}} = P_{m_{t,n}} \times (1-\tfrac{x}{2})\\
&S_{CB_{t,n}} = \begin{cases} 1 & P_t > P_{h_{t,n}} \\ -1 & P_t < P_{l_{t,n}} \end{cases}
\end{aligned} \tag{5.4}$$

$P_{m_{t,n}}$ is the mid point between the maximum and minimum prices over the last n days
$P_{h_{t,n}}$ is the top of the channel
$P_{l_{t,n}}$ is the bottom of the channel
$max([P_{t-n}, ..., P_{t-1}])$ is the maximum price for the days $t-n$ to $t-1$
$min([P_{t-n}, ..., P_{t-1}])$ is the minimum price for the days $t-n$ to $t-1$
Positions are held for c days ignoring all signals in that time.

Sullivan et al. (1999) used $n = 200$, $x = 0.15$ and $c = 50$ days. See table 5.2.

5.5 Sullivan et al. (1999) Results

Tables 1 and 2 (pages 1661 and 1662 of Sullivan et al. (1999)) are reproduced here in table 5.1. They list the best rules and the periods where they performed best as reported by Sullivan et al. (1999).

Detailed data for each rule and how it performed in simulation, the number of trades etcetera is missing from Sullivan et al. (1999).

The authors claim that "In all four subperiods we find again that the best trading rule outperforms the benchmark".

[7] Without fixing the description to be what we think a channel rule should really be. We have no basis on which we can change the descriptions from Sullivan et al. (1999).

Sample		Best Rule, criterion	
Period	Dates	Mean return	Sharpe Ratio
Sub–period 1	1897–1914	5–day support & resistance, 0.005 band, 5–day holding period	20–day channel rule 0.075 width 5–day holding period
Sub–period 2	1915–1938	5–day moving average	5–day moving average 0.001 band
Sub–period 3	1939–June 1962	2–day on–balance volume	2–day moving average 0.001 band
Sub–period 4	July 1962–1986	2–day on–balance volume	2–day moving average 0.001 band
Sub–period 5	1987–1996	Filter rule, 0.12 position initiation 0.10 position liquidation	200–day channel rule 0.150 width 50–day holding period
90 years	1897–1986	5–day moving average	5–day moving average 0.001 band
100 years	1897–1996	5–day moving average	5–day moving average 0.001 band

Table 5.1: The "historically best performing trading rule"…"with respect to mean return criterion" and "with respect to mean Sharpe ratio" from Sullivan et al. (1999, pp. 1661–2)

5.6 Reproduction

We attempted to reproduce the results of Sullivan et al. (1999). We used the algorithms, in the previous section, that Sullivan et al. (1999) found effective by either of their two measures in any period.

The reproductions went smoothly for the moving average rules. For the Filter and Channel breakout rules, however, we could not get results that we would call satisfactory or even interesting. It could be worth exploring these rules in more depth, going beyond what Sullivan et al. (1999) described and looking for more precise definitions of the rules, but that is beyond our scope.

5.6.1 The Data, Rules and Objective Criteria

We obtained daily data for the Dow Jones Industrial Average index from Yahoo.com[8], and we use the "Adj Close" field. Using this source we can obtain data for 1928 to 2010. Sullivan et al. (1999), however, used data from 1897 to 1996. This limits the areas of Sullivan et al. (1999) that we can study.

[8] http://finance.yahoo.com/q/hp?s=^DJI+Historical+Prices

In Sullivan et al. (1999) the data were divided into five sub–periods described in section 5.1.1. Given the later start and finish of our data series we will use the same sub–periods where we can. The 20–Day Channel rule and the 5–Day Support and Resistance rule were found by Sullivan et al. (1999) to be best only for the sub–period 1897–1914. We have none of those data so we ignore those two rules. So our sub–periods are as described in table 5.2.

Period	Rule	Comment
1928–1996	5–Day Moving Average	The 5–Day Moving Average rule was found by Sullivan et al. (1999) to be best using the mean return criterion for 90 and 100–year periods starting in 1897 and finishing in 1986 and 1996 respectively
1928–1996	5–Day Moving Average with 0.001 band	The 5–Day Moving Average rule with a 0.001 band was found to be best using the Sharpe ratio for the same periods as above
1939–1986	The 2–Day Moving Average rule with a 0.001 band.	This rule was found to be best for the two consecutive periods that we have combined using the Sharpe ratio criterion
1939–1986	The 2–Day On Balance Volume.	This rule was found to be best for the two consecutive periods that we have combined using the mean return criterion
1987–1996	200–Day Channel Rule 0.15 width and 50–day holding period	This rule was found to be best, in this period, using the Sharpe ratio criterion.
1987–1996	Filter Rule 0.12 initiation and 0.1 liquidation	This rule was found to be best, in this same period, using the mean return criterion.
1997–2010	5–Day Moving Average with and without a 0.001 band	This period is not in Sullivan et al. (1999), but the 5–day moving average rules (with and without the band) are used across all the data in Sullivan et al. (1999) so we used both 5–day moving average rules on hold out samples. We found that with out the band the 5–day moving average rule makes a loss in this period.

Table 5.2: Periods and rules trialed to reproduce some of the results of Sullivan et al. (1999). The data the rules were tested on was the daily Dow Jones Industrial Average closing price.

We will use “buy and hold the index” as a benchmark where we need benchmarks for calculating break even transaction costs and profitability.

The rules generate trading signals in $\{-1, 0, 1\}$ where $-1 \implies$ short, $0 \implies$ neutral (i.e., in the risk free asset) and 1 $\implies$ long positions.

To assess feasibility of each rule we will find the break even transaction cost (as a proportion of the market price of the security at the time of the trade). While this is a rough measure of feasibility it benefits from being simple and tractable. There are more advanced models of transaction costs that we could use (see for example Trethewey and Crack (2010)), but break even transaction costs prove to be illuminating in several cases as we shall see.

We do not consider the issue of practicality. Can the signals generated can be acted on in time? It is not obvious how we could check this. There is the question of how an index could be traded before the advent of modern derivatives, the overheads involved in managing large portfolios that track the index etcetera. Whilst these are interesting questions they are beyond our scope.

5.6.2 5–Day Moving Average

Sullivan et al. (1999) found a simple 5–day moving average to be best for the periods 1915–1938 and the 90 years and 100 years from 1897 when using mean return as the criterion. We simulate trading using these rules and the Dow Jones index data.

The description in Sullivan et al. (1999) of the moving average rule is sufficient to reproduce their simulations. However we do not have all the data that Sullivan et al. (1999) used, so we cannot reproduce their simulations exactly. Our data series starts in 1928 so we made the decision to do the simulation in two tranches: 1928–1996 which is as much of the Sullivan et al. (1999) data as we have, and 1997–2010 as a holdout sample.

On the face of it the returns from this strategy are tremendous. In table 5.3 and figure 5.1 we report the results of simulating trading of the signals generated by the rule from 1928 to 1996. Over 17,132 trading days the wealth of a trader following this strategy grows by about 82,000 times. The mean daily return is better than three times the return of the index.

This is too good to be true, and as we shall see it is not so good and not so true.

	Index	Simulated Trading
Number of Trades	N/A	3,950
Total Return	25.87	362,647.77
(Geometric) Mean Daily Return	1.921×10^{-4}	7.475×10^{-4}
Standard Deviation	1.14%	1.08%
17,132 trading days		

Table 5.3: Trading the Dow Jones Industrial Average index using a 5–Day Moving Average signal for 1928–1996

5.6.3 5–Day Moving Average with a 0.001 Band

Sullivan et al. (1999) found a simple 5–day moving average to be best for the periods 1915–1938, and the 90–years and 100–years from 1897 when using the Sharpe ratio as the objective criterion.

These simulations are very similar to those in the previous section so we will follow the same procedures. The results are in table 5.4 and figure 5.2.

There are slightly more trades with the band than with out. This is because without the band the rule is effectively always in the market, long or short[9]. With the band, when the price is close to the moving average the algorithm is out of the market. So in a falling market without the band there is one trade from long to short, with the band it is long to flat to short. So there is an extra trade with the band[10].

The purpose of the band is partly to stop frequent trading. Small oscillations in price around the moving average signal will not cause frequent trades. But as we have observed it does not mean there will be fewer trades over all. It would be interesting to see how the size of the band effects the number of trades.

5.6.4 2–Day Moving Average with a 0.001 Band

Sullivan et al. (1999) found that the 2–Day Moving Average, 0.001 Band rule had the best performance for the two adjoining periods: 1939–1962 and 1962–1986. We have combined the two periods and we consider this rule over the period 1939–1986. We can

[9] Without the band the rule is only out of the market when the moving average equals the price. This happens once for the 2–Day moving average (24 May 1957) and not at all for the 5–Day moving average.

[10] This is a matter of definition. Going long to short could be described as 2 trades which would double the number of trades with no band.

	Index	Simulated Trading
Number of Trades	N/A	4,439
Total Return	25.87	121,695.11
(Geometric) Mean Daily Return	1.921×10^{-4}	6.837×10^{-4}
Standard Deviation	1.14%	1.06%
17,132 trading days		

Table 5.4: Trading the Dow Jones Industrial Average index using a 5–Day Moving Average signal and a band of 0.001 for 1928–1996. There are more trades using the band than without the band (see table 5.3) because with the band the rule trades out of the market between being long and short or vice verse

do this as there is no break between the two periods and both periods were evaluated using the same criterion.

This 2–Day Moving Average Band algorithm is very similar to the 2–Day Moving Average Band described in section 5.6.3 and it is clear how it can be implemented from the description provided in Sullivan et al. (1999).

The results of the simulated trading of this algorithm are in figure 5.3 and table 5.5. As with the previous 5–Day moving average the results are too good to be true. (See section 5.7 below).

	Index	Simulated Trading
Number of Trades	N/A	5,783
Total Return	11.34	45,731.45
(Geometric) Mean Daily Return	2.087×10^{-4}	8.913×10^{-4}
Standard Deviation	0.81%	0.76%
12,045 trading days		

Table 5.5: The 2–Day Moving Average rule with a 0.001 Band on the DJIA for 1939–1986

5.6.5 2 Day On Balance Volume Rule

Sullivan et al. (1999) found this rule was best using the mean return criterion for the same two adjoining periods as the 2-day moving average rule above (January 1939–June 1962, July 1962–December 1986) We have again combined the two periods, each roughly 20 years, to use one longer period of about 40 years.

The performance of the rule is illustrated in figure 5.4 and table 5.6. As in the above examples this seems too good to be true, and again it is.

	Index	Simulated Trading
Number of Trades	N/A	5,436
Total Return	11.34	233,371.78
(Geometric) Mean Daily Return	2.087×10^{-4}	1.027×10^{-3}
Standard Deviation	0.81%	0.80%
12,045 trading days		

Table 5.6: The performance of simulating trading the 2–Day On Balance Volume rule 1939–1986

5.6.6 200–Day Channel Rule with 0.150 Width and a 50–Day Holding Period

Sullivan et al. (1999) found this rule to be the best of the trialled rules for the 1987–1996 period using Sharpe ratio criterion.

From the description in Sullivan et al. (1999) we could not design a rule to do better than buy and hold of the index with zero transaction costs. See figure 5.5 and table 5.7. We discuss some of the difficulties we had with this rule in more detail in section 5.6.8.

	Index	Simulated Trading
Number of Trades	N/A	14
Total Return	2.35	0.14
(Geometric) Mean Daily Return	4.778×10^{-4}	5.130×10^{-5}
Standard Deviation	1.04%	0.26%
2,529 trading days		

Table 5.7: Trading the Dow Jones Industrial Average index using a 200–day channel rule with 0.150 width and a 50–day holding period for the period 1987–1996

5.6.7 Filter Rule with 0.120 Initiation, 0.100 Liquidation

In Sullivan et al. (1999) the filter rule was found to perform best for the 1987–1996 period using the mean return criterion.

However our simulation indicates far from being too good to be true this rule is quite dismal and performs much worse than buy and hold. See figure 5.6 and table 5.8. The mean daily return, while positive, is much worse than the benchmark. There is an improvement in volatility, but we do not think that is much compensation for returns so much worse than the index.

	Index	Simulated Trading
Number of Trades	N/A	591
Total Return	2.35	0.18
(Geometric) Mean Daily Return	4.778×10^{-04}	6.623×10^{-05}
Standard Deviation	1.04%	0.89%
2,529 trading days		

Table 5.8: Trading the Dow Jones Industrial Average index using a filter rule with 0.120 initiation, 0.100 liquidation 1987–1996

5.6.8 Ambiguity in the Rules: The Channel Rule Examined

It is unclear from Sullivan et al. (1999) exactly how the "channel" is defined. It is clear when a channel exists but the rule has a signal to go long generated when the price is above the channel and a signal to go short generated when the price is below the channel. So although we defined a rule in equation 5.4 (page 78) it is not clear that that is the only definition.

In equation 5.4 we defined the channel and channel rule as:

$$\begin{aligned} &P_{m_{t,n}} = \tfrac{(max([P_{t-n}..P_{t-1}])+min([P_{t-n}..P_{t-1}]))}{2} \\ &P_{h_{t,n}} = P_{m_{t,n}} \times (1+\tfrac{x}{2})\text{the top of the channel} \\ &P_{l_{t,n}} = P_{m_{t,n}} \times (1-\tfrac{x}{2})\text{the bottom of the channel} \end{aligned} \tag{5.5}$$

But the channel could be defined:

$$\begin{aligned} P'_{h_{t,n}} &= min(P_{t-n}...P_{t-1}) \times (1+x) \\ P'_{l_{t,n}} &= max(P_{t-n}...P_{t-1}) \times (1-x) \end{aligned} \tag{5.6}$$

Then the signal could be defined as:

$$S_t = \begin{cases} +1 & P_t > P'_{h_{t,n}} \\ -1 & P_t < P'_{l_{t,n}} \end{cases} \tag{5.7}$$

An alternative that is plausible given the description in Sullivan et al. (1999) is to define the channel around the mid point between the max and min over $[P_{t-n}..P_{t-1}]$. So the channel would be defined around the mid point P_m:

$$P_m = \frac{max(P_{t-n}..P_{t-1}) + min(P_{t-n}..P_{t-1})}{2} \tag{5.8}$$

The width of the channel would be $P_m \times x$. So the top of the channel (P_h) and the bottom (P_l) would be:

$$\begin{aligned} P_h &= P_m \times (1 + \tfrac{x}{2}) \\ P_l &= P_m \times (1 - \tfrac{x}{2}) \end{aligned} \tag{5.9}$$

Then the long signal would be generated for $P_t > P_h$ and a short for $P_t < P_l$. This is symmetric and unambiguous, but more complex.

Additionally while there is not much authoritative literature on channel breakouts what we can find describes the channels as being roughly parallel to the price curves, which is not apparent from Sullivan et al. (1999)'s description in which the channels are parallel to the x–axis. See for example Caginalp and Balenovich (1996).

5.6.9 Conclusion on Reproduction

At this point we are in the unfortunate position of being unable to reproduce Sullivan et al. (1999)'s results for the filter rule (section 5.6.7) or the channel rule (section 5.6.6). So in keeping with our comments in section 2.1, we will discontinue investigation of these rules. This leaves the 2–Day Moving Average with 0.001 Band, the 2–Day On Balance Volume, the 5–Day Moving Average with 0.001 Band and the 5–Day Moving Average no band rules as the rules from Sullivan et al. (1999) that we have data for and we can reproduce.

This is frustrating and we cannot help but think that our implementation lacks something compared to Sullivan et al. (1999)'s. But we are limiting ourselves to using the information supplied by, or referenced from, Sullivan et al. (1999). It is up to the authors of Sullivan et al. (1999) to provide clear details of what it was they were studying. We cannot believe that the algorithms for the filter or channel rules that we built using information provided by Sullivan et al. (1999) are the same as theirs. We could have done our own research to look for alternative descriptions of these rules, but that, too, is beyond our scope.

5.7 Feasibility (Break Even Transaction Costs)

Sullivan et al. (1999) did not apply any transaction costs to their simulations. But there are an enormous number of transactions that some of the rules generate. For example we can see in table 5.3 that over the period of our simulation there were nearly 4,000 trades, about 42 a year or nearly one a week,[11] for the 5–day moving average rule.

One approach is to ask "what is the break even transaction cost for the rule?". We can then judge if the transactions costs are low enough to be feasible. A related question is "what are the expected transaction costs?".

Determining transaction costs is not easy. Transaction costs consist of a portion of the bid–ask spread plus the trading commissions, at least. To this must be added the price impact of trading, taxes and sundry other idiosyncratic costs to the trading agent. There have been several estimates of transaction costs over time. In Lesmond et al. (1999) it is estimated that the minimum average transaction cost was 1.2% on the New York Stock Exchange in the years from 1963 to 1990[12]. It is tempting to assume that transaction costs have reduced over time, thus for the portion (roughly half) of our sample which predates 1963 could be assumed to be higher than 1.2%. However in Gehrig and Fohlin (2006) transaction costs were estimated for the Berlin stock exchange for 1880–1910, using the same technique developed in Lesmond et al. (1999). Gehrig and Fohlin (2006) found the transaction costs were as low as 0.53%. In Sullivan et al. (1999) for the early

[11] It is interesting that the rule trades at roughly the same frequency as the length of data it uses.

[12] Being the average transaction cost for the largest size decile. For the smallest size decile the average was 10.3%.

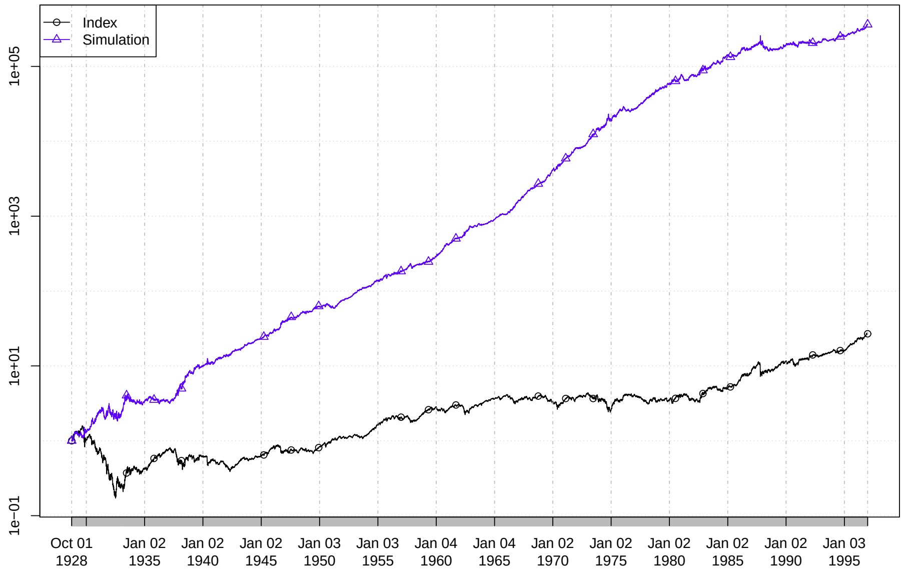


Figure 5.1: Trading the Dow Jones Industrial Average index using a 5–Day Moving Average signal (Y–Axis is log scale) for 1928–1996

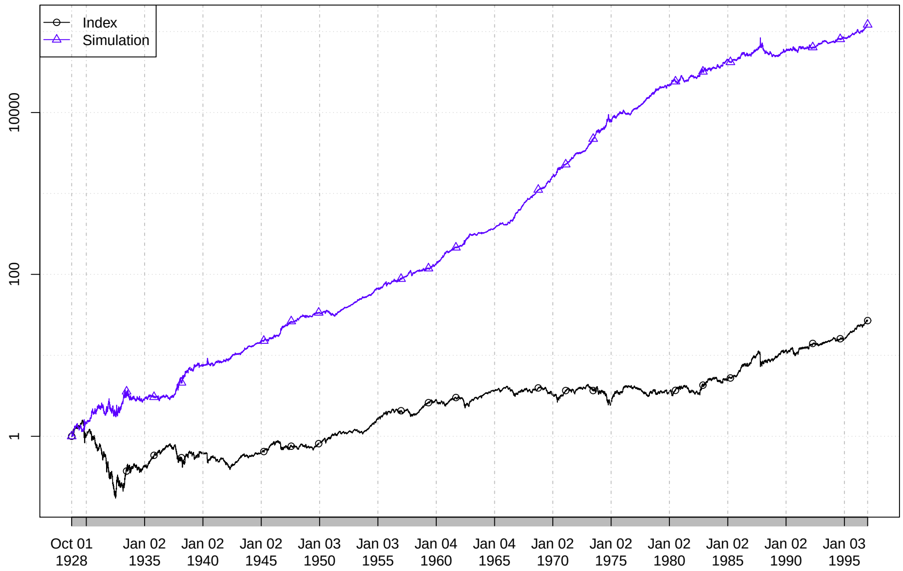


Figure 5.2: Trading the Dow Jones Industrial Average index using a 5–Day Moving Average signal with a 0.001 band (Y–Axis is log scale) for 1928–1996

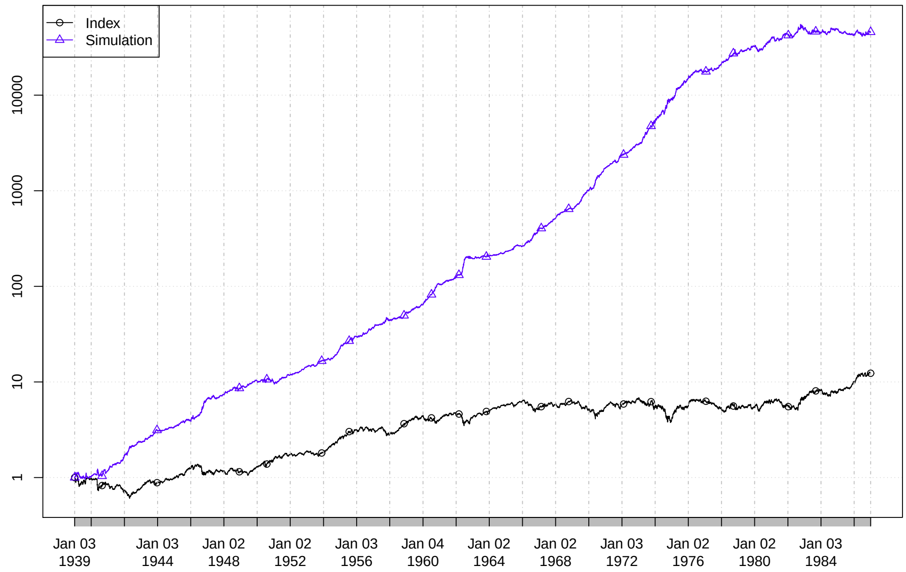


Figure 5.3: The 2–Day Moving Average rule with a 0.001 Band on the DJIA for 1939–1986. (The Y–Axis is log scale).

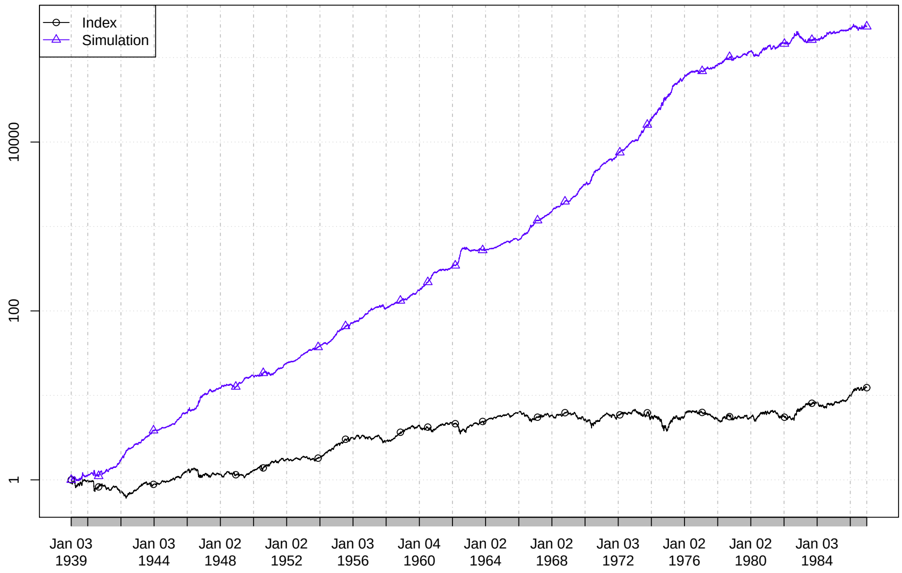


Figure 5.4: Trading the Dow Jones Industrial Average index using a 2–Day On Balance Volume rule for 1939–1986. (The Y–axis is Log scale).

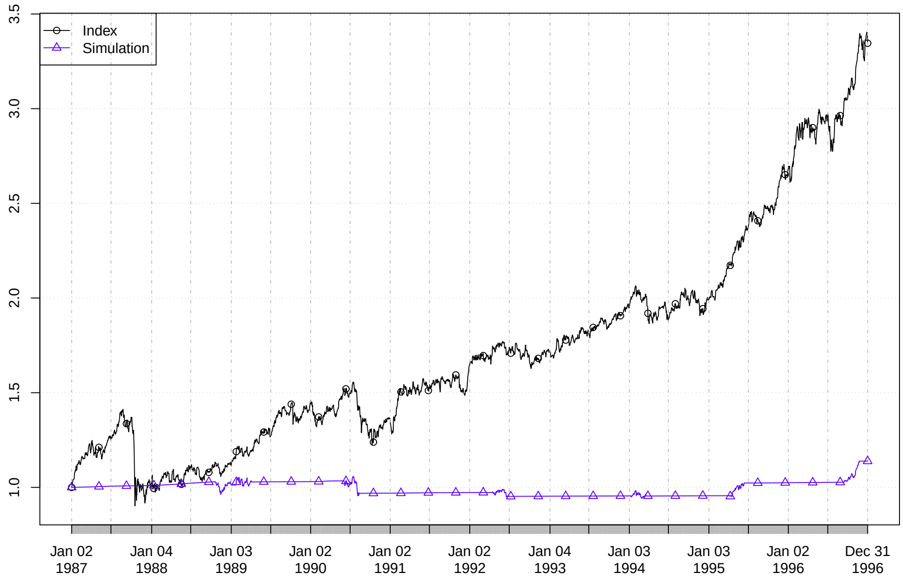


Figure 5.5: Trading the Dow Jones Industrial Average index using a 200–day channel rule with 0.150 width and a 50–day holding period for the period 1987–1996. This is an illustration of the problems we had reproducing the results from Sullivan et al. (1999).

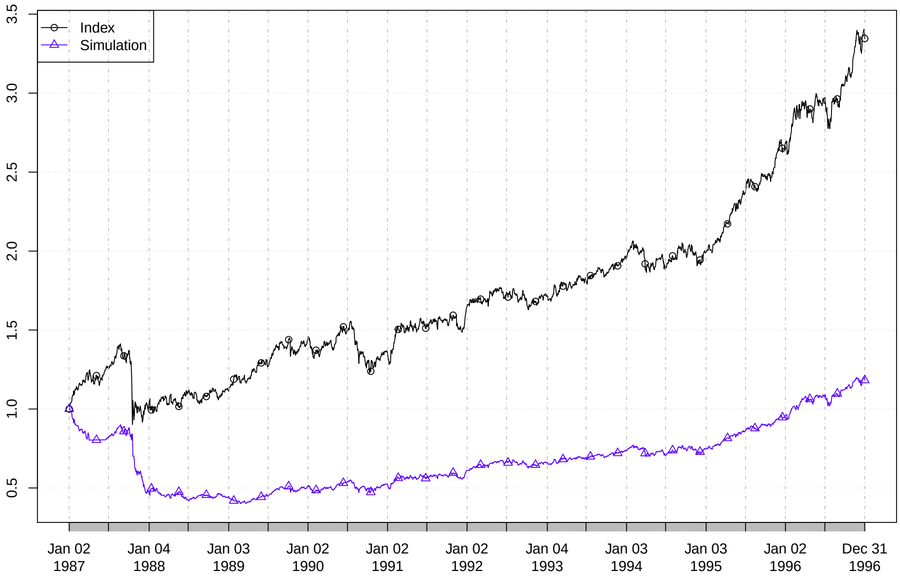


Figure 5.6: Trading the Dow Jones Industrial Average index using a filter rule with 0.120 initiation, 0.100 liquidation 1987–1996

part of the sample there are formidable barriers to implementing the strategy[13]. So we are comfortable that saying a minimum break even transaction cost of 0.5%, 50 basis points, is a conservative estimate. If the break even transaction costs are less than 50 basis points the trading rules are not feasible.

[13] As we discussed in section 2.8 there were no ETFs or index funds available before the 1970s.

5.7.1 Results

In section 5.6.9 we gave up investigating the filter and channel rules as we could not come close to reproducing the results from Sullivan et al. (1999). So we will study the remaining rules listed in table 5.2 page 80 that we can reproduce, excluding the 5–Day moving average rule on the data from 1996–2010 that we will use as a hold out sample.

		Break even TC	
Period	Rule	Net	Gross
1928–1996	5–Day Moving Average No Band	0.24%	0.32%
1928–1996	5–Day Moving Average with 0.001 Band	0.22%	0.31%
1939–1986	2–Day Moving Average with 0.001 Band	0.18%	0.23%
1939–1986	2–Day On Balance Volume	0.23%	0.18%

Table 5.9: The break even transaction costs that we calculated for the algorithms from Sullivan et al. (1999) that we were able to reproduce. The transaction costs are calculated net where break even is getting the same return as the index, and gross, where break even is returning 0. The highest break even transaction cost is 32 basis points. In section 5.7 we decided that 50 basis points would be a conservative estimate of the minimum break even transaction costs for feasible technical analysis rules.

5.7.2 Conclusion from Feasibility

None of the rules is feasible once transaction costs are considered. We find it disappointing that the Sullivan et al. (1999) authors did not do more to consider transaction costs in their paper. Finding the break even transaction cost is a very simple procedure and it is very illuminating.

Not considering transaction costs for a trading strategy is a form of data snooping. It leads to spurious results. Likely transaction costs are part of the data, and ignoring them biases the results.

There are clearly some interesting properties of the DJIA index that are worth pursuing, properties that make the moving average rules look so attractive, but they are, again, beyond our scope.

Nevertheless we will apply our tools to see how the remaining strategies work. But we will conclude that the feasibility test indicates that the results in Sullivan et al. (1999) are data snooping.

5.8 Statistical Tests

The statistical tests for the promising rules reported in Sullivan et al. (1999) are in table 5.10. The null hypothesis tests are for the mean returns of the rule being greater than zero and/or greater than the buy and hold returns on the index.

We use the single sample t–test and the Wilcoxon Signed Rank test when comparing the algorithm's returns with zero and the t–test for two independent samples and the Mann–Whitney U test when comparing the algorithm's returns with the index's returns.

The results for the four rules at the top of the table are consistent, there is strong evidence to reject the null hypothesis and accept the alternative. It does appear for those four rules that they perform better than nothing and they perform better than the index. The graphs of these rules in figure 5.4 page 89 and figure 5.3 page 89 bear this out.

The remaining two rules have much more interesting statistical properties. These are the two 5–Day moving average rules (with and without a band). Here the statistical tests diverge.

The results of the standard t–test and the Wilcoxion test disagree for the case where the alternative is "Greater than Index" where the t–test rejects and the Wilcoxion does not reject the null.

There are 24 results (six rules, two alternative criteria and two tests) in total, of which three do not reject the null at the 5% level. Given that, it is not unbelievable that the cases which fail to reject the null are due to random variation. So there is no evidence of data snooping using Statistical Significance.

5.9 Neighbourhood Analysis

The neighbourhood analysis results for the four algorithms from Sullivan et al. (1999) that we can simulate are tabulated in table 5.11.

There is a greater than ten percentage points change in the number of trades for the 5–Day moving average algorithm (without a band) around 5 days. There is also a greater than ten percentage point change in the mean return around the band for the 2–day moving average algorithm. For all the algorithms there is a large difference in the total

	Alternative Greater than			
	Zero		Index	
Rule	t–test	Wilcoxion	t–test	Wilcoxion
2 Day on Balance Volume 1939-01-03 to 1962-06-29	0.000	0.000	0.000	0.000
2 Day on Balance Volume 1962-07-02 to 1986-12-31	0.000	0.000	0.000	0.000
2 Day Moving Average 0.001 Band 1939-01-03 to 1962-06-29	0.000	0.000	0.000	0.009
2 Day Moving Average 0.001 Band 1962-07-02 to 1986-12-31	0.000	0.000	0.000	0.000
5 Day Moving Average 1928-10-01 to 2010-10-27	0.000	0.000	0.000	0.013
5 Day Moving Average 0.001 Band 1928-10-01 to 2010-10-27	0.000	0.000	0.000	**0.067**

Table 5.10: Statistical significance for Sullivan et al. (1999) rules. The numbers are *p*–values from the t–test (single sample t–test when comparing the algorithm's returns with zero and t–test for two independent samples when comparing the algorithm's returns with the index's returns) and Wilcoxon test (the Wilcoxon Signed rank test) when comparing with zero and the Mann–Whitney U test otherwise. The two results in the second panel (the 5–day moving average rules with and without a band) for comparing the algorithm's returns with those of the index are curious in the degree to which the t–test and the Wilcoxon tests disagree. Given that there are 24 tests it is not inconceivable that the divergence between the t–test and Wilcoxon test are due to random variation.

return, except for the return of the 3–Day OBV when compared to the 2–Day OBV. Here the daily returns are remarkably close. (The 1–Day OBV is meaningless as there are no trades triggered. The 1–day moving average of the OBV is the same as the OBV, as there are no signals).

For the two algorithms that use a band a small change in the band caused small changes in the number of trades. In both cases the mean daily return was higher for the smaller band. So too was the volatility. Since both were evaluated using Sharpe ratios (see table 5.2 page 80) we should compare Sharpe Ratios. As a quick approximation the mean daily return over the volatility from table 5.11 are in table 5.12.

For the 2–Day MA we can see that the small change in the band caused a large change in the objective measure used in Sullivan et al. (1999) as well as return and volatility. The changes in the objective measures were less dramatic for the 5-Day MA. But this is evidence of data snooping[14].

For the 5-Day MA and 2–Day OBV algorithms, with out a band, Sullivan et al. (1999)

[14] The Sharpe ratios are dramatic.

used mean return as the objective measure. Adjusting the periods (days over which a moving average is calculated) for these two algorithms to create a neighbourhood requires that we make a major change. The 4–Day MA algorithm's period is only 80% of the 5–Day period and the 6–Day algorithm is 120% of the 5–Day period. The difference is even more dramatic for the 2–Day OBV. Still the changes in performance were much less dramatic for the 2–Day OBV and 5–Day MA algorithms than for the 2-Day MA where we explored the neighbourhood in the band. But given the size of the neighbourhood we would be reluctant to draw too many conclusions using this technique in this case, because the changes are not small.

So using Neighbourhood Analysis around the MA period parameter for MA rules we cannot make a conclusion about data snooping.

Algorithm	Var	Number of Trades	Total Return	Mean Daily Return	Volatility	Break–Even TC
2–Day Moving Average **(Band)** 1939–1986	0.0005	5,725 (99%)	151,538 (331%)	0.099% (111%)	0.78%	0.24%
	0.001	5,783	45,731	0.089	0.76%	0%
	0.0015	5,565 (96%)	13,992 (31%)	0.079% (89%)	0.74%	0.23%
5–Day Moving Average **(Band)** 1928–1996	0.0005	4,215 (95%)	209,770 (172%)	0.072% (105%)	1.07%	0.32%
	0.001	4,439	121,695	0.068	1.06%	0%
	0.0015	4,611 (104%)	78,636 (65%)	0.066% (96%)	1.05%	0.31%
5–Day Moving Average **(MA Period)** 1928–1926	4–Day	4,652 (118%)	311,178 (86%)	0.074% (99%)	1.09%	0.27%
	5–Day	3,950	362,648	0.075	1.08%	0.32%
	6–Day	3,504 (89%)	154,604 (43%)	0.07% (93%)	1.07%	0.34%
2–Day On Balance Volume Average **(MA Period)** 1939–1986	1	–	–	–	–	–
	2	5,436	233,372	0.103	0.8%	0.23%
	3	5,411 (100%)	217,992 (93%)	0.102% (99%)	0.8%	0.23%
	4	3,825 (70%)	6,496 (3%)	0.073% (71%)	0.8%	0.23%

Table 5.11: Neighbourhood Analysis for the algorithms that we could reproduce. Displaying how the number of trades, total and mean return, volatility and the gross break even transaction cost for the algorithms and neighbours.
The first column names the algorithm. In **(bold)** is the parameter we change.
The second column describes the values of that parameter.
The next five columns contain the results of the simulations. The percentages in brackets for the number of trades, total and mean return are the percentage of the main result that was in the neighbour.
Note that the 1–day moving average (a neighbour for the 2-Day moving average) makes no trades. This is because the 1-Day moving average of a daily series is the series itself. Since the two series (the index and the 1–day average) are identical the 1–day MA is never greater (smaller) than the index to give a long (short) signal.

Algorithm	Band	Return/Volatility
2–Day Moving Average	0.0005	0.1269
	0.0010	0.1171
	0.0015	0.1068
5–Day Moving Average	0.0005	0.0673
	0.0010	0.0642
	0.0015	0.6286

Table 5.12: Neighbourhood Analysis in the band for the two Moving Average algorithms. The mean daily return over the volitility from table 5.11.

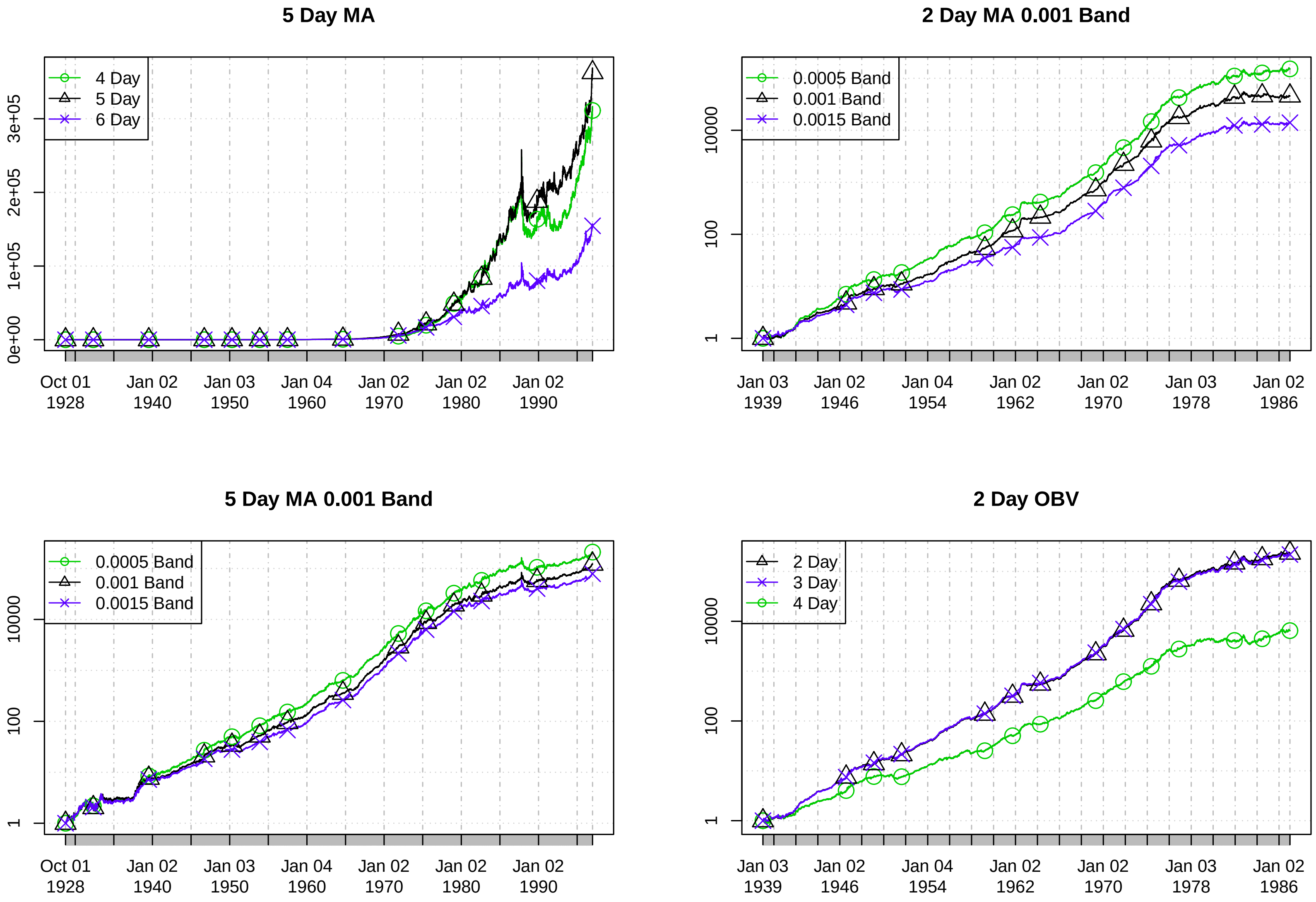


Figure 5.7: Neighbourhood Analysis for the algorithms that we could reproduce. The x-axis is the value of the $1 portfolio invested in the index following the signals of the algorithms, ignoring transaction costs. Given the long time period, where a small divergence in mean returns can create a large divergence in returns three of the algorithms' portfolios remained very close in value.
For the 5–day moving analysis there is greater divergence. Of the four algorithms here this is the only one that shows signs of data snooping.

5.10 Other Markets

In this section we test the rules, that we can reproduce, on other markets.

If the results in Sullivan et al. (1999) are due to data snooping we do not expect to see similar performance in other markets. If they were not due to data snooping we expect to see similar results Especially in related markets.

From section 5.6.9 we know we that the rules we can reproduce are:

- 2–Day Moving Average with 0.001 Band (1939–1986)
- 2–Day On Balance Volume (1939–1986)
- 5–Day Moving Average with 0.001 Band (1897–1996)
- 5–Day Moving Average no band (1897–1996)

For the other markets to test we will use the markets from section 3.2.4. (In 3.2.4 we used the Dow Jones Industrial average, which is the main market here so it is excluded).

These are (with the starting date, the correlation with the DJIA and the risk free index used):

- SPTSX Toronto Stock Exchange Composite Index (from 1977, 95.2%, risk free rate is the Canadian 3 month treasury)
- CCMP The NASDAQ composite index (from 1971, 95.0%, risk free is the USE 3 month treasury)
- HSI The Hang Seng Index. (from 1986, 88%, risk free rate is 0[15])
- MEXBOL The Mexican IPC index (from 1991, risk free rate is 0[15])
- NKY The Nikkei–225 Stock Average. (from 1984, risk free rate is mean Japanese discount rate.)

[15] For the MEXBOL (Mexican) and HSI (Hong Kong) we did not apply a risk free rate, using zero instead. In these two markets there is so much uncertainty that there is no such thing as a risk free rate. The MEXBOL is a Mexican index and over this period there was a lot of instability in the Mexican state. The HSI is from Hong Kong and there has been both structural uncertainty (with the 1997 transfer of sovereignty) and state risk in that the current administration rules at the whim of the Chinese state, and the Chinese state is not a transparent organisation, nor has it a track record of stability, yet. So in these two cases we felt that zero was the correct risk free rate.

The markets are listed, in table 5.13, in descending order of correlation with the DJIA index that is used by Sullivan et al. (1999).

In figures 5.8 to 5.15 the performance of the rules is shown in the other markets. Summary statistics for each market and algorithm are in table 5.14.

5.10.1 Break Even Transaction Costs in Other Markets

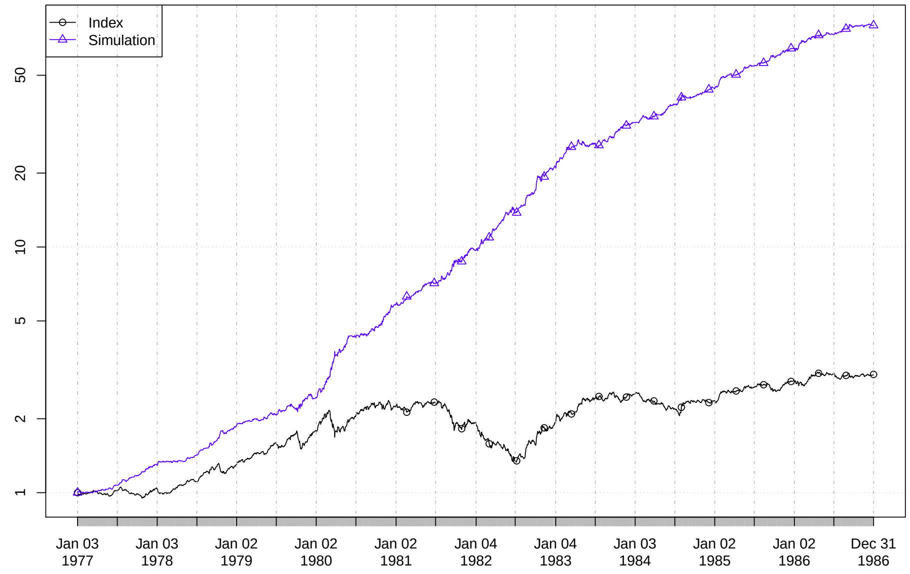
Given the instances of good apparent performance of the algorithms in other markets (CCMP – 5–Day MA with and without a band and SPTSX – 2–Day OBV) we felt it would be illuminating to look at the break even transaction costs for these markets. This we do in table 5.13.

We see no sign of feasible break even transaction costs for any of the markets except the 2–day OBV rule using the SPTSX index.

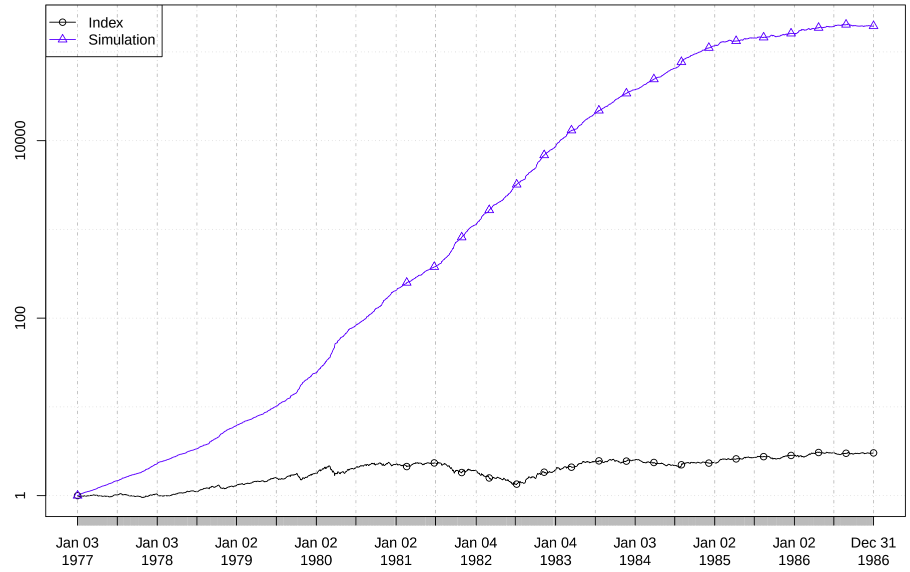
Disregarding transaction costs, these three examples have exponential returns. But after a simple adjustment for transaction costs the results are much less impressive.

Market	Algorithm	Net Break Even Transaction Cost	Date Covered
CCMP	5–Day Moving Average 0.001 Band	0.30%	1971-02-05 to 2010-10-27
SPTSX		0.27%	1977-01-03 to 2010-10-27
NKY		0.00%	1984-01-04 to 2010-10-27
HSI		0.38%	1986-12-31 to 2010-10-27
MEXBOL		0.21%	1991-11-08 to 2010-10-27
CCMP	5–Day Moving Average	0.27%	1971-02-05 to 2010-10-27
SPTSX		0.27%	1977-01-03 to 2010-10-27
NKY		0.00%	1984-01-04 to 2010-10-27
HSI		0.35%	1986-12-31 to 2010-10-27
MEXBOL		0.23%	1991-11-08 to 2010-10-27
CCMP	2–Day Moving Average 0.001 Band	0.48%	1971-02-05 to 1986-12-31
SPTSX		0.40%	1977-01-03 to 1986-12-31
NKY		0.00%	1984-01-04 to 1986-12-26
HSI		NA	
MEXBOL		NA	
CCMP	2–Day On Balance Volume	0.00%	1971-02-05 to 1986-12-31
SPTSX		1.10%	1977-01-03 to 1986-12-31
NKY		0.00%	1984-01-04 to 1986-12-26
HSI		NA	
MEXBOL		NA	

Table 5.13: The break even transaction costs for other markets. There is an uneven coverage for the periods used in Sullivan et al. (1999) for these markets. The date range each data set encompasses is in the last two columns. Only one of the break even transaction costs (2–Day OBV for SPTSX at 1.1%) is better than the feasibility cutoff of 0.5%. This indicates that, in almost all cases, the algorithms described in Sullivan et al. (1999) and that we were able to reproduce would not have been feasible in markets other than the ones used in Sullivan et al. (1999).

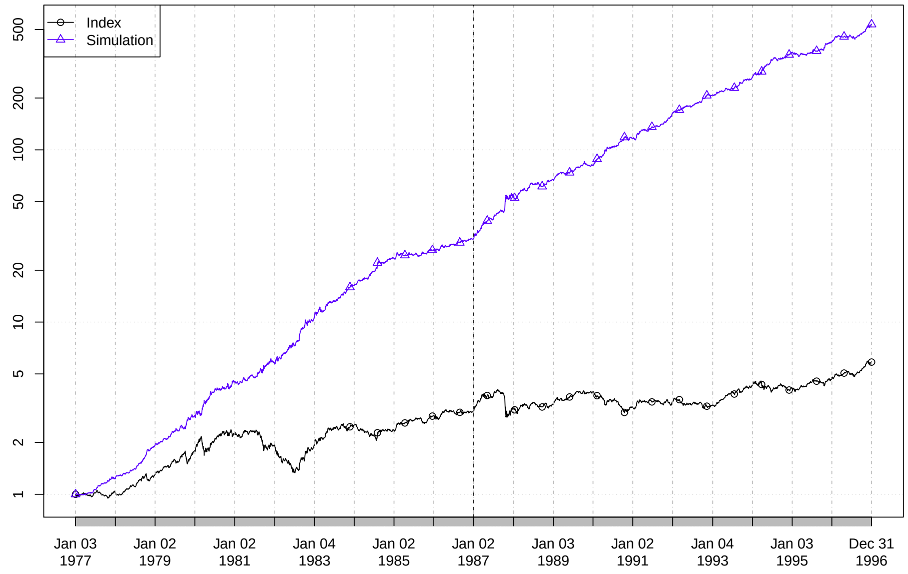


(a) 2–Day MA 0.001 Band

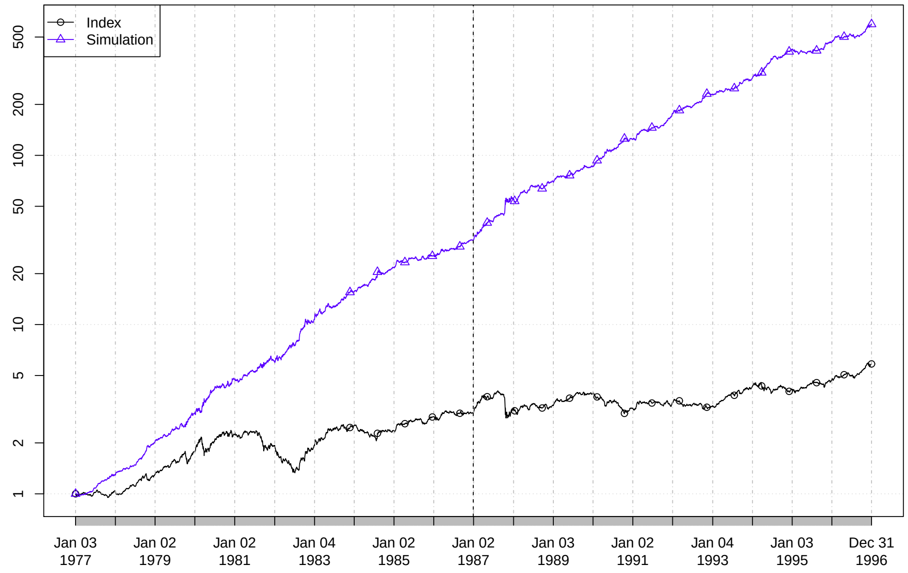


(b) 2–Day OBV

Figure 5.8: The 2–Day MA with Band and 2–Day OBV rules applied to the Toronto Stock Exchange Composite Index. The simulations all do better than the index. The 2–Day OBV has very good returns (the y–axis uses a logrithmic scale).

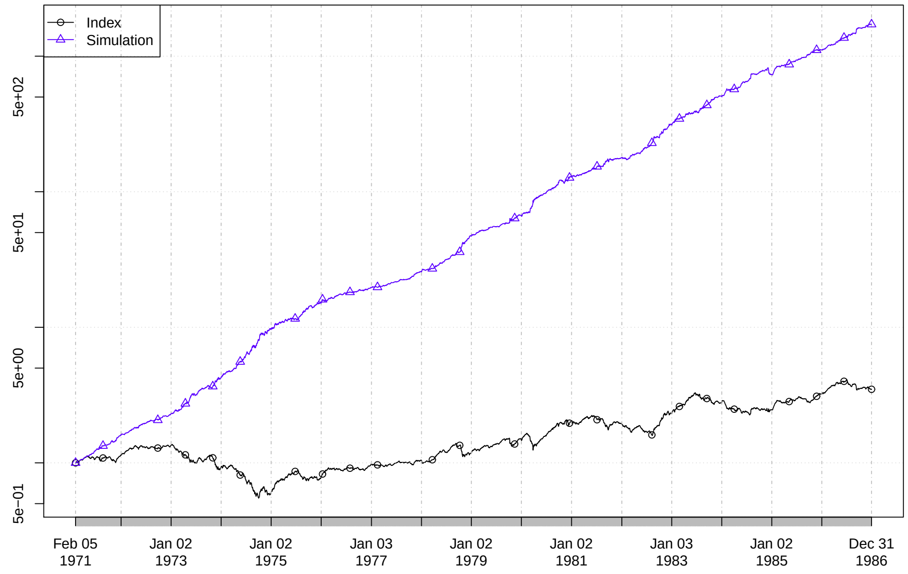


(a) 5–Day MA 0.001 Band

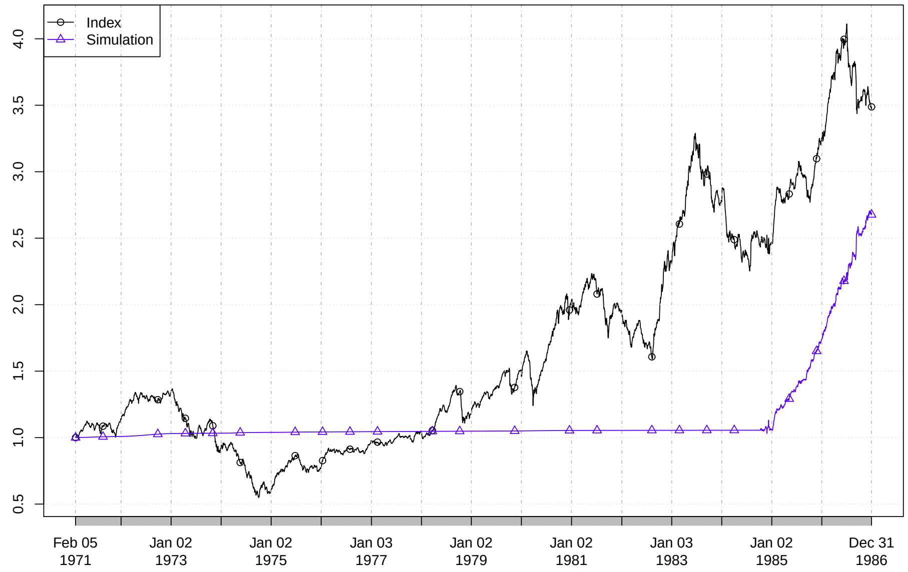


(b) 5–Day MA No Band

Figure 5.9: The 5–Day MA with and without a band rules applied to the Toronto Stock Exchange Composite Index. The dotted vertical lines indicate the extent of the plots in figure 5.8.The simulations all do better than the index

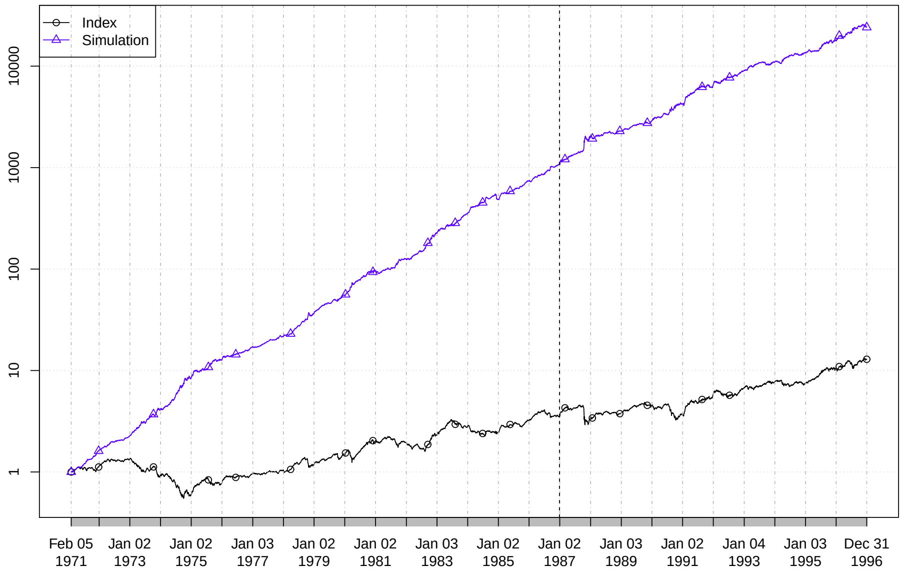


(a) 2–Day MA 0.001 Band

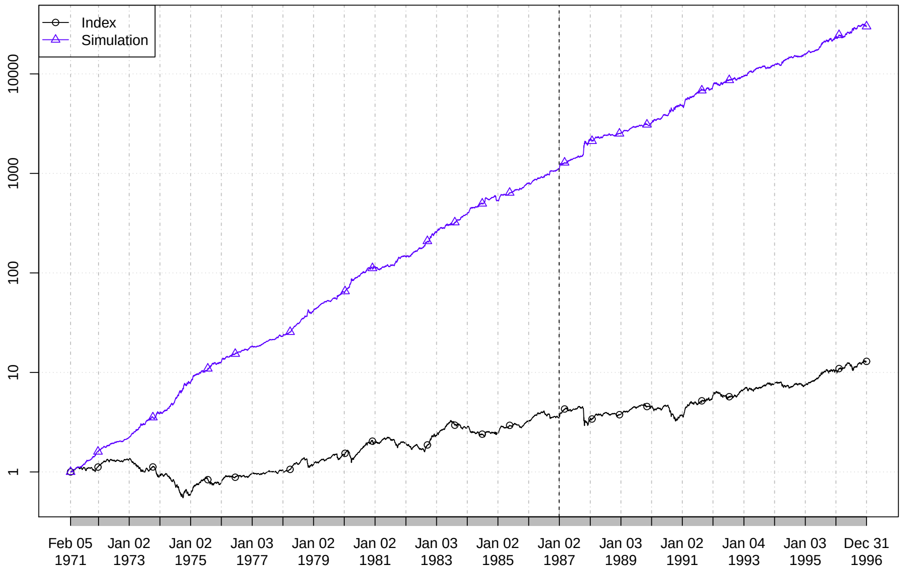


(b) 2–Day OBV

Figure 5.10: The 2–Day MA with Band and 2–Day OBV rules applied to the Nasdaq Composite Index. There are no trades for the 2–Day OBV before 1985. After 1985 the 2–Day OBV rule seems to do well.

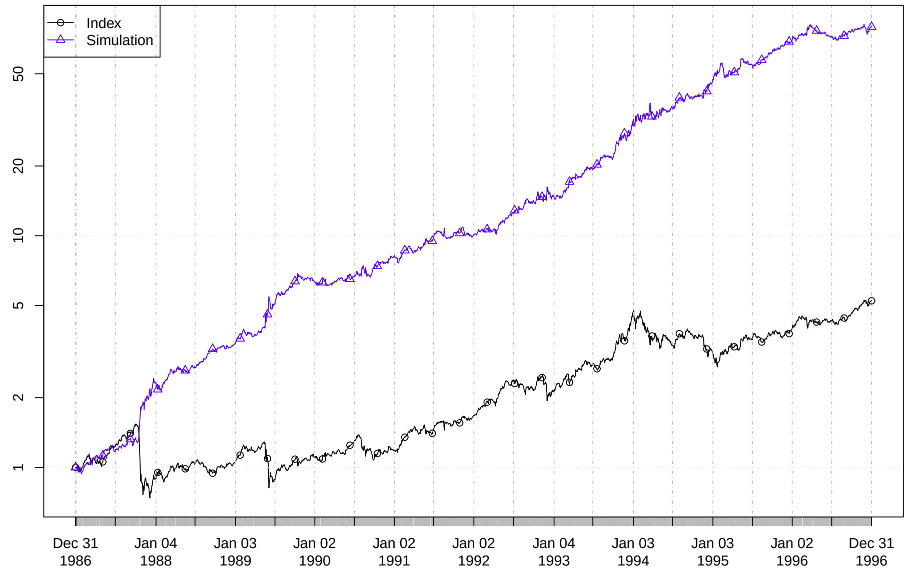


(a) 5–Day MA 0.001 Band

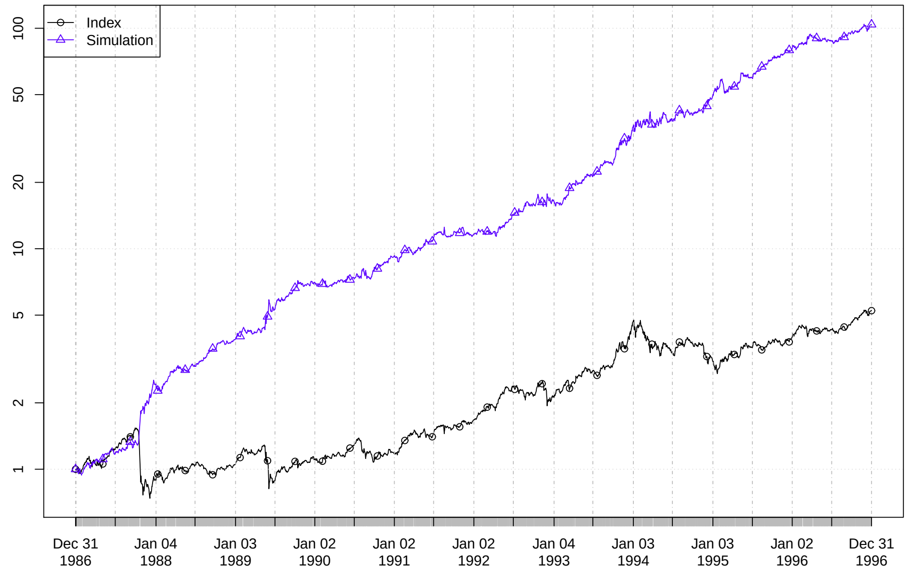


(b) 5–Day MA No Band

Figure 5.11: The 5–Day MA with and without a band rules applied to the Nasdaq Composite Index (CCMP). The simulations all do better than the index. The dotted vertical lines indicate the extent of the plots in figure 5.10. Both of these rules have very good returns. (The y–axes use a logrithmic scale).

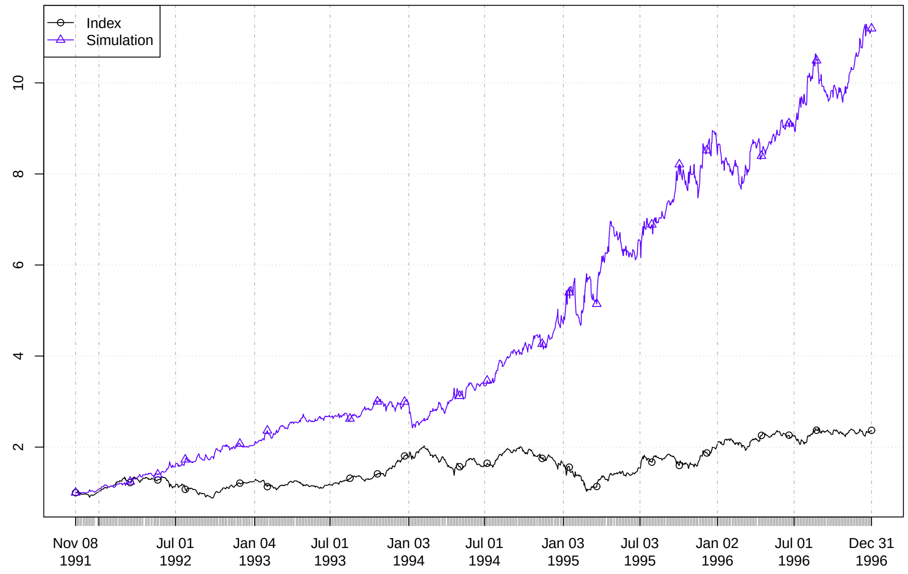


(a) 5–Day MA 0.001 Band

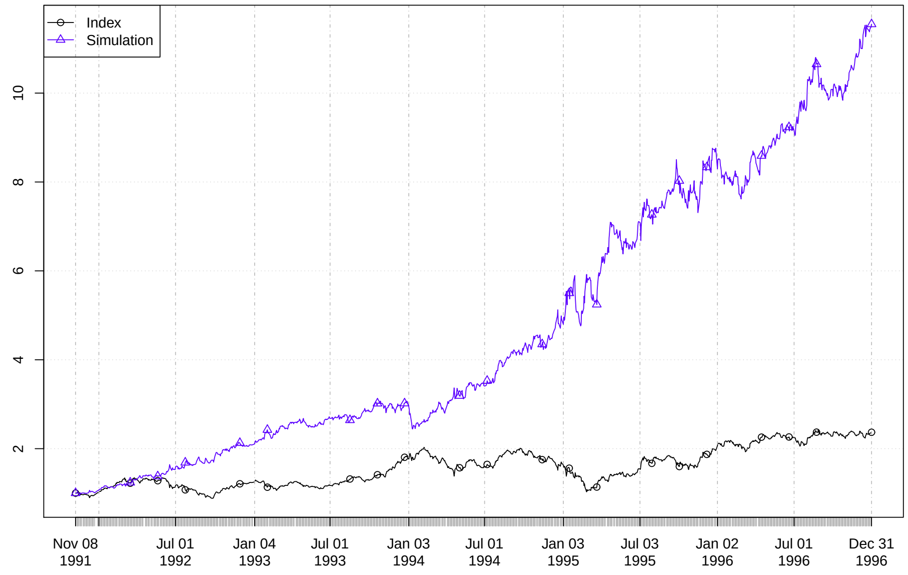


(b) 5–Day MA No Band

Figure 5.12: The 5–Day MA with and without a band rules applied to the Hang Seng Index. The simulations all do better than the index.

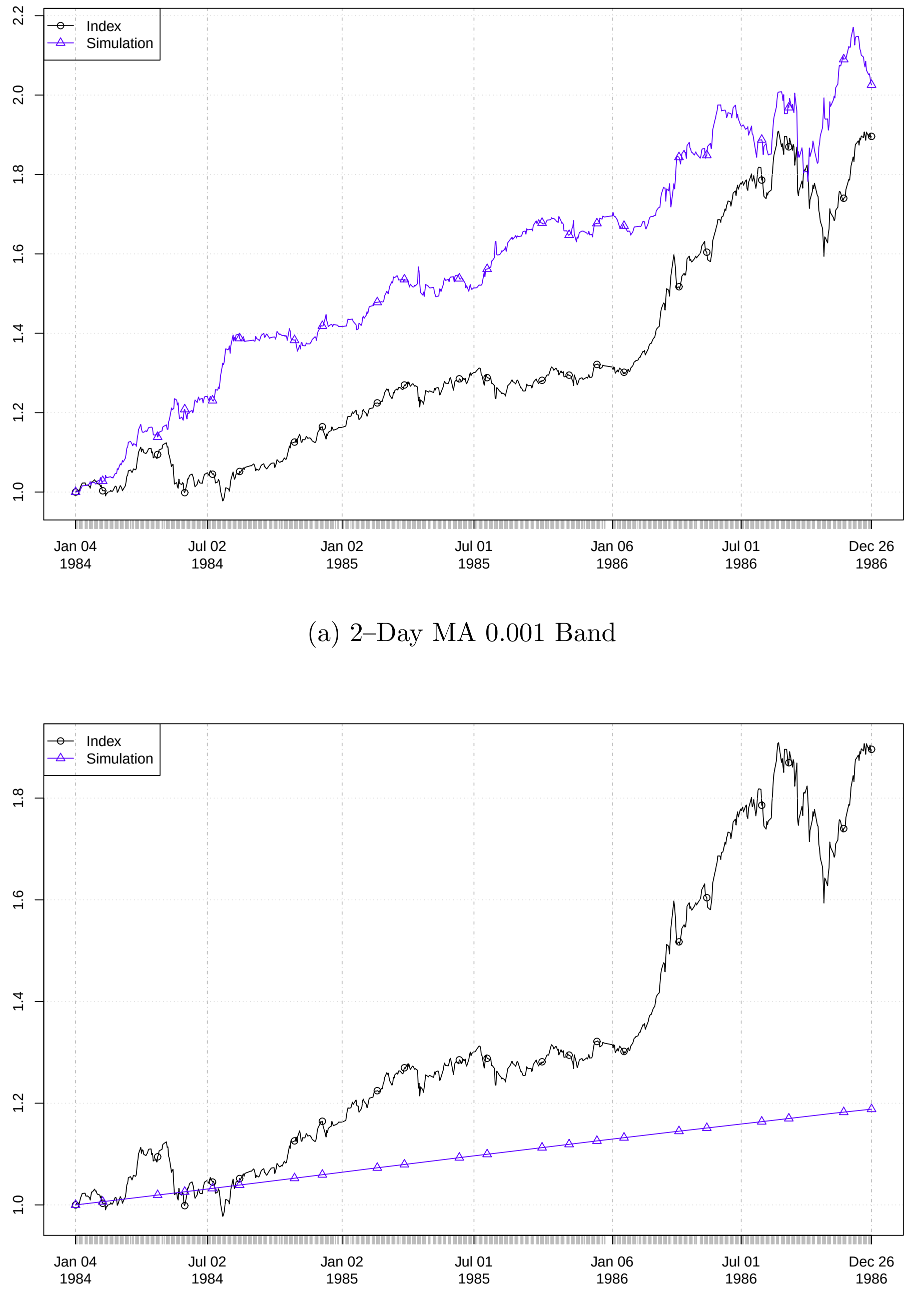


(a) 5–Day MA 0.001 Band



(b) 5–Day MA No Band

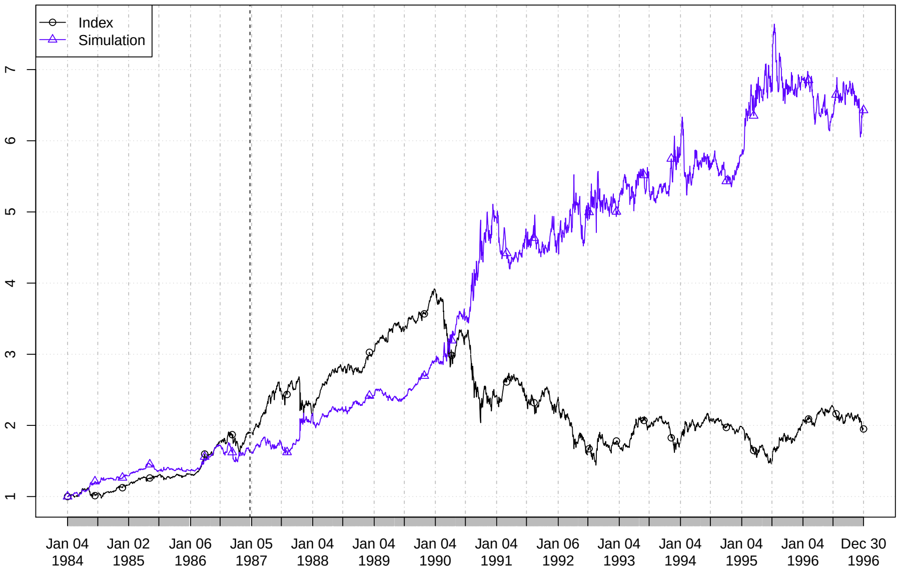
Figure 5.13: The 5–Day MA with and without a band rules applied to the Mexico IPC Index. The simulations all do better than the index.



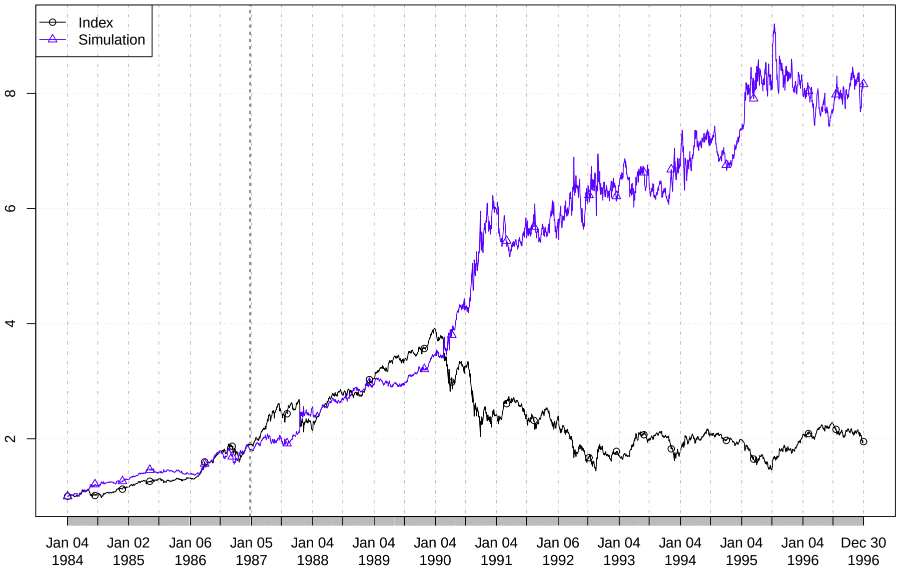
(a) 2–Day MA 0.001 Band

(b) 2–Day OBV

Figure 5.14: The 2–Day MA with Band and 2–Day OBV rules applied to the Nikkei Index (NKY). The 2–Day OBV has no trades and grows at the risk free rate.



(a) 5–Day MA 0.001 Band



(b) 5–Day MA No Band

Figure 5.15: The rules applied to the Nikkei Index. The simulations all do better than the index (NKY). The dotted vertical lines indicate the extent of the plots in figure 5.14.

Function	Start	End	Number of Trades	Index Total Return	Simulation Total Return	Index Volatility (Std. Dev.)	Simulation Volatility (Std. Dev.)	Daily Mean Index return	Daily Mean Simulation return
SPTSX									
5–Day MA	1977-01-03	1996-12-31	986	4.856	594.320	7.44×10^{-03}	6.98×10^{-03}	3.51×10^{-04}	1.27×10^{-03}
5–Day MA 0.001 Band	1977-01-03	1996-12-31	1135	4.856	534.023	7.44×10^{-03}	6.79×10^{-03}	3.51×10^{-04}	1.25×10^{-03}
2–Day MA 0.001 Band	1977-01-03	1986-12-31	1052	2.030	78.840	7.79×10^{-03}	7.12×10^{-03}	4.40×10^{-04}	1.74×10^{-03}
2–Day OBV	1977-01-03	1986-12-31	1008	2.030	195600.300	7.79×10^{-03}	5.99×10^{-03}	4.40×10^{-04}	4.85×10^{-03}
CCMP									
5–Day MA	1971-02-05	1996-12-31	1265	11.910	29992.340	8.19×10^{-03}	7.62×10^{-03}	3.91×10^{-04}	1.58×10^{-03}
5–Day MA 0.001 Band	1971-02-05	1996-12-31	1395	11.910	24101.720	8.19×10^{-03}	7.47×10^{-03}	3.91×10^{-03}	1.54×10^{-03}
2–Day MA 0.001 Band	1971-02-05	1986-12-31	1708	2.488	1716.619	7.50×10^{-03}	6.70×10^{-03}	3.11×10^{-04}	1.86×10^{-03}
2–Day OBV	1971-02-05	1986-12-31	211	2.488	1.677	7.50×10^{-03}	2.43×10^{-03}	3.11×10^{-04}	2.45×10^{-04}
HSI									
5–Day MA	1986-12-31	1996-12-31	527	4.238	103.056	1.64×10^{-02}	1.47×10^{-02}	6.68×10^{-04}	1.88×10^{-03}
5–Day MA 0.001 Band	1986-12-31	1996-12-31	574	4.238	78.672	1.64×10^{-02}	1.45×10^{-02}	6.68×10^{-04}	1.77×10^{-03}
MEXBOL									
5–Day MA	1991-11-08	1996-12-31	265	1.369	10.550	1.70×10^{-02}	1.60×10^{-02}	6.78×10^{-04}	1.92×10^{-03}
5–Day MA 0.001 Band	1991-11-08	1996-12-31	289	1.369	10.194	1.70×10^{-02}	1.59×10^{-02}	6.78×10^{-04}	1.90×10^{-03}
NKY									
5–Day MA	1984-01-04	1996-12-30	769	0.950	7.160	1.29×10^{-02}	1.22×10^{-02}	2.09×10^{-04}	6.56×10^{-04}
5–Day MA 0.001 Band	1984-01-04	1996-12-30	832	0.950	5.429	1.29×10^{-02}	1.21×10^{-02}	2.09×10^{-04}	5.81×10^{-04}
2–Day MA 0.001 Band	1984-01-04	1986-12-26	371	0.896	1.026	8.25×10^{-03}	7.67×10^{-03}	8.73×10^{-04}	9.63×10^{-04}
2–Day OBV	1984-01-04	1986-12-26	0	0.896	0.188	8.25×10^{-03}	1.52×10^{-04}	8.73×10^{-04}	2.35×10^{-04}

Table 5.14: Details of “Other Market” simulations

5.10.2 Summary

	Other Market (Index)				
Data Snooping Algorithm	SPTSX	CCMP	HSI	MEXBOL	NKY
2–Day MA Band	✓	✓	✓	N/A	✓
2–Day OBV	✓	**x**	✓	N/A	**x**
5–Day MA Band	✓	**x**	✓	✓	✓
5–Day MA No Band	✓	✓	✓	✓	✓

Table 5.15: Most but not all of the other markets and algorithms did well. In this table ✓ indicates the simulation did better than the index and **x** indicates the simulation did worse than the index.

The results of Other Markets are detailed in table 5.14 summarised in table 5.15 and figures 5.8 to 5.15.

The two and five day moving average rules all do better than the index.

The 2–Day OBV on the NKY, CCMP (Nasdaq) does worse than the index. It generated no signals at all for the NKY.

The 2–Day OBV does better than the index only on SPTSX.

The 2–Day MA rule for NKY does do better than the index but not by much, and has a break even transaction cost that is 0% to two decimal places.

The 5–Day MA algorithm, with and without a band look good when plotted (figure 5.15), especially after 1996. But like all the algorithms used on the NKY the break even transaction costs are very close to 0.

All four of the algorithms on the SPTSX look good when plotted (figures 5.8 and 5.9). The mean daily returns are all an order of magnitude better than the index. The break even transaction cost of the 2–Day OBV (1.1%) is the only break even transaction cost we found that is better than the 50 basis points that we have set as our cut off for feasibility. The 2–Day MA break even transaction cost, at 0.4% below our cutoff but is better than the break even transaction costs for the 5–Day MA algorithms, both of which are 0.27%.

On the Nasdaq (CCMP) the two and five day moving average algorithms all look good when plotted (figures 5.10 and 5.11). The 2–Day OBV improved its performance after 1985, but till then did not generate many, if any, signals. The break even transaction cost for the 2–Day moving average algorithm (on CCMP) at 0.48% came close to our feasibility limit, but did not quite make it.

For both the Hang Seng and MEXBOL indexes we could only test the 5–Day moving average algorithm because the indexes do not over lap the periods used for the 2–Day OBV or the 2–Day MA algorithms. When plotted the algorithms look to perform well (figures 5.12 and 5.13). The mean daily returns are an order of magnitude better than the index. But the break even transaction costs are all below our 50 basis point feasibility cutoff.

The 2–Day MA with band and the 2–Day OBV on the NASDAQ (CCMP) (figures 5.10 and 5.14) perform very differently from the performance on the DJIA where they both did very well. This is on a market that is highly correlated with the DJIA. We see the same difference between the 2–Day OBV and Moving Average in the NKY index as in the CCMP. This in the index with the least correlation to the DJIA. Given that 2–Day MA and 2–Day OBV both did well on the same time period on the DJIA this is evidence of data snooping.

5.11 Holdout Sample

The data for the Sullivan et al. (1999) study finish in 1996. This leaves us with 14 years of data (1997–2010) after the end of their study to use as a holdout sample. Some of the rules tested in Sullivan et al. (1999) were only claimed to be effective for sections of the period covered, so it would be unreasonable to judge those rules on periods that no effectiveness is claimed for.

But the two 5–Day Moving Average rules were found to be effective over the complete period. So this rule should be effective for the 14 years following the century where that rule was found to be amongst the most effective.

In table 5.16 and figure 5.17 we can see that in the 15 years following the study the 5–Day Moving Average rule did not continue to reliably outperform the index. In fact it had a negative return over the period, and only marginally better volatility.

We tested the 5–Day Moving Average with 0.001 band on data from 1997 to 2010. Similarly as we can see in table 5.16 and figure 5.17, this rule also stopped performing well at the same time the Sullivan et al. (1999) study ended. This is evidence of data snooping(or of the market having learned).

	Index	Simulated Trading
Number of Trades	N/A	981
Total Return	0.73	-0.67
(Geometric) Mean Daily Return	1.571×10^{-04}	-3.158×10^{-04}
Standard Deviation	1.27%	1.21%
3,479 trading days		

Table 5.16: Holdout sample for 5–Day Moving Average and 0.001 Band rule for 1997 – October 2010

	Index	Simulated Trading
Number of Trades	N/A	908
Total Return	0.73	-0.67
(Geometric) Mean Daily Return	1.571×10^{-04}	-3.186×10^{-04}
Standard Deviation	1.27%	1.22%
3,479 trading days		

Table 5.17: Holdout sample for 5–Day Moving Average rule for 1997 – October 2010

Neither rule performs in the holdout sample (1997–2010) as well as it it did in the main sample (1928–1996).

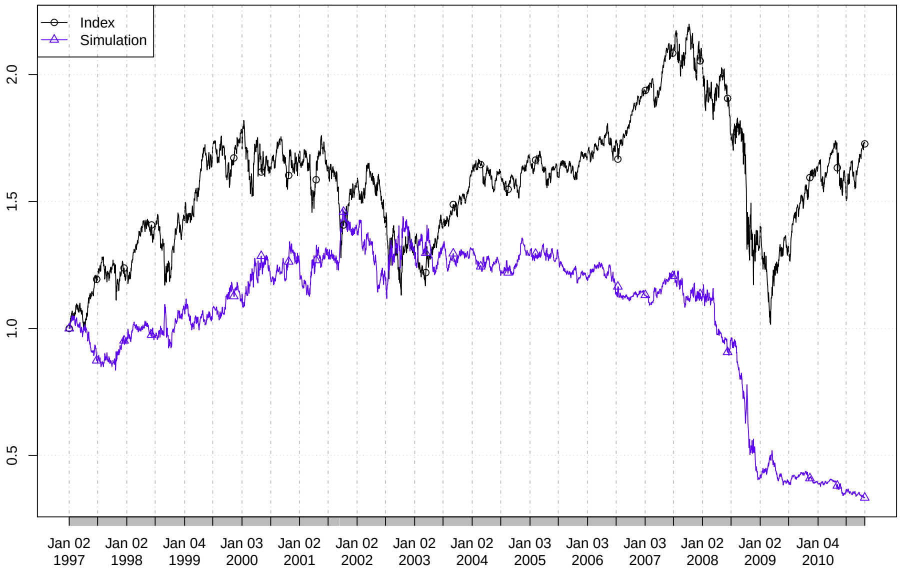


Figure 5.16: Holdout sample for 5–Day Moving Average and 0.001 Band rule for 1997 – October 2010

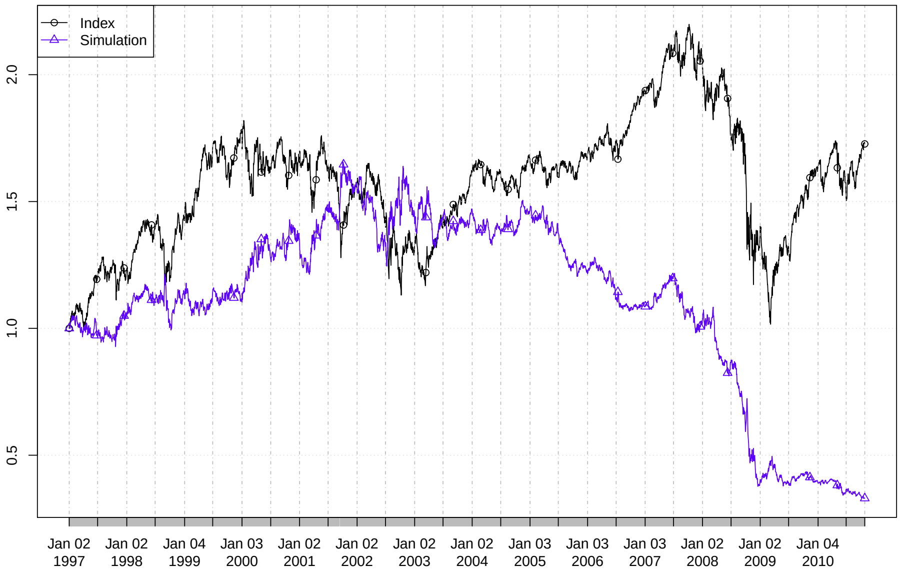


Figure 5.17: Holdout sample for 5–Day Moving Average, no band, rule for 1997 – October 2010

5.12 Conclusions on Sullivan et al. (1999)

Data Snooping Test	No Evidence of Data Snooping	Inconclusive	Evidence of Data Snooping
Statistical Significance	✓		
Holdouts			✓
Neighbourhood Analysis		✓	
Other Markets			✓
Feasibility			✓

Table 5.18: The results of our data snooping tests on the data from Sullivan et al. (1999).

Our results for detecting data snooping in Sullivan et al. (1999) are summarised in table 5.18. Sullivan et al. (1999) used an algorithm to detect data snooping. Given that we found evidence of data snooping in their results we cannot find that it is effective.

Also we cannot take the results seriously given the lack of consideration of transaction costs in Sullivan et al. (1999), and the failure of the best algorithms therein to perform with a simple and conservative allowance for transaction costs.

Recall that the objection to the effectiveness of technical trading rules, based on the EMH, depends crucially on feasibility. Given the results in section 5.7 it is clear that the rules are not useful as trading signals, unless transaction costs can be brought down to very low levels, less than 50 basis points per trade. Given the periods that these rules were found to be "effective" we think it is unlikely that such low transaction costs could have been available. If we can classify ignoring transaction costs as a type of data snooping, we can definitely say Sullivan et al. (1999) were data snooping.

A question remains: Have the Sullivan et al. (1999) authors uncovered some underlying properties of the index? Given the fact that the rules only apply to carefully defined periods (for the 5–Day moving averages the affect disappeared the year after the study) it is unlikely. They appear to have described the index rather than the underlying properties. I.e., they were data snooping. It is a possibility that after the publication of Sullivan et al. (1999), traders discovered these anomalies in the market and they were arbitraged away. This assumes that traders could lower their transaction costs or improve on the rules.

The most disappointing aspect of this study was that we could not reproduce Sullivan et al. (1999)'s results using the information they provided. This did mean that there were properties observed by the authors of Sullivan et al. (1999) that we could not observe.

Our next disappointment was the lack of investigation of transaction costs. Whilst it is very difficult to accurately model transaction costs we were able to use a very simple procedure that placed boundaries on the transaction costs. Those boundaries ruled out any of the rules we could simulate from being feasible.

The superior performance of the moving average rules, albeit infeasible for traders, was interesting. It would be interesting to investigate why they were so effective, and directly compare different moving average models of the markets. Is there some spurious thin–trading induced serial correlation in index returns from 1897 through 1996? That is beyond our scope.

Our analysis of Sullivan et al. (1999) using holdout samples and transaction costs is not original. Bajgrowicz and Scaillet (2009) makes similar observations to us regarding the (in)feasibility of the rules in Sullivan et al. (1999) due to transaction costs and the failure of even the most successful rules when studied after the time period covered in Sullivan et al. (1999).

5.12.1 Bootstrapping and Other False Discoveries

Sullivan et al. (1999) introduced an algorithm tailored to detect data snooping. It draws on the "Stationary Bootstrap" developed in Politis and Romano (1994) to considerably expand the amount of data available to test the technical trading rules on.

Sullivan et al. (1999) was a follow on paper to Brock et al. (1992). The earlier paper examined many of the same rules on on the Dow Jones Industrial Averages index as Sullivan et al. (1999) did. The innovation of Sullivan et al. (1999) is to use the bootstrapping algorithm (fully documented in the companion paper White (2000) and termed the "Reality Check") to detect data snooping. But we see 3 major problems with this innovation.

Firstly; in Brock et al. (1992) the best algorithm found stopped working straight after the end of the study. Thus it failed to work out of sample which we argue is a sign of data snooping. We see the same lack of performance in hold out samples in Sullivan et al.

(1999). But the claim is Sullivan et al. (1999) is that the "Reality Check" has prevented data snooping. To us this is empirical evidence that the "Reality Check" does not work.

The next two problems have to do with how the bootstrapped samples are constructed. The method is that outlined in Politis and Romano (1994). It involves taking randomly selected sub-sequences of random length from the original series and concatenating them to create a new series that has the same statistical properties as the original series but is effectively new data. This is used hundreds of times to repeatedly test the technical trading algorithms.

There is a problem in that the stationary bootstrap of Politis and Romano (1994) is designed for use on strictly stationary data. It is trivial to see why this must be so. If we are concatenating randomly selected sequences to create a sequence with the same properties then if the original series was not stationary the bootstrapped series will have inconsistencies around the joins of the sub sequences. But the returns of the DJIA are not stationary. At the very least there are periods of low volatility and periods of high volatility. The bootstrapping algorithm is applied to the outputs of the rules, and we can see no reason to accept that the output of the algorithms is stationary when they applied to non–stationary data.

Fundamentally there is another, more subtle, problem with this approach. Financial data (such as a time series of index returns) are not drawn from a distribution. It is random in an algorithmic sense[16] but it is not random in the sense that it is sampled from a distribution. Financial data are the product of a process. The process is a self–aware one. It is the actions of all traders in the market, who are all working together, in communication with each other and optimising their behaviour. Thus all the data are idiosyncratic. A bootstrapped sequence is not the output of the same process, or anything like it. Even if we cannot tell the difference by examining it.

To put it another way. If the idiosyncratic part of financial returns are due to a random process then there would be no information in historical prices that could be used to help predict future prices. The EMH would be correct and technical analysis rules would be ineffective. In that case (ignoring problems with non stationarity associated

[16] See section 2.3.7 page 12

with changes in macro–economic variables and conditions) bootstrapping could make sense. A return series from one time segment would be much like another. If however the idiosyncratic part of financial returns are due to some (possibly, probably, unknown) process then: the EMH is false, it is possible that technical analysis rules could be effective and bootstrapping cannot work as it would erase the structure imposed by the process.

Chapter 6

Conclusion

Detecting data snooping is hard.

We have investigated six techniques for detecting if a particular result is due to data snooping, seven if ignoring feasibility is counted as data snooping. None of the methods we addressed can on their own give a definitive answer that a result is not due to data snooping, but together they can be helpful.

In section 4 we studied Faber's algorithm that, anti–hoc, we felt sure was due to data snooping. But it turned out to be extremely unclear. In fact if the algorithm in Faber (2009) was presented using data snooping, (a better market and a better moving average, see section 4.6) it would have performed better. In section 5 we dealt with a set of results that we decided was likely due to data snooping. Yet it was presented in the context of a study that explicitly ruled out data snooping.

Of all the techniques we used we feel the most satisfactory was using holdout data. But even if such data are available the huge multitude of technical analysis rules that can be conceived weakens the effectiveness of using holdout data.

We did not get a chance to evaluate "story telling". None of the examples we tested were presented in the context of a story. But we feel it is appropriate to have an economic basis for technical analysis rules (that is a story). But it is not a scientific test, so it is hard to evaluate.

The least useful of the tools we looked at was statistical significance. This is probably due to the fact that the process of data snooping is, often, to find significance, even where there is none. It is concerning that p–values from simple tests continue to be used when the assumptions underlying those tests (independence and constant distribution) do not apply to financial markets in the general case (Lo and MacKinlay, 2001).

The examples we have used all use daily data. Given that we can define technical analysis rules almost without limit, and there are 2,500 trading days per decade, there will always be serious problems with data snooping if rules are optimised by back testing on such (daily) data. But if a technical analysis rule was successful in back testing up to a day ago, and fails tomorrow, is that due to data snooping in rule discovery or is it due to changes in the nature of the market (e.g., traders learning)?

It is clear that the economy generally and the financial system in particular is prone to periodicity. Periods of contraction, expansion and stagnation . Whilst defining the temporal borders of these periods is problematic, there exist different market modes and structural breaks.

So we must conclude that using such data to test technical analysis rules data snooping bias is a persistent risk and it is unlikely that there is any effective single solution to the problem. The best that we can do is be aware of the risk of data snooping and to report how we have dealt with the risk as part of our analysis.

Bibliography

D Aldous. Probability Approximations via the Poisson Clumping Heuristic. Springer, New York., 1989.

J. Scott Armstrong. Significance tests harm progress in forecasting. International Journal of Forecasting, 23(2):321 – 327, 2007. ISSN 0169-2070. doi: DOI:10.1016/j.ijforecast.2007.03.004. URL http://www.sciencedirect.com/science/article/B6V92-4NMKVBF-1/2/e7883972840b8ed65c7b59bb5ed7b3c1.

Thomas M. Arnold and Timothy F. Crack. A Practical Guide to GMM (with Applications to Option Pricing). SSRN eLibrary, 1999. doi: 10.2139/ssrn.268828.

George S. Atsalakis and Kimon P. Valavanis. Forecasting stock market short-term trends using a neuro-fuzzy based methodology. Expert Systems with Applications, 36(7): 10696 – 10707, 2009. ISSN 0957-4174. doi: DOI:10.1016/j.eswa.2009.02.043. URL http://www.sciencedirect.com/science/article/pii/S0957417409001948.

Pierre Bajgrowicz and Olivier Scaillet. Technical Trading Revisited: False Discoveries, Persistence Tests, and Transaction Costs. SSRN eLibrary, 2009.

Jedrzej P. Bialkowski, Ahmad Etebari, and Tomasz P. Wisniewski. Piety and Profits: Stock Market Anomaly During the Muslim Holy Month. SSRN eLibrary, 2009.

Zvi Bodie, Alex Kane, and Alan J. Marcus. Investments. McGraw Hill, 2005.

John C. Bogle. The first index mutual fund: A history of vanguard index trust and the vanguard index strategy. http://www.vanguard.com/bogle_site/lib/sp19970401.html, 2006.

Tim Bollerslev, Y Chou Ray, and Kenneth F Kroner. Arch modeling in finance : A review of the theory and empirical evidence. Journal of Econometrics., 52:5–59, 1992.

C. Alan Boneau. The effects of violations of assumptions underlying the t test. Psychological Bulletin, 57(1), 1960.

Jamers V. Bradley. Robustness? British Journal of Mathematical and Statistical Psychology, 31, 1978.

W. Brock, J. Lakonishok, and B. LeBaron. Simple technical trading rules and the stochastic properties of stock returns. The Journal of Finance, 47(5):1731–1764, 1992.

Philip Brown and Jason Mitchell. Culture and stock price clustering: Evidence from the peoples' republic of china. Pacific-Basin Finance Journal, 16 (1-2):95 – 120, 2008. ISSN 0927-538X. doi: DOI:10.1016/j.pacfin.2007.04.005. URL `http://www.sciencedirect.com/science/article/B6VFF-4NN0W57-5/2/403be1f39a5a84668482606d39dfae82`. Behavioral Finance in Asia.

G. Caginalp and D. Balenovich. Trend-based asset flow in technical analysis and securities marketing. Psychology and Marketing, 13(4):407–444, 1996. ISSN 1520-6793. doi: 10.1002/(SICI)1520-6793(199607)13:4⟨405::AID-MAR5⟩3.0.CO;2-E. URL `http://dx.doi.org/10.1002/(SICI)1520-6793(199607)13:4<405::AID-MAR5>3.0.CO;2-E`.

Cristian Calude. Information and Randomness. Springer–Verlag, 1994.

Kalok Chan, Allaudeen Hameed, and Wilson Tong. Profitability of momentum stragegies in the international equity markets. Journal of Financial and Quantitative Analysis, 35(02):153–172, 2000. doi: 10.2307/2676188. URL `http://dx.doi.org/10.1017/S0022109000009133`.

T. Chavarnakul and D. Enke. Intelligent technical analysis based equivolume charting for stock trading using neural networks. Expert Systems with Applications, 34(2): 1004–1017, 2008. ISSN 0957-4174.

Timothy F. Crack. A classic case of data snooping for classroom discussion. The Journal of Financial Education, 25:92 – 97, 1999.

Ilia D. Dichev and Troy D. Janes. Lunar cycle effects in stock returns. Journal of Private Equity, 6:8–22, 2003.

DrSuess. One Fish, Two Fish, Red Fish, Blue Fish. Harper Collins, 1960. "...for these things are fun, and fun is good!". The best advice I have ever had from a doctor.

Mebane T. Faber. A Quantitative Approach to Tactical Asset Allocation. Journal of Wealth Management, Spring 2007, 2009. URL `http://ssrn.com/abstract=962461`.

Eugene Fama and Marshall Blume. Filter rules and stock-market trading. Journal of Business, 39:226–241, 1966.

Michael F. Ferguson and Richard L. Shockley. Equilibrium "anomalies". The Journal of Finance, 58(6):pp. 2549–2580, 2003. ISSN 00221082. URL `http://www.jstor.org/stable/3648203`.

Thomas Gehrig and Caroline Fohlin. Trading costs in early securities markets: The case of the berlin stock exchange 18801910. Review of Finance, 10:587–612, 2006. doi: 10.1007/s10679-006-9010-y.

Peter Reinhard Hansen. A test for superior predictive ability. Journal of Business and Economic Statistics, 23(4):365–380, 2005. doi: 10.1198/073500105000000063. URL `http://pubs.amstat.org/doi/abs/10.1198/073500105000000063`.

Boyd Holcomb. Market timing: You can't trade that way! `http://www.turtletrader.com/market-timing.html`, 2011.

Po-Hsuan Hsu and Chung-Ming Kuan. Reexamining the profitability of technical analysis with data snooping checks. Journal of Financial Econometrics, 3(4):606–628, 2005. doi: 10.1093/jjfinec/nbi026. URL `http://jfec.oxfordjournals.org/content/3/4/606.abstract`.

Ross Ihaka and Robert Gentleman. The r project for statistical computing. `http://www.r-project.org`, 12 2010.

John P. A. Ioannidis. Why most published research findings are false. PLoS Med, 2 (8):e124, 08 2005. doi: 10.1371/journal.pmed.0020124. URL http://dx.doi.org/10.1371%2Fjournal.pmed.0020124.

A. Johanson and D. Sornette. Stock market crashes are outliers. Eur. Phys. J. B, 1: 141–143, 1998.

Pei Kuang, Michael Schröder, and Qingwei Wang. Illusory Profitability of Technical Analysis in Emerging Foreign Exchange Markets. SSRN eLibrary, 2010.

Markus Leippold and Harald Lohre. Data snooping and the global accrual anomaly. Applied Financial Economics, 22(7):509–535, 2012. doi: 10.1080/09603107.2011.631892. URL http://www.tandfonline.com/doi/abs/10.1080/09603107.2011.631892.

David A. Lesmond, Joseph P. Ogden, and Charles Trzcinka. A new estimate of transaction costs. The Review of Financial Studies, 12(5):1113–1141, 1999. doi: 10.1093/rfs/12.5.1113.

Andrew W. Lo and A. Craig MacKinlay. Data–snooping biases in tests of financial asset pricing models. The Review of Financial Studies, 3(3):431–467, 1990.

Andrew W. Lo and A. Craig MacKinlay. A Non-Random Walk Down Wall Street. Princeton University Press, 2001.

Andrew W. Lo, Harry Mamaysky, and Jiang Wang. Foundations of technical analysis: Computational algorithms, statistical inference, and... Journal of Finance, 55(4):1705 – 1765, 2000. ISSN 00221082. URL http://search.ebscohost.com/login.aspx?direct=true\&db=bth\&AN=3435701\&site=ehost-live\&scope=site.

Malik Magdon-Ismail, Amir F. Atiya, Amrit Pratap, and Yaser S. Abu-Mostafa. On the maximum drawdown of a brownian motion. Journal of Applied Probability, 41(1): 147–161, 2004.

Burton G. Malkiel. A Random Walk Down Wall Street. W. W. Norton and Comapny, 1999.

Donald N. McCloskey. The loss function has been mislaid: The rhetoric of significance tests. American Economic Review, 75(2):201–206, 1985.

United States Naval Meteorology and Oceanography Command. http://www.usno.navy.mil/USNO/astronomical-applications/data-services/phases-moon, 2010.

Randall Munroe. http://imgs.xkcd.com/comics/significant.png, 2011.

Dimitris N. Politis and Joseph P. Romano. The stationary bootstrap. Journal of the American Statistical Association, 89(428):pp. 1303–1313, 1994. ISSN 01621459. URL http://www.jstor.org/stable/2290993.

Roger Porkess. Dictionary of Statistics. Collins, 2004.

Allen Poteshman and Vitaly Serbin. Clearly irrational financial market behavior: Evidence from the early exercise of exchange traded stock options. The Journal of Finance, 58(1):37–70, 2003.

S. James Press. A compound events model for security prices. Journal of Business, 40(3):317 – 335, 1967. ISSN 00219398. URL http://search.ebscohost.com.ezproxy.otago.ac.nz/login.aspx?direct=true&db=bth&AN=4585653&site=ehost-live&scope=site.

Min Qi and Yangru Wu. Technical trading-rule profitability, data snooping, and reality check: Evidence from the foreign exchange market. Journal of Money, Credit and Banking, 38:2135 – 2158, 2005.

Michael S. Rashes. Massively confused investors making conspicuously ignorant choices (MCI-MCIC). The Journal of Finance, 56(5):1911–1927, 2001. ISSN 1540-6261. doi: 10.1111/0022-1082.00394. URL http://dx.doi.org/10.1111/0022-1082.00394.

Federal Reserve. Historical data. http://www.federalreserve.gov/releases/h15/update/, 08 2010.

J.C. Rodriguez. Measuring financial contagion: A copula approach. Journal of Empirical Finance, 14(3):401–423, 2007. ISSN 0927-5398.

Joseph P. Romano and Michael Wolf. Stepwise multiple testing as formalized data snooping. Econometrica, 73(4):1237–1282, 2005. ISSN 00129682. URL `http://www.jstor.org/stable/3598821`.

David J Sheskin. Parametric and Nonparametric Statistical Procedures. Chapman and Hall/CRC, 3 edition, 2004.

S&P 500. `http://www.standardandpoors.com/indices/sp-500/en/us/?indexId=spusa-500-usduf--p-us-l--`, 2012.

Ryan Sullivan, Allan Timmermann, and Halbert White. Data-snooping, technical trading rule performance, and the bootstrap. The Journal of Finance, 54(5):1647–1691, 1999. ISSN 00221082. URL `http://www.jstor.org/stable/222500`.

R. Tajaddini and T.F. Crack. Do momentum-based trading strategies work in emerging currency markets? Journal of International Financial Markets, Institutions and Money, 22(3):521–537, 2012.

Nassim N. Taleb. The Black Swan: The Impact of the Highly Improbable. Random House, 1 edition, April 2007. ISBN 1400063515. URL `http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20\&path=ASIN/1400063515`.

Sam Trethewey and Timothy Falcon Crack. Price momentum in the new zealand stock market: a proper accounting for transactions costs and risk. Accounting & Finance, 50(4):941–965, December 2010.

Roger Waters. Brain damage. Long Player Record, March 1973.

Halbert White. A reality check for data snooping. Econometrica, 68(5):1097–1126, 2000. ISSN 00129682. URL `http://www.jstor.org/stable/2999444`.

Jack W. Wilson and Charles P. Jones. An analysis of the s&p 500 index and cowless extensions: Price indexes and stock returns, 1870–1999. The Journal of Business, 75 (3):505–533, 2002.

Diethelm Wuertz and Rmetrics. fbasics: Rmetrics - markets and basic statistics. `http://cran.r-project.org/web/packages/fBasics/index.html`, 02 2010.

Kathy Yuan, Lu Zheng, and Qiaoqiao Zhu. Are investors moonstruck? lunar phases and stock returns. Journal of Empirical Finance, 13(1):1 – 23, 2006. ISSN 0927-5398. doi: DOI:10.1016/j.jempfin.2005.06.001. URL http://www.sciencedirect.com/science/article/B6VFG-4J08VB7-3/2/668d2fab23ad1684570294ff6f17fc3b.

Appendix A

Statistical Test Assumptions

Single Sample t–test Sheskin 2004, pp 135

Hypothesis evaluated. Does the sample come from a population in which the mean equals a specific value?

Assumptions:

- Sample is randomly selected, no sample's value depends on the value of another.
- Underlying distribution is normal
- True variance is unknown[1]

t–test for two independent samples Sheskin 2004, pp 375

Hypothesis evaluated. Do two independent samples represent two populations with different mean values?

Assumptions:

- Sample is randomly selected
- Underlying distribution is normal
- True variance is unknown[1]
- The two samples have the same variance

[1] If the variance is known the t–test is not invalid but the z–test is better.

t–test for two dependant samples (i.e., paired two–sample test) Sheskin 2004, pp 575

Hypothesis evaluated. Do two dependant samples represent populations with different means?

Assumptions:

- Sample is randomly selected
- Underlying distribution is normal
- True variance is unknown[1]
- The two samples have the same variance

Wilcoxon Signed Ranks test Sheskin 2004, pp 189

Hypothesis evaluated. Does sample come from a population in which the median equals a specified value?

Assumptions[2]

- Sample is randomly selected
- interval/ratio data
- The underlying distribution is symmetrical

Mann–Whitney U test Sheskin 2004, pp 423

Hypothesis evaluated. Do the independent samples represent two populations with different median values?

Assumptions[2]

- Sample is randomly selected
- Samples independent of each other
- The underlying distributions are identical in shape[3].

[2] According to Sheskin there is an additional assumption that is not adhered to and that is that the underlying distribution is continuous.

[3] Sheshkin argues that this assumption implies homogeneous variance .

Wilcoxon matched pairs Signed ranks test Sheskin 2004, pp 609

Hypothesis evaluated. Do two dependant samples represent two different populations?

Assumptions[2]

- Sample is randomly selected
- The original data are interval/ratio data.
- The underlying distributions are identical in shape[3].

Appendix B

Code to Repeat the Results from Crack (1999)

"evs.dat" is a data file of returns. EWRETD, VWRETD and S&P500 respectively. "moon.prn" is a record of lunar phases. The data were downloaded from the XXXXX website. In the 10 years since this was done there has been a small change to the data on one day. The results here are from the original data on the original paper.

A: Get that webs
Quantify the diffe

The first few lines of "evs.dat":

```
19500104 0.015487 0.0108 0.011340020059674
19500105 0.00832 0.004112 0.0047365393466734
19500106 0.006124 0.002286 0.00294898473815719
19500109 0.006944 0.001546 0.0058720074772225
19500110 -0.001199 -0.003662 -0.0029316936858268
```

The first few lines of "moon.prn":

```
19650101 0.01
19650102 0.00
19650103 0.01
19650104 0.03
19650105 0.07
19650106 0.13
```

The code...

```
rm(list=ls())
## Fit a model to the data as in Crack's paper
library(xts)
library(timeSeries)
library(Hmisc)
## Load the data, split into New/Full and Interphase, calculate mean
## and SD for both

## Location of CRSP data

## Data file name
dfn <- "evs.dat"

get.moon.data <- function(){
  ## Load moon data from the disk.  Return a xts object of all the
  ## data.  As of 2010-07-02 read the data from the Crack's original
  moon <- read.table("moon.prn")
  moon.xts <- xts(moon[,2],
                  order.by=as.Date(as.character(moon[,1]), format="%Y%m%d"))
  return(moon.xts)
}

get.returns <- function(date.range, which="Equally", method="compound"){

  ## As of 2010-07-02 the data is all in one file, of returns.  EWRETD
  ## and VWRETD are from CRSP,  S&P500 is from Yahoo

  ret <- read.table(dfn)

  ## The first column is the date YYYYMMDD, the second is EWRETD, next
  ## VWRETD then S&P500 in the 4th column
  c <- 0
  if(which=="Equally"){
    c <- 2
  }else if(which == "Value"){
    c <- 3
  }else if(which == "S&P"){
    c <- 4
  }else{
    stop(which, "unknown")
  }
  ret.xts <- xts(ret[,c],
                 order.by=as.Date(as.character(ret[,1]), format="%Y%m%d"))
  ## Filter to the date range we want
  ret.xts <- ret.xts[date.range,]
  names(ret.xts) <- c("returns")
  return(ret.xts)
}

lunar.returns <- function(ret.xts, moon.xts, alpha){
  ## Using the alpha separate returns into those near new and full
  ## moons (NF) and the "interphase" (IP) returns.  Return them in two
  ## xts objects in a list

  ## Only care about days we trade, so filter out all others
  moon.xts <- moon.xts[index(ret.xts),]

  ## Get the IP dates. 'moon.IP.xts' will be TRUE in Interphase period
  ## and 'mmon.NF.xts will be true in the other periods
  moon.IP.xts <- alpha<=moon.xts & moon.xts<(1-alpha)
  moon.NF.xts <- moon.xts<alpha | (1-alpha) <=moon.xts

  ## Get the trading dates for IP and NF periods.  This is very
  ## R-centric syntax but also very cute!  'moon.IP.idx' will hold
  ## just the dates where there was trading and it was an "interphase"
  ## period.  Similarly 'moon.NF.idx'.
  moon.IP.idx <- index(moon.IP.xts[moon.IP.xts])
  moon.NF.idx <- index(moon.NF.xts[moon.NF.xts])

  ## Get the returns in tne NF and IP periods and name them
  returns.IP <- ret.xts[moon.IP.idx]
  returns.NF <- ret.xts[moon.NF.idx]

  l.ret <- list(returns.IP, returns.NF)
  names(l.ret[[1]]) <- "returns.IP"
  names(l.ret[[2]]) <- "returns.NF"

  return(l.ret)
}

doit <- function(ret.xts, moon.xts, date.range, alpha, dy){

  ## Get the IP and NF returns

  ## From Crack's code.  This log stuff.  *FIXME* Why?
  Rt <- log(1+ret.xts)
  l.r <- lunar.returns(Rt, moon.xts, alpha)
  returns.IP <- l.r[[1]]
  returns.NF <- l.r[[2]]

  ## The mean and SD...
  rIP <- returns.IP
  rNF <- returns.NF

  mean.IP <- mean(rIP)
  SD.IP <- sd(rIP)
  mean.NF <- mean(rNF)
  SD.NF <- sd(rNF)
```

```
## Testing for statistical significance
x <- as.numeric(rIP)
y <- as.numeric(rNF)

## one--sided where the alternative is tht IP returns are greater
## than NF.
t.t.1 <- t.test(x, y, alternative="greater")
w.t.1 <- wilcox.test(x, y, alternative="greater") # mu = 0, paired =
                                                  # FALSE, exact =
                                                  # NULL, correct =
                                                  # TRUE, conf.int =
                                                  # FALSE,
                                                  # conf.level =
                                                  # 0.95,

## two--sided where the alternative is that the means of IP are not
## the same as the means of NF
t.t.2 <- t.test(x,y)
w.t.2 <- wilcox.test(x,y)

## GMM Test.  Copy Crack's code.  Use same variable names

NNF <- length(rNF)
NI <- length(rIP)
T <- NNF+NI
muNF <- mean.NF
muI <- mean.IP
## Creat deltaNF, 1 when in NF, 0 in IP.  Also deltaI vice versa
mni <- merge(rNF, rIP)
mni[!is.na(mni)] <- 1
mni[is.na(mni)] <- 0
deltaNF <- mni[,1]
deltaI <- mni[,2]

sigmaNF2 <- sum((rNF-mean.NF)^2)/NNF
sigmaI2 <- sum((rIP-mean.IP)^2)/NI

theta <- c(muNF, muI, sqrt(sigmaNF2), sqrt(sigmaI2))

Gamma <- matrix(ncol=4, nrow=4)
Gamma[1,] <- c(-NNF/T, 0, 0, 0)
Gamma[2,] <- c(0, -NI/T, 0, 0)
Gamma[3,] <- c(0, 0, -2*NNF*theta[3]/T, 0)
Gamma[4,] <- c(0, 0, 0, -2*NI*theta[4]/T)

f <- matrix(ncol=T, nrow=4)
f[1,] <- deltaNF*(Rt-theta[1])
f[2,] <- deltaI*(Rt-theta[2])
f[3,] <- deltaNF*((Rt-theta[1])^2 - theta[3]^2)
f[4,] <- deltaI*((Rt-theta[2])^2 - theta[4]^2)
```

```
  WHITE <- (f%*%t(f))*1/T
  NW <- matrix(0, ncol=ncol(WHITE), nrow=nrow(WHITE))
  m <- 200
  for(j in 1:m){
    phij <- (1/T) * f[,(j+1):T] %*% t(f[,(1:(T-j))])
    NW <- NW +  (1-(j/(m+1))) * (phij+t(phij))
  }

  NW <- NW+WHITE
  OMEGA <- NW
  VGMM <- solve(t(Gamma)%*%solve(OMEGA)%*%Gamma)
  SE <- sqrt(diag(VGMM)/T)

  ## H0 muNF=muI
  R <- t(c(1, -1, 0, 0))
  Rtheta <- R%*%theta
  VRtheta <- R%*%(VGMM/T)%*%t(R)
  statmu <- Rtheta/sqrt(VRtheta)

  ## H0: sigmaNF=sigmaI
  R <- t(c(0, 0, 1, -1))
  Rtheta=R%*%theta;
  VRtheta=R%*%(VGMM/T)%*%t(R)
  statsig=Rtheta/sqrt(VRtheta);

  ## Prepare results to return
  df.ret <- data.frame("alpha"=alpha,
                       "NNF"=NNF,
                       "NI"=NI,
                       "muNF"= muNF*dy,
                       "SE.NF"= SE[1]*dy,
                       "muI"= muI*dy,
                       "SE.I"= SE[2]*dy,
                       "statmu"=statmu,
                       "p.statmu"=pnorm(statmu),
                       "sigmaNF"=sqrt(sigmaNF2*dy),
                       "SE.S.NF"=SE[3]*sqrt(dy),
                       "sigmaI"=sqrt(sigmaI2*dy),
                       "SE.S.I"=SE[4]*sqrt(dy),
                       "statsig"=statsig,
                       "p.statsig"=1-pnorm(statsig)
                       )
  return(df.ret)
}

test.2 <- function(date.range){

  ## Reproduce Crack's results
```

```
alphas <- c(.2, .25, .3, .35, .4)
N <- length(alphas)
df.e <- data.frame("alpha" = vector(length=N, mode="numeric"),
                   "NNF" = vector(length=N, mode="numeric"),
                   "NI" = vector(length=N, mode="numeric"),
                   "muNF" = vector(length=N, mode="numeric"),
                   "SE.NF" = vector(length=N, mode="numeric"),
                   "muI" = vector(length=N, mode="numeric"),
                   "SE.I" = vector(length=N, mode="numeric"),
                   "statmu" = vector(length=N, mode="numeric"),
                   "p.statmu" = vector(length=N, mode="numeric"),
                   "sigmaNF" = vector(length=N, mode="numeric"),
                   "SE.S.NF" = vector(length=N, mode="numeric"),
                   "sigmaI" = vector(length=N, mode="numeric"),
                   "SE.S.I" = vector(length=N, mode="numeric"),
                   "statsig" = vector(length=N, mode="numeric"),
                   "p.statsig" = vector(length=N, mode="numeric"))
df.v <- data.frame("alpha" = vector(length=N, mode="numeric"),
                   "NNF" = vector(length=N, mode="numeric"),
                   "NI" = vector(length=N, mode="numeric"),
                   "muNF" = vector(length=N, mode="numeric"),
                   "SE.NF" = vector(length=N, mode="numeric"),
                   "muI" = vector(length=N, mode="numeric"),
                   "SE.I" = vector(length=N, mode="numeric"),
                   "statmu" = vector(length=N, mode="numeric"),
                   "p.statmu" = vector(length=N, mode="numeric"),
                   "sigmaNF" = vector(length=N, mode="numeric"),
                   "SE.S.NF" = vector(length=N, mode="numeric"),
                   "sigmaI" = vector(length=N, mode="numeric"),
                   "SE.S.I" = vector(length=N, mode="numeric"),
                   "statsig" = vector(length=N, mode="numeric"),
                   "p.statsig" = vector(length=N, mode="numeric"))

ret.v.xts <- get.returns(date.range, which="Value", method="compound")
ret.e.xts <- get.returns(date.range, which="Equally", method="compound")
moon.xts <- get.moon.data()
for(i in 1:N){
  alpha <- alphas[i]
  v <- doit(ret.v.xts, moon.xts, date.range, alpha, 253)
  e <- doit(ret.e.xts, moon.xts, date.range, alpha, 253)
  for(n in names(v)){
    df.v[[n]][i] <- v[[n]]
    df.e[[n]][i] <- e[[n]]
  }
}

## Construct the file name for the LaTeX output
fn.date <- paste(substring(date.range, 1, 4),
                 substring(date.range, 6, 7),
                 substring(date.range, 9, 10),
                 substring(date.range, 13, 16),
                 substring(date.range, 18, 19),
                 substring(date.range, 21, 22),
                 sep="")
fn.e <- paste("Fragments/Lunar.", fn.date,".e.tex", sep="")
fn.v <- paste("Fragments/Lunar.", fn.date,".v.tex", sep="")

## Label for caption
cap.date.1 <- substr(date.range, 1, 10)
cap.date.2 <- substr(date.range, 13, 23)

## Display frame. 11 columns.  N * 2 rows.  One for data the next
## for T-stats
disp.e <- matrix("", ncol=11, nrow=(N*2))
disp.v <- matrix("", ncol=11, nrow=(N*2))

d.names <- c("alpha","NNF","NI","muNF","muI","statmu","p.statmu","sigmaNF","sigmaI"
d.stats <- c("","","","SE.NF","SE.I","","","SE.S.NF","SE.S.I")
stat.col <- c(4,5,8,9)

## Amount to round data
nr <- 4

for(i in 1:N){
  dr <- (i-1)*2+1
  ds <- dr+1
  for(j in 1:length(d.names)){
    n <- d.names[j]
    val.e <- round(df.e[[n]][i], nr)
    val.v <- round(df.v[[n]][i], nr)
    disp.e[dr,j] <- paste("$",val.e,"$",sep="")
    disp.v[dr,j] <- paste("$",val.v,"$",sep="")
  }
  for(n in stat.col){
    val.e <- round(df.e[[d.stats[n]]][i], nr)
    val.v <- round(df.v[[d.stats[n]]][i], nr)
    disp.e[ds,n] <- paste("$(",val.e,")$",sep="")
    disp.v[ds,n] <- paste("$(",val.v,")$",sep="")
  }
}
l <- latex(disp.e,
           file=fn.e,
           title=’’,

           colheads=c(
             ’$\\alpha$’,
             ’$N_{NF}$’,
             ’$N_{I}$’,

             ’$\\mu_{NF}\\times253$’,
               '$\\mu_{I}\\times253$',
               'Z--stat',
               '($p$ val.)',

               '$\\sigma_{NF}\\times\\sqrt{253}$',
               '$\\sigma_{I}\\times\\sqrt{253}$',
               'Z--stat',
               '($p$ val.)'),

             caption=paste("Results for the CRSP EWRETD (equally weighted)",
               "index.  From ",cap.date.1,"to",cap.date.2),

             caption.loc="bottom",
             align=c("|ccc|cccc|cccc|"),
             ##landscape=TRUE,
             rowname=NULL,
             where="p",
             label=paste("tab:.e.",fn.date,sep=""),
             size="footnotesize"
             )

  l <- latex(disp.v,
             file=fn.v,
             title='',
             colheads=c(
               '$\\alpha$',
               '$N_{NF}$',
               '$N_{I}$',
               '$\\mu_{NF}\\times253$',
               '$\\mu_{I}\\times253$',
               'Z--stat',
               '($p$ val.)',
               '$\\sigma_{NF}\\times\\sqrt{253}$',
               '$\\sigma_{I}\\times\\sqrt{253}$',
               'F--stat', '($p$ val.)'),
             align=c("|ccc|cccc|cccc|"),
             caption=paste("Results for the CRSP VWRETD (value weighted)",
               "index.  From ",cap.date.1,"to",cap.date.2),
             caption.loc="bottom",
             ##landscape=TRUE,
             ##append=TRUE,
             where="p",
             label=paste("tab:.v.",fn.date,sep=""),
             size="footnotesize",
             rowname=NULL)
}

all.dates.test <- function(){
  date.range <- "1977-01-01::1996-12-31"
  test.2(date.range)
  date.range <- "1962-07-02::1976-12-31"
  test.2(date.range)
  date.range <- "1997-01-01::2006-12-29"
  test.2(date.range)
}
```

B.1 Original Matlab Code Crack (1999)

```
clear;

load moon.prn;  % the USNO data
load evs.dat;  % CRSP E, V, S indices
data=evs;

% load data.dat;
% data=data(:,1:8);
% date1=min(moon(:,1));ind1=min(find(data(:,1)>=date1));

date1=770101;ind1=min(find(data(:,1)>=date1));

date2=max(moon(:,1));
ind2=max(find(data(:,1)<=date2));
data=data(ind1:ind2,:);


tic
obs=[];
% k=0;
% for i=1:length(data)
% for j=1:length(moon)
% if(data(i,1)==moon(j,1))
% k=k+1;
% obs(k,:)=[data(i,:) moon(j,2)];
% end
% end
% end

k=1;
for i=1:length(data)
 obs(k,:)=[data(i,:) moon(find(data(i,1)==moon(:,1)),2)];
 k=k+1;
end


toc
T=length(obs);
```

```
A=size(obs);
K=A(1,2);
% plot(100*exp(cumsum(log(1+obs(2:T,3)))),diff(obs(:,4)))
plot(obs(2:T,2),sign(diff(obs(:,K))),’.’)
waxc=obs(find(sign(diff(obs(:,K)))>0&obs(2:T,K)<.50),2);
waxg=obs(find(sign(diff(obs(:,K)))>0&obs(2:T,K)>.50),2);
wang=obs(find(sign(diff(obs(:,K)))<0&obs(2:T,K)>.50),2);
wanc=obs(find(sign(diff(obs(:,K)))<0&obs(2:T,K)<.50),2);

% test diff of two means
clear
s=2;
k=1;
for alpha=[0.10 0.15 0.20 0.25 0.30 0.35 0.40]
 partial=obs(find((obs(:,K)>=alpha)&(obs(:,K)<(1-alpha))),s);
 newfull=obs(find((obs(:,K)<alpha)|(obs(:,K)>=(1-alpha))),s);
 n=newfull;
 p=partial;
 N(k,:)=[length(n) length(p)];
 Z1(k)=(mean(n)-mean(p))/sqrt(var(n)/length(n)+var(p)/length(p));
 k=k+1;
end
[Z1’ N]
plot(Z1)
pause

% test same pop
k=1;
for alpha=0.01:0.01:0.49
 partial=obs(find((obs(:,K)>=alpha)&(obs(:,K)<(1-alpha))),2);
 newfull=obs(find((obs(:,K)<alpha)|(obs(:,K)>=(1-alpha))),2);
 n=newfull;
 p=partial;
 [length(n) length(p)];
 Z2(k)=(mean(n)-mean(p))/sqrt((length(n)*var(n)+length(p)*var(p))/(length(p)+length(n
 k=k+1;
end
plot(Z2)
pause

F=[];
A=[];
P=[];
V=[];
N=[];
AC=[];
M=[];
% test same var
k=1;
```

```
for alpha=0.01:0.01:0.49
 partial=obs(find((obs(:,K)>=alpha)&(obs(:,K)<(1-alpha))),s);
 newfull=obs(find((obs(:,K)<alpha)|(obs(:,K)>=(1-alpha))),s);
 n=newfull;
 p=partial;
 n(find(abs(n)>10))=[];
 p(find(abs(p)>10))=[];
 [length(n) length(p)];
 F(k)=var(n)/var(p);
 A(k)=alpha;
 AC(k,:)=[autocorr(n) autocorr(p)];
 P(k)=cdf(’F’,F(k),length(n)-1,length(p)-1);
 V(k,:)=[std(n) std(p)];
 M(k,:)=[mean(n) mean(p)];
 N(k,:)=[length(n) length(p)];
 k=k+1;
end
[A’ F’ P’ N/1000 V]
plot(A’,[F’ P’])

error(’tim’)
gibbous=obs(find(obs(:,K)>=0.5),2);
crescent=obs(find(obs(:,K)<0.5),2);
[mean(gibbous) mean(crescent)]

% Txc=length(waxc);
 Tng=length(wang);
 % Txg=length(waxg);
 Tnc=length(wanc);

[mean(waxc) mean(waxg) mean(wanc) mean(wang)]

g=gibbous;
c=crescent;

Z=(mean(g)-mean(c))/sqrt(var(g)/length(g)+var(c)/length(c))

plot(obs(:,K),obs(:,4),’.’)
set(gca,’ylim’,[-0.05 0.05])

% TRADING STRATEGIES
alpha=0.15
 partial=obs(find((obs(:,K)>=alpha)&(obs(:,K)<(1-alpha))),2);
 newfull=obs(find((obs(:,K)<alpha)|(obs(:,K)>=(1-alpha))),2);
 n=newfull;
 p=partial;
 [length(n) length(p)];
 F(k)=var(n)/var(p);
 A(k)=alpha;
 P(k)=cdf(’F’,F(k),length(n)-1,length(p)-1);
 V(k,:)=[std(n) std(p)];
 N(k,:)=[length(n) length(p)];
 k=k+1;
end
[A’ F’ P’ N/1000 V]


[N,X]=hist(p,50);
N=N/sum(N);
bar(X,N)
set(gca,’xlim’,[-0.04 0.04])
set(gca,’ylim’,[0 0.35])
```

Appendix C

Other Markets Stat Tests

Date Range	Level	Vol Paired	Vol UP	Ret Paired	Ret UP	Sharpe Paired	Sharpe UP
	t–test						
1972::2009	0.000	0.000	0.001	0.326	0.399	0.218	0.264
1972::1981	0.000	0.020	0.028	0.311	0.382	0.270	0.244
1982::1991	0.000	0.015	0.093	0.562	0.529	0.860	0.659
1992::2001	1.000	0.086	0.118	0.534	0.519	0.245	0.283
1999::2008	0.000	0.024	0.029	0.167	0.256	0.222	0.264
	Mann–Whitney test						
1972::2009	0.000	0.000	0.001	0.309	0.602	0.913	0.541
1972::1981	0.000	0.003	0.038	0.103	0.381	0.752	0.534
1982::1991	0.000	0.002	0.062	0.661	0.625	0.850	0.602
1992::2001	1.000	0.002	0.140	0.588	0.698	0.820	0.534
1999::2008	0.000	0.005	0.012	0.437	0.320	0.715	0.432

Table C.1: Statistical Tests on the CCMP. The first column, "Level", is the p–value of the test of the level of Faber's timing portfolio (P_T) vs. the index (P_I). $H_0 : P_T \leq P_I$. The second two columns are the p–values of the tests of the timing portfolio volatility (V_T) Vs. the index volatility (V_I). The "Vol Paired" column is a paired test with $H_0 : V_T \geq V_I$ and the "Vol UP" $H_0 : \mu_{V_T} \geq \mu_{V_I}$. The fifth and sixth columns are similar tests of returns. They compare the returns of the timing portfolio R_T with the returns of the index. Column "Ret Paired" tests $H_0 : R_T \leq R_I$ and column "Ret UP" tests $H_0 : \mu_{R_T} \leq \mu_{R_I}$. The last two columns test the Sharpe ratios, S_T and S_I for the timing algorithm Sharpe ratio and the index Sharpe ratio respectively. Column "Sharpe Paired" tests $H_0 : S_T \leq S_I$ and the column "Sharpe UP" tests $H_0 : \mu(S_T) \leq \mu(S_I)$.

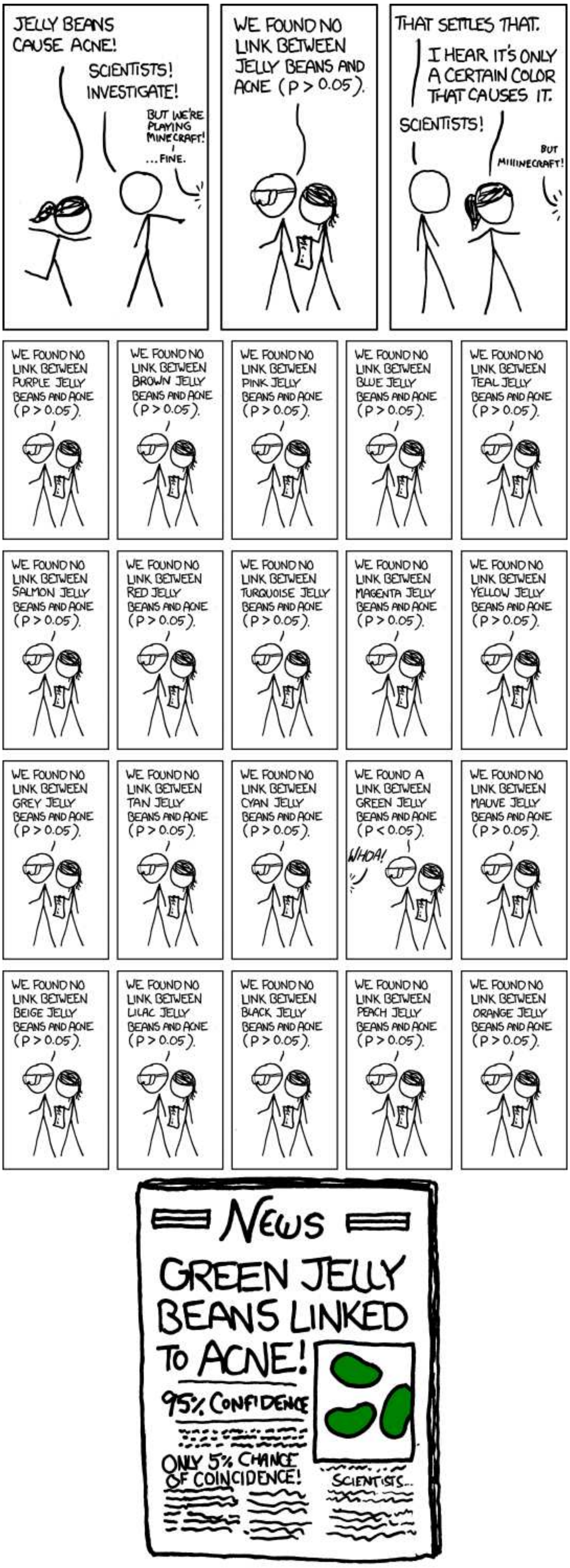

JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Figure C.1: Green Jelly Beans Cause Cancer. Munroe (2011)
licensed under a Creative Commons Attribution–NonCommercial 2.5 License.